UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Investigating the molecular basis of radiation resistance in proteins & Oxidative modifications of SARS-CoV-2 $\mathrm{M^{pro}}$

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg

vorgelegt von
Henry Vincent Gieseler

Hamburg
2023

Gutachter/innen der Dissertation:	Prof. Dr. Arwen R. Pearson
	Dr. Thomas R. Schneider
Zusammensetzung der Prüfungskommission:	Prof. Dr. Arwen R. Pearson
	Dr. Thomas R. Schneider
	Prof. Dr. Daniela Pfannkuche
	Prof. Dr. Dr. Henry Chapman
	Prof. Dr. Henning Tidow
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Daniela Pfannkuche
Datum der Disputation:	07.12.2023
Vorsitzender Fach-Promotionsausschuss PHYSIK:	Prof. Dr. Günter H. W. Sigl
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang J. Parak
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

“What we observe is not nature itself, but nature exposed to our method of questioning.”
– *Werner Heisenberg, Physics and Philosophy: The Revolution in Modern Science*

Acknowledgements

First and foremost I am extremely grateful to my supervisors Prof. Arwen Pearson and Dr. Thomas Schneider for all their help and advice with this PhD. While I am still not sure whether I like or hate unpredictable wobbly proteins that form (more or less) solid crystals, mind-bending symmetry operations or the dark magic of biochemistry, it was absolutely worth it and an enjoyable experience to study them. My gratitude extends to CUI, Universität Hamburg and the Max Planck Institute for Structure and Dynamics which all funded me during my PhD and made my studies possible.

I must thank Dr. Sam Horrell from whom I not only learnt a lot about radiation damage and structure refinement but also the art of combining memes and science fiction references with actual science and how much fun that is.

Special thanks to the technical team for their invaluable support and kindness. No matter what the problem was, you found a solution and made the impossible possible. Most and foremost, I would like to thank Susanne Meier. Without your help, the glove box aftermath would have been a real nightmare for me - still haunts me. So, THANK YOU !

All the hardships of a PhD would have been much worse without my friends and colleagues. In addition to the many inspiring discussions and fun evenings, I knew could always count on you if I needed help with anything. I would like to specially thank Marty Rogers, Helen Ginn and Dimitris Triandafillidis who have all helped me to bring my project forward in one or the other way.

Finally, but most importantly, I would like to thank my family and Lea von Soosten for all the support and love that allowed me to finish this work without ever having to roll a willpower saving throw.

Eidesstattliche Versicherung / Declaration on oath

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.

I, Henry Gieseler, declare that this thesis titled, ” Investigating the molecular basis of radiation resistance in proteins & Oxidative modifications of SARS-CoV-2 M^{pro}” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- The dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (Doctoral Office Physics) for archiving are identical.

Hamburg, 27.09.2023
Place, Date

H. Giesel
Signed

Abstract

The nature of biological radiation resistance is a question that has fascinated and inspired researchers since the first discovery of extremely radiation resistant organisms like *Deinococcus radiodurans*. To date many tightly coordinated mechanisms have been identified that lead to the emergence of the phenotype of extreme radiation resistance but only in the last decade was it found that proteome protection constitutes survival not DNA protection. This thesis aims to provide a molecular understanding of how protein sequence and structure evolved to mitigate the damaging effects of ionising radiation in the environment. To systematically study whether the amino acid composition (primary sequence) is the source of a protein radiation resistance free amino acids, specifically tryptophan, has been soaked into multiple protein crystals and the average response to radiation insult has been compared with their apo counterparts. By collecting dose series measurements and using online UV/Vis spectroscopy the effect on global and specific damage rates has been analysed, showing that for lysozyme, thaumatin, AcNiR and 4HB1 no significant radio protective effect could be achieved by supplementing a protein crystal with tryptophan. Interestingly, one 4HB1 crystal that had been soaked with 100 mM tryptophan survived a total absorbed dose of 182.3 MGy (half dose of 64.43 MGy). Although the exact conditions leading to this result were not reproducible, this result constitutes an unprecedented case of extreme radiation tolerance in an intense X-ray beam and was hence included in the analysis.

A bioinformatics study has been performed, which analysed the amino acid composition radiation hard bacteria and compared them with 8000 bacteria proteomes. The amino acid distribution of radiation resistant organisms showed no common bias towards a particular amino acid or combination of amino acids. Clustering a subset of 200 proteomes from each domain showed that the phylogenetic filiation can be predicted from the amino acid composition, the phenomenon of extreme radiation resistance however can not be predicted. This result supports the conclusion that there is no single amino acid or combination thereof that are the source of a proteins radiation resistance.

Oxidative modifications are not uncommon in cysteine proteases and have been shown affect and even inhibit enzymatic activity. Recent structures of the SARS-CoV-2 main protease (M^{pro}) show indications for similar modifications despite the fact that the enzyme is naturally expressed in the cytosol which is considered to be of reducing nature. To determine whether this modification is an artefact of the purification strategy and what impact such a modification on the enzyme activity as well as recent active-site drug screening efforts would have, M^{pro} was purified under aerobic conditions (as reported by most studies), aerobic conditions without the use of reducing agents and anaerobic conditions. X-ray diffraction data of M^{pro} from both aerobic purifications indicate oxidative modifications of the active site Cys145. Using mass spectrometry, we could show that in the presence of reducing agents M^{pro} is only oxidised when the effectiveness of reducing agents decays, e.g. during long crystallisation periods but not during the purification itself. Without reducing agents at latest after 12 days M^{pro} molecules can be expected to contain sulfenic acid ($-SO$) and sulfinic acid ($-SO_2$) modifications at the active site Cys145. As a result the oxidised enzyme has a specificity constant approximately 50 % lower than unmodified M^{pro} for the substrate Ac-Abu-Tle-Leu-Gln-AMC. By purifying and crystallising M^{pro} under anaerobic conditions this study shows that the oxidation of the enzyme can be avoided and is therefore likely an artefact of the *in vitro* enzyme processing.

Zusammenfassung

Die Natur der biologischen Strahlenresistenz ist eine Frage, die Forscher seit der Entdeckung extrem strahlenresistenter Organismen wie *Deinococcus radiodurans* fasziniert und inspiriert. Bis heute wurden viele eng koordinierte Mechanismen identifiziert, die zur Entstehung des Phänotyps der extremen Strahlenresistenz führen, aber erst im letzten Jahrzehnt wurde festgestellt, dass der Proteomschutz wichtiger für das Überlebendes Organismus ist als der DNA-Schutz. Ziel dieser Arbeit ist es, auf molekularer Ebene zu verstehen, wie sich Proteinsequenz und -struktur entwickelt haben, um die schädlichen Auswirkungen ionisierender Strahlung in der Umwelt abzuschwächen. Um systematisch zu untersuchen, ob die Aminosäurenzusammensetzung (Primärsequenz) die Ursache für die Strahlungsresistenz eines Proteins ist, wurden freie Aminosäuren, insbesondere Tryptophan, zu mehrere Proteinkristallen hinzugefügt und die durchschnittliche Reaktion auf eine Strahlenbelastung mit ihren Apo-Gegenstücken verglichen. Anhand von Dosisreihenmessungen und Online-UV/Vis-Spektroskopie wurden die Auswirkungen auf die globalen und spezifischen Schädigungsraten analysiert. Dabei zeigte sich, dass für Lysozym, Thaumatin, AcNiR und 4HB1 keine signifikante Strahlenschutzwirkung durch die Ergänzung eines Proteinkristalls mit Tryptophan erzielt werden konnte. Interessanterweise überlebte ein 4HB1-Kristall, der mit 100 mM Tryptophan getränkt worden war, eine absorbierte Gesamtdosis von 182.3 MGy (halbe Dosis von 64.43 MGy). Obwohl die genauen Bedingungen, die zu diesem Ergebnis führten, nicht reproduzierbar waren, stellt dieses Ergebnis einen noch nie dagewesenen Fall von extremer Strahlungstoleranz in einem intensiven Röntgenstrahl dar und wurde daher in die Analyse einbezogen.

Es wurde eine bioinformatische Studie durchgeführt, in der die Aminosäurezusammensetzung strahlenresistenter Bakterien analysiert und mit 8000

Bakterienproteomen verglichen wurde. Die Aminosäureverteilung der strahlenresistenten Organismen zeigte keine gemeinsame Tendenz zu einer bestimmten Aminosäure oder Kombination von Aminosäuren. Das Clustering einer Teilmenge von 200 Proteomen aus jeder Domäne zeigte, dass die phylogenetische Abstammung aus der Aminosäurezusammensetzung vorhergesagt werden kann, das Phänomen der extremen Strahlenresistenz jedoch nicht. Dieses Ergebnis unterstützt die Schlussfolgerung, dass es keine einzelne Aminosäure oder eine Kombination davon gibt, die die Quelle für die Strahlungsresistenz eines Proteins ist.

Oxidative Modifikationen sind bei Cysteinproteasen keine Seltenheit und es wurde gezeigt, dass sie die enzymatische Aktivität beeinflussen und sogar hemmen können. Aktuelle Proteinstrukturen der SARS-CoV-2 Hauptprotease (M^{pro}) weisen auf ähnliche Modifikationen hin, obwohl das Enzym natürlicherweise im Zytosol exprimiert wird, das als reduzierend gilt. Um festzustellen, ob diese Modifikation ein Artefakt der Aufreinigungsstrategie ist und welche Auswirkungen eine solche Modifikation auf die Enzymaktivität sowie auf aktuelle Bemühungen zur Entwicklung von Inhibitoren des aktiven Zentrum hätte, wurde M^{pro} unter aeroben Bedingungen aufgereinigt (analog zu den meisten Studien), aerobe Bedingungen ohne Einsatz von Reduktionsmitteln und anaerobe Bedingungen. Röntgenbeugungsdaten von M^{pro} aus beiden aeroben Aufreineinigungen weisen auf oxidative Modifikationen des katalytischen Cys145 im aktiven Zentrums hin. Mittels Massenspektrometrie konnte gezeigt werden, dass M^{pro} in Gegenwart von Reduktionsmitteln nur dann oxidiert wird, wenn die Wirksamkeit der Reduktionsmittel nachlässt, z.B. während langer Kristallisationszeiten, jedoch nicht während der Aufreinigung selbst. Ohne Reduktionsmittel ist spätestens nach 12 Tagen zu erwarten, dass M^{pro}-Moleküle Sulfensäure ($-SO$) und Sulfinsäure ($-SO_2$)-Modifikationen am katalytischen Cys145 enthalten.

Infolgedessen weist das oxidierte Enzym eine etwa 50 % niedrigere Spezifitätskonstante als unmodifiziertes M^{pro} für das Substrat Ac-Abu-Tle-Leu-Gln-AMC auf. Durch die Aufreinigung und Kristallisation von M^{pro} unter anaeroben Bedingungen zeigt diese Studie, dass die Oxidation des Enzyms vermieden werden kann und daher wahrscheinlich ein Artefakt der *in vitro* Proteinhandhabung ist.

Contents

List of Figures

List of Tables

Part I

Tryptophan as a radical scavenger in macromolecular crystallography

Chapter 1

Introduction & Objectives

1.1 Introduction

The nature of biological radiation resistance is a question that has fascinated researchers for decades. Whether it is for cancer radiotherapy[1], space exploration[2, 3], nuclear waste management[4, 5] or X-ray crystallography[6], understanding how ionising radiation affects biological matter is crucial.

Ionising radiation carries enough energy to overcome the binding energy of electrons and if absorbed by an atom or molecule, causes the ionisation thereof. In cells this radiation damage causes oxidative stress, genotoxic effects such as DNA single and double strand breaks and, depending on the dose, eventually cell death.

Nevertheless, some organisms have evolved mechanisms to endure extreme radiation insult and protect themselves from the damaging effects. The most prominent in this regard is the bacterium *Deinococcus radiodurans*. It is known for its extraordinary radiation resistance as it is able to survive exposures of up to 5 kGy without loss of viability[7]. So far many tightly coordinated mechanisms have been identified, for example, high cellular concentration of manganese ions[8, 9], defence mechanisms against oxidative stress[10] and highly efficient DNA repair mechanisms that together form the survival kit of *D. radiodurans*. But only recently it was discovered that damage to the proteome correlates much better with cell death than damage to the genome, indicating that proteome protection is required for survival (and not DNA protection)[11]. Similar results were obtained from X-ray diffraction experiments on nucleoprotein complexes where the DNA component was determined to be far more resistant to specific damage than the protein[12].

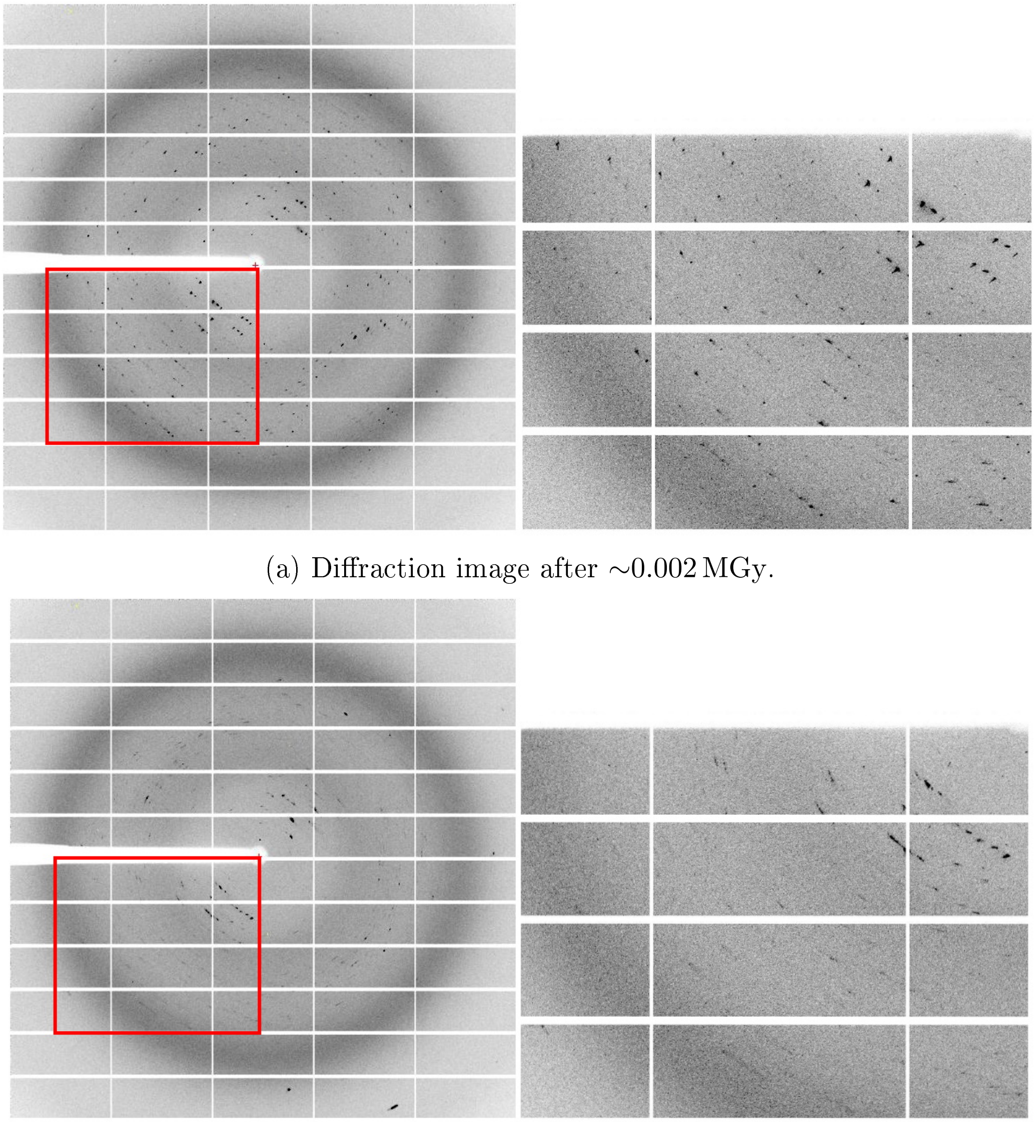
In protein crystallography radiation damage has always been a challenge as it hampers the interpretation of structural features and their biological relevance or even prevents structure determination, for example, by multi wavelength anomalous diffraction (MAD)[13] or of particularly susceptible proteins such as holoferritin. This led to the imperative to develop a better understanding of the underlying radiation damage pathways. In this regard significant progress has been made in the past decades. The most noteworthy achievements include but are not limited to the development of tools for dose estimation[14], the definition of a dose limit beyond which significant structural changes can be expected[15] and the identification of particularly susceptible residues and specific damage[16, 17].
However, the fundamentally diverse nature of proteins (chemically and structurally) poses a major challenge in radiation damage studies as two different proteins may respond differently when exposed to the same dose. Thus, knowledge from one protein system can not always be generalised to another protein system, making systematic studies of, for exmaple, the effect of environment on specific damage, difficult. Current strategies aim to avoid and, where not possible, minimise radiation damage, because it remains hard to predict, model and correct for. Among all experimental techniques that aim to minimise radiation damage such as cyrocooling, helical scans or serial crystallography, the use of small molecules with radioprotective properties

(radical scavengers) is the most questionable in terms of effectiveness[18, 19]. The problem is two-fold. Firstly, the improvement is often not more than a factor of two and may vary depending on the metric used. Secondly, scavengers are most of the time only effective in the buffer system they were tested in and are not transferable to other systems due to their substantially different radiation chemistry.

In contrast to mitigation strategies there is a rapidly increasing number of structures determined with complementary methods where radiation damage is less of a concern, such as neutron scattering, NMR and electron microscopy. Nevertheless, as of today X-ray crystallography remains the method of choice[20] for structure determination. Therefore, especially with the upcoming of 4th generation synchrotrons[21, 22] with even brighter beams and the trend back to room temperature data collection[23–26], tracking, quantifying and understanding radiation damage will be of even more importance for future crystallographers.

1.2 Preliminary Results

During my Master project I created a synthetic protein scaffold, comprised of a limited amino acid alphabet excluding most damageable residues in order to create an as inert background as possible (see Chapter 2.2.4). This scaffold can be used to test the effect of different amino acids on the radiation robustness of the protein when reintroduced into the crystal. However, after collecting sequential low-dose X-ray diffraction data the apo protein was found to be surprisingly susceptible to radiation damage (Figure 1.1).



(a) Diffraction image after ~0.002 MGy.

(b) Diffraction image after 2.6 MGy.

Figure 1.1: Diffraction images collected from a native 4HB1 crystal. The red box marks the magnified area. (a) First image of the 1st data set and (b) First image of the 3rd data set.

In an initial screening, carried out by Dr. Sam Horrell, different amino acids (histidine, aspartate, serine, tryptophan) were soaked into 4HB1 crystals in an attempt to re-introduce these amino acids to the protein crystal and improve its radiation hardness. Aspartate and histidine had minor effects (less than a factor 2 improvement) on the crystal lifetime, serine had no effect. However, one crystal soaked with 100 mM tryptophan appeared to survive a dose of 182.3 MGy while maintaining reasonable diffraction which represented an approximately 90 fold increase in crystal lifetime compared to the apo variant. This was even more impressive because a protein crystal with an absorbed dose of 30 MGy is expected to have suffered from significant damage which will be discussed later in more detail (Section 2.1.5). Further experiments aimed to determine whether this effect is universal for all proteins or just due to the unique properties of the synthetic protein scaffold. A follow up experiment on lysozyme crystals showed that specific damage to disulphide bonds was suppressed when tryptophan is soaked into the crystal. Figure 1.2 shows the spectroscopic signal of the disulphide anion radical (a precursor of disulphide bond breakage) with increasing dose.

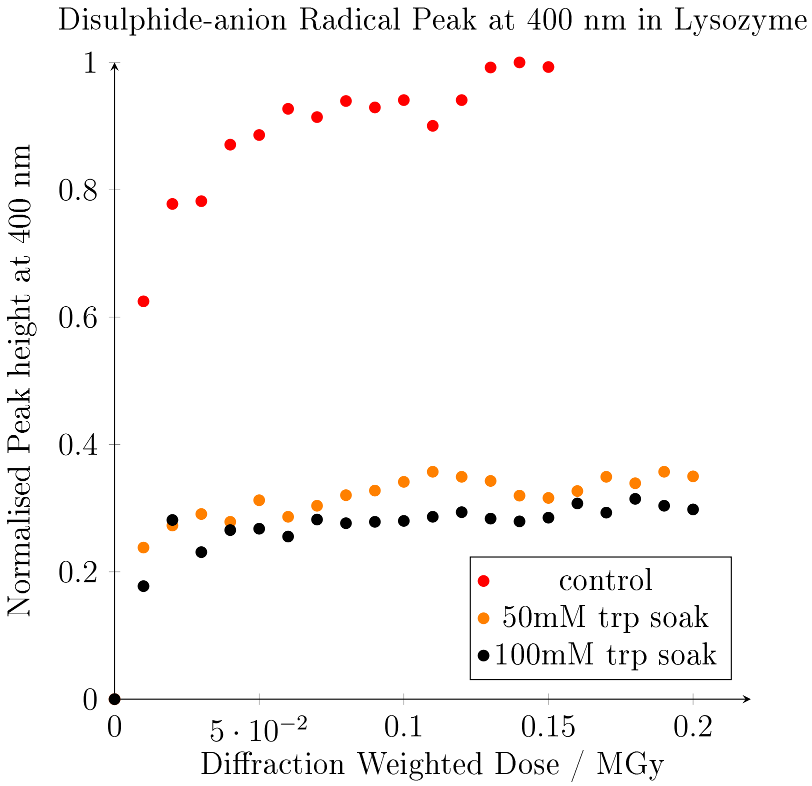


Figure 1.2: Evolution of the disulphide-anion radical peak at 400 nm in Lysozyme during X-ray irradiation. Each data point is normalised to the highest value in the control series. Collected at FIP 14.12.17.

These initial single crystal observations suggested that the addition of tryptophan to a crystal's mother liquor as a radical scavenger is viable strategy to mitigate the effects of radiation damage on a global and atomic level. These initial results were used as a starting point for this thesis.

1.3 Objectives

The aim of this thesis is to provide a molecular understanding of how protein sequence and structure evolved to mitigate the damaging effects of UV and ionising radiation in the environment. Specifically the role of the amino acid composition (primary sequence) is investigated to understand whether the primary sequence gives a protein inherent radiation robustness. Preliminary experiments suggested that tryptophan seems to be a promising radio-protecting additive given that it is also well known for its free-radical scavenging capabilities[27]. Therefore, this thesis focused on the characterisation of tryptophan's radio-protective properties. Building upon the preliminary results, one major objective was to gain statistical confidence by systematically studying the protective effect with sufficient crystals to either confirm or contradict the preliminary single crystal observations. The overall protective effect was determined by soaking the free amino acid into a crystal before carrying out a dose series measurement. The decay of diffraction quality, measured by the total intensity after each data set, was used as a metric to quantify the extend of global damage. Additionally, the effect on specific damage to disulphides and metal centres was investigated by online UV/Vis spectroscopy. A broad range of chemically and functionally different proteins including hen egg white lysozyme, thaumatin and achromobacter cycloclastes copper-nitrite reductase, were used for these experiments. In addition the radiation robustness of the synthetic four helix bundle protein (PDB: 4HB1, see Chapter 2.2.4) was investigated which represents a special case due to its lack of most damageable residues.

To complement the experimental findings a bioinformatics study was conducted with the aim to compare the amino acid composition of radiation hard organisms with the typical distribution of their respective domain to reveal potential deviations that may be linked to the phenotype of extreme radiation resistance.

Chapter 2

Theory & Background

2.1 Radiation damage in macromolecular X-ray crystallography

Macromolecular crystallography relies on ionising radiation to achieve the necessary sub-nanometer resolution to resolve protein structures in great detail. However, as the name implies, ionising radiation carries enough energy to overcome the binding energy of electrons in the sample. Radiation damage occurs when energy from the incident photon is deposited partly or completely within the crystal.

This means, that the X-ray photons used to interrogate the sample are also altering it simultaneously. The outcome is an altered structure which is affected by radiation damage.

This section will discuss the underlying physical principles of this interaction, its consequences, how radiation damage is observed, measured and possible mitigation strategies.

2.1.1 Interaction of photons with matter

The interaction of photons with matter can occur via four different processes, namely photoabsorption, Compton scattering, Thomson scattering and pair production. For the special case of a protein crystal, approximately 98 % of the incident photons pass through the crystal without interacting at all (assuming a photon energy of 12.4 keV). The remaining 2 % interact via

1. **Photoelectric effect** 84 %
   The photoelectric effect is an inelastic scattering event in which the energy of the incident photon is fully transferred to an inner shell electron which gets ejected from the atom. This secondary electron carries enough energy (12 keV) to cause approximately 500 additional ionisation events within a few micrometers of the primary absorption location, assuming a binding energy of 25 eV[28]. Additionally the ionised atom can lose an outer shell electron to fill the lower shell hole in an Auger-process.
   The photoelectric effect has the largest contribution to radiation damage.

2. **Compton scattering** 8 %
   Similar to the photoelectric effect, Compton scattering describes an inelastic scattering event. Here, the incident photon is scattered off an outer shell electron. During this process the photon transfers a part of its energy to the electron which recoils from the point of impact,causing the ionisation of the atom. The photon is in turn deflected onto a new path with a new energy equal to the incident photon energy minus the sum of the electron's kinetic energy and binding energy.

3. **Thomson (Rayleigh) scattering** 8 %
   Thomson scattering is elastic scattering (no energy transfer) and gives rise to the observable diffraction pattern in X-ray crystallography.

4. **Pair production (negligible for energies used in MX)**
   Pair production describes the process in which a photon creates an electron-positron pair near a nucleus. Since the photon energy needed has to be at least equal to the total rest mass energy of both particles, this effect is negligible for energies used in most macromolecular X-ray diffraction experiments.

The cross sections of these interactions are dependent on the photon energy and atomic number Z of the interacting atom with the total cross section σ_{tot} being the sum of all contributions:

$$\sigma_{tot} = \sigma_{pe} + \sigma_{inel} + \sigma_{el} + \sigma_{pp}. \tag{2.1}$$

For carbon, the most common element in a protein structure after hydrogen, the cross section for each interaction and the total cross section σ_{tot} are plotted against the photon energy in Figure 2.1.

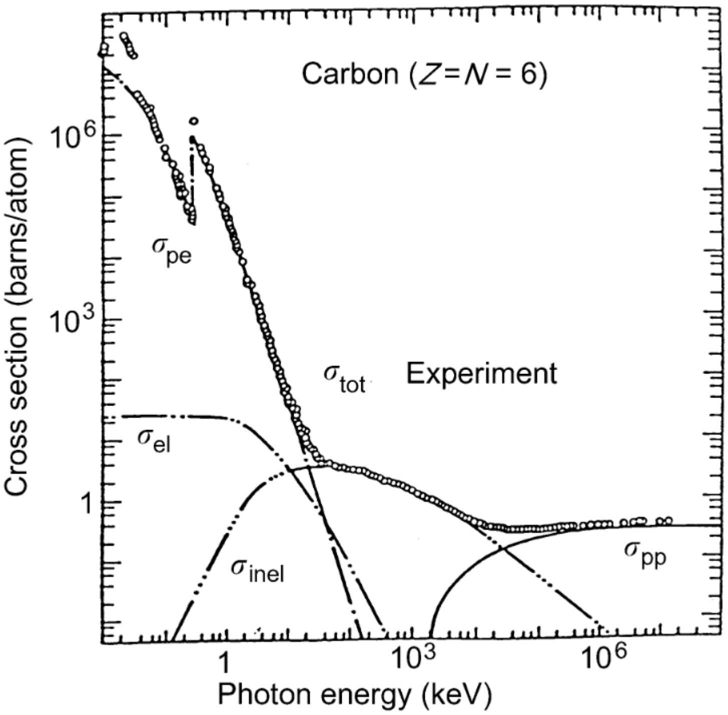


Figure 2.1: Theoretical cross sections for photon interactions with carbon showing the contributions of photoelectric σ_{pe}, elastic (Rayleigh) σ_{el}, inelastic (Compton) σ_{inel}, and pair-production cross sections σ_{pp} to the total cross sections σ_{tot}. Also shown is experimental data (open circles) from Gerstenberg & Hubbell (1982)[29].

It can be seen that cross section for elastic scattering σ_{el} is almost constant up to 1 keV photon energy and then slowly decreases. However, at 1 keV energy the contribution of the photoelectric effect σ_{pe} is several orders of magnitudes larger. To reduce the contribution of damaging effects while maintaining the highest level of elastic scattering, macromolecular X-ray diffraction experiments are usually performed between 5 keV and 15 keV[30].

2.1.2 Primary Damage

When energy is deposited within the crystal by an inelastic scattering event, the resulting ionisation of an atom can be described as a primary damage event. Consequently, primary damage depends only on the photon energy and number of absorbed photons and is inherent to all methods that use ionizing radiation. Because inelastic scattering processes occur simultaneously to elastic scattering processes, primary radiation damage cannot be avoided.

2.1.3 Secondary Damage

Secondary damage is caused by free radicals resulting from the primary damage event or secondary ionisations. The reaction of radical species with the protein may cause bond breakage, hydrogen abstraction and covalent addition of secondary species to side chains or the protein backbone. Frequently observed are metal centre reductions, disulphide bond breakage and decarboxylation events, which are covered in more detail in Section 2.1.6.
In protein crystals the radiolysis of water is the main source of radical species because protein crystals contain a significant amount of solvent compared to small molecule crystals. The radiolysis of water produces a wide range of different radiolytic products resulting in a cascade of possible reactions:

$$\begin{aligned}
&\mathrm{H_2O} \xrightarrow{\mathrm{h}\nu} \mathrm{H_2O^{\bullet +}} + \mathrm{e^-}\\
&\mathrm{H_2O} \xrightarrow{\mathrm{h}\nu} \mathrm{H_2O^{\star}} \longrightarrow \mathrm{{}^{\bullet}OH} + \mathrm{H^{\bullet}}\\
&\mathrm{e^-} + \mathrm{nH_2O} \longrightarrow \mathrm{e^-_{aq}}\\
&\mathrm{e^-_{aq}} + \mathrm{H^+} \longrightarrow \mathrm{H^{\bullet}}\\
&\mathrm{H_2O^{\bullet +}} + \mathrm{H_2O} \longrightarrow \mathrm{H_3O^+} + \mathrm{{}^{\bullet}OH}
\end{aligned}$$

The radical species produced are highly reactive, can react with each other producing molecules like H_2O_2 and H_2, or react with the proteins backbone or side chains. It is important to note that recombination of electron-loss and electron-gain centers is always in competition with charge separation through migration.

The two most prominent radiolytic products are the solvated electron e^-_{aq} and the hydroxyl radical $^\bullet OH$. Solvated electrons are the major damaging species at cryotemperatures as they can still migrate quantum mechanically via a tunneling mechanism[31]. It was shown that electrons can travel a significant distance along the protein backbone until they are trapped by an electron affine moiety[32]. The carbonyl group of a peptide bond represents a major trapping centre for solvated electrons which can lead to main chain scission (Equation 2.2).

$$\mathrm{P{-}CONHCH(R){-}P} + e^-_{aq} \longrightarrow \mathrm{P{-}CONH_2} + {}^\bullet\mathrm{CH(R){-}P} \qquad (2.2)$$

In contrast to solvated electrons, hydroxyl radicals are strong oxidising agents. They have been found to react with α-carbons of the main chain. In oxygenated solutions this results in oxidative degradation and backbone scission (Equation 2.3).

$$\mathrm{P{-}CONHCH(R){-}P} + \mathrm{OH}^\bullet + \mathrm{O_2} \longrightarrow \mathrm{P{-}CONH_2} + \mathrm{RCO{-}P} + \mathrm{Products} \qquad (2.3)$$

Other targets include unsaturated aromatic residues like tryptophan or histidine, which in case of the former causes a ring opening of the indole moiety. For these damaging processes the proximity of the hydroxyl radical to a protein site is the key factor that controls if and where the damage occurs. As such they are distributed stochastically over the whole protein with the protein surface having the highest probability. However, as they do not target a specific site, they are very rarely observed in electron density maps due to the fact that they average out.

Any damaging species can also react with solvent components to either form new reactive products or immobilise/stabilise them. As each protein requires a unique crystallisation/cryo-cocktail which adds a lot of chemical complexity to the system, the radiation chemistry can become extremely complex. This complexity and variety of different crystallisation/cryo-cocktails makes the prediction of exact reaction pathways very hard, if not impossible. Thus, secondary damage varies with the nature of the solvent, temperature and the presence or absence of free-radical scavengers that affect the mobility and reactivity of the radiolytic products[33].

2.1.4 Dose

The dose D is defined as the energy deposited per unit mass of sample expressed in SI units of Gray $[Gy = J/kg]$. However, the dose cannot be directly measured but instead is estimated using knowledge of beam characteristics, crystal characteristics and experimental details[15]. These characteristics include photon energy, beam flux, beam size, and two-dimensional profile, crystal volume, crystal morphology, unit cell size, protein atomic contents, number of amino acids, solvent composition as well as the exposure time per image and the total number of images.

Raddose-3D is a software package developed by *Zeldin et al.*[14] that permits the estimation of the absorbed dose using the parameters mentioned above. It distinguishes between the maximum dose, the average dose and the diffraction weighted dose. The maximum dose reflects the worst case and is the highest dose any crystal voxel has received, whereas the average dose is the average across all illuminated crystal voxels. Depending on the beam profile (top hat or gaussian) these two values can deviate quite a bit from each other. The diffraction weighted dose accounts only for the dose absorbed by crystal voxels that still contribute to the diffraction pattern and is therefore the most accurate for determining the impact on the diffraction data.

Because protein crystals are only 10 to a few 100 µm in size, the typical dose a protein crystal will receive during a standard dataset collection is in the MGy regime.

2.1.5 Dose Limits for MX

In 1990 Henderson deduced from observations made in electron diffraction experiments at 77 K, that a protein crystal would lose half of its diffracting power after being exposed to a dose of 20 MGy, the "Henderson limit"[34]. This initial suggestion for a dose limit $D_{1/2}$ was later experimentally measured by *Owen et al.* to be 43 MGy for macromolecular crystallography at 100 K[35]. The final electron density maps, however, showed significant damage to certain amino acids (specific damage) and a more conservative experimental dose limit of $D_{0.7} = 30\,\text{MGy}$ was recommended, the "Garman limit". It may still be the case that a crystal "dies" before reaching the dose limit because the dose as a metric takes only the physical processes into account when X-rays interact with a crystal (primary damage) but not any radiation chemistry related to the sample composition[15].

2.1.6 Specific Damage

Specific damage refers to changes in a protein's atomic structure due to the reaction with different radical species or solvated electrons. These changes occur before global damage becomes apparent.
At cryotemperatures, radiation damage can be observed in a well defined and predictable order due to the susceptibility of certain protein residues to radiation damage.

1. Generally most susceptible are metalloproteins due to the high cross-section of their bound metal ion. Even at very low doses (10 - 100 kGy) these metal ions get reduced by solvated electrons e^-_{aq}[36]. As a conse-

quence, the coordination geometry of the metal ion might change causing a reorganisation of the local environment. Because metal ions often have an essential role in the function of a protein (e.g. charge transfer, catalysis, organisation of residues, etc.), crystallographers must be particularly wary of radiation damage to these sites before assigning biological function or relevance to a structural feature[37].

2. Disulphide bonds are the second most susceptible motifs in proteins[16]. Within a protein sulfur has the highest electron affinity ($-\Delta H = 200\,\text{kJ/mol}$) and hence electrons that travel along the protein backbone or solvated electrons are readily trapped by disulphide bonds. Upon reduction by an electron, a disulphide-radical anion forms which causes an elongation of the bond by up to $\sim 0.7\,\text{Å}$. The formation of this intermediate can be followed spectroscopically as the UV/Vis-absorption at 400 nm will rise with increasing disulphide-anion-radical concentration[38]. The disulphide-radical anion can undergo spontaneous and reversible bond breakage (Equation 2.4).

   $$\text{RS-SR} \xrightarrow{+e^-} \text{RS}^{\bullet-}\text{-SR} \rightleftharpoons \text{RS}^{\bullet-} + \text{SR} \tag{2.4}$$

   While the required dose for the formation of disulphide-radical anion is in the kGy regime, structural changes such as bond ruptures become only visible in the electron density at higher doses, typically several 100 kGy.

3. At higher doses (approx. 3 MGy) the acidic residues aspartate and glutamate may lose their definition and ultimately suffer decarboxylation

in a two step reaction:

$$\mathrm{R{-}(CH_2)_n{-}CO_2^{-}} \xrightarrow{h\nu} \mathrm{R{-}(CH_2)_n{-}CO_2^{\bullet}{+}\,e^{-}} \tag{2.5}$$

$$\mathrm{R{-}(CH_2)_n{-}CO_2^{\bullet}} \longrightarrow \mathrm{R{-}(CH_2)_{n\text{-}1}{-}CH_2^{\bullet}{+}\,CO_2} \tag{2.6}$$

> Carbon dioxide is released, but at cryotemperatures is trapped inside the crystal. It is believed that the formation of gas inside the crystal is the driving force of unit cell expansion and the main cause for loss of high resolution information with increasing dose. Meents *et al.*[39] could show that the majority of gas produced ($> 80\,\%$) is H_2 and originates from organic compounds present in the irradiated sample and not directly from water.

The exact mechanism of site specific radiation susceptibility is still unclear. It has been shown that within the same crystal two identical chemical groups can reproducibly experience radiation damage at different rates. This phenomenon is called preferential specific radiation damage. Although several mechanism have been proposed to explain these observations, such as solvent accessibility, electric field lines, local chemical environment or higher absorption cross sections of heavy atoms, for each there is both evidence and counter examples.

This inherent unpredictability means that in most cases the structure needs to be solved before damaged sites can be identified.

2.1.7 Global damage

Global damage affects the crystal lattice and is observed in reciprocal space. The most visible sign is the gradual fading of the diffraction pattern with high resolution reflections disappearing first as the absorbed dose increases. Other signs manifest in the data as decreasing intensities I, increasing noise ($\sigma(I)$) and therefore decreasing $I/\sigma(I)$, an increasing Wilson B-factor, increasing unit cell dimensions, worse merging R-factors and often increasing mosaicity. Not all of these observations are equally good metrics to measure the extent of global radiation damage. An increase in mosaicity or unit cell dimensions can sometimes be observed but due to their infrequent observation and variability, even within fragments of the same crystal are not reliable indicators. The decrease of signal to noise ratio $I/\sigma(I)$ might appear worse than it actually is because the associated noise level $\sigma(I)$ is also increasing with increasing dose.

Three metrics have been proposed to be plotted as a function against the absorbed dose D[40].

1. The total intensity of the nth data set divided by the total intensity of the first data set I_n/I_1.
2. The relative scaling B-factor $B_{rel} = B_n - B_1$, which is the difference in Wilson B-factor from the nth data set to the first data set.
3. The pairwise R-factor R_d between identical and symmetry related reflections occurring on different diffraction images.

However, it should be noted that these three indicators may yield inconsistent results for analysis of the same data as shown by *De la Mora* [41]. This poses an yet unresolved issue in systematic radiation damage studies.

2.1.8 Radiation Damage Mitigation Strategies

Cryo-cooling

In the first decades of protein crystallography radiation damage was the major factor that prevented successful structure determination. At room temperature secondary damage is time dependent and with several hours exposure required for a complete data set, crystal degradation was severe. This problem was largely alleviated by the introduction of cryo-cooling. Flash cooling the crystal in liquid nitrogen below 100 K immobilized most secondary damaging species produced during irradiation and prolonged the life time of a protein crystal in an X-ray beam significantly (approx. 70-fold)[42]. However, with the introduction of 3rd generation synchrotrons and increasing photon fluxes, the problem of radiation damage re-emerged even at cryo-temperatures.

Serial Crystallography

Serial Crystallography is a method in which many crystals contribute to one full data set. Depending on the exact method only a small wedge or as little as one image is collected from each crystal and later recombined with others to obtain a full data set. Using serial crystallography the dose can be distributed evenly across multiple crystals yielding a "low dose" structure. This method of data collection is promising for proteins which are particularly susceptible to radiation damage (e.g. metalloproteins). Serial crystallography often used for time resolved experiments which involve some kind of perturbation of the crystallised molecules for example ligand soaking or induced dynamics. In this context small crystals offer a number of advantages. Because small crystals have much less unit cells compared to larger crystals, such perturbations can be applied more uniformly and much faster while simultaneously creating less strain resulting from changes in crystal lattice dimensions[43].

An extreme example is the utilization of an X-ray Free Electron Laser (XFEL) source to collect a serial data set. Here, an intense X-ray pulse is fired at a stream of microcrystals, each yielding one diffraction image on a femtosecond time scale before the crystal is destroyed by the pulse ("diffraction before destruction")[44, 45].

Radical Scavenger

Another strategy to mitigate secondary damage is the addition of small molecules to the cryobuffer or crystallisation solution. These compounds react with radical species to form a more stable or less motile product and thereby neutralise the damaging potential of the radical before they reach the protein. In electron spin resonance (ESR) spectroscopy scavengers like 5,5-dimethylpyroline-N-oxide (DPMO) and 2,2,6,6-tetramethyl-4-piperidone (TEMP) are widely used as spin traps but have not been investigated in protein crystallography so far.

The first mention of radical scavengers in protein crystallography was by *Zaloga* & *Sarma* 1974 who co-crystallised radiation sensitive IgG immunoglobin molecules with styrene monomers. They observed an improved crystal lifetime of up to 10-fold at room temperature measured by the intensity of a single reflection[46].

The chemical properties of the scavenger molecule dictates what type of radical it can scavenge. As such, many cryoprotectant agents are already efficient hydroxyl radical and H-atom scavengers (e.g. glycerol, ethylene glycol, PEG, glucose or other sugars) while acetone and transition metal compounds are acceptors for free electrons. However, supplying high enough scavenger concentrations for a measurable impact can be difficult, depending on the buffer system, protein system and solubility of the scavenger. For example, it was calculated that at room temperature a 1 M scavenger concentration is needed to reduce the migration track length of hydroxyl radicals $OH\cdot$ to approximately 1 nm[47]. The usage of transition metal compounds is especially problematic as high concentrations would significantly increase the absorption cross-section and thereby worsen radiation damage.

Since the first reports many other molecules with potential scavenging capability have been proposed but only few systematically tested with mixed results. *Holton* [48] suggested an at least twofold increase of $D_{1/2}$ as a benchmark for judging the effectiveness of radiation damage mitigation strategies, which only very few scavengers surpassed. Among the most promising radial scavengers are ascorbate [49], nicotinic acid[50] and benzoquinone[51]. However, a scavenger that was found effective in one buffer system might not produce similar results in another, as some buffer components react strongly with radiolytic products. Additionally, the effectiveness of a scavenger when judged by different metrics may yield disagreeing results, making drawing meaningful conclusions even more difficult. Lastly, due to variability of radiation vulnerability between otherwise similar crystals, statistically meaningful results can only be obtained when multiple crystals are examined[18]. Due the general lack thereof and the disagreement about the utility of scavengers within the literature, they are rarely used as a tool to mitigate secondary damage[19].

2.2 Protein Model Systems for MX Radiation Damage Studies in this Thesis

2.2.1 Hen Egg White Lysozyme

Lysozyme is one of the best studied and characterised proteins due to a few very beneficial properties. It crystallises in a variety of crystallisation buffers and pH ranges, it is stable and easy to handle at room temperature and reliably yields large, well diffracting crystals. Many researchers have therefore used the 129 amino acid long (14.3 kDa) protein as their first target for method development or in proof-of-principle studies.

The above mentioned qualities make lysozyme also an attractive subject for systematic radiation damage studies because large quantities of high quality data can be produced. The decay of high resolution reflections, for example, can be monitored to track global radiation damage. In addition lysozyme has eight cysteine residues which all take part in its four disulphide bonds and nine acidic residues which can be used to asses the level of specific damage.

2.2.2 Thaumatin

Thaumatin is a plant derived 207 amino acid long (22.2 kDa) protein containing eight disulphide bonds. It has been in the focus of research due the fact that it is a taste-active (e.g. incredibly sweet) protein. Considering its availability and ease of handling thaumatin has become a model system for many crystallographic studies including radiation damage studies[52, 53].

2.2.3 Achromobacter Cycloclastes Copper-Nitrite Reductase

Achromobacter cycloclastes copper-nitrite reductase (AcNiR) is a 334 amino acid (36.6 kDa) long protein that forms a homotrimer and is involved in the global denitrification pathway. It reduces nitrite to nitric oxide[54].

$$\mathrm{NO_2^-} + \mathrm{e^-} + 2\,\mathrm{H^+} \longrightarrow \mathrm{NO} + \mathrm{H_2O} \tag{2.7}$$

The reaction is catalysed by two bound copper ions, a type I copper electron-transfer site and a catalytic type II copper site[55]. As described in Chapter 2.1.6, metal sites are particularly vulnerable to radiation damage which makes AcNiR an interesting and useful model system. Specific damage to one of the metal sites can be investigated spectroscopically by following the UV/Vis absorption at 450 nm. The absolute peak is caused by the interaction of a methionine sulphur atom with a copper atom, in a copper type I configuration[56]. Loss of the 450 nm peak is therefore correlated to any chemical or structural change that would interfere with this interaction.

2.2.4 The Four Helix Bundle Protein

The four helix boundle protein (PDB 4HB1) is a synthetic protein which was *de novo* designed by *Stroud et al.* to self-assemble into a four helix bundle[57]. The protein is made from a limited amino acid alphabet using only seven out of the possible 22 natural occurring amino acids, specifically, glycine, lysine, leucine, serine, alanine, glutamine and glutamic acid. Each of the four helices is 24 amino acids long and has an identical sequence, in which hydrophobic (alanine and leucine) and hydrophilic (glutamine) residues are arranged in a primary pattern that leads to a helical secondary structure and the assembly of a four helix bundle (tertiary structure) with a hydrophobic core and a hydrophilic surface. The helices are connected via loops comprising 3, 4 and 3 glycines. Together with the two N-terminal serines 4HB1 has in total 108 amino acids (11.8 kDa).

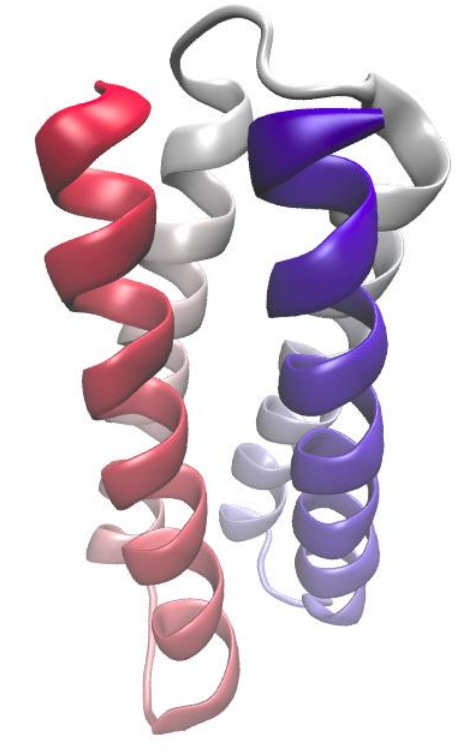


Figure 2.2: Cartoon representation of the four helix bundle protein structure.

The lack of aromatic amino acids, disulphide bonds, metal binding sites as well as the presence of hardly any charged residues provides a uniform chemical background against which a variety of different mutations can be introduced to create different chemical environments.

The absence of almost all known specific damage sites and its reduced amino acid alphabet is a unique property among all proteins. The four helix bundle 1 is therefore an interesting target and ideal scaffold for systematic radiation damage studies.

2.3 Evolution of Extreme Radiation Resistance

The earth is fairly well protected from ionising radiation by its thick atmosphere and its magnetic field. It is no surprise that the average annual dose any organism is exposed to is approximately six orders of magnitude lower than what is experienced by a protein crystal during data collection of a single data set. For example, the average annual dose from background radiation in Germany is around 2.4 mGy. Despite the comparatively low levels of ambient radiation on earth all life has developed some form of protection against ionizing radiation or its damaging effects.

The pigmentation of human skin, for example, is a direct response to exposure to damaging UV radiation[58]. The associated pigment melanin is able to dissipate 99.9 % of the absorbed energy by ultrafast internal conversions of electronically exited states to vibrational states[59]. As a second line of defence several mechanisms are employed to mitigate secondary damage by reactive oxygen species (ROS) and reactive nitrogen species (NOS). On a cellular level these species are usually kept under tight control due to their damaging potential. However, exposure to ionizing radiation causes increased oxidative stress. To combat oxidative stress the human body uses enzymes like catalase and superoxide dismutase as well as small molecules such as glutathione, vitamin C and E. Additionally, amino acids and their derivatives constitute another family of compounds that function as free radical scavengers and antioxidants. The molecule that has been most widely investigated in this regard is N-acetyl-5-methoxytryptamine (melatonin), a tryptophan derivative[60].

Despite all these different mechanisms humans are by far not the most radiation robust species. Across the tree of life multiple organisms have evolved means to withstand extreme radiation insult. Compared to humans, for which an immediate exposure to a dose of 6 Gy is lethal, these organisms can survive up to 10 kGy with only minimal loss of viability.

In terms of natural selection the evolution of extreme resistance to ionizing radiation is the most difficult phenotype to rationalize as there is no natural occurring environment that would exert sufficient selection pressure. Instead, it is hypothesised that the effects of ionizing radiation are similar to other physiological stresses such as desiccation, to which organisms adapted[61, 62].

2.3.1 Deinococcus radiodurans

Since its discovery in 4 kGy γ-irradiated sterilised food by *Anderson et al.*[63] in 1956, the bacterium *Deinococcus radiodurans* has been well known for its extraordinary radiation resistance. Strikingly, it was shown that its genome is not more resistant *per se* as it suffers the same amount of DNA double strand breaks as non-resistant bacteria[64]. Instead, researchers could show that the fully fragmented genome was fully restored 3 hours post irradiation[65, 66], indicating a highly efficient DNA repair mechanism.

Additionally, other mechanisms such as the redundancy of genomic information through multiple genome copies[67], high cellular Mn(II) content[8, 9] and defence mechanisms against oxidative stress[10] are also likely to contribute to the survival kit of *D. radiodurans*.

Strikingly, researchers also found that cell death correlates far better with damage to the proteome rather than DNA damage and this is caused primarily by oxidative damage with consequential loss of enzymatic activities including DNA repair [11, 68]. This indicated that proteome protection is neccessary for survival and not DNA protection alone.

Chapter 3

Materials & Methods

3.1 4HB1 Protein Production and Purification

3.1.1 Buffers and Solutions

- **Binding Buffer**: 150 mM NaCl, 20 mM Imidazole, 50 mM Tris pH 8
- **Elution Buffer**: 150 mM NaCl, 300 mM Imidazole, 50 mM Tris pH 8
- **Dialysis Buffer**: 50 mM NaCl, 50 mM Tris pH 8

3.1.2 Cloning & Transformation

Since 4HB1 contains four helices with identical amino acid sequence, the gene was designed to allow individual mutation of a specific helix by choosing different sequences of coding nucleotide triplets for each amino acid within each helix while respecting preferential codon usage of *E. coli*. The construct also contains a N-terminal His_6-tag followed by a TEV cleavage site (ENLYFQ↓S) which after tag removal produces the desired N-terminal serine. The 4HB1 gene was cloned into the pET-24d(+) expression vector using XhoI and BamHI restriction sites (Figure 3.1).

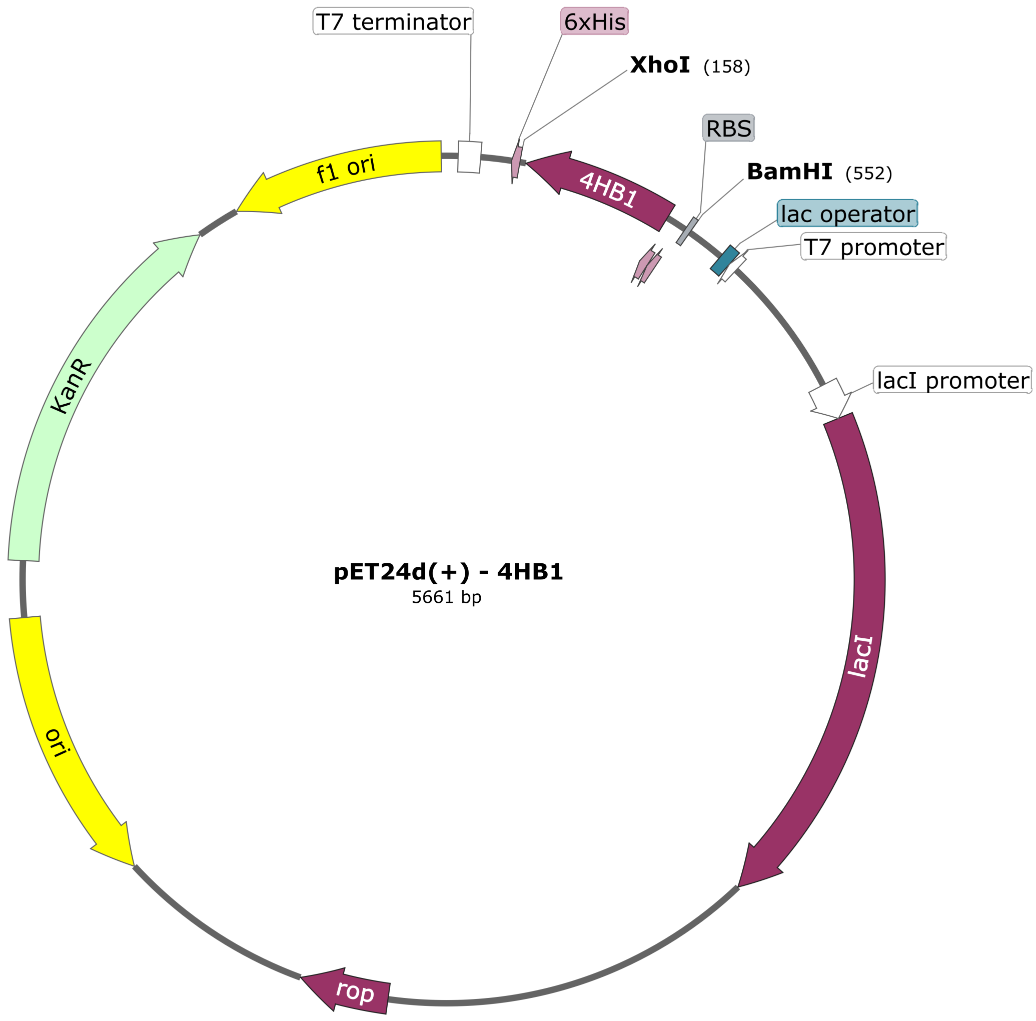


Figure 3.1: Expression vector map pET-24d(+) with the 4HB1 insert.

An aliquot of competent *E. coli* strains XL10 and BL21(DE3) was thawed on ice and each tube incubated with 100 ng of the pET-24d(+)-4HB1 vector for 20 min. The cells were subsequently transformed using the heatshock method at 42 °C for 40 s and afterwards placed on ice for 2 min. After addition of 300 µL LB media to each tube, the tubes were incubated at 37 °C, 180 rpm for 40 min. The cells were then plated onto LB agar plates supplemented with 25 µg/mL kanamycin and allowed to grow overnight at 37 °C.

3.1.3 Cell Culture

3 L culture flasks with 1 L LB media were supplemented with 50 µg/mL kanamycin and inoculated with 2 mL starter culture. The cultures were grown at 37 °C and 200 rpm using an INFORS HT Multitron standard incubator until an $OD_{600} = 0.5$ was reached. Protein expression was induced by adding IPTG to a final concentration of 300 µM and incubation at 37 °C, 200 rpm continued overnight.

3.1.4 Cell Lysis

The cell culture was transferred to six 1 L centrifuge bottles (500 mL each) and balanced with water or cell culture. Cells were harvested by centrifugation using a Sorvall Lynx 6000 centrifuge with the F9-6x1000 LEX rotor at 7000 rpm for 40 min at 4 °C and subsequently resuspended in 10 mL binding buffer. PMSF was added to the cell suspension to a final concentration of 1 mM and sonicated for 10 min with the following settings:

- sonicate for 5 s
- pause for 55 s

The lysate was transferred to two 35 mL centrifuge bottles, balanced and centrifuged at 12.000 rpm (F14-14x50cy rotor) for 40 min at 4 °C. The supernatant was collected and filtered through a 0.45 µm syringe filter.

3.1.5 Ni-Affinity Chromatography

Prior to use a 5 mL Hi-Trap column was stripped and recharged. It was then washed with six column volumes (CV) water and equillibrated with six CV binding buffer. The supernatant (SUP) was loaded onto the column and the flowthrough (FT) collected. The column was washed with six CV binding buffer and two 15 mL fractions (W1 and W2) were collected. Any bound protein was subsequently eluted by applying 15 mL elution buffer to the column and three 5 mL fractions (E1-E3) were collected. Finally, the column was washed with six CV water for further use.

3.1.6 SDS-PAGE

SDS-PAGE gels were prepared according to the following recipe:

Table 3.1: Recipe for four 15 % Tris-Glycine gels.

running gel		stacking gel	
1.5 M Tris buffer pH 8.8	5.25 mL	1 M Tris buffer pH 6.8	584 µL
40 % *(v/v)* bis-/acrylamide	7.88 mL	40 % *(v/v)* bis-/acrylamide	693 µL
H_2O	7.66 mL	H_2O	5.66 mL
10 % *(w/v)* SDS	210 µL	10 % SDS *(w/v)*	70 µL
TEMED	7 µL	TEMED	7 µL
15 % *(w/v)* APS	140 µL	15 % APS *(w/v)*	70 µL

10 µL of each sample was mixed with 10 µL 2x SDS loading dye and boiled at 80 °C before 3 µL of ultra-low range marker was loaded to the first well and 10 µL of loading dye/sample mix were loaded per well. Each SDS-PAGE was run at 180 V for 45 min and afterwards stained with InstantBlue for 15 min.

3.1.7 Dialysis

Fractions containing protein were pooled and dialysed in a 3.5 kDa MWCO SnakeSkin® dialysis tubing against 2 L dialysis buffer. Any air bubbles were removed from the tube before it was sealed and left for dialysis for 2 h at room temperature with slow stirring.

3.1.8 His-tag Cleavage

The dialysed sample was transferred to a new falcon tube and 2 mL TEV protease (1 mg/mL) and EDTA to a final concentration of 0.5 mM added. The cleavage mix was incubated at 30 °C and 170 rpm overnight. Any precipitation was spun down before continuing.

3.1.9 Reverse Ni-Affinity Chromatography

The column was washed with six CV water and afterwards equilibrated with six CV binding buffer. Then, the cleaved protein (CP) was loaded onto the column and three 3 mL fractions flow through (FT1-FT3) were collected. The column was washed with two CV binding buffer and two 6 mL fractions (W1 and W2) were collected before any remaining protein was eluted with 15 mL elution buffer. Three 5 mL fractions (E1-E3) were collected and the column was washed with six CV water.

3.1.10 Protein Concentration

A 5 kDa MWCO concentrator was used to concentrate the protein solution to 200 µL using a HERAEUS MEGAFUGE 40R with a Tx-1000 rotor at 4000 rpm in 10 min intervals. To avoid accumulation the protein was gently pipetted up and down after each interval.

3.1.11 4HB1 Concentration Estimation

Due to the unique chemical composition of 4HB1 (e.g. absence of aromatic residues), the protein concentration could not be determined using conventional UV/Vis absorption measurements at 280 nm, Bradford, BCA or Folin-

Lowry assays.

Instead a concentration series of lysozyme in water with 5, 2.5, 1, 0.5, 0.1 0.05 mg/mL and of 4HB1 with 100 % and 50 % was prepared for comparison on an SDS-gel. The purified protein was aliquoted, flash cooled with liquid nitrogen and stored at −80 °C.

3.2 Crystallisation

All proteins were crystallised using the hanging drop vapour diffusion setup with 1000 µL reservoir volume and 2 µL drops. 4HB1 was prepared as described above, AcNiR was provided by Dr. Sam Horrell and lysozyme and thaumatin purchased as a powder and dissolved in their respective crystallisation buffer, listed in Table 3.2.

Before flash cooling, crystals were transferred for 3 s to a protein specific cryobuffer (Table 3.3) which contained up to 100 mM tryptophan depending on the experiment.

Table 3.2: Crystallisation Conditions

Lysozyme	1.7 M NaCl, 25% *(v/v)* Ethylene glycol, 50 mM sodium acetate pH 4.7
Thaumatin	1.2 M Na/K tartrate, 15% *(v/v)* Ethylene glycol, 0.1 M Bis-Tris propane pH 6.8
AcNiR	1.4 M Ammonium sulphate, 50 mM sodium acetate pH 4.8
4HB1	65% saturated Ammonium sulphate, 3% *(v/v)* Isopropanol 0.1 M Tris pH 8.6

Table 3.3: Cryo buffers for selected proteins

Lysozyme	1.7 M NaCl, 25% *(v/v)* Ethylene glycol, 50 mM sodium acetate pH 4.7, X mM tryptophan
Thaumatin	1.2 M Na/K tartrate, 25% *(v/v)* Ethylene glycol, 0.1 M Bis-Tris propane pH 6.8, X mM tryptophan
AcNiR	3 M Ammonium sulphate, 50 mM sodium acetate pH 4.8, X mM tryptophan
4HB1	same as crystallisation buffer

3.3 4HB1 Dose series

3.3.1 Experimental setup

For this experiment 4HB1 crystals were soaked in a cryobuffer supplemented with 100 mM tryptophan. Extensive vortexing and heating to 60 °C was necessary to dissolve tryptophan in the cryobuffer. The samples were cryo-cooled and shipped to the ESRF (11th November 2017).
The experiment was carried out on beamline ID30A-3 at 100 K with a $15 \times 15\ \mu m^2$ gaussian beam, 12.8 keV photon energy and a flux 1.81×10^{13} ph/s at 100 % transmission.
A dose series with 10 data sets at 1 % transmission, 10 data sets at 5 % transmission, 10 data sets at 20 % transmission and 2 data sets at 100 % transmission was collected from one crystal (32 data sets total). For later dose calculations the crystal dimensions were measured to be $50 \times 25 \times 25\ \mu m^3$. Each data set was collected with a total oscillation range of 135°, 0.15° oscillation and 0.1 s exposure time per image and 900 images total.

3.3.2 Data processing

Spot finding, indexing in $P6_522$ and integration was done using the XDS program package[69]. Pointless/Aimless from the CCP4i2 suite[70] was used for data reduction and to cut the data by $CC_{1/2} > 0.5$ followed by MOLREP[71] for molecular replacement. Two Coot[72]/REFMAC5[73] cycles were performed to obtain a final structure for each data set in the dose series. Dose calculations were performed with RADDOSE-3D[14].

A $F_o - F_o$ difference map from DS1 and DS31 was generated using the "Calculate unusual map coefficients" task in CCP4i2 with the option selected to scale them by matching the second data set to first data set.

To calculate the half life of the crystal a custom script was written (see Appendix A.3.1) which calculates the total intensity of each data set from the corresponding ASCII.HKL file and plots the values against the diffraction weighted dose in a semi logarithmic plot. This yields a linear relationship from which the slope (decay constant) was calculated and further the half life.

For comparison with the dose series, calculated electron density maps with different resolution limits were generated according to the documentation from James Holton by adapting his example script[74]. Perfect phases and amplitudes (R-factor 0.0 %) were calculated from the atomic positions of the refined model from data set 2 (R-factor 0.0 %). The resolution limits were imposed by applying an overall B-factor to the map using the empirical equation $B = 79 \times (resolution/3)^2$. This relationship seems to reflect the resolution limit to which a map, calculated with this B, can be cut-off without distortion.

3.4 X-ray crystallography with online UV/Vis spectroscopy on Lysozyme, Thaumatin and AcNiR

3.4.1 Experimental setup

Two experiments were conducted at the ESRF beamline BM30A (FIP) on the 20th of April 2018 and 25th of June 2018. An online UV/Vis-spectrometer was used to record spectral changes during irradiation with X-rays at 100 K. Both experiments were carried out with a top hat beam, 12.65 keV photon energy and 4.97×10^9 ph/s. Only beam size, crystal sizes and tryptophan concentration varied between the experiments (Table 3.4).

Table 3.4: Beam size and crystal sizes for all tested lysozyme crystals for the experiment on the 20.04.18 and 25.06.18.

Experimental parameters 20.04.18	
Beam size [μm^2]:	150 x 300
Crystal size [μm^3]:	300 x 300 x 200 Ctrl 1
	300 x 150 x 100 Ctrl 2
	300 x 300 x 200 Ctrl 3
	300 x 300 x 150 50mM Trp 1
	300 x 150 x 100 50mM Trp 2
	300 x 300 x 175 50mM Trp 3
Experimental parameters 25.06.18	
Beam size [μm^2]:	300 x 300
Crystal size [μm^3]:	200 x 150 x 150 Ctrl 1
	300 x 150 x 150 Ctrl 2
	250 x 200 x 250 Ctrl 3
	200 x 200 x 250 Ctrl 4
	200 x 100 x 100 100mM Trp 1
	200 x 150 x 150 100mM Trp 2
	300 x 200 x 100 100mM Trp 3
	300 x 150 x 200 100mM Trp 4

Each crystal was rotated in the UV/Vis beam to find the spot where the spectrum looked best. A reference spectrum was collected and the crystal dimension measured for dose calculations. The X-ray shutter was opened and the crystal exposed to X-rays while UV/Vis spectra were continuously recorded with 100 ms integration time and 20 scans averaged. The crystals were not rotated during exposure. As soon as no further spectral changes were occurring, data collection was terminated and the next crystal mounted.

3.4.2 Data processing

Doses were calculated for each crystal using RADDOSE-3D[14]. A custom script (see Appendix A.3.2) was written to extract wavelength, absorption

and average dose values (exposed region) for the first and every 75th spectrum afterwards for each dose series. The absorption values were used to calculate a set of difference spectra to the first spectrum. A low pass filter was applied to reduce the noise (Appendix A.1 and A.2).
A second script (see Appendix A.4) was used to extract the absorption values at 400 nm from each difference spectrum (dose series) with their corresponding dose value. These were normalised with respect to the highest absorption value within each dose series, averaged with corresponding time points from other crystals and standard deviations calculated for control crystals and tryptophan soaks respectively. Crystals which had very noisy data (even after filtering) were excluded from averaging. Average absorption values were plotted against average doses.
The script was adjusted for AcNiR to extract 450 nm absorption values instead.

3.5 Bioinformatic proteome survey

For the bioinformatic survey proteome data deposited in the UniProt databank[75] was used. Proteomes were selected based on two search criteria. Firstly, the proteome must be tagged with the "reference proteome" keyword meaning that it is constituting a representative cross-section of the taxonomic diversity within UniProtKB. Secondly, the completeness of genomic data in terms of expected gene content must be at least 50 % expressed by the BUSCO score (Benchmarking Universal Single-Copy Ortholog). All proteomes fulfilling these criteria were downloaded and subsequently their amino acid composition determined using a custom script (Appendix A.4.1). The dataset was then sorted by domain and a list with organism names and their respective amino acid composition created. This data was then used to generate a histogram plot and a heatmap (Appendix A.4.2).
A second dataset was created containing only a subset of 200 organisms from

each domain which was used as the input for cluster4x[76]. The data were organised in a way that cluster4x would use the occurrence of each amino acid to construct a 20-dimensional vector (each amino acid representing one dimension) for singular value decomposition.

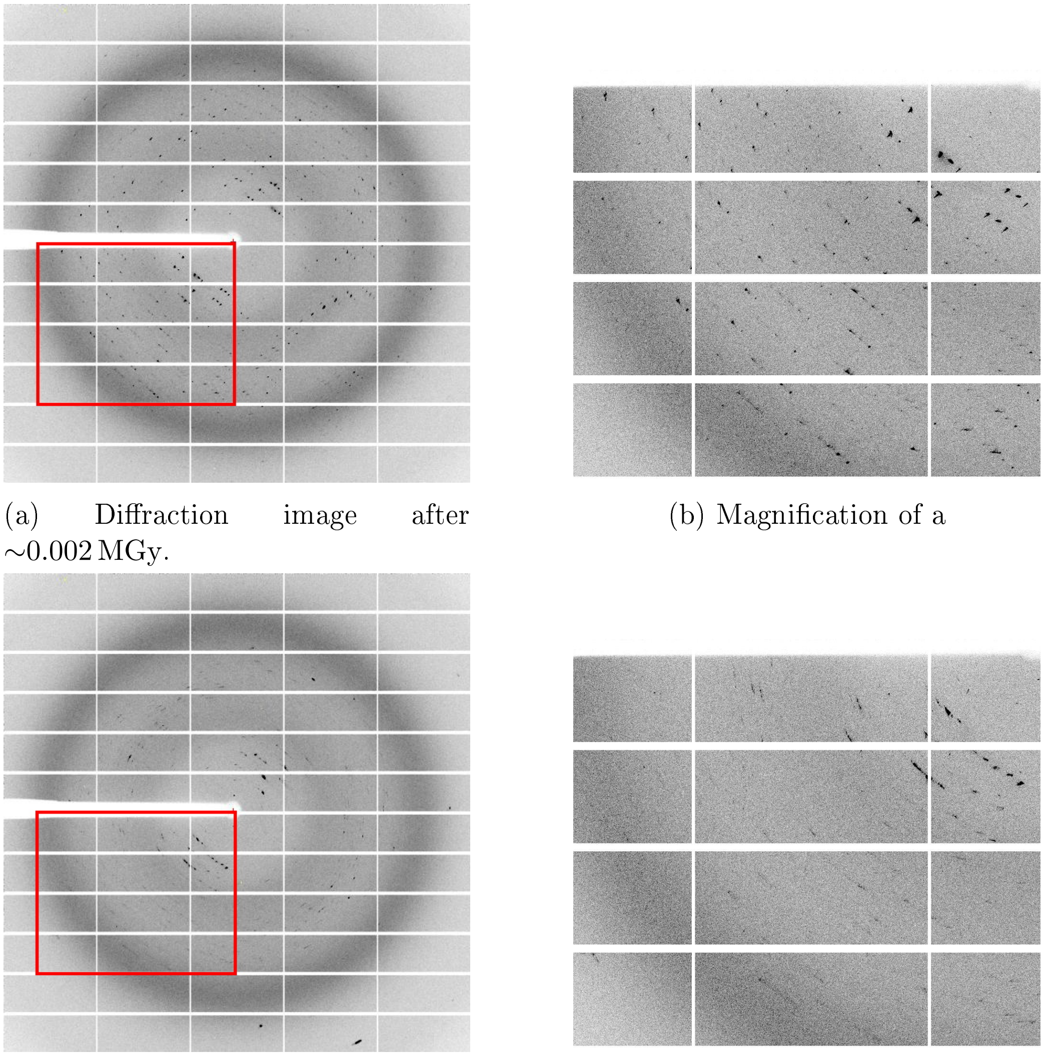
Chapter 4

Results & Discussion

4.1 How Does Tryptophan Affect Global Damage?

4.1.1 4HB1

4HB1 is a protein that lacks most radiation susceptible residues like disulphide bonds, bound metal ions or aromatic amino acids. It was hence expected that 4HB1 would show only high dose global damage and no low dose specific damage. Counter-intuitively preliminary results showed that the protein is surprisingly susceptible to radiation damage even at 100 K. Figure 4.1a shows the first diffraction image of the first data set as a reference from a native 4HB1 crystal. The diffraction spots are slightly mosaic but extend to roughly 2.3 Å. In Figure 4.1c the first diffraction image of the third data set (after exposure to 2.6 MGy) can be seen.



(a) Diffraction image after $\sim$0.002 MGy.

(b) Magnification of a

(c) Diffraction image after 2.6 MGy.

(d) Magnification of c

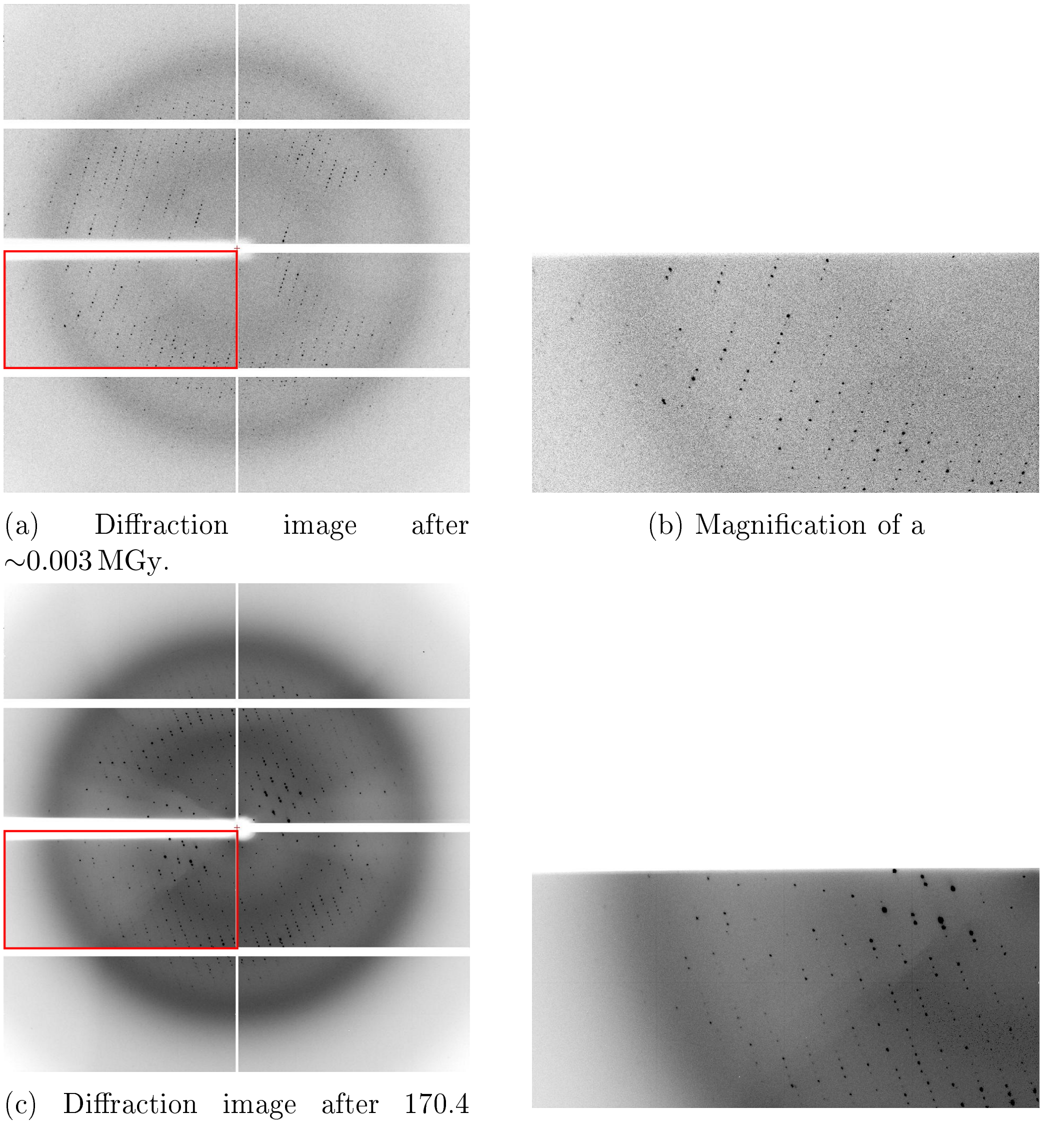
Figure 4.1: Diffraction images collected from a native 4HB1 crystal at 100 K. The red box marks the magnified area. (a) Diffraction image of the 1st data set and (b) a magnified section of it. (c) Diffraction image of the 3rd data set and (d) a magnified section of it.

When compared, it becomes apparent that the diffraction quality in the later data set has degraded considerably as a result of global radiation damage. The high resolution reflections have disappeared or became significantly weaker, the remaining reflections appear very mosaic and anisotropically distributed.

When this experiment was repeated with more crystals, all of them showed a high degree of mosaicity and limited high resolution signal (3 Å at best), indicating that the crystal lattice is not well ordered. This could be caused by weak crystal contacts of the protein molecules in between neighbouring unit cells and it raises the question whether the radiation susceptibility of 4HB1 is due its unique composition or due to a fragile crystal lattice. It also shows that either the crystallisation conditions or the cryo-protection were not sufficiently optimised resulting in poor crystal quality.

Interestingly, crystals that were soaked with varying concentrations of tryptophan (10 mM - 30 mM) showed a similar pathology. Analogous to the unsoaked crystals the degradation of the crystalline lattice occurred within a few MGy. From this it can be concluded that 4HB1 crystals are equally susceptible to global radiation damage regardless whether they had been soaked with tryptophan or not, with one exception.

One crystal soaked with 100 mM tryptophan showed good diffraction quality from the first data set onwards and maintained it surprisingly well even at high doses. This can be seen in Figure 4.2a which shows an image from the first data set and Figure 4.2c an image from the 30st data set. It can be seen that despite an absorbed diffraction weighted dose of 170.4 MGy the diffraction spots are well defined and the visible resolution has suffered only marginally, indicating that the crystal lattice is still intact. The later image appears darker because the dose series was collected with increasing photon flux each 10th data set to cover a wider dose range.



(a) Diffraction image after ~0.003 MGy.

(b) Magnification of a

(c) Diffraction image after 170.4 MGy.

(d) Magnification of c

Figure 4.2: Diffraction images collected from a 4HB1 crystal soaked with 100 mM tryptophan at 100 K. The red box marks the magnified area. (a) Diffraction image of the 1st data set and (b) a magnified section of it. (c) Diffraction image of the 30st data set and (d) a magnified section of it.

The radiation robustness against global damage was quantified by collecting successive data sets over the same region of reciprocal space and tracking the total intensity I_{tot} after each data set. To allow easy comparison between different crystals, the summed intensity was normalized to 1.0 for the first data set $I_{tot} = I_D/I_0$ and plotted against the diffraction weighted dose. To describe the exponential decay of the crystals diffracting power equation 4.1 can be used, where N is the intensity, t is the dose and λ is the decay rate constant.

$$\frac{dN}{dt} = -\lambda N \tag{4.1}$$

Separation of variables yields

$$\frac{dN}{N} = -\lambda dt \tag{4.2}$$

and after integration

$$lnN = -\lambda t + C \tag{4.3}$$

This linear relationship was used to fit the data in a semi-logarithmic plot, where the decay rate constant λ could be extracted as the slope of the fitting curve and the intensity at zero dose C as the intercept with the y-axis. Rearranging equation 4.3 and evaluating C at $t = 0$ leads to equation 4.5 which was used to fit the data on a linear scale.

$$N = e^C e^{-\lambda t} \tag{4.4}$$

$$N(t) = N_0 \times e^{-t\lambda} \tag{4.5}$$

The half-life $t_{1/2} = \frac{ln(2)}{\lambda}$ (here half-dose $D_{1/2}$) is the dose required for the diffraction intensities to fall off to 50 % of its initial value. This metric was used to measure the radiation robustness of a crystal. A plot showing the exponential diffraction decay of a 4HB1 crystal soaked with 100 mM tryptophan can be seen in Figure 4.3. The total intensity follows an exponential decay as the dose increases. Overall, this crystal survived the collection of 30 data sets (as determined by the last indexable dataset) with a total absorbed dose of 182.3 MGy and a half-dose of 64.43 MGy.

Compared to all other tested 4HB1 crystals (soaked and apo), which were considered "dead" after approximately 2.6 MGy, the 100 mM tryptophan soaked crystal was able to tolerate a roughly 70-fold higher dose before diffraction quality degraded too much to be processable. Unfortunately the exact conditions which yielded this result were not reproducible.

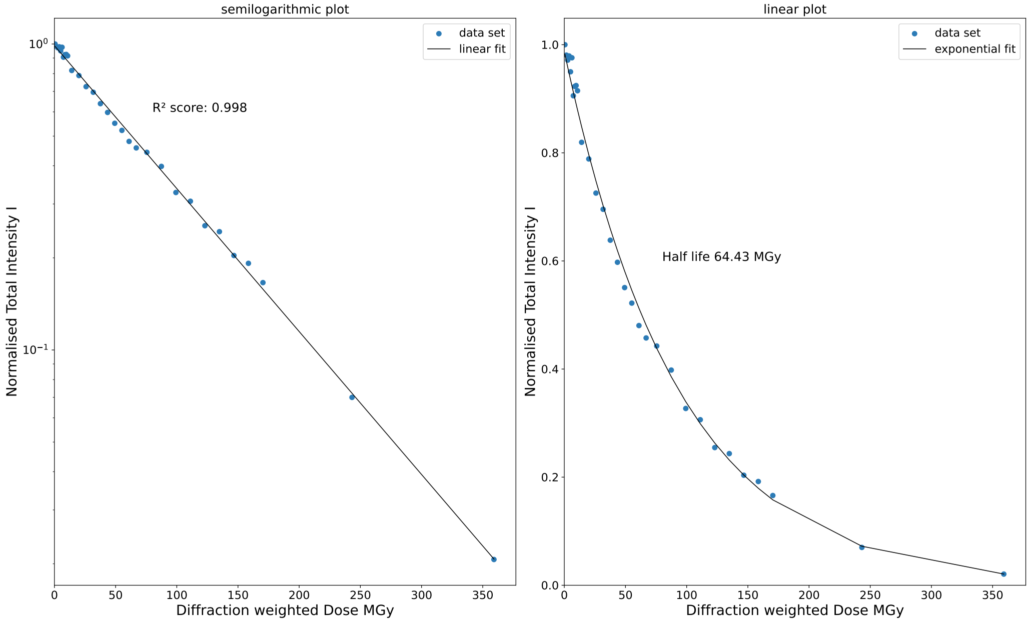


Figure 4.3: Diffraction decay of a 4HB1 crystal soaked with 100 mM tryptophan at 100 K on a semi-logarithmic scale (left) and a linear scale (right).

4.1.2 Lysozyme

To assess what impact tryptophan has on other proteins than 4HB1 lysozyme was chosen as a second test system. The level of global damage was measured for multiple native and tryptophan soaked lysozyme crystals by using the half-dose $D_{1/2}$ as a metric. Figure 4.4 and 4.5 showing the diffraction decay with increasing dose exemplary for one native and one soaked lysozyme crystal, respectively at 100 K.

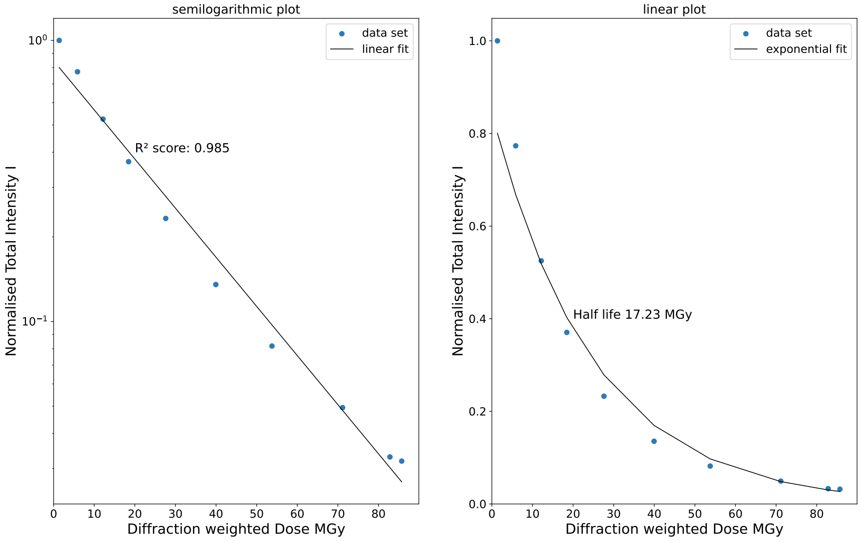


Figure 4.4: Diffraction decay of a native lysozyme crystal on a semi-logarithmic scale (left) and a linear scale (right).

It can be seen that the total intensity of both crystals follows an exponential decay with good agreement. Furthermore, it can be seen that the $D_{1/2}$ for the tryptophan soaked crystal is 2.82 MGy (16.4 %) smaller than for the native crystal, indicating an even lower radiation robustness than the native crystal.

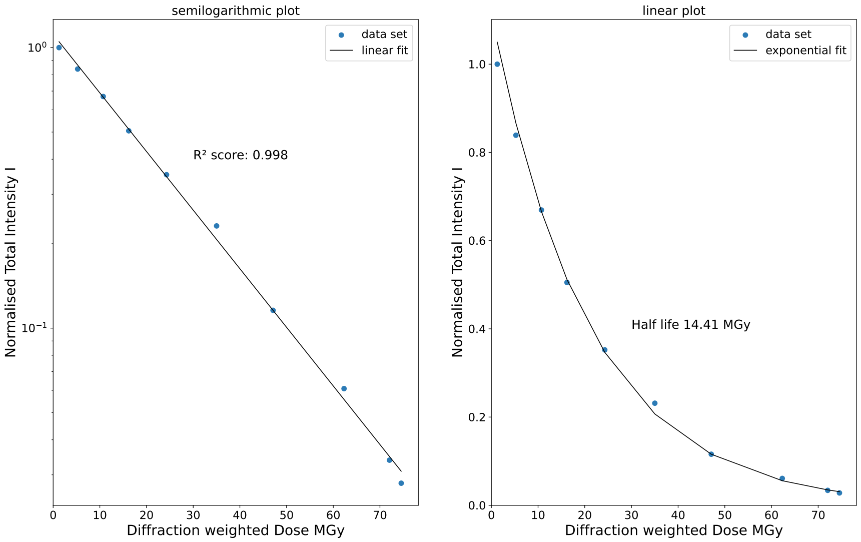


Figure 4.5: Diffraction decay of a lysozyme crystal soaked with 20 mM tryptophan at 100 K on a semi-logarithmic scale (left) and a linear scale (right).

Crystal-to-crystal variations can influence the results, which is why drawing conclusions from a single observation should be avoided. To gain more statistical confidence in the result the experiment was repeated with more crystals and the $D_{1/2}$ values for native and soaked crystals are reported in Table 4.1.

Table 4.1: Half-dose values of four native lysozyme crystals and three lysozyme crystals soaked with 20 mM tryptophan at 100 K.

Half-dose $D_{1/2}$ native crystals [MGy]	Half-dose $D_{1/2}$ of tryptophan soaked crystals [MGy]
13.79	14.41
17.23	15.74
13.68	12.77
13.49	

Two observations can be made. Firstly, all tested lysozyme crystals reach the half-dose far below the Garman limit of 30 MGy. Nevertheless, the results agree well with previous studies which report a lower limit for $D_{1/2}$ of 10 MGy for lysozyme[33] and 13 MGy for myrosinase crystals [17].

Secondly, the individual $D_{1/2}$ values all lie close together with an average $D_{1/2}$ for native lysozyme crystals of 14.54 ± 1.79 MGy and 14.30 ± 1.49 MGy for tryptophan soaked crystals. This suggests that tryptophan has no significant effect on the radiation resistance of lysozyme crystals against global damage.

4.2 How Does Tryptophan Affect Specific Damage?

4.2.1 4HB1

Specific damage to 4HB1 is expected to be minimal as it only has eight glutamates which are susceptible to decarboxylation. To assess the extent of specific damage a $F_o - F_o$ difference map between the 1st and the 30th data set was generated. It shows negative difference electron density centred around many oxygen atoms of the protein (Figure 4.6, red mesh). Negative difference electron density usually indicates the loss of electrons and therefore radiation damage. However, it should be noted that this can also be an artefact of the map calculation or scaling procedure. Since the intensities of the high dose data set decreased by over 80 % relative to their initial values (reference Figure 4.3), the associated structure factors are likely affected by large systematic errors such as significantly worse signal to noise ratio, which lead to an underestimation of their true value. It is therefore more likely that the observed negative difference density is a consequence of large systematic errors associated with the extremely high dose of 182.3 MGy in data set 30 rather than an indicator of radiation damage.

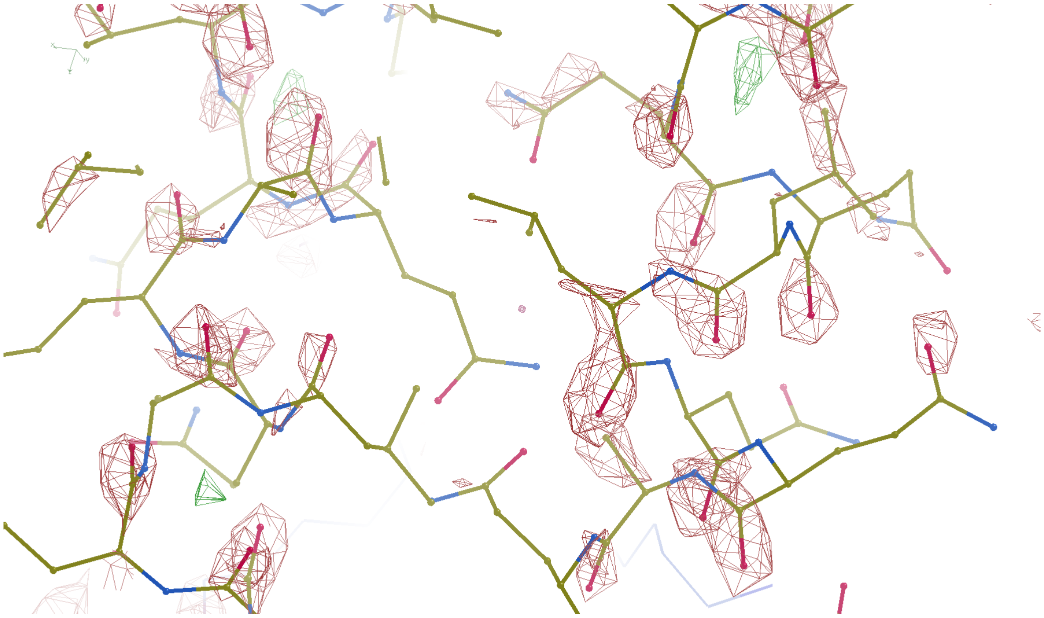
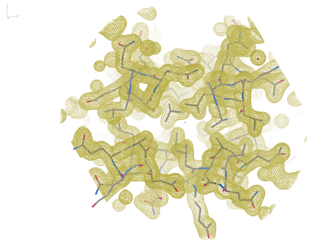
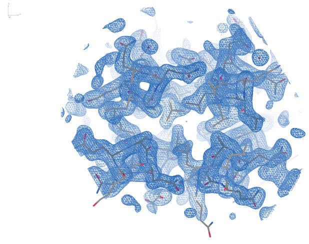
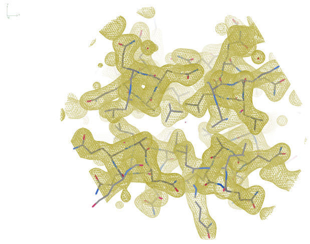
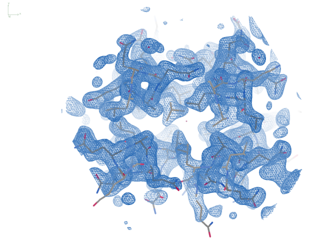
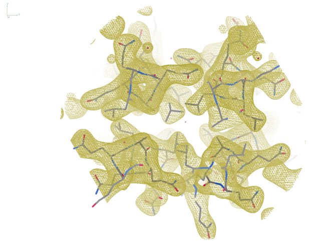
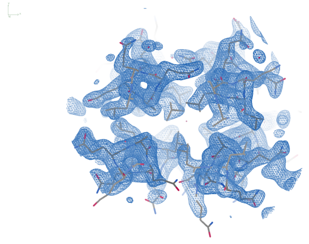
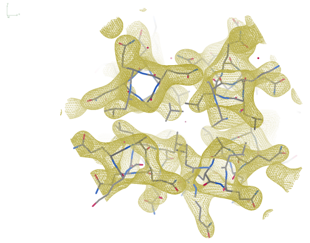
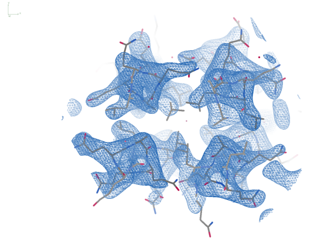


Figure 4.6: $F_{o(D=182.3\,\mathrm{MGy})} - F_{o(D=0.5\,\mathrm{MGy})}$ difference map contoured at 3 rmsd of a 4HB1 crystal soaked with 100 mM tryptophan at 100 K.

Choosing a different data set with a lower dose (smaller systematic errors) produced a difference map with no negative difference density, e.g. no visible specific damage. To get an understanding for the quality of the electron density map at different doses, the experimental maps were compared with calculated perfect maps (Table 4.2). The first data set matches the appearance of a perfect map with an imposed resolution cut-off of 2 Å. With increasing dose the electron density map progressively loses its definition, as expected. And yet, after an absorbed dose of 182.3 MGy, the electron density still matches the appearance of a 3.5 Å calculated map. That the map quality is preserved, despite the high dose, may be explained by phase bias considering the relative contribution of the phase information (coming from the model) to the map in comparison to the much lower quality intensities (coming from the data).

Table 4.2: Comparison of calculated $2F_c - F_c$ perfect maps (yellow) with experimentally determined $2mF_o{-}DF_c$ electron density maps (blue) rendered at 1 rmsd.

Calculated map	Experimentally determined map
 2 Å map	 DS1 (0.5 MGy)
 2.5 Å map	 DS15 (37.6 MGy)
 3 Å map	 DS20 (66.8 MGy)
 3.5 Å map	 DS30 (182.3 MGy)

4.2.2 Lysozyme

To assess whether tryptophan slows down specific damage in lysozyme, the reduction of its four disulphide bonds was monitored spectroscopically during irradiation, specifically the absorption increase at 400 nm caused by the formation of disulphide-anion radicals.

A possible way to interpret the data is to consider that the characteristic transient arose as a results of equation 2.4, where the number of electrons e^- is proportional to the dose. At low doses, the number of electrons is small compared to the number of disulphide bonds and hence the absorption increases rapidly and approximately linearly as all disulphides become ionised. With increasing dose the number of disulphides that have not yet reacted will decrease, causing the absorption increase to decelerate. Eventually the number of electrons becomes larger than the number of disulphide bonds, so that everything that can react has done so. At that point, an increase in dose will not cause an increase in absorption and consequently leads to a saturation. The data presented in Figure 4.7 shows the average absorption increase at 400 nm of native and tryptophan soaked crystals for two separate experiments.

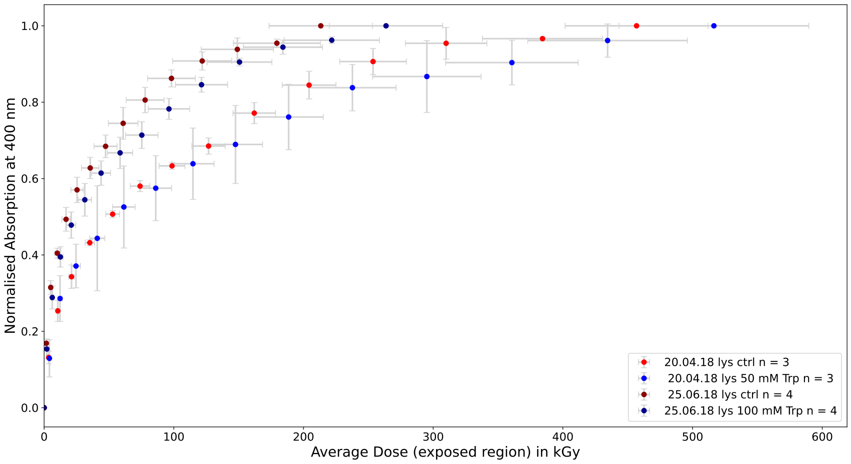


Figure 4.7: Average UV/Vis absorption at 400 nm of lysozyme crystals with increasing dose of two separate measurements. n denotes the number of averaged crystals for each group.

The last data point in each dose series was normalised to 1.0 to allow comparison between different measurements. This reveals that there is a systematic difference between dose series collected on different days but with almost identical experimental parameters. If these differences were solely apparent in the dose series of the tryptophan soaked crystals this differences could be explained by an concentration dependent effect of tryptophan. However, also the control dose series match the difference. A closer look at the experimental parameters reveals that for the earlier experiment the beam size was smaller than the crystal sizes whereas in the later experiment the beam size was in all cases similar in size or larger than the crystal (see Table 3.4). What can be concluded is that experimental results that were produced under otherwise identical parameters deviate due to beam-to-crystal ratio differences. Consequently, great care must been taken when attempting to average results from experiments that were produced under seemingly identical conditions.

Looking at any of the two experiments it is also clear that there is no significant difference between native and soaked crystals. If tryptophan has a protective effect, a slower onset of specific damage is expected. Although very slightly, such a trend can be seen for the measurement from the 25.06.2018. However, the difference between native and soaked crystals is well within the error and therefore not significant. The increasing error, particularly in the dose estimation, is a result of varying average doses of the individual crystals caused by different crystal sizes.

4.2.3 Thaumatin

Specific damage to disulphide bonds was monitored in thaumatin crystals analogously to lysozyme. The average absorption of the tryptophan soaked crystals raises slightly slower than for the native crystals, indicating a small benefit when tryptophan is added (Figure 4.8).

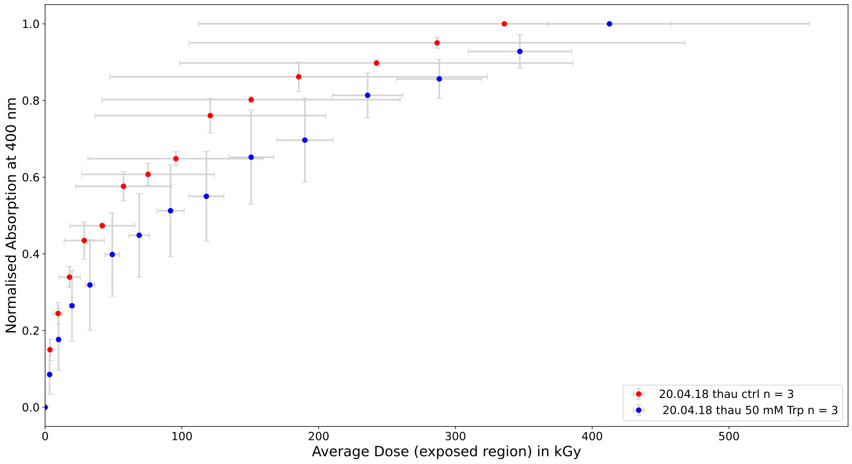


Figure 4.8: Average UV/Vis absorption at 400 nm of thaumatin crystals with increasing dose.

However, the effect is smaller than a factor of two and the large dose errors, originating from substantially different crystal sizes, make it difficult to draw conclusions with confidence.

4.2.4 AcNiR

The protection of metal centres is particularly difficult due to their high susceptibility. AcNiR, a protein with no disulphide bonds but two bound copper ions, was used to test tryptophan's effectiveness against specific damage to metal sites. Figure 4.9 shows the average change of absorption at 450 nm with increasing dose of native and tryptophan soaked crystals.

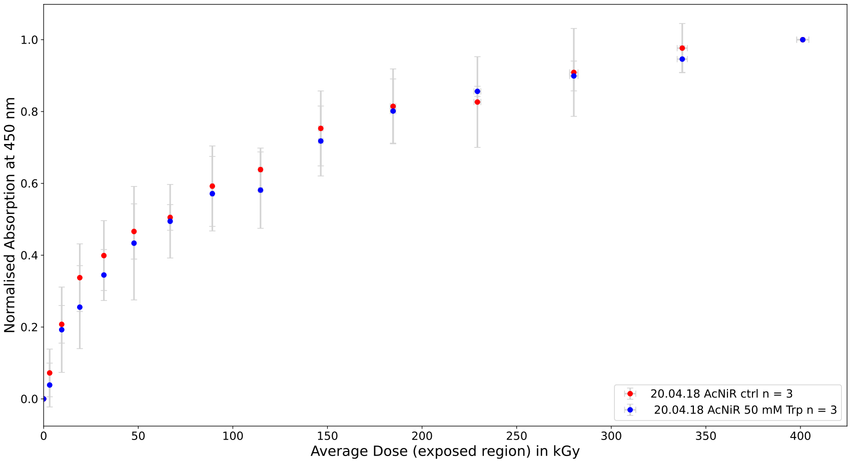


Figure 4.9: Average UV/Vis absorption at 450 nm of AcNiR crystals with increasing dose.

The dose error for this measurement is small due to comparable crystal sizes. Both samples behave the same way and take damage equally fast, indicating that 50 mM tryptophan does not prevent or slow down the reduction of metal sites due to radiation damage.

4.3 Re-evaluation of Preliminary Results

4.3.1 Spectroscopy Lysozyme

In contrast to the spectroscopic data presented so far the preliminary data measured before this work clearly showed a protective effect when tryptophan is soaked into the crystal (Figure 1.2). The graph was produced by normalising all data points with respect to highest value. However, this is not a suitable normalisation if the crystals are to be compared as crystal-to-crystal variations will affect the results. In fact, upon reevaluation of the preliminary data it became clear that the absorption values at 400 nm are affected by and correspond to the individual crystal dimension (Table 4.3).

Table 4.3: Crystal dimensions of used lysozyme crystals.

	x [µm]	y [µm]	z [µm]
100 mM Trp	100	150	100
50 mM Trp	100	150	100
Apo 4HB1	250	250	250

According to Beer-Lambert law the absorption is proportional to the concentration C, the path length l and the molar extinction coefficient ϵ (Equation 4.6).

$$A = C \cdot l \cdot \epsilon \tag{4.6}$$

In an online UV/Vis setup the path length l is a fixed distanced partly occupied by the crystal, the surrounding liquid and air. Assuming that the absorbing species is the part of the crystal, a larger crystal will effectively increase the concentration along the measured path length and lead to a higher absorption value. This can be corrected for if the exact shape and orientation of the crystal with respect to the spectrometer is known. However, these parameters are at best estimates because the crystal alignment with spectrometer and X-ray beam is very finicky and never perfect and the crystal shape is commonly approximated as a simple box.

Therefore, to avoid the impact of crystal-to-crystal variations each data point is instead normalised to the highest value within that series. Every other spectroscopic data presented so far has been normalised that way. When applied to the preliminary data this yields Figure 4.10 in which the apparent difference between control and soaked crystals disappears.

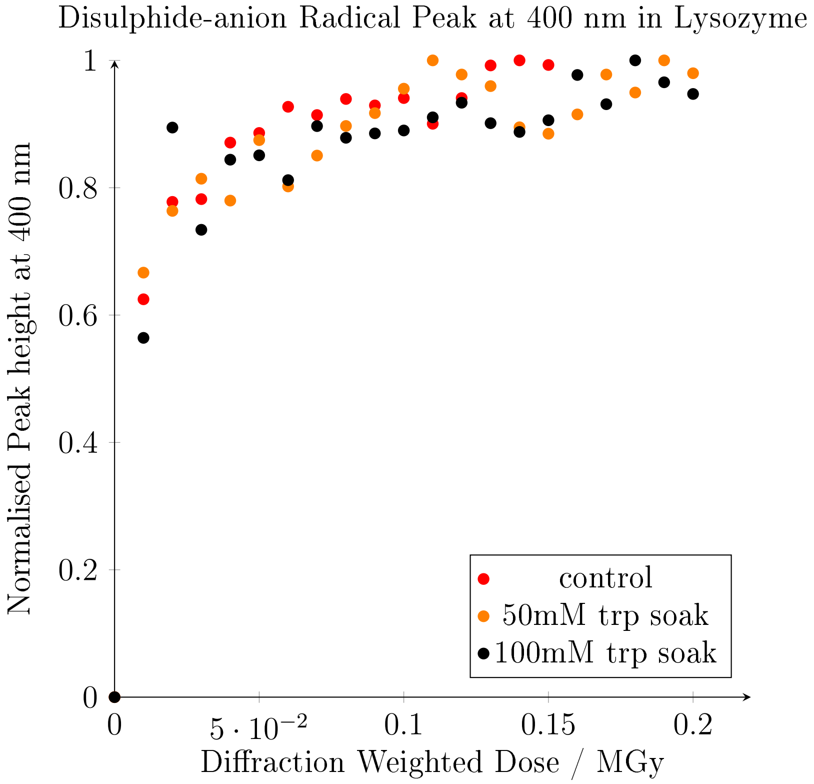


Figure 4.10: Evolution of the disulphide-anion radical peak at 400 nm in Lysozyme during X-ray irradiation. Collected at FIP 14.12.17.

With the corrected normalisation the measurements are in agreement with the other data, indicating that the addition of tryptophan has no protective effect against radiation induced damage to disulphide bond.

4.4 Proteome Survey

4.4.1 Radiation Resistant Bacteria

To determine whether there is evolutionary evidence that the inherent radiation resistance of proteins is correlated with unusually high or low occurrences of one amino acid or a set of amino acids within a proteome, the amino acid distribution of over 10.000 proteomes was examined. In this regard the domain of bacteria and archaea are particularly interesting because the largest number of known radiation resistant organisms belong to them. Among the three domains archaea are the least well studied and from the 10.000 reference proteomes only 200 belong to archaea. Therefore the analysis focused on 8000 bacteria proteomes for better statistical significance. Firstly, histograms for the occurrence of each amino acid were extracted from the proteome data.

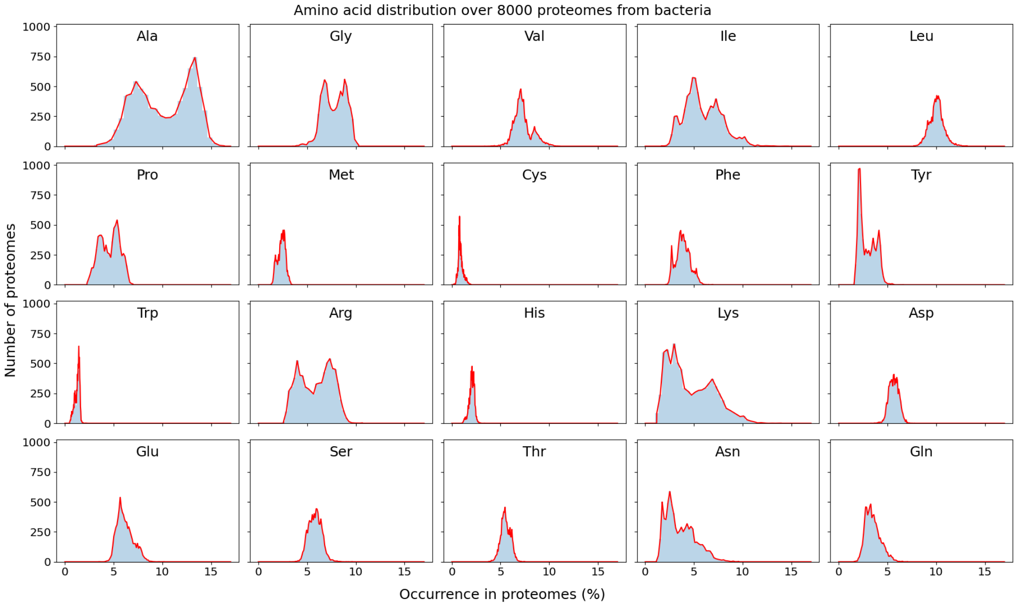


Figure 4.11

A constant bin size according to Knuth's rule was chosen to avoid loss of the fine structure (not enough bins) or that heights of individual bins are affected by sampling errors (too many bins). Knuth's rule minimizes the error of the histogram's approximation to the data by maximising the optimal bin size M of the costfunction

$$F(M|x, I) = n\log(M) + \log\Gamma(\frac{M}{2}) - M\log\Gamma(\frac{1}{2}) - \log\Gamma(\frac{2n+M}{2}) + \sum_{k=1}^{M}\log\Gamma(n_k + \frac{1}{2}) \tag{4.7}$$

where Γ is the Gamma function, n is the number of data points and n_k is the number of measurements in a bin[77]. Secondly, all histograms were combined in form of a heatmap (Figure 4.12), where the y-axis shows the occurrence of a particular amino acid (in %) within a proteome and the colour indicates the frequency for that percentage to appear within the 8000 proteomes.

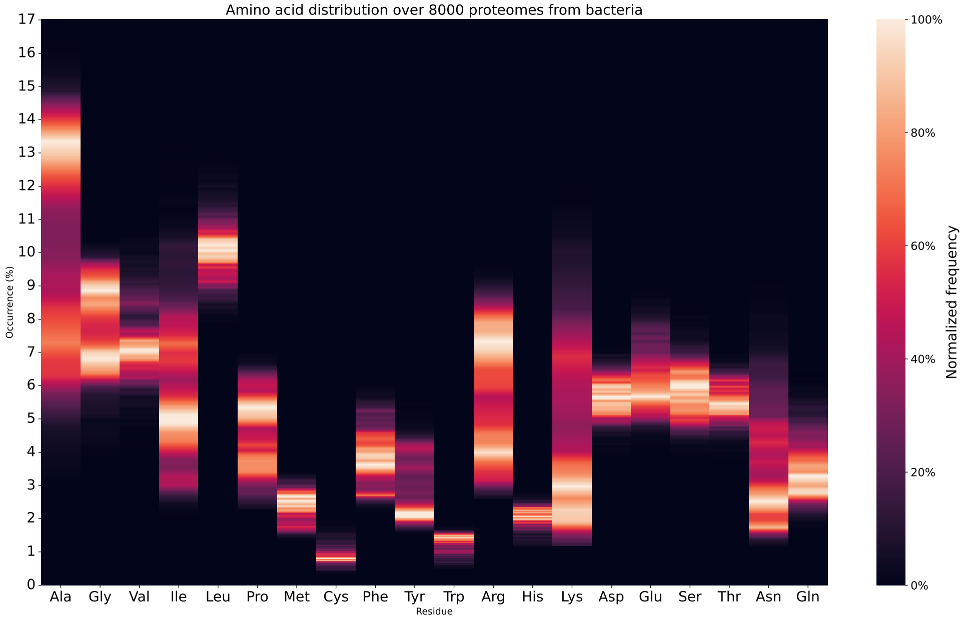


Figure 4.12: Amino acid distribution of 8000 bacteria proteomes.

From this plot the distribution of each amino acid can be analysed individually or directly compared with that of another amino acid. For example, glycine has a maximum occurrence around 9 % and 6.8 %. Also it seems that bacteria are quite flexible in the usage of some amino acids (e.g. alanine) as the ranges are relatively broad compared to for example cysteine or tryptophan.
However, the true power of this graph is that it provides a background against which the amino acid distribution of radiation resistant bacteria can be plotted. Any deviation from this background may indicate an evolutionary adaptation linked to radiation resistance, if it were to appear in multiple radiation resistant organisms. Table 4.4 gives a selection of radiation resistant bacteria that where used for this comparison.

Table 4.4: Radioresistant bacterial organisms with their corresponding D_{10} (dose at which only 10 % of a population survives). Note that D_{10} values vary significantly depending on the growth conditions[78].

Species	Phylum	Class	Average D_{10} Dose
Rubrobacter radiotolerans	Actinobacteria	Rubrobacteria	12 kGy
Deinococcus radiodurans	Deinococcus Thermus	Deinococci	10 kGy
Geodermatophilus obscurus	Actinobacteria	Actinobacteria	9 kGy
Modestobacter marinus	Actinobacteria	Actinobacteria	6 kGy

The amino acid distribution of their proteomes was determined and plotted on top of the background to find any common deviation (Figure 4.13).

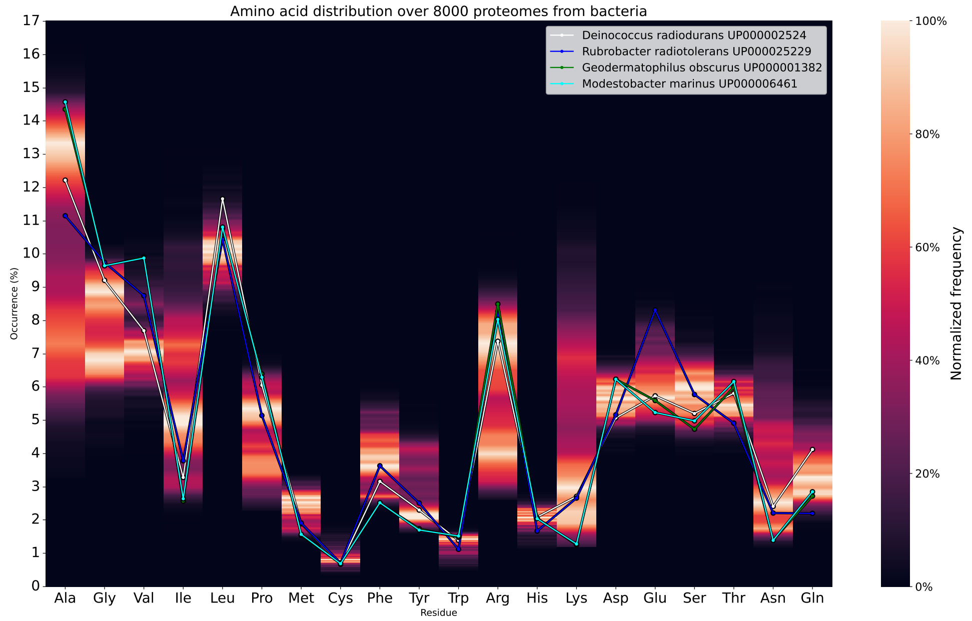
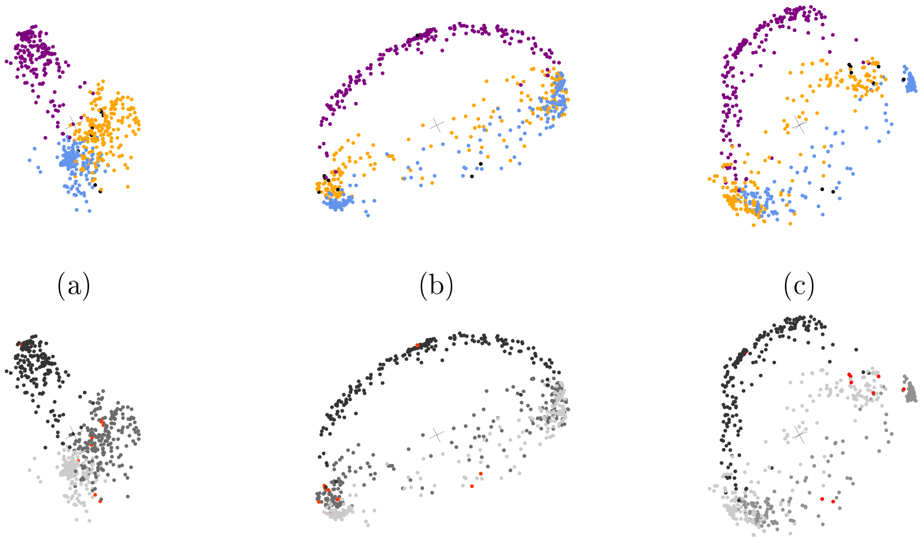


Figure 4.13: Amino acid distribution of 8000 bacteria proteomes overlaid with the distribution of four radiation resistant bacteria.

Because they belong to the same phylogenetic family, *Modestobacter marinus* and *Geodermatophilus obscurus* display a similar amino acid distribution except for glutamate, serine and arginine. Overall, it appears that there is no obvious significant deviation that all radiation resistant bacteria have in common. Also, for the sampled species there is no correlation between the magnitude of deviation from the typical amino acid distribution and their radiation hardness measured by D_{10}.

4.4.2 Clustering

Rather than interpreting the data by looking for deviations from the optimal amino acids distribution, a different approach was also taken. Here, the amino acid composition of 200 proteomes from each domain were taken and clustered using cluster4x[76]. Cluster4x was originally designed to handle multi-data-set diffraction data and cluster them by similarity, but the underlying algorithms work perfectly for proteome data as well. The resulting plot is shown in Figure 4.14. Eukaryotic proteomes (green), bacterial proteomes (orange) and archaea (blue) form distinct clusters (Figure 4.14a).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.14: Singular value decomposition of amino acid compositions plotted along arbitrary axes. Each dot represents one proteome with green being proteomes from eukaryotic organisms, orange bacteria, blue archaea and black being radiation resistant organisms. (a-c) shows the same data from different perspectives. (d-f) highlights radiation resistant proteomes in red.

Evaluating the same plot along a different axis (4.14b) reveals that even within a cluster there is a separation which can be attributed to amino acid differences between different phyla. For example, the eukaryotic cluster can be subdivided into regions containing only proteomes from crustaceans, mosses, vertebrates, etc. However, radiation resistant organisms (highlighted red in 4.14e - 4.14f) do not form a distinct cluster. Instead, they sit within their respective domain e.g. tardigrades within the eukaryotic domain, *Thermococcus radiotolerans* and *Thermococcus gammatolerans* within the archaea

domain and within the bacteria domain the radiation resistant organisms mentioned in Table 4.4.
This shows, that the amino acid composition holds the necessary information to determine its phylogenetic filiation down to the phylum level but it can not predict the phenotype of extreme radiation resistance.

Chapter 5

Conclusion

From the data presented it can be concluded that the amino acid composition (primary sequence) is not the origin of a proteins inherent radiation resistance. In the context of an X-ray diffraction experiment this means that supplementing a protein crystal with free amino acids (here exclusively shown for tryptophan) does not improve the radiation robustness of the crystal. Specifically, the protective effect on copper type I in AcNiR crystals was determined by online UV/Vis spectroscopy. Metal centres are the most susceptible motive in a protein structure and the data presented here showed that soaking tryptophan into AcNiR crystals did not protect the bound copper ions against specific damage (Figure 4.9). Lysozyme and thaumatin were subjected to X-ray irradiation to determine whether tryptophan can protect disulphide bonds against specific damage but no significant effect on the rate of specific damage compared to apo crystals was found (Figure 4.7 and 4.8). Global radiation damage was assessed by measuring the total intensity I_{tot} after each successive data set. For apo lysozyme crystals this resulted in an average half dose of 14.54 ± 1.79 MGy and for tryptophan soaked crystals in an average half dose of 14.30 ± 1.49 MGy, showing that there is no significant radiation protecting effect (Table 4.1).

Analysing the radiation robustness of the four helix bundle protein it was found that the protein is surprisingly susceptible regardless of whether the crystal had been soaked with tryptophan or not. However, all crystals showed a high degree of mosacity and limited high resolution signal (3 Å at best) to begin with. It is therefore unclear whether the radiation susceptibility is caused by a fragile lattice or its unique amino acid composition.

Interestingly, one 4HB1 crystal that had been soaked with 100 mM tryptophan survived a total absorbed dose of 182.3 MGy (half dose of 64.43 MGy, Figure 4.3). In comparison to all other tested 4HB1 crystal (soaked or apo) this result shows an improvement by a factor of approximately 70 with respect to total absorbed dose. Analysis of a difference electron density map revealed that at this dose the protein molecules suffered from severe backbone fragmentation (Figure 4.6) and yet producing a processable diffraction pattern. Although the exact conditions leading to this result were not reproducible, this result constitutes an unprecedented case of extreme radiation tolerance in an intense X-ray beam and was hence included in the analysis.

To summarise, testing different protein systems, tryptophan has no significant effect (threshold of a factor 2 improvement as suggested by James Holton [48]) on global or specific damage rates (with one not reproducible exception).

A complementary bioinformatics study was performed and the amino acid composition of 8000 bacteria proteomes were analysed (Figure 4.13). The amino acid distribution of radiation resistant organisms showed no common bias towards a particular amino acid or combination of amino acids. Clustering a subset of 200 proteomes from each domain showed that the phylogenetic filiation can be predicted from the amino acid composition, the phenomenon of extreme radiation resistance however can not be predicted (Figure 4.14). This result supports the conclusion that there is no single amino acid or combination thereof that are the source of a proteins radiation resistance.

Part II

Oxidation of SARS-CoV-2 M^{pro}

Chapter 6

Introduction

In December 2019 a new coronavirus (SARS-CoV-2) emerged in Wuhan China and caused a global pandemic which had drastic consequences on all aspects of life. In response the scientific community reacted in an outstandingly rapid and effective way pushing the development of suitable therapeutics and vaccines.

While other coronaviruses that are pathogenic to humans cause only mild clinical symptoms there are two notable exceptions in recent history: the severe respiratory syndrome coronavirus (SARS-CoV) in 2002 with 8098 global cases (9.6% fatality rate)[79, 80] and the Middle East respiratory syndrome coronavirus (MERS-CoV) in 2017 with 1,493 global cases (more than 35% fatality rate)[81]. In contrast, SARS-CoV-2 is not particularly lethal but its infectiousness surpassed its predecessors accounting for 172 million infections worldwide 1.5 years after the outbreak[82].

Due to the urgent need for protection measures against SARS-CoV-2 vaccine development processes were sped up where possible resulting in the first approved vaccine in the EU only one year after the first reported case[83]. Typically, the development of a new vaccines can take up to 15 years[84] which raised concerns regarding safety and efficacy as well as over public acceptance of the new SARS-CoV-2 vaccines. On the other hand, the global vaccination efforts had critical impact on the Covid-19 pandemic, reducing the number of severe cases, hospitalisations and deaths. [85, 86]

Similarly fast, drug screenings against other key viral proteins such as the the main protease (M^{pro})[87–91] and the papain-like protease (PL^{pro})[92–95] were conducted, resulting in thousands of articles reported in the first few month[96]. While the rapid publication of many studies ensured that new evidence is shared in a timely manner, which is particularly important during a worldwide health crisis, researchers also raised concerns regarding the accelerated pace of COVID-19 publishing. Many studies and trials had poor quality, were too small or poorly designed to be helpful, merely adding to the COVID-19 noise[97–99]. Additionally, despite the best efforts of researchers mistakes may happen when research progresses at such breath-taking speed and subtle details may be overlooked.

One such detail is an active site modification of M^{pro} which was first addressed by *Kneller and co-workers*[100] and identified as a peroxysulfenic cysteine modification that occurred at physiological pH. They hypothesised that the active site thiolate reacted with ambient oxygen when crystals were allowed to grow for a long time. Similar oxidative modifications ($-SO_2^-$, $-SO_3^-$) have been observed for transmissible gastroenteritis coronavirus (TGEV) M^{pro}[101] and are considered to inactivate the enzyme. In general M^{pro} has an unusually high number of cysteine residues for a viral protein. Some studies suggest that these cysteines are part of a redox regulation mechanism via N-O-S/S-O-N-O-S bridge formation [102] and others hypothesise that the oxidation of surface cysteines potentially protects the active site from oxidative damage[103]. While it is not clear what the specific biological implications of these modifications are they may have significant impact on the ongoing drug discovery efforts.

6.1 Objectives

Despite the fact that M^{pro} is expressed *in vivo* in the cytosol, which is considered to be of reducing nature, many published structures show signs of oxidation at the active site cysteine Cys145. It is therefore unclear whether this modification is an artifact of the *in vitro* purification procedure or whether this is a natural state of the enzyme. Oxidation of active site cysteines is not uncommon due to their high reactivity and have been shown for other cysteine proteases such as papain[104] or TGEV M^{pro}[101]). How such a modification would affect the enzyme's activity and recent active-site drug screening efforts is also an open question.

To answer this question, this study determined the level of M^{pro} oxidation when purified in accordance with most published protocols using a 1 mM concentration of reducing agent throughout the purification by X-ray crystallography and mass spectrometry. The results were compared to M^{pro} purified under (1) anaerobic conditions and (2) conditions where no reducing agents were used. Lastly, the impact of potential oxidative modifications on enzyme activity for all three M^{pro} samples was determined using a tetra-peptide substrate containing a fluorescent and UV/Vis-active tag.

Chapter 7

Theory & Background

7.1 Structure of SARS-CoV-2 Virus Particle

SARS-CoV-2 is an enveloped, positive-sense, single-stranded RNA virus of the genus Betacoronavirus. The surface of the virus is covered with membrane (M), envelope (E) and spike (S) proteins (Figure 7.1). The envelope and membrane proteins are small membrane proteins that are essential for virus assembly and budding. whereas the spike protein mediates receptor binding and the membrane fusion process. Spike protein mutations occur frequently, which may increase both binding to ACE2 receptors and entry efficiency[105]. Each virus particle contains the condensed viral RNA genome which is protected by nucleocapsid proteins (N).

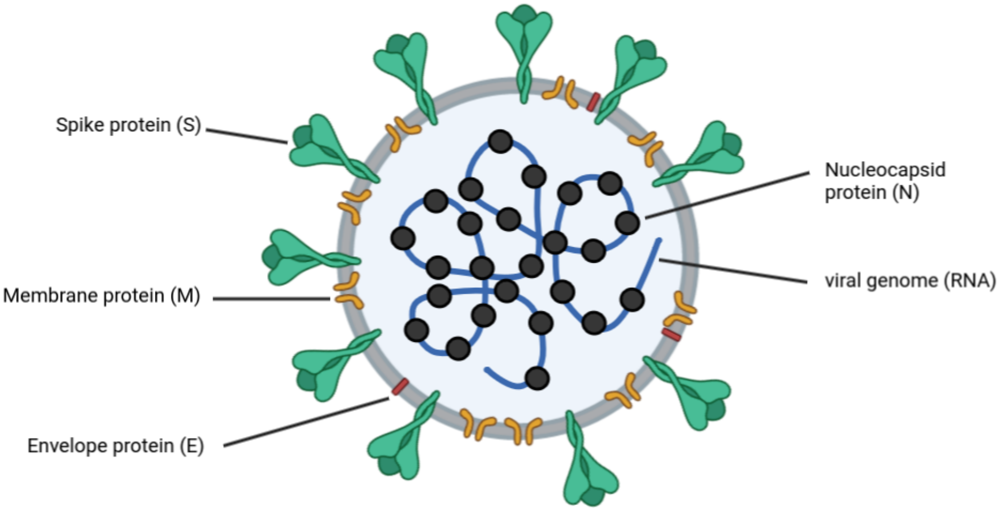


Figure 7.1: Schematic representation of a SARS-CoV-2 virus particle. Created with BioRender[106].

7.1.1 The Viral Life Cycle

The first step of the viral infection is the entry of a virus particle into the host cell. Two complementary pathways, endocytosis or fusion, are likely. In both cases the cell entry of SARS-CoV-2 is mediated by the glycoprotein spike which binds to the membrane bound angiotensin-converting enzyme 2 (ACE2) receptor on the host cell. Upon receptor engagement spike reveals a previously hidden cleavage site. The transmembrane protease serine 2 (TMPRSS2)[107, 108] on the cell surface or cathepsin L in the endosomal compartment, cleaves spike into S1 (extracellular) and S2 (transmembrane) subunits[109]. Subsequently, the ACE2-S1 fusion protein is released allowing the S2 subunit to undergo dramatic conformational changes, anchoring itself to the host cell membrane and pulling the virus and cell membrane together, initiating fusion pore formation. The virus then releases its genomic RNA into the cytosol of the host cell for viral RNA and protein synthesis.

The viral RNA is approximately 30 kb long. The 3'-third of the genome encodes the viral envelope proteins including the membrane (M), envelope (E) and spike (S) protein which contain N-terminal signal sequences for ER translocation and the nucleocapsid protein (N) which is in contrast to the others translated on free ribosomes. Additionally, accessory proteins which are not essential for virus replication but have a role in pathogenesis[110] are also encoded on this genome part. The other two thirds of the genome code for non-structural proteins (Nsps) which are not included in the virus particle but are vital for RNA-replication and viral proliferation[111]. Nsp1-16 are expressed as two polyproteins pp1a and pp1ab. The mature and functionally active Nsps are released from the polyprotein only after proteolytic processing by the two viral proteases M^{pro} (Nsp5) and PL^{pro} (Nsp3). M^{pro} cleaves the viral pp1ab polyprotein at 11 sites, while the PL^{pro} cleaves at 3 sites, generating functional non-structural proteins which form the essential replicase-transcriptase complex for RNA-synthesis.

Only three hours post-infection intracellular membranes are modified to form double-membrane vesicles (DMV) in which RNA-synthesis takes place[112, 113]. Full length genomic RNA copies and subgenomic strands are produced through discontinuous transcription. Positive mRNA strands are then exported into the cytosol for translation to viral proteins. The assembly of new virions is initiated when viral RNA coated with nucleocapsid proteins bud into the endoplasmic reticulum Golgi intermediate compartment (ERGIC). The structural proteins (M, E, S) located on the ERGIC membranes are incorporated into the envelope as the virion forms. Finally, progeny virus particles are released from the cell via exocytosis.

7.2 SARS-CoV-2 Main Protease (M^{pro})

SARS-CoV-2's main protease (M^{pro}) belongs to the class of cysteine proteases. Its name "main protease" is indicative of its crucial role during proteolytic processing of the polyproteins which enables the viral replication and transcription machinery. It is also referred to as chymotrypsin-like protease ($3CL^{pro}$) due to its similarity in substrate specificities and core structural homology with the 3C proteases seen in picornaviruses[114].

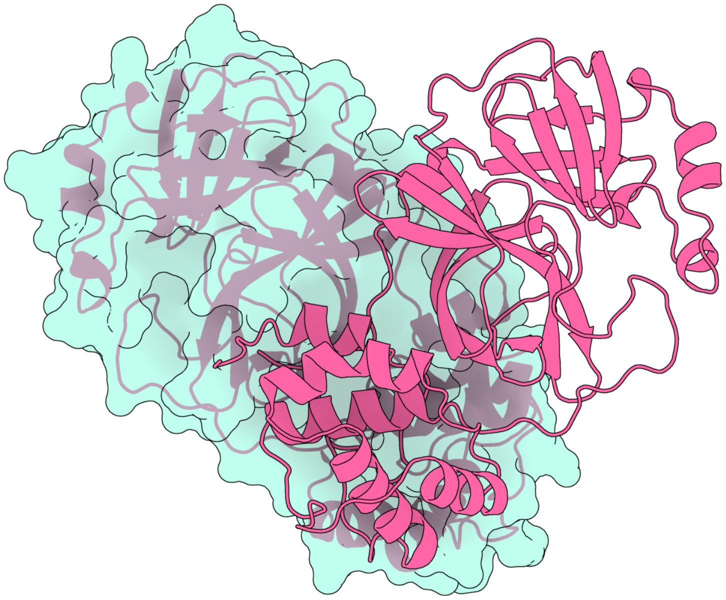


Figure 7.2: Cartoon representation of the M^{pro} dimer. The protein surface is indicated in teal for one protomer. Created with 3D Protein Imager [115].

M^{pro} is encoded by Nsp5 and is released from pp1a/pp1ab by autoproteolysis to form the mature enzyme. Mature M^{pro} cleaves the polyprotein downstream at 11 sites with the consensus recognition sequence x-Leu-Gln↓(Ser, Ala, Gly, Val)-x (x = any residue; ↓ indicates the cleavage site). Despite the emergence of new SARS-CoV-2 variants M^{pro} shows a very high degree of structural conservation especially around the catalytic dyad and the substrate-binding site[116, 117]. The protease naturally forms a functional homo-dimer where each protomer is composed of three domains (Figure 7.2).

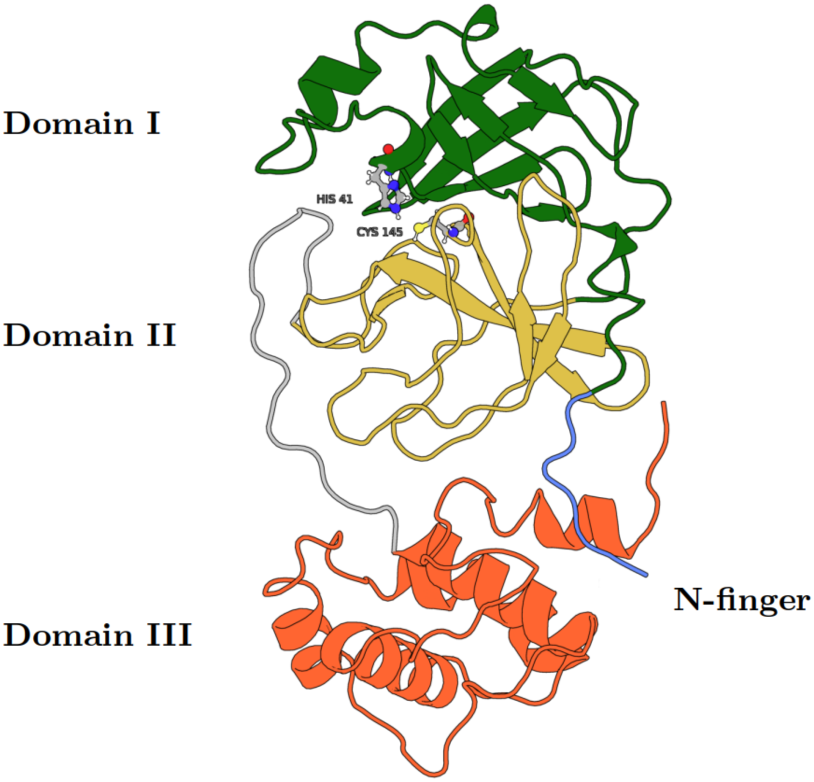


Figure 7.3: Cartoon representation of the M^{pro} monomer (PDB 7AR5). Domain I (green), domain II (yellow), domain III (orange), the N-finger (blue) and the catalytic dyad (His41 and Cys145) are shown. Created with 3D Protein Imager [115].

Domains I and II (residues 8–101 and 102–184) each contain an anti-parallel β-barrel and form a chymotrypsin-like fold harbouring the active site in a cleft between the two domains[114, 118](Figure 7.4). A long loop region (residues 185–200) connects Domain II with the C-terminal Domain III (residues 201–306) which consists of a globular cluster of 5 α-helices (Figure 7.3). Although much less is known about this third extra domain, it has been shown to stabilise the chymotrypsin-like fold and to be directly involved in the critical dimerisation step[119, 120]. Another key structural feature is the N-terminal finger (residue 1-7) which directly interacts with the other protomer. These N-terminal residues are considered to have an important role in the proteolytic activity of the enzyme as manipulation (mutation or deletion) of the N-terminal residues has been shown to significantly reduce the enzymatic activity[121–123].

7.2.1 Structure and Reaction mechanism of the Active Site

The active site consists of a catalytic dyad composed of the nucleophilic Cys145 and the imidazole ring of His41 which acts as a general base. In contrast to other cysteine or serine proteases that contain a third catalytic residue, in M^{pro} a buried water molecule occupies this place[114].

The sites cut by M^{pro} all include a hydrophobic residue (Leu, Phe, or Val) at the P2 position, a conserved Gln at the P1 position and a small amino acid (Ser, Ala or Gly) at the P1' position (P and P' denote the residues placed before and after the scissile bond, respectively). The substrate cleavage is believed to follow a multi-step mechanism (Figure 7.5)[124–126]. Firstly, the Cys145 side chain proton is abstracted by the imidazole ring of His41, creating an nucleophilic thiolate. Upon substrate binding the activated Cys145 then attacks the amide bond of the substrate (Figure 7.5 step 1). This generates an oxyanion which is stabilised by the oxyanion hole formed by the backbone amides of Gly143, Ser144, and Cys145. The formation of the oxyanion hole is crucial for the enzyme's kinetics and can be used as a marker for activity [127, 128]. In a second step the N-terminal peptide product is released by proton abstraction from His41. The intermediate thioester is then hydrolysed, releasing the C-terminal peptide product and restoring the

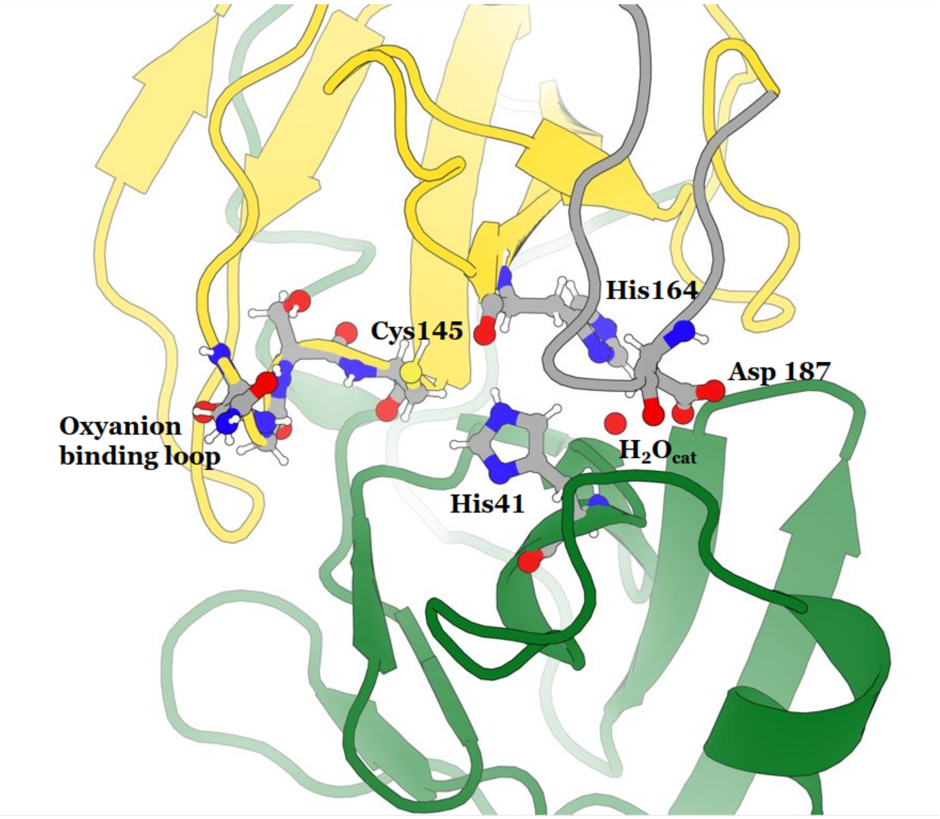


Figure 7.4: Active site of M^{pro}. Domain I is shown in green and domain II in yellow. The catalytically important oxyanion binding loop, Cys145, His41 and H_2O_{cat} form the active site in between the two domains. The buried water H_2O_{cat} is coordinated by His41, His164 and Asp187.

catalytic dyad (Figure 7.5 step 3 and 4).

7.2.2 M^{pro} as a Drug Target

Due to its vital role in the viral replication cycle M^{pro} is considered to be a promising target for the development of therapeutics against COVID-19 infections[88, 129]. The most common strategy is the development of active site inhibitors. These compounds often mimic natural peptide substrates and bind covalently or non-covalently to the active site. SARS-CoV-2 M^{pro}'s

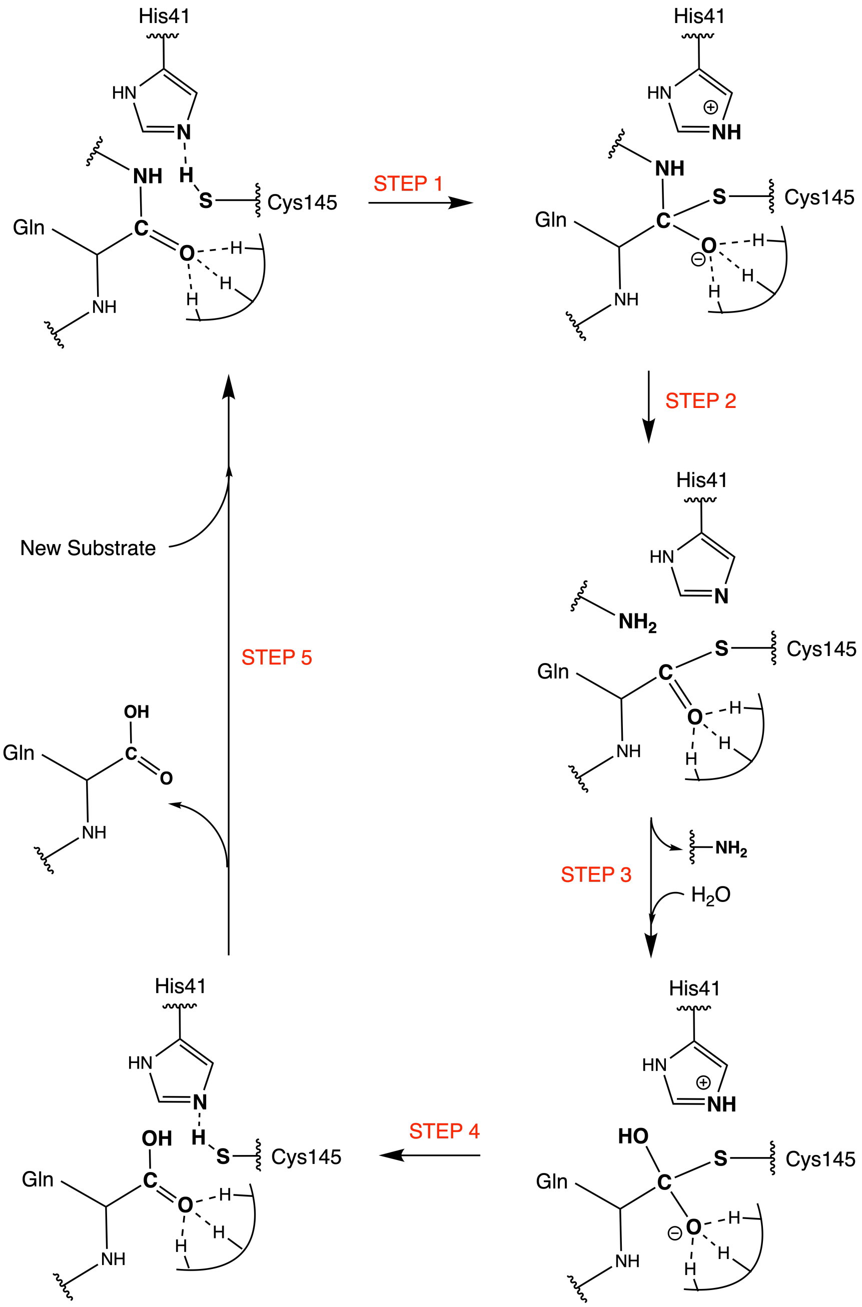


Figure 7.5: Reaction mechanism of M^{pro}.

amino acid sequence shares 96% identity with that of SARS-CoV, with differences at 12 residues between the two viruses [117]. These differences are localised exclusively at active-site distal regions, making it likely that a potential active site drug would not only be effective in SARS-CoV-2 but also against other coronaviruses. Additionally M^{pro}'s nucleophilic active site cysteine allows the design of covalent inhibitors that provide increased inhibition duration and potency[130]. Furthermore, M^{pro} has no human homologue or known overlapping substrate specificity with any human protease making off-target effects unlikely. Nevertheless, it was shown that certain inhibitors can affect host cathepsins and mimic antiviral efficiency without directly causing target inhibition. [131].

Another class of inhibitors are allosteric inhibitors that either affect protein folding, stability, dimerisation behaviour or interactions with host or viral proteins [132]. The formation of the homo-dimer, for example is critical for M^{pro}'s activity[133, 134]. A potential dimer formation inhibitor to target this is the non-active site cysteine Cys300 which sits at the dimerisation interface between the two protomers. Recently it was shown that glutathionylation of Cys300 can reversibly block M^{pro}'s dimerisation and thereby significantly reduce its activity [135]. For Murine hepatitis virus (MHV), a close relative of SARS-CoV, it was shown that M^{pro} associates with numerous components of the replicase complex. Although the exact mechanism is not understood, it was found that modifications of nsp3 and nsp10 would negatively affect the activity of M^{pro}, opening a new route for allosteric drug designs[136].

A third category of inhibitors employs redox active substances to inactivate the enzyme. M^{pro} appears to be quite susceptible to oxidation, in particular the active site cysteine Cys145 shows evidence of oxidation to sulfenic acid (Mono-oxidised), sulfinic acid (Di-oxidised) and sulfonic acid (Tri-oxidised). Oxidation to the later two species is considered to be irreversible and abolishes enzyme activity [137]. Interestingly, M^{pro} was found to possess a redox switch which protects the redox-vulnerable catalytic cysteine by reversibly forming a disulphide bond between Cys145 and Cys117 under transient oxidative stress conditions [138]. Additionally, the stepwise formation of a NOS (nitrogen-oxygen-sulfur) or SONOS (sulfur-oxygen-nitrogen-oxygen-sulfur) bridge between Cys22, Cys44 and Lys61 seems to aid the stability of the protein [139, 140]. These reversible redox-modifications temporarily lower the activity of M^{pro} but protect the enzyme from irreversible over-oxidation. Using non-redox active crosslinkers to mimic the redox switching could therefore be a potential novel approach to design M^{pro} inhibitors.

7.3 Cysteine Oxidation in Proteins

Within the thiol group the sulphur atom is electron-rich and as such a good nucleophile, particularly when deprotonated to a thiolate (RS^-). It is readily oxidised and oxidations states ranging from $-II$ to $+VI$ are possible. Cysteine residues can therefore partake in a variety of redox reactions which provides versatile functionality to the residue. Whether as reactive species in active sites, metal binding, involvement in stabilising a protein's structure or being a site for redox regulation via post translational modifications, cysteine has many biological roles[141]. The best known post-translational modification is the disulphide bridge R−SS−R which is essential for stabilisation of the tertiary structure in many proteins.

Other post-translational cysteine modifications such as the reversible oxidation to sulfenic acid (R−SOH) are less well understood and evidence for their biological significance has accumulated relatively recently [142, 143]. This is due to the fact that sulfenic acid modifications in proteins are unstable to acid hydrolysis and have no distinguishing spectroscopic features[144]. Their detection therefore relied on direct structural evidence. It was shown that R−SOH are excellent electrophilic centres that are well suited for participation in oxidative catalysis, as sensors for oxidative stress and for regulating some transcription regulators[145]. However, oxidation to sulfinic acid can

also cause the reversible inhibition of a protein as was shown for protein tyrosine phosphatase's active site cysteine (PTP) [146, 147].
R$-$SOH can be further oxidised to the more stable sulfinic R$-$SO$_2$H and sulfonic acid R$-$SO$_3$H forms (Equation 7.1) which are considered to be irreversible steps. It is estimated that ∼5 % of cellular protein cysteines occur in one of the two forms [148].

$$\mathrm{RSH} \xrightarrow{[\mathrm{O}]} \mathrm{RSOH} \xrightarrow{[\mathrm{O}]} \mathrm{RSO_2H} \xrightarrow{[\mathrm{O}]} \mathrm{RSO_3H} \tag{7.1}$$

The oxidising species [O] can either be hydrogen peroxide or molecular oxygen, which are both biological oxidants. These reactive oxygen species (ROS) have important roles in cellular redox signalling and in the innate immune response. Cellular ROS are usually kept under tight control due to their damaging potential. But particularly during an immune reaction an imbalance of cellular antioxidants and oxidants (oxidative stress condition) can occur, resulting in proteins with oxidative modifications. For the most part these proteins are not repaired and must be removed by proteolytic degradation[149].

7.4 Enzyme Kinetics

The kinetics of an enzyme following Michaelis-Menten kinetics can be understood in the form of a schematic reaction where an enzyme E and a substrate S form an intermediate complex ES in a reversible reaction where k_1 and k_{-1} are the rate constants for the forward and reverse reaction (Equation 7.2). In the second step the enzyme-substrate complex irreversibly releases the product P with the rate constant k_{cat} which is defined as the maximum number of chemical conversions of substrate molecules per second that a single active site will execute for a given enzyme concentration $[E]_0$.

$$\mathrm{E + S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_{cat}} \mathrm{E + P} \tag{7.2}$$

While product formation is in fact not strictly irreversible, this is a necessary assumption in order to yield a tractable analytic solution. Additionally, the assumption of a steady-state equilibrium is necessary, meaning that the concentration of ES is constant because its rate of formation is balanced by its destruction. These assumptions are valid when (a) the concentration of substrate is much larger than the concentration of product ($[S] >> [P]$), which is the case in most *in vitro* assays and (b) when the rate measurements are restricted to a short time interval where concentration of substrate does not greatly change ($[S] >> [E]$). This allows the rate of product formation to be expressed as the Michaelis-Menten equation:

$$v = \frac{d[P]}{dt} = \frac{V_{max} \cdot [S]}{K_M + [S]} \tag{7.3}$$

where v is the reaction velocity, V_{max} is the maximal reaction velocity, $[S]$ is the substrate concentration and K_M the Michaelis constant. K_M describes the concentration at which the reaction velocity is at half maximum. Low K_M values indicate a high substrate affinity for the enzyme, meaning that the rate will approach V_{max} with lower substrate concentrations $[S]$ than those reactions with large K_M. When $[S] << K_M$ the reaction velocity $v = k_{cat}[E]_0\frac{[S]}{K_M}$ scales linearly with substrate concentration $[S]$ (first-order kinetics). On the other hand, when $[S] >> K_M$ the reaction velocity asymptotically approaches its maximum rate $V_{max} = k_{cat}[E]_0$ and becomes independent of $[S]$ (zero-order kinetics). The activity of an enzyme can also be affected by binding of molecules to sites other than the active site which is called allosteric regulation. The binding of an effector molecule to an allosteric site can either cause a reduction of activity (allosteric inhibition) or an increase in activity (positive cooperativity). To account for potential allostery of enzymes with multiple binding sites Equation 7.3 can be modified to yield the Hill-equation

$$v = \frac{d[P]}{dt} = \frac{V_{max} \cdot [S]^n}{K_{0.5}^n + [S]^n} \tag{7.4}$$

where n is the Hill-constant which describes the degree of interaction between ligand binding sites. When $n = 1.0$ there is no cooperativity between the binding sites and the equation simplifies to the Michaelis-Menten equation. When $n > 1.0$ binding of a first ligand molecule to an active site increases the affinity binding of ligands to the second active site, and indicates cooperative binding. Conversely, when $n < 1.0$ binding of of a first ligand molecule decreases the affinity for the binding of a second ligand and indicates negative cooperativity.

Chapter 8

Materials & Methods

8.1 Protein Purification

8.1.1 Buffers and Solutions

- **Buffer A**: 20 mM Tris pH 7.8, 150 mM NaCl, 20 mM Imidazole, 1 mM DTT

- **Buffer B**: 20 mM Tris pH 7.8, 150 mM NaCl, 500 mM Imidazole, 1 mM DTT

- **Dialysis Buffer**: 20 mM Tris pH 7.8, 150 mM NaCl, 1 mM DTT

- **SEC Buffer**: 220 mM Tris pH 7.8, 150 mM NaCl, 1 mM EDTA, 1 mM TCEP

- **Crystallisation Buffer**: 100 mM Bis-Tris pH 6.5, 175 mM $LiSO_4$, 17.5 % (*w/v*) PEG 3350

- **Cryo Buffers**: 100 mM Bis-Tris pH 6.5, 175 mM $LiSO_4$, 12 % *(w/v)* PEG 3350, 5 %/10 %/15 %/20 % *(v/v)* glycerol

8.1.2 Cloning

Two constructs for M^{pro} expression were used:

- PGEX-6P-1 (GE Healthcare) made available by Rolf Hilgenfeld
- PGEX-4T-1 (GE Healthcare) made available by Sebastian Günther (see Figure 8.1)

Both constructs code for M^{pro} with an N-terminal M^{pro} cleavage site (SAVLQ↓SGFRK) connected to a GST-tag (cleavable by autoproteolysis). The C-terminus has a modified PreScission protease cleavage site (SGVTFQ↓GP) connected to a His_6-tag (Hilgenfeld) or His_8-tag (Günther). This generates authentic N and C-termini after tag removal.

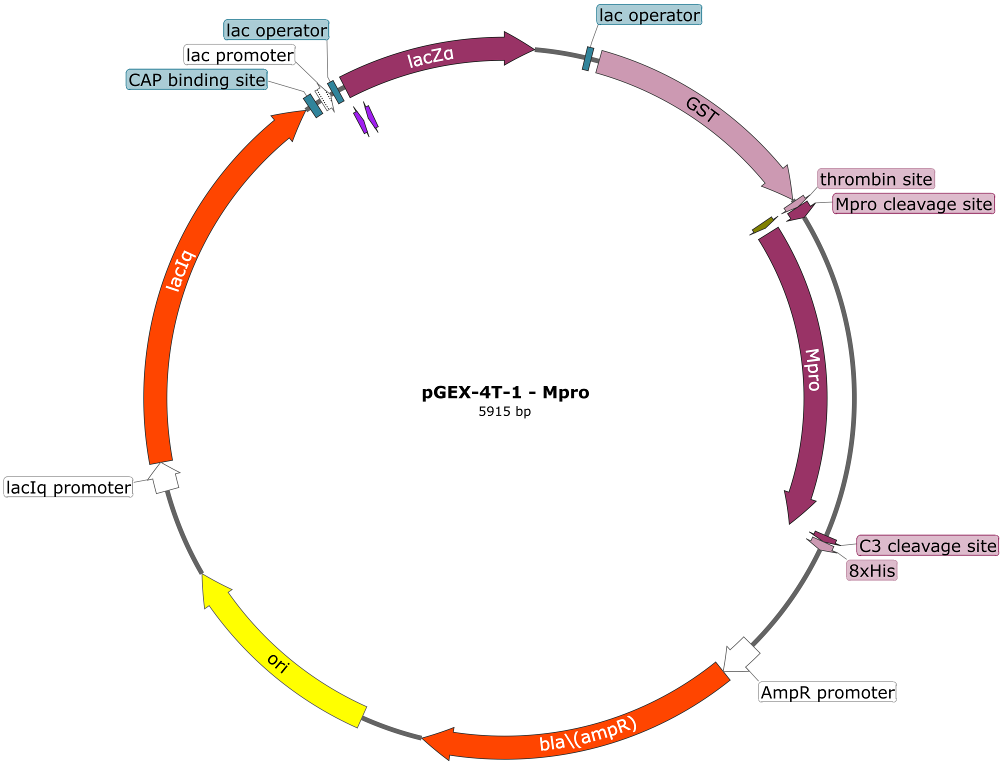


Figure 8.1: Expression vector pGEX-4T-1 for SARS-CoV-2 M^{pro}.

The PGEX-6P-1 plasmid was transformed into *E. coli* BL21 gold (DE3) and the PGEX-4T-1 plasmid into *E. coli* Rosetta II cells using electroporation.

8.1.3 Cell Culture

6 x 1 L LB media supplied with 100 µg/mL ampicillin were inoculated with 10 mL starter culture. The cell cultures were grown at 37 °C and 180 rpm until an OD_{600} of 0.5 was reached. Expression was induced by adding isopropyl-D-thiogalactoside (IPTG) to a final concentration of 0.5 mM. After overnight incubation at 27 °C and 180 rpm, cells were harvested by centrifugation at 9000 rpm (Sorvall Lynx 6000, F9-6x1000 LEX rotor) for 15 min at 4 °C. The cell pellet was resuspended with a 1:5 ratio (w/v) in Buffer A.

8.1.4 Cell Lysis

The cell suspension was divided into 40 mL volumes in falcon tubes and placed on ice. Each tube was sonicated for 6 min, with 3 cycles at 75 % power on ice using a Bandelin Sonoplus system. The lysate was spun down at 14 000 rpm (Sorvall Lynx 6000, F14-14x50cy rotor) for 45 min at 4 °C. The Supernatant was decanted off and further cleared by sequential filtering through 0.8 µm and 0.45 µm syringe filters.

8.1.5 Ni-Affinity Chromatography

A 5 mL Hi-Trap column was washed with 4 column volumes (CV) water prior to equilibration with 8 CV Buffer A. The cleared supernatant was loaded onto the column and the flow through collected. The column was then washed with 4 CV Buffer A and 4 x 5 mL fractions (Wash 1 - Wash 4) were collected. M^{pro} was eluted from the column with 4 CV Buffer B and 4 x 5 mL fractions (Elution 1 - Elution 4) were collected.

8.1.6 His-tag Removal and Dialysis

Fractions containing the M^{pro} were pooled. The concentration was determined with a Nanodrop using $\varepsilon = 33.6\ mM^{-1}cm^{-1}$ and a molecular weight of 33.8 kDa. 1 mg PreScission protease was added to the protein solution for each 5 mg of M^{pro}. The cleavage mixture was then transferred to a Spectra/Por® dialysis tubing with a MWCO of 6-8 kDa and dialysed against 5 L of dialysis buffer at 4 °C, stirring and overnight.

8.1.7 Reverse Ni-Affinity Chromatography

Precipitate that formed during dialysis was centrifuged down and the cleared protein solution was continued with as described in 8.1.5. This time M^{pro} is expected to be in the flow through an wash fractions whereas the elution fractions contain the cleaved His-tag and the PreScission protease. Subsequently the purity of M^{pro} was assessed with an SDS-PAGE gel.

8.1.8 Protein Concentration

Fractions containing the cleaved M^{pro} were pooled and the buffer exchanged with SEC buffer using an Amicon® centrifugal concentrator with a MWCO of 10 kDa. Finally, M^{pro} was concentrated to 20 mg/mL using a HEREAUS MEGAFUGE 40R centrifuge with the Tx-1000 rotor at 4000 rpm and 30 µL aliquots flash cooled in liquid nitrogen for later use.

8.1.9 Modifications for Purification under Anaerobic Conditions

For the anaerobic preparation all buffers were degassed under vacuum and subsequently sparged with nitrogen gas. This process was repeated twice.
After centrifugation of the lysate the supernatant was immediately transferred into a glove box which was operated under a constant nitrogen atmo-

sphere (max. 10 ppm O_2). All subsequent purification steps (as described in section 8.1.5 to section 8.1.7) were carried out under nitrogen atmosphere. Prior to the final concentration of pure M^{pro}, TCEP pH 7.0 was added to the protein solution to a final concentration of 20 mM. A 10 kDa MWCO concentrator was prepared under nitrogen atmosphere, sealed and then removed from the glove box for the final concentration.

8.1.10 Modifications for Aerobic Purification without Reducing Agent

For the aerobic preparation without reducing agent all buffers were prepared without the addition of DTT or TCEP. After the reverse Ni-affinity chromatography all fractions containing M^{pro} were pooled and left stirring at 4 °C for 5 days to allow complete oxidation of the protein by atmospheric oxygen. The protein solution was centrifuged at 4000 rpm, 4 °C for 5 min to remove any precipitated protein and subsequently concentrated.

8.2 Crystallisation

8.2.1 Seed Preparation

M^{pro} was crystallised at 5 mg/mL in 9 µL drops in a hanging drop vapour diffusion setup over a 1000 µL reservoir. The drops contained protein solution and crystallisation buffer mixed in a 1:1 ratio.

20 µL of reservoir solution was added to drops containing clusters of M^{pro} crystals before a glass pipette with a molten tip was used to crush them to seeds. This solution was then diluted 1:200 and 1:300 with reservoir solution to obtain the first generation seed solution.

The quality of seeds was improved by repeating this procedure until 4th generation seeds were obtained.

8.2.2 Seeded Crystallisation of M^{pro}

Using hanging drop vapour diffusion setup M^{pro} was crystallised over a 1000 µL reservoir. Crystallisation drops contained 4.5 µL protein solution at 5 mg/mL, 4.5 µL crystallisation buffer with 12 % *(w/v)* PEG 3350 and 0.5 µL of 200x or 300x seed solution.

8.2.3 Cryo-protection

Cryo-protection was achieved by sequentially transferring crystals into synthetic cryo buffers containing 5 %, 10 %, 15 % and 20 % *(v/v)* glycerol respectively prior to flash cooling in liquid nitrogen..

8.3 X-ray Crystallography

8.3.1 Data collection

For the anaerobic M^{pro} samples data were collected on the 28.04.2021 on beamline ID30B (ESRF) at 100 K with a 20×20 μm^2 gaussian beam, 12.7 keV photon energy and a flux 2.07×10^{12} ph/s (1.4 % transmission). Following an initial characterisation, individual automatic data collection strategies were used for each crystal to collect single datasets using a Pilatus3 6M detector. For aerobically purified M^{pro} (with and without reducing agents) data were collected on the 23.02.2022 on ID23-2 (ESRF) at 100 K. Here, helical scans were used utilise the full size of the crystal compared to the relatively small beam size of 8×25 μm^2 (gaussian beam profile). Data were collected with 14.2 keV photon energy, a flux of 6×10^{10} ph/s (15 % transmission), 0.1° oscillations, for a total of 1990 images on an EigerX9M detector.

8.3.2 Anisotropic refinement

Data sets collected were indexed and integrated with XDS[69] and further processed with Aimless[70] for data reduction. Then, images with $R_{merge} > 0.4$ were identified and the data set reprocessed excluding those images in a new XDS run. This was done for multiple crystals and the ASCII.HKL files obtained this way were scaled together to a common global scale using XSCALE[69] to produce an unmerged composite data set. For data reduction the composite data set was uploaded to the Staraniso web server [150] and an anisotropic resolution cut-off at local $I_{mean}/\sigma_{I_{mean}}$: 1.2 was applied. Subsequently, sftools[151] was used according to the instructions on the Staraniso website to enable downstream programs the correct handling of the data set. A *de novo* generated R_{free}-set was to added to composite data set using cad[152]. Finally, the model was refined in multiple REFMAC5[73]/Coot[72] cycles.

8.4 Mass spectrometry

Prior to mass spectrometry (MS) analysis, fresh aliquots of M^{pro} were buffer exchanged into 5 % (v/v) acetonitrile in water using Amicon® centrifugal concentrators with a MWCO of 10 kDa and subjected to UPLC using the ACQUITY UPLC BEH C18 1.7 µm column at 55 °C. The following UPLC method was used (Table 8.1) with 0.1 % *(v/v)* formic acid in water as buffer A and acetonitrile with 0.1 % *(v/v)* formic acid as buffer B.

Samples were analysed using a Waters Xevo G2-XS QTof platform. The system was operated in positive sensitivity mode using a cone voltage of 40 V and electrospray ionisation. The source temperature was maintained at 120 °C and a desolvation temperature of 44 °C was applied to assist desolvation. No collision energy was applied. Time of flight data was continuously recorded over 10 min in an m/z range between 400 to 3000 Da with a 1 s scan

Table 8.1: UPLC method.

	Flow (ml/min)	%A	%B
1 min	0.4	95	5
1.1 min	0.2	95	5
3.5 min	0.2	20	80
4 min	0.2	20	80
4.1 min	0.4	5	95
4.5 min	0.4	5	95

time. Mass spectral data was deconvoluted to zero-charge spectra using the MaxEnt1 algorithm of the MassLynx software.

8.5 Activity Assay

M^{pro} was diluted to a concentration of 25 µM with assay buffer (20 mM Tris, 100 mM mM NaCl, 1 mM DTT, 1 mM EDTA pH 7.3). The substrate Ac-Abu-Tle-Leu-Gln-AMC (7-Amino-4-methylcoumarin) was purchased from Biosynth Carbosynth (product number: FA178674) and dissolved in DMSO to generate a 25 mM stock solution. From this stock solution a 1 mM substrate working solution in assay buffer was made which was used to prepare multiple samples containing a different substrate concentrations ranging from 1 µM to 800 µM. Above 850 µM the substrate becomes insoluble.

The activity measurement was performed using a Varian Cary50 Bio UV/Vis spectrophotometer at 20 °C and 700 µL Hellma quartz cuvettes 104-002B-QG (product number: 104.002B-QG). Each sample was prepared in a quartz cuvette and the UV/Vis-spectrum was measured until a stable baseline was acquired. Subsequently, M^{pro} was added to a final concentration of 1 µM and rapidly mixed. The absorption change at 380 nm for each substrate concentration was measured for 2 min collecting one data point every 0.5 sec.

Chapter 9

Results & Discussion

9.1 Active site analysis of published M^{pro} structures

Drug development relies on accurate high resolution structural data. For SARS-CoV-2 the main protease was identified as a promising drug target and a major focus of early COVID-19 related research. In this section 5 published M^{pro} structures and their corresponding electron density are analysed with respect to an active site modification that might have been missed.

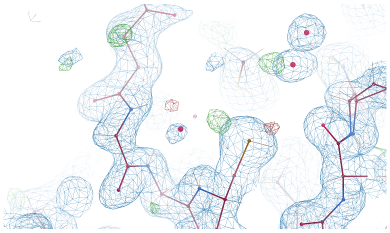
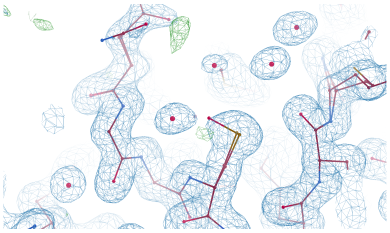
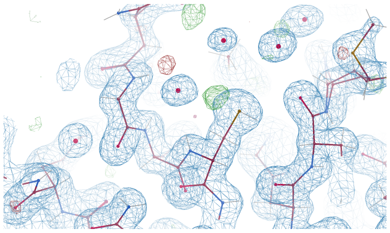
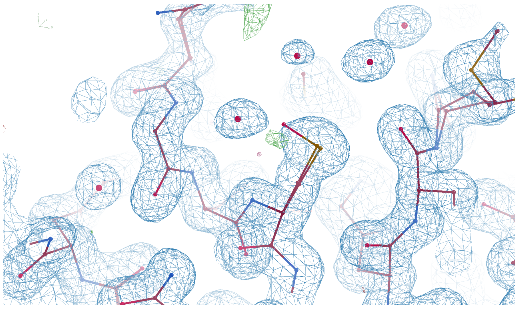
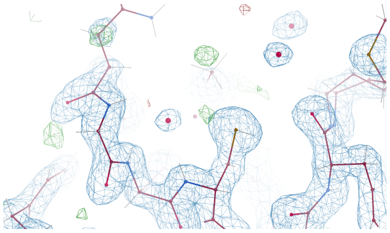
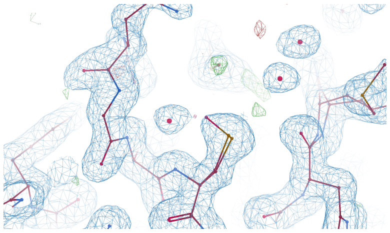
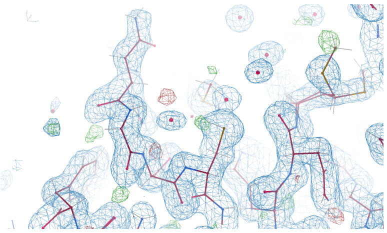
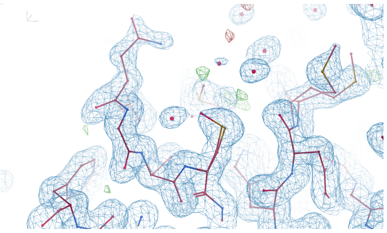
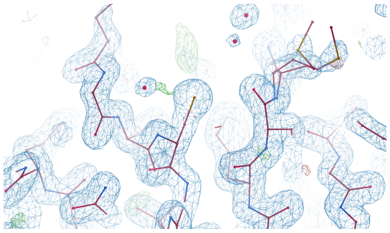
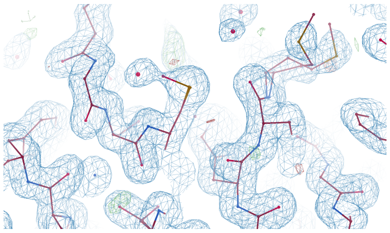
These structures were among the first published, high resolution, apo structures of M^{pro}. All data sets have a comparable resolution ranging from 1.4 Å to 1.6 Å but were produced under different experimental parameters including different crystallisation conditions, pH and acquisition temperatures. For example, the structure 7MHG is a result of data produced at 240 K and pH of 7.0, 7AR5/7AR6 are results of data collected at 100 K at a pH of 7.5 to 7.8 and 7JPY is a result of data produced at 120 K at a pH of 8.0. The pH is an important factor as its determines the charge state of a molecule. For SARS-CoV M^{pro}, that shares 96 % homology with the SARS-CoV-2 enzyme, the pKa of Cys145 and His41 was experimentally determined to be 8.0 ± 0.3 and 6.3 ± 0.1, respectively[153, 154]. Therefore, it is expected that both catalytic residues are in the less reactive uncharged state.

Nevertheless, in all cases additional positive difference electron density is present at the active site cysteine, indicating a modified cysteine residue (shown in Table 9.1). The first column shows the published structure with additional positive difference density which can not be explained by a water molecule because the distance to the sulphur atom would be too small.

Since the additional positive difference density appears in all studies it is unlikely noise. Considering the high reactivity of the active site cysteine a possible explanation could be an oxidative modification. However, it is then surprising that this feature also appears at pH's where the catalytic dyad is expected to be in the less reactive uncharged form, indicating a very high susceptibility. When cysteine 145 is modelled as sulfenic acid the structures fit their density much better. This is illustrated in the second column which shows a re-processed version where the cysteine has been modelled with an alternative conformation representing a mono-oxidised cysteine (no further occupancy refinement done). There are three plausible scenarios that could explain an oxidative modification: (1) The modification is a result of secondary radiation damage to the highly reactive site and hence an artifact of the data collection. (2) The modification is a result of oxidation during the aerobic purification by ambient oxygen despite the use of 1 mM reducing agents and hence an artifact of the purification strategy. (3) The modification is part of the enzyme's natural state and has possibly an important role in catalysis and the regulation thereof.

The first scenario is unlikely because the difference density also appears at cryo-temperatures where damaging secondary species are largely immobilised. If the third scenario is true M^{pro} should have this modification even if it is purified and crystallised under anaerobic conditions. Lastly, if the second scenario is true the modification should be present when no reducing agents are used, may be present with 1 mM reducing agents but should not be present when M^{pro} is purified under anaerobic conditions.

Table 9.1: Comparison of M^{pro} active sites of published structures with structures where cysteine 145 was replaced by a sulfenic acid residue. $2mF_o - DF_c$ electron density is shown at a $1rmsd$ level and $mF_o - DF_c$ difference density at a $3rmsd$ level.

PDB code	published structure	rebuilt with CSO
7AR5 [88]		
7AR6 [88]		
7MHG[155]		
7MHF[155]		
7JPY[156]		

9.2 M^{pro} Purification and Crystallisation

M^{pro} was purified under anaerobic conditions, aerobic conditions in presence of 1 mM reducing agents and aerobic conditions without reducing agents. Exemplary for aerobically purified M^{pro} Figure 9.1 shows an SDS-PAGE gel with the results of the first purification step. The absence of a band at 61.9 kDa, which would correspond to the GST-M^{pro}-His$_6$ construct, indicates that the produced M^{pro} is active and able to autoproteolytically cleave itself off the GST-tag. The resulting Mpro-His$_6$ has a size of 35.0 kDa and corresponding bands can be seen in the elution fractions (lane 11 to 14).

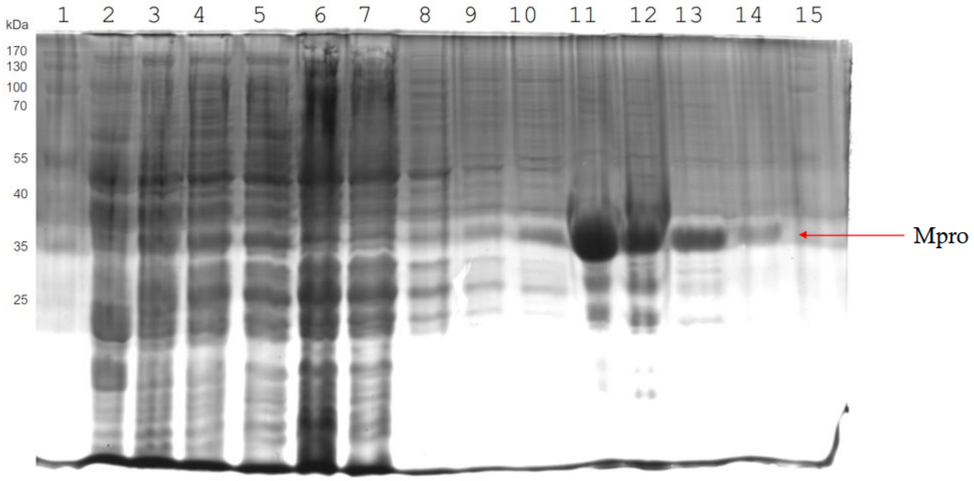


Figure 9.1: 12 % SDS Tris-Tricine PAGE gel of aerobically purified M^{pro} after Ni-affinity chromatography. Lane 1: PageRuler prestained Protein Ladder, lane 2: uninduced cells, lane 3: induced cells, lane 4: lysate, lane 5: supernatant, lane 6: flow through, lane 7: wash 1, lane 8: wash 2, lane 9: wash 3, lane 10: wash 4, lane 11: elution 1, lane 12: elution 2, lane 13: elution 3, lane 14: elution 4.

After His-tag removal the sample was subjected to a reverse Ni-affinity chromatography and the resulting gel is shown in Figure 9.2. Lane 2 shows the pooled fractions containing M^{pro} before the cleavage and lane 3 after the cleavage. Most notably, a strong band at 24 kDa appeared in lane 3 which can be assigned to the His-tagged PreScission Protease used for the tag removal. The released M$^{\text{pro}}$ has a size of 33.8 kDa and is only marginally smaller than its His-tagged counterpart. A difference can hardly be made out due to the high concentrations present. In the elution fractions (lane 9 to 12) the remaining impurities, the PreScission Protease and a small amount of uncleaved M$^{\text{pro}}$ can be seen. However, as lane 4 to 8 show, the majority of M$^{\text{pro}}$ molecules are cleaved off their tags and therefore elute in the flow through and wash fractions.

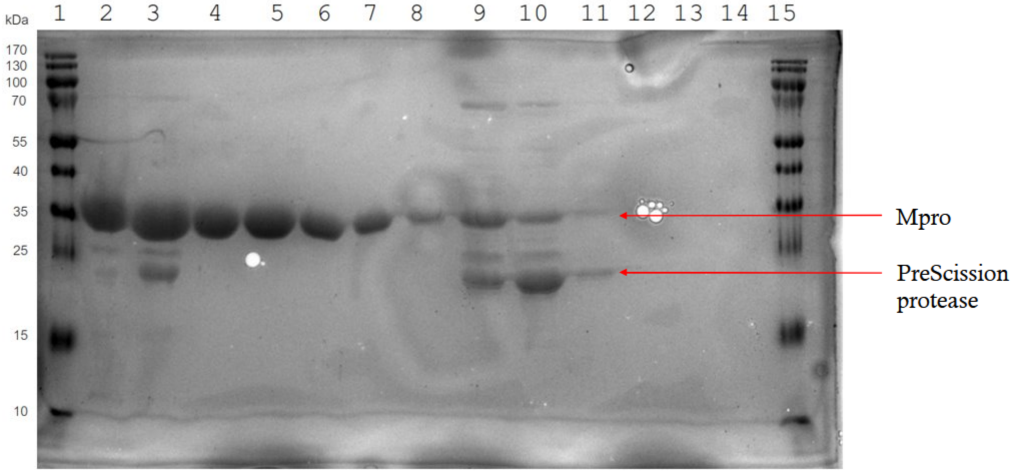
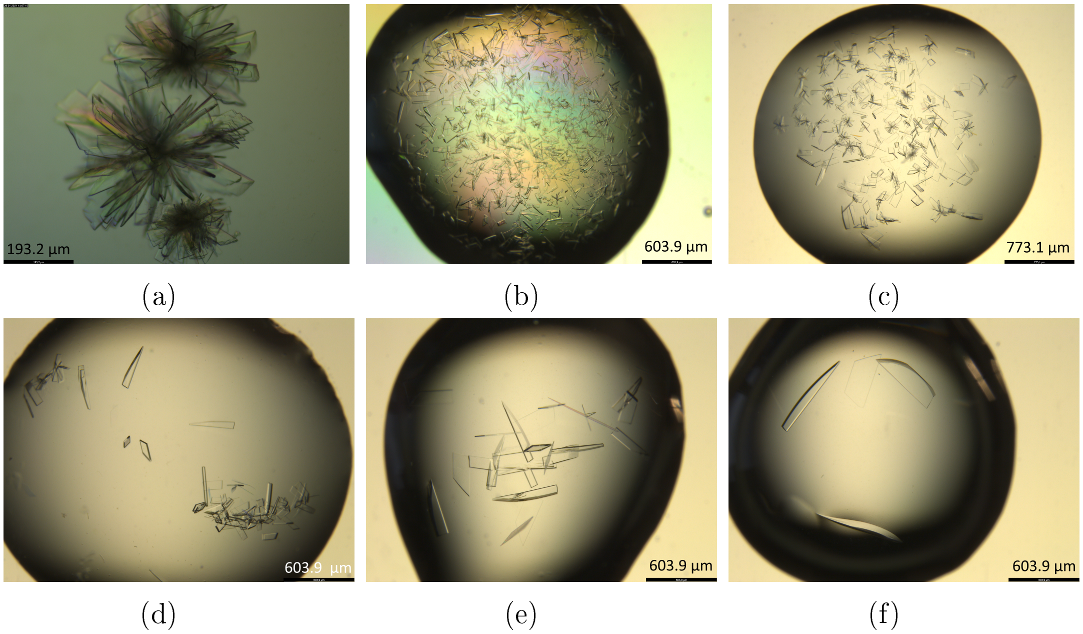


Figure 9.2: 12 % SDS Tris-Tricine PAGE gel of aerobically purified M$^{\text{pro}}$ after reverse Ni-affinity chromatography. Lane 1: PageRuler unstained Protein Ladder, lane 2: pre-cleavage, lane 3: post-cleavage, lane 4: flow through, lane 5: wash 1, lane 6: wash 2, lane 7: wash 3, lane 8: wash 4, lane 9: elution 1, lane 10: elution 2, lane 11: elution 3, lane 12: elution 4.

These fractions were pooled and used as pure M$^{\text{pro}}$ for subsequent experiments. Both the anaerobic and aerobic purification without reducing agents yielded comparable results which can be found in Appendix A.3 to A.6.



(a) (b) (c)

(d) (e) (f)

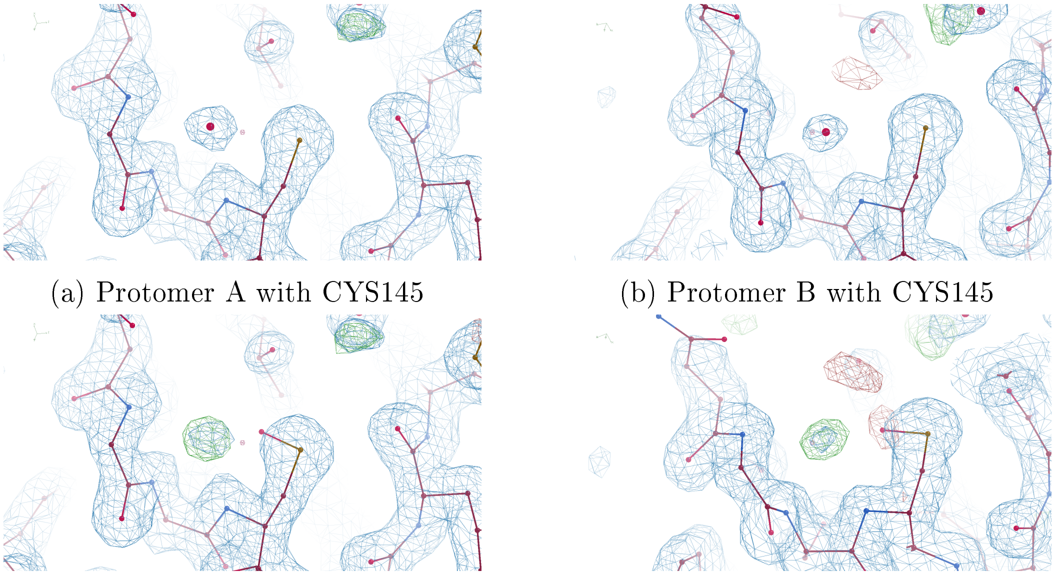
Figure 9.3: Crystal images of M^{pro} showing crystals (a) unseeded (b) seeded with first generation seeds (c) seeded with second generation seeds (d) seeded with second generation seeds and precipitant adjustment (e) seeded with third generation seeds and (f) seeded with fourth generation seeds.

M$^{\text{pro}}$ crystallises in multifaceted clusters unsuitable for X-ray diffraction (Figure 9.3a). Microseeding was employed to grow large single crystals. Figure 9.3b to 9.3e shows the successive improvement of crystal quality with each new generation of seeds. To reduce the number of crystals per drop the precipitant concentration was lowered from 17.5 % *(w/v)* PEG 3350 to 12 % *(w/v)*. The most optimised iteration of crystals (Figure 9.3f) were growing to 600 x 100 x 25 µm^3 in size over a few days. In contrast, when no reducing agents were used throughout the purification, crystals were fewer, much smaller and generally worse diffracting. This indicates that the protein is significantly affected by the lack of reducing agents and therefore oxidative pressure which limits crystal formation, crystal growth and lattice order.

9.3 X-ray crystallography

To assess the effect of different levels of oxidative pressure on the protein structure X-ray diffraction data were collected from the three M$^{\text{pro}}$ samples. Previously reported structures (Chapter 9.1) have been predominately indexed in C121 and I121. In space group C121 the asymmetric unit contains only one protomer, meaning that the functional dimer will consist of two crystallographically related and thus identical monomers. Here, each data set is indexed in P12$_1$1 in which the asymmetric unit comprises both protomers, allowing them to be modelled independently and revealing differences between them. Data collection and refinement statistics for anaerobic M$^{\text{pro}}$ are reported in Appendix A.1, for aerobic M$^{\text{pro}}$ in Appendix A.2 and for aerobic M$^{\text{pro}}$ without reducing agents in Appendix A.3.

Starting with the anaerobic M^{pro} data (Figure 9.4), each image shows the active site and its catalytically relevant residues, starting from left to right with the buried catalytic water H_2O_{cat}, Histidine 41, Cysteine 145 and the oxyanion binding loop. Figure 9.4a and 9.4b show that in both protomers the active site cysteine has no associated additional positive difference electron density. Instead the oxyanion binding hole is occupied by a water molecule that forms hydrogen bonds with the backbone amides of the oxyanion loop. When cysteine 145 is modelled as sulfenic acid (Figure 9.4c and 9.4d) and the water molecule is removed two things can be observed. Firstly, the sulfenic acid moeity does not move into the space where the water was placed and instead produces negative difference density as seen in protomer B. This suggests that no oxidative modification is present in either of the protomers. Secondly, despite the presence of an additional oxygen atom in the vicinity, positive difference density shows up where the water was placed giving evidence for its presence.



(a) Protomer A with CYS145

(b) Protomer B with CYS145

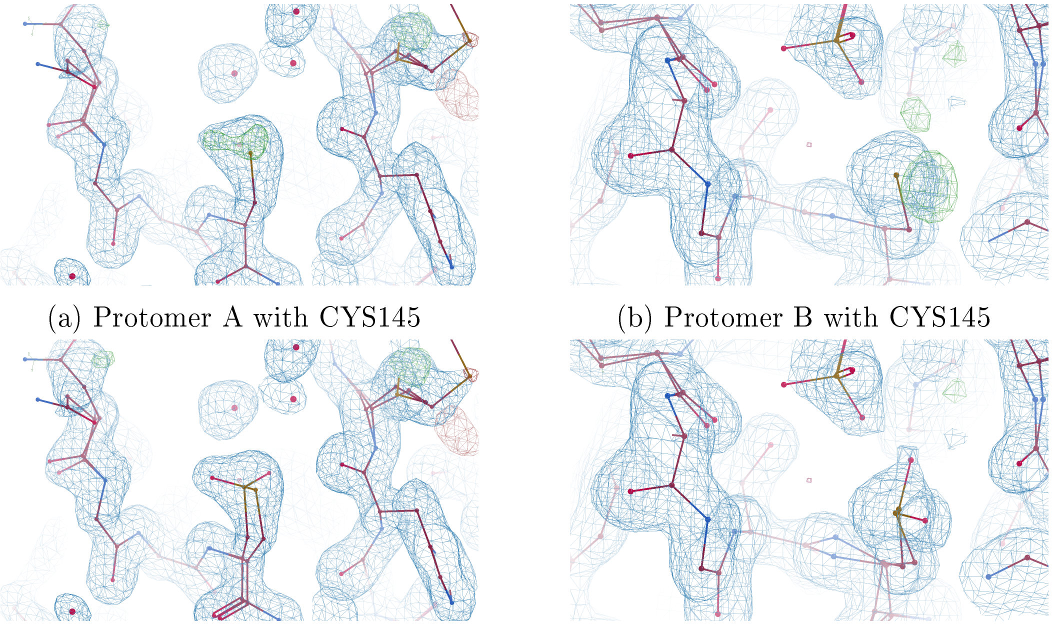
(c) Protomer A with CSO145

(d) Protomer B with CSO145

Figure 9.4: $2mF_o - DF_c$ electron density map of **anaerobically** purified M$^{\text{pro}}$, rendered at $1rmsd$ and $mF_o - DF_c$ difference density contoured at $3rmsd$. (a, b) showing the active site when residue 145 is modelled as cysteine for protomer A and B, respectively. (c, d) shows the active site when residue 145 is modelled as sulfenic acid for protomer A and B, respectively.

The situation is different for aerobically purified M$^{\mathrm{pro}}$ (Figure 9.5). Here, the active site cysteine in both protomers shows additional positive electron density around the S_γ atom and an empty oxyanion binding hole, suggesting a modified cysteine. The density can be interpreted in two ways. Either as a sulfinic acid residue with low occupancy or as a sulfenic acid residue with two conformers, the later of which is shown in Figure 9.5c for protomer A and Figure 9.5d for protomer B. This fully explains the electron density and suggests that Cys145 in aerobically purified M$^{\mathrm{pro}}$ is oxidised even when reducing agents are used throughout the purification and during crystallisation. However, the exact modification (mono-oxidised or di-oxidised) can not be unambiguously determined from the X-ray diffraction data alone.

When no reducing agents are used during aerobic purification, unsurprisingly, a similar result is obtained. If the active site is built with a cysteine, additional positive difference density appears next to the S_γ atom, indicating an insufficient model (Figure 9.6a and 9.6b). This density disappears when the cysteine is replaced by a sulfenic acid residue (Figure 9.6c and 9.6d). Here the density indicates only one conformation. However, considering the resolution of this data set is 0.8 Å worse than in the previous two it is possible that the information for a second conformation is simply missing in this data set.



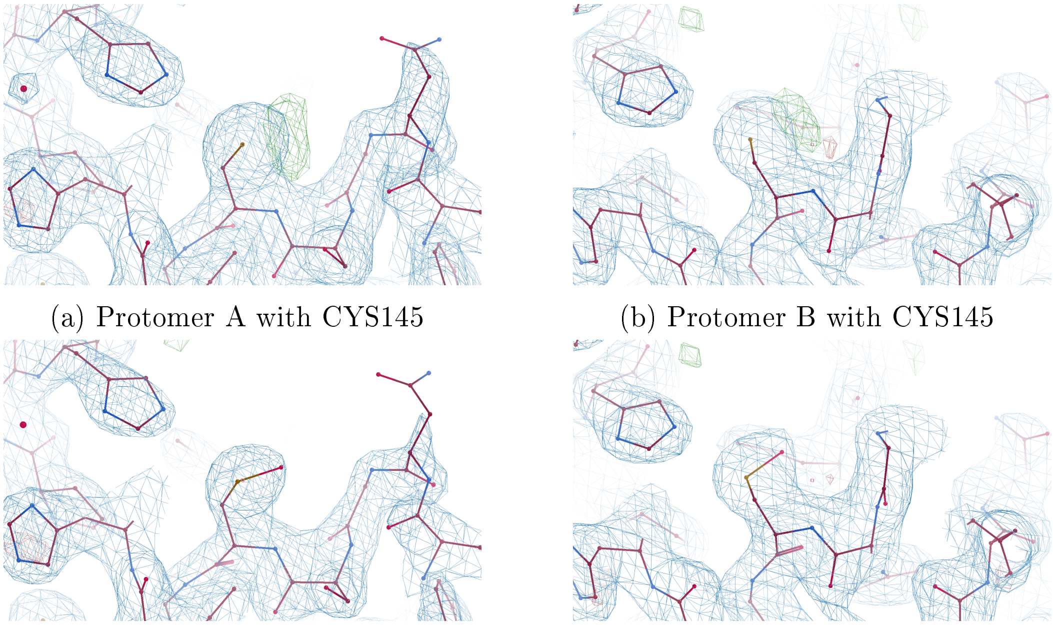
(a) Protomer A with CYS145

(b) Protomer B with CYS145

(c) Protomer A with CSO145

(d) Protomer B with CSO145

Figure 9.5: $2mF_o - DF_c$ electron density map of **aerobically** purified M$^{\text{pro}}$, rendered at $1rmsd$ and $mF_o - DF_c$ difference density contoured at $3rmsd$. (a, b) showing the active site when residue 145 is modelled as cysteine for protomer A and B, respectively. (c, d) shows the active site when residue 145 is modelled as sulfenic acid for protomer A and B, respectively.



(a) Protomer A with CYS145

(b) Protomer B with CYS145

(c) Protomer A with CSO145

(d) Protomer B with CSO145

Figure 9.6: $2mF_o - DF_c$ electron density map of **aerobically (w/o reducing agents))** purified M$^{\text{pro}}$, rendered at $1rmsd$ and $mF_o - DF_c$ difference density contoured at $3rmsd$. (a, b) showing the active site when residue 145 is modelled as cysteine for protomer A and B, respectively. (c, d) shows the active site when residue 145 is modelled as sulfenic acid for protomer A and B, respectively.

To summarise, the presented X-ray diffraction data indicate an oxidised active site Cys145 regardless of whether reducing agents were used or not for both aerobic purifications. Conversely, when M^{pro} is purified and crystallised under anaerobic conditions no indication of an active site modification can be observed.

9.4 Mass spectrometry

To confirm the interpretation of electron density maps, M^{pro} was further analysed with mass spectrometry. The expected average mass for unoxidised M^{pro} is 33 796.64 Da. Each additional oxygen atom would increase the average mass by 16 Da. The zero-mass spectrum of aerobically purified M^{pro} with reducing agents is shown in Figure 9.7.

Two major peaks are visible which can be assigned to a dehydrated species (I) and the average mass (II). The presence of dehydrated species such as succinimide formation from aspartic acid or pyroglutamic acid formation from glutamic acid is [157]) potentially linked to the electrospray process as these peaks appear in all M^{pro} samples. However, no attempt is made to determine the exact modification of these species.

Interestingly, the zero-mass spectrum of aerobically purified M^{pro} with reducing agents shows no peaks for oxidised species despite their evidence in the X-ray diffraction data.

When this is compared to the zero mass spectrum of aerobically purified M^{pro} without reducing agents (Figure 9.8) again the major peak (IV) can be observed and some peaks with lower abundance at smaller masses (II, III) that can be interpreted as different dehydrated species. Additionally, the spectrum also contains several new peaks at higher masses, which can be assigned to different oxygenated species (V - VII) as their mass match the addition of one oxygen atom(V), two oxygen atoms (VI) and three oxygen atoms with three additional protons (VII).

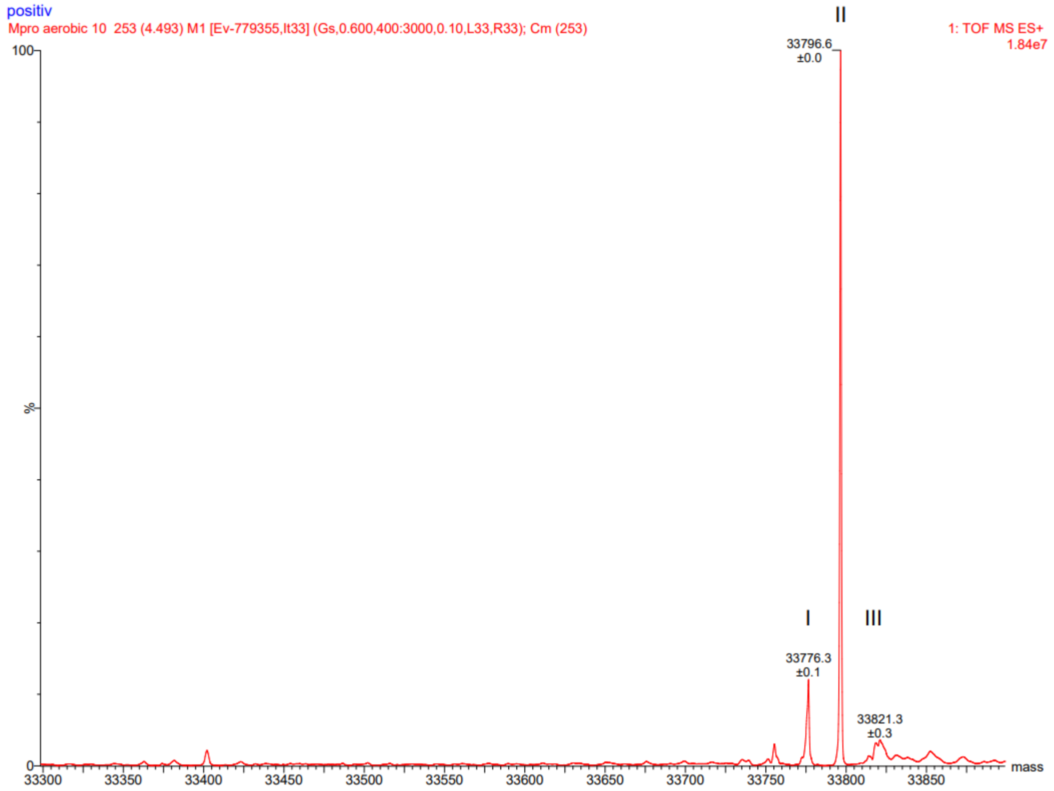


Figure 9.7: Zero-charge mass spectrum of aerobically purified M^{pro} with reducing agents.

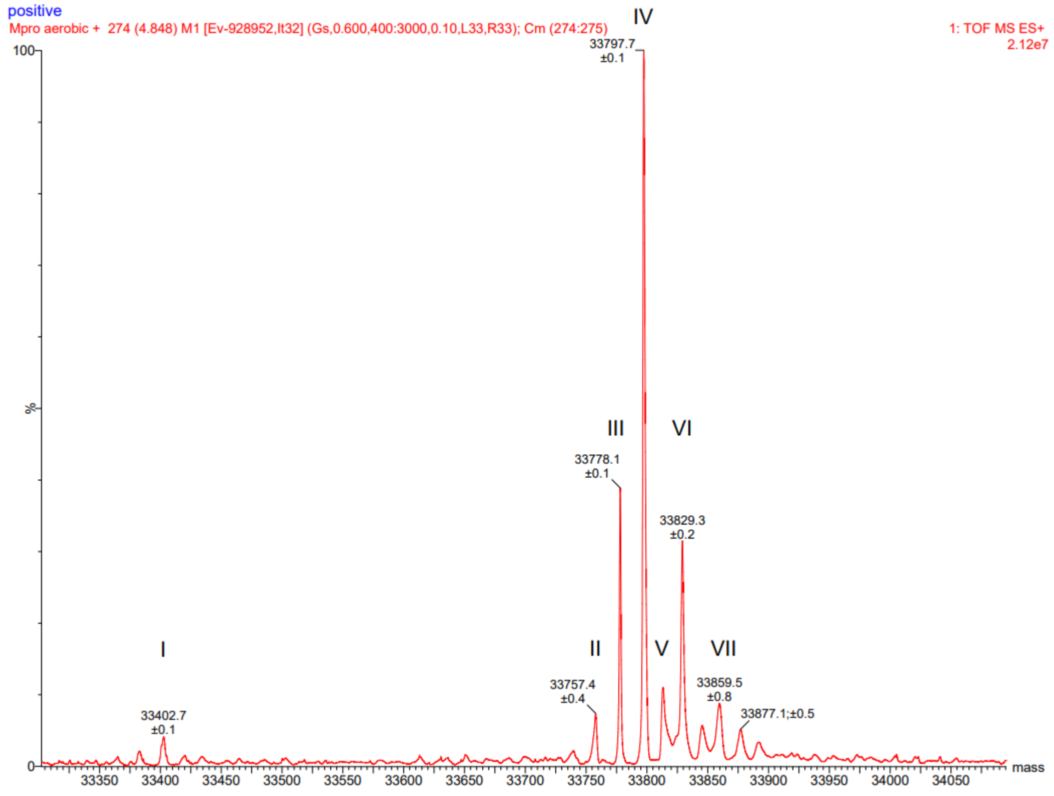


Figure 9.8: Zero-charge mass spectrum of aerobically purified M^{pro} without using reducing agents.

However, due to the relatively low abundance of this third species, and hence larger associated error, the interpretation is ambiguous.
Lastly a small peak (I) with a Δm of -394.32 Da can be seen which would be consistent with the possible loss of the three N-terminal residues TFQ (394.43 Da). This is an important finding considering that the N-terminal residues have an essential role in enzyme activity and loss thereof would significantly reduced the enzymatic activity[121–123]. Considering that the protein was incubated an additional 5 days after the purification to allow full oxidation by ambient oxygen, it is not surprising that signs of protein degradation such as the loss of N-terminal residues become apparent.

To recapitulate, the electron density maps of both samples ($\mathrm{M_{pro}}$ with and without reducing agents) show signs of oxidation. However, the mass spectra suggest that before crystallisation only the sample that has been purified without reducing agents contains oxidative modifications. Consequently, when reducing agents are used the oxidation must occur during crystallisation. If this interpretation is correct $\mathrm{M^{pro}}$ should have no modifications when purified and crystallised under anaerobic conditions and indeed the zero charge mass spectrum for anaerobically purified $\mathrm{M^{pro}}$ has no peaks that would suggest an oxidative modification (Figure (9.9).

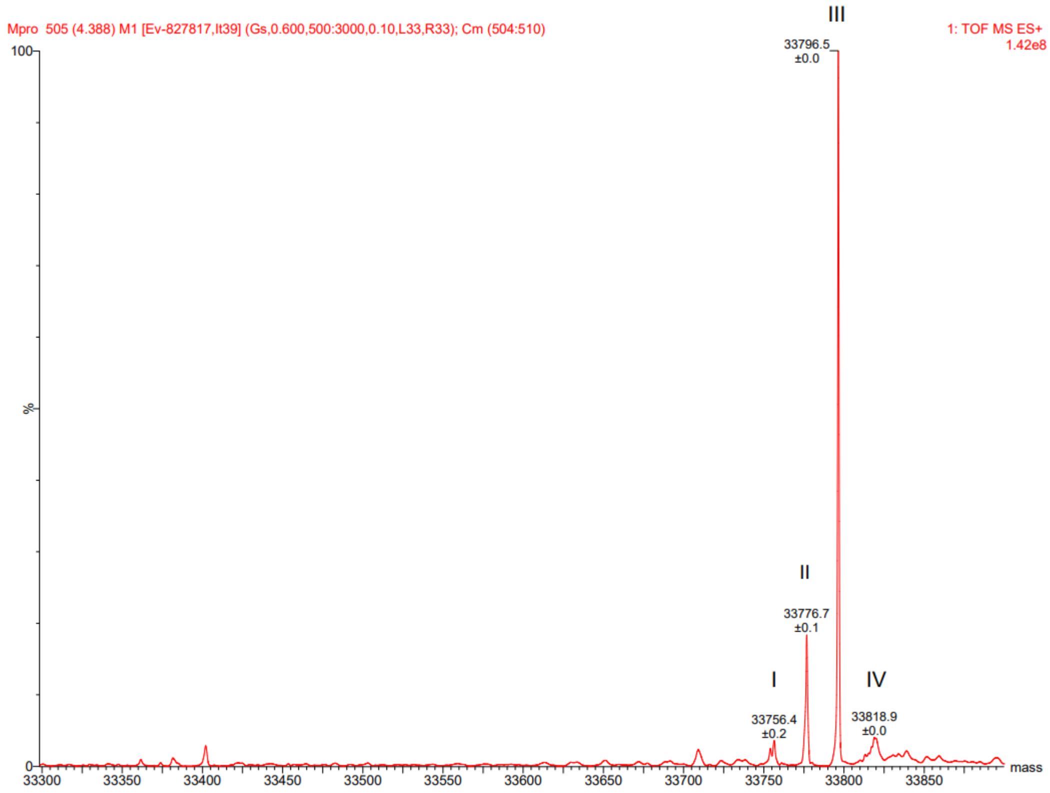


Figure 9.9: Zero-charge mass spectrum of anaerobically purified M^{pro}.

Instead peaks for two dehydrated species can be found (I, II), a peak for the average mass (III) and a sodium adduct (IV). For better comparability all samples,their major peaks and assignments are summarised in Table 9.2. Lastly, to confirm that the previous peak assignment for the M_{pro} sample without reducing agents is correct and really belongs to oxidative species, M^{pro} was incubated with H_2O_2, a strong oxidising agent, to obtain the fully oxidised protein. Figure 9.10 shows the deconvoluted zero-charge mass spectrum of M^{pro} incubated with 1 mM H_2O_2.

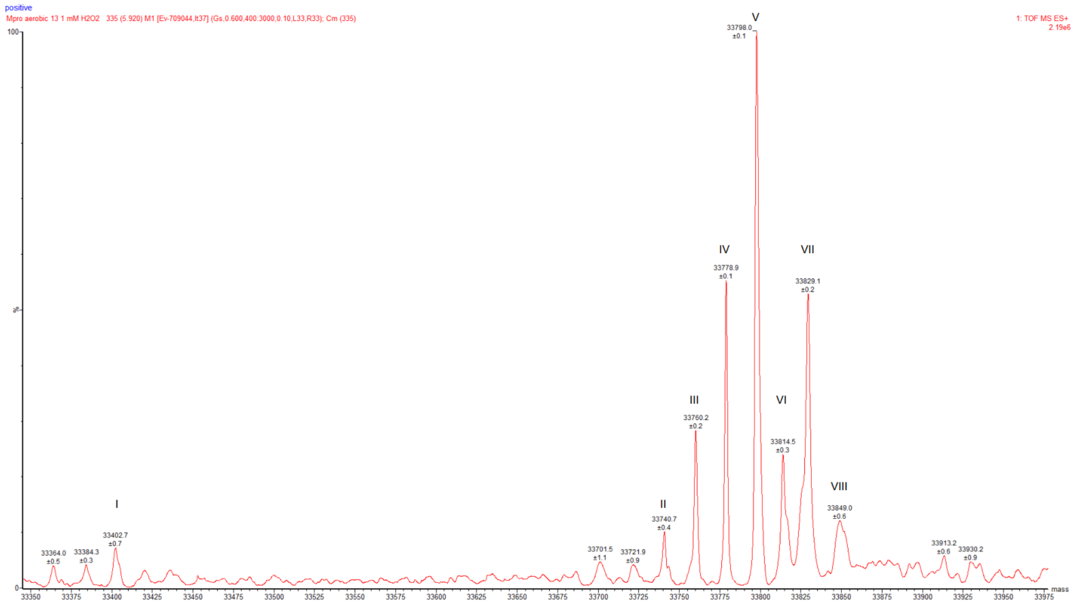


Figure 9.10: Zero-charge mass spectrum of M^{pro} treated with 1 mM hydrogen peroxide.

The main peak (V) corresponds to the expected average weight for unoxidised M^{pro} with a 1.36 Da difference, indicating an additional proton. To higher masses three peaks can be observed (VI - VIII). These peaks have the same mass as the the previously observed oxidised species, confirming the interpretation.

Below the average mass of M^{pro} three smaller peaks can be observed (II - IV) with decreasing abundance indicating the presence of dehydrated species as well as a fourth peak (I) representing the loss of the three N-terminal residues.

This leads to the conclusion that M^{pro} is not oxidised during the purification as long as reducing agents are present. However, during a prolonged crystallisation period the effectiveness of reducing agents will diminish over time gradually increasing the oxidation pressure for the enzyme. Additionally, other buffer components, such as PEG, have been shown to degrade when in exposed to light or when solubilised, creating peroxides and aldehydes that produce an oxidative environment[158].

Table 9.2: List of all major mass peaks with possible modifications.

M^{pro} treated with 1 mM H_2O_2			
Observed peak mass [Da]	Possible modification	Expected mass [Da]	Mass error [Da]
33402.07 ± 0.7	−TFQ	33402.2	−0.1
33740.7 ± 0.4	$-3\,H_2O$	33742.6	−1.9
33760.2 ± 0.2	$-2\,H_2O$	33760.6	−0.4
33778.9 ± 0.1	$-H_2O$	33778.6	+0.3
33798.0 ± 0.1	+H	33797.6	+0.4
33814.5 ± 0.3	$+H_2O$	33814.6	−0.1
33829.1 ± 0.2	+2 O + H	33828.6	−0.5
33849.0 ± 0.6	+3 O	33844.6	+4.4
Aerobically purified M^{pro} without reducing agents			
Observed peak mass [Da]	Possible modification	Expected mass [Da]	Mass error [Da]
33402.7 ± 0.1	−TFQ	33402.2	+0.5
33757.4 ± 0.4	$-2\,H_2O$	33760.6	−3.2
33778.1 ± 0.1	$-H_2O$	33778.6	−0.5
33797.7 ± 0.1	+H	33797.6	+0.1
33812.6 ± 0.3	+O	33812.6	0.0
33829.3 ± 0.2	+2 O + H	33828.6	−0.3
33859.5 ± 0.8	+4 O	33860.6	−1.1
Aerobically purified M^{pro}			
Observed peak mass [Da]	Possible modification	Expected mass [Da]	Mass error [Da]
33776.3 ± 0.1	$-H_2O$	33778.6	−2.3
33796.6 ± 0.0	av. mass	33796.6	0.0
Anaerobically purified M^{pro}			
Observed peak mass [Da]	Possible modification	Expected mass [Da]	Mass error [Da]
33756.4 ± 0.2	$-2\,H_2O$	33760.6	−4.2
33776.7 ± 0.1	$-H_2O$	33778.6	−2.1
33796.5 ± 0.0	av. mass	33796.6	−0.1
33818.9 ± 0.0	+Na	33718.6	+0.3

9.5 Activity assay

To determine how potential active site modifications affect the binding behaviour of substrates and the catalytic turnover of the enzyme an activity assay was used. The proteolytic activity of aerobically and anaerobically purified M^{pro} was examined using the tetrapeptide Ac-Abu-Tle-Leu-Gln-AMC[159] which contains two non-natural amino acids (Abu is L-2-aminobutyric acid; Tle is L-tert-leucine) and mimics a natural substrate but possesses higher binding affinity (Figure 9.11). The oligopeptide has a C-terminal amide link to a fluorescent dye (7-amino-4-methylcoumarin, AMC) which is released upon cleavage by M^{pro}.

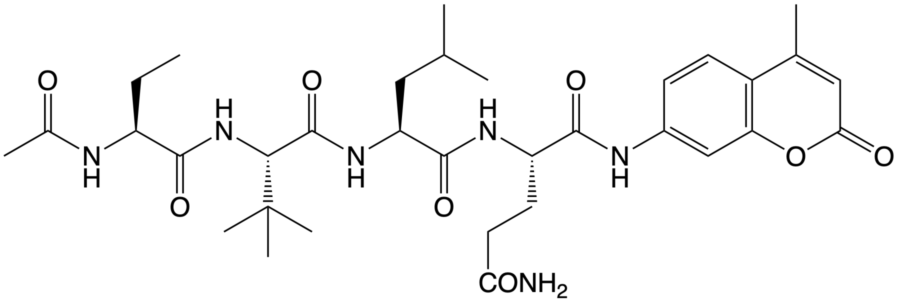


Figure 9.11: Chemical structure of the synthetic substrate peptide Ac-Abu-Tle-Leu-Gln-AMC with enhanced binding affinity to M^{pro}. A fluorescent AMC-group is attached to the oligopeptide via an amide link which is cleaved upon processing by M^{pro}.

The release of the AMC group generates a fluorescent signal (Ex 380 nm / Em 455 nm) and a change in the UV/Vis spectrum. This process was followed spectroscopically by recording the change of UV/Vis absorption at 380 nm (Figure 9.12).

Initial reaction velocities were calculated over the first two minutes by linear regression and converted to $[\mu M \cdot s^{-1}]$ with a molar extinction coefficient of $\epsilon = 19.000\ M^{-1}cm^{-1}$ for AMC. Each experiment was done in duplicate and repeated for different substrate concentrations ranging from 1 µM to 850 µM.

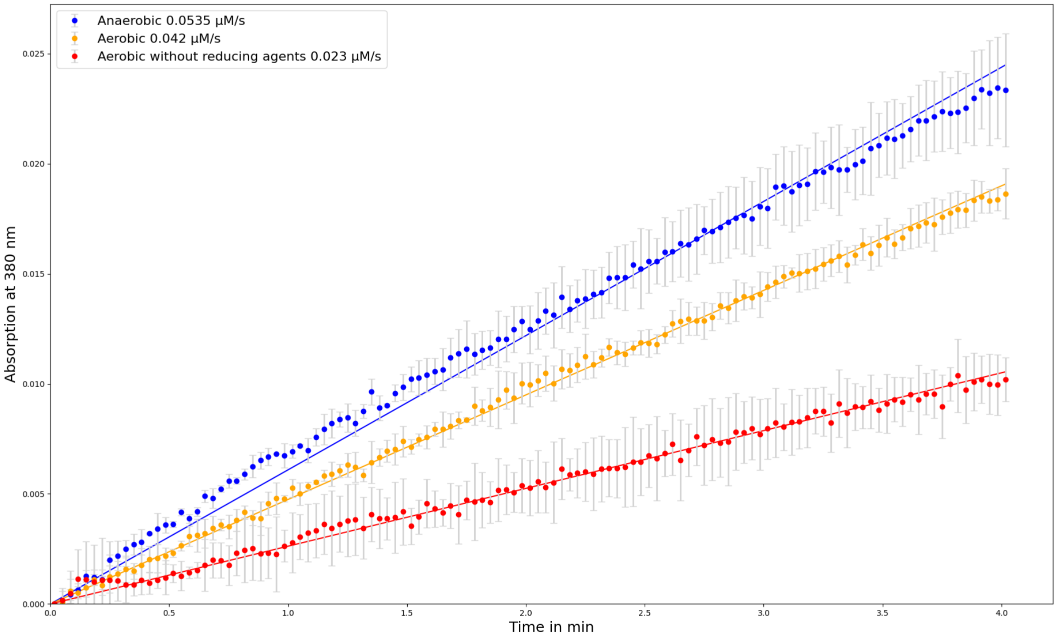


Figure 9.12: Absorption change at 380 nm produced by the release of AMC in a reaction mix of 200 µM Ac-Abu-Tle-Leu-Gln-AMC with 1 µM M$^{\text{pro}}$.

Above 850 µM the substrate became insoluble. Figure 9.13 shows the initial reaction velocity plotted against the substrate concentration.

The Hill-Equation (Equation 7.4) was used to fit the parameters $K_{0.5}$, n and V_{max} to the data. It can be seen that the anaerobic and aerobic data points just barely started to leave the region where they behave approximately linear. When no reducing agents are used all data points are within the linear region. Rut et al.[159] pointed out that SARS-CoV-2 M$^{\text{pro}}$ exhibits low activity toward tetrapeptide substrates which explains the relatively weak binding of Ac-Abu-Tle-Leu-Gln-AMC to M$^{\text{pro}}$ and the high substrate concentrations necessary to reach saturation.

Since the data points do not sufficiently sample the upper substrate concentration range where the curve is expected to reach saturation, $K_{0.5}$ and V_{max} become artificially large. This is particularly obvious for the kinetic parameters obtained for the M$^{\text{pro}}$ sample that has been purified aerobically without reducing agents (summarised in Table 9.3). Therefore, here the focus will be on the linear region. Within that region the slope is equal to $V_{max}/K_{0.5}$ from which the specificity constant $k_{cat}/K_{0.5}$ can be calculated by substituting V_{max} with

$$V_{max} = k_{cat} \cdot [E_0] \tag{9.1}$$

which leads to:

$$k_{cat}/K_{0.5} = slope/[E_0] \tag{9.2}$$

This yields a specificity constant of 526.02 $\mathrm{M^{-1}s^{-1}}$ and 524.85 $\mathrm{M^{-1}s^{-1}}$ for the anaerobically and aerobically purified M$^{\text{pro}}$. Despite being smaller, the obtained values for the specifity constant are generally in the same ballpark as the previously reported specificity constant of $859 \pm 57\,\mathrm{M^{-1}s^{-1}}$ for Ac-Abu-Tle-Leu-Gln-ACC[159]. When no reducing agents are used the specificity constant decreases to 223.44 $\mathrm{M^{-1}s^{-1}}$ which represents a reduction by approximately a factor of 2 in comparison to the other two samples.

These results are in agreement with the previous measurements and support the hypothesis that M$^{\text{pro}}$ is not oxidised as long as reducing agents are present but as soon as their effectiveness starts to decay (e.g. during crystallisation).

Table 9.3: Kinetic parameters of M$^{\text{pro}}$ determined by fitting the Hill-equation to the data shown in Figure 9.13.

	$K_{0.5}$[µM]	V_{max}[µM $\cdot s^{-1}$]	$k_{cat}/K_{0.5}$[$\mathrm{M^{-1}s^{-1}}$]	Hill coefficient n
Anaerobic	1175.81	0.78	526.02	1.06
Aerobic	719.6	0.497	524.85	1.04
Aerobic w/o reducing agents	1612919414.86	44476.05	223.44	0.87

Interesting is also that the Hill coefficient for the aerobic and anaerobic sample suggests positively cooperative binding, even though it is weak. In contrast, when no reducing agents are used the Hill coefficient decreases below 1.0, indicating negatively cooperative binding. In other words, binding of the first substrate molecule by protomer 1 negatively affects the binding affinity of protomer 2. Considering the mass spectrum which showed a mixture of non-, mono- and di-oxidised species and the refined occupancies from the X-ray data it can be estimated that more than 50 % of the enzyme molecules are oxidised. Consequently each dimer is very likely to contain at least one oxidised protomer. Therefore, the decreased binding affinity for the second substrate can be explained by the presence of an oxidative modification in the second protomer.

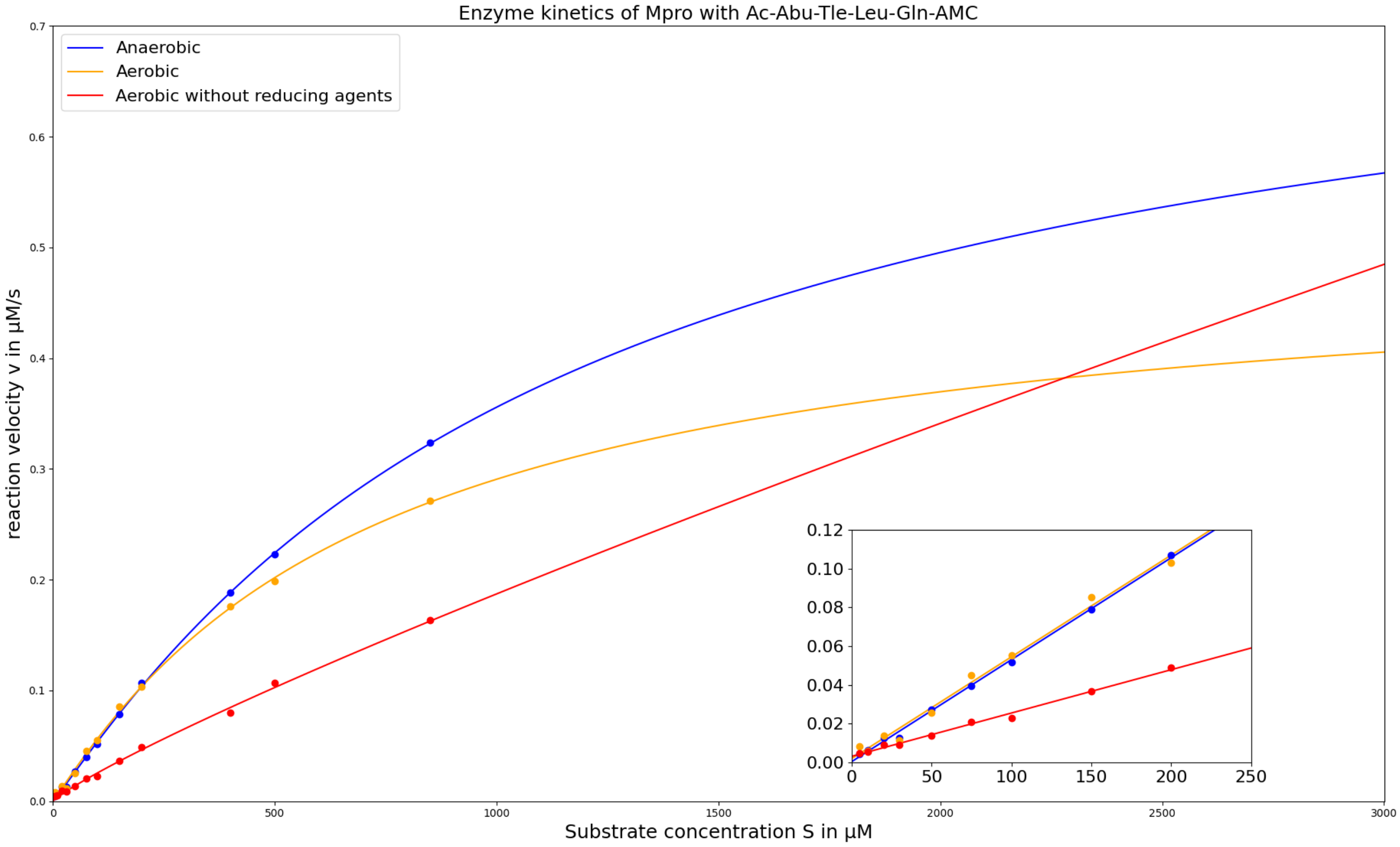


Figure 9.13: Enzyme kinetics of M^{pro} with Ac-Abu-Tle-Leu-Gln-AMC. Measurements were done in duplicates. The insert shows a magnification of the lower substrate concentration range.

Chapter 10

Conclusion

This study showed that many published M$^{\mathrm{pro}}$ structures have unexplained positive difference density around the the active site cysteines S_γ atom which indicate a post-translational modification (Table 9.1). The fact that this unmodelled density can be observed in several independent experiments rules out the possibility of it being noise and further shows that this modification has gone unnoticed by many studies.

With the aim to determine whether this modification is an artefact of the purification strategy, M$^{\mathrm{pro}}$ was purified under aerobic conditions (as reported by most studies), aerobic conditions without the use of reducing agents and anaerobic conditions. Initial crystals clusters were successively optimised by microseeding the crystallisation drops to yield large, well diffracting ($\sim$ 1.6 Å), single crystals (Figure 9.3). Notably, the sample that had been purified without reducing agents yielded much smaller, more heterogeneous and worse diffracting crystals ($\sim$ 2.6 Å). Analysis of X-ray diffraction data revealed that both aerobic samples showed signs of oxidation around the active site cysteine (Figure 9.5 and 9.6). Interestingly, with mass spectrometry this could only be confirmed for the aerobic preparation with no reducing agents (Table 9.2) showing the presence of sulfenic ($-$SO) and sulfinic acid ($-$SO$_2$)

modifications, the later of which is considered to be irreversible. The second aerobic sample (where reducing agents were used) must therefore have been oxidised after the purification, e.g. during crystallisation. This is supported by activity measurements which showed that an aerobic purification with reducing agents and an anaerobic purification strategy yield M^{pro} with almost similar specificity constants for Ac-Abu-Tle-Leu-Gln-AMC (Table 9.3). In contrast when no reducing agents are used in an aerobic purification the enzyme is oxidised over 12 days by ambient oxygen causing a reduction of substrate specificity by approximately 50 %. Twelve days is also a reasonable estimate for the time it would take from the setup of the crystallisation drop to eventually harvesting the crystal for a suitable beamtime. During that period the crystallisation buffer components such as PEG and reducing agents may decay, creating an increasingly oxidising environment, which explains why the aerobic M^{pro} structure presented here as well as many published structures show signs of oxidation despite the fact that reducing agents were used during the purification.

Although accurate kinetic constants could not be determined, this study showed that the oxidation of M^{pro}'s active site Cys145 has a strong impact on its enzymatic behaviour (Figure 9.13). The presented data suggests that the oxidation only occurs after the purification when reducing agents are used or an anaerobic purification strategy is employed. In consequence, activity and binding experiments of ongoing drug development campaigns are not expected to be affected by potential oxidative modifications when solubilised enzyme is used and the presence of fresh reducing agents is maintained. However crystal based experiments, especially soaking experiments with "old" crystals (older that 12 days but likely also before), may significantly be affected leading to wrong conclusions regarding the binding effectiveness of potential active site inhibitors. M^{pro} protein structures in complex with ligands are therefore best obtained by co-crystallisation or by soaking experiments under anaerobic conditions.

Bibliography

1. Howell, S. Resistance to apoptosis in prostate cancer cells. *Molecular urology* **4,** 225–9 (2000).
2. Ott, E. *et al.* Molecular repertoire of Deinococcus radiodurans after 1 year of exposure outside the International Space Station within the Tanpopo mission. *Microbiome* **8,** 1–16 (2020).
3. On the Forward Contamination of Europa, T. G., Board, S. S., Council, N. R., *et al. Preventing the forward contamination of Europa* (National Academies Press, 1900).
4. Brim, H. *et al.* Engineering Deinococcus radiodurans for metal remediation in radioactive mixed waste environments. *Nature biotechnology* **18,** 85–90 (2000).
5. Lange, C. C., Wackett, L. P., Minton, K. W. & Daly, M. J. Engineering a recombinant Deinococcus radiodurans for organopollutant degradation in radioactive mixed waste environments. *Nature biotechnology* **16,** 929–933 (1998).
6. Garman, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D: Biological Crystallography* **66,** 339–351 (2010).
7. Cox, M. M. & Battista, J. R. Deinococcus radiodurans—the consummate survivor. *Nature Reviews Microbiology* **3,** 882–892 (2005).

8. Sharma, A. *et al.* Across the tree of life, radiation resistance is governed by antioxidant Mn2+, gauged by paramagnetic resonance. *Proceedings of the National Academy of Sciences* **114,** E9253–E9260 (2017).
9. Daly, M. J. *et al.* Accumulation of Mn (II) in Deinococcus radiodurans facilitates gamma-radiation resistance. *Science* **306,** 1025–1028 (2004).
10. Slade, D. & Radman, M. Oxidative stress resistance in Deinococcus radiodurans. *Microbiology and molecular biology reviews* **75,** 133–191 (2011).
11. Daly, M. J. A new perspective on radiation resistance based on Deinococcus radiodurans. *Nature Reviews Microbiology* **7,** 237–245 (2009).
12. Bury, C. *et al.* Radiation damage to nucleoprotein complexes in macromolecular crystallography. *Journal of synchrotron radiation* **22,** 213–224 (2015).
13. Ravelli, R. B. & McSweeney, S. M. The 'fingerprint'that X-rays can leave on structures. *Structure* **8,** 315–328 (2000).
14. Zeldin, O. B., Gerstel, M. & Garman, E. F. RADDOSE-3D: time-and space-resolved modelling of dose in macromolecular crystallography. *Journal of applied crystallography* **46,** 1225–1230 (2013).
15. Bury, C. S., Brooks-Bartlett, J. C., Walsh, S. P. & Garman, E. F. Estimate your dose: RADDOSE-3D. *Protein Science* **27,** 217–228 (2018).
16. Weik, M. *et al.* Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences* **97,** 623–628 (2000).
17. Burmeister, W. P. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallographica Section D: Biological Crystallography* **56,** 328–341 (2000).
18. Nowak, E., Brzuszkiewicz, A., Dauter, M., Dauter, Z. & Rosenbaum, G. To scavenge or not to scavenge: that is the question. *Acta Crystallographica Section D: Biological Crystallography* **65,** 1004–1006 (2009).

19. Allan, E. G., Kander, M. C., Carmichael, I. & Garman, E. F. To scavenge or not to scavenge, that is STILL the question. *Journal of synchrotron radiation* **20,** 23–36 (2013).

20. *Number of Released PDB Structures per Year* `https://www.rcsb.org/stats/all-released-structures`.

21. Raimondi, P. ESRF-EBS: The extremely brilliant source project. *Synchrotron Radiation News* **29,** 8–15 (2016).

22. Schroer, C. G. *et al.* The synchrotron radiation source PETRA III and its future ultra-low-emittance upgrade PETRA IV. *The European Physical Journal Plus* **137,** 1312 (2022).

23. Fischer, M. Macromolecular room temperature crystallography. *Quarterly Reviews of Biophysics* **54,** e1 (2021).

24. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462,** 669–673 (2009).

25. Keedy, D. A. *et al.* Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR. *Structure* **22,** 899–910 (2014).

26. Gotthard, G. *et al.* Specific radiation damage is a lesser concern at room temperature. *IUCrJ* **6,** 665–680 (2019).

27. Pérez-González, A., Alvarez-Idaboy, J. R. & Galano, A. Free-radical scavenging by tryptophan and its metabolites through electron transfer based processes. *Journal of Molecular Modeling* **21,** 1–11 (2015).

28. O'Neill, P., Stevens, D. L. & Garman, E. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. *Journal of synchrotron radiation* **9,** 329–332 (2002).

29. Ďurovič, S., Krishna, P. & Pandey, D. in *International Tables for Crystallography* 213 (Kluwer Academic Publishers, Dordrecht, The Netherlands, Oct. 2004).

30. Dickerson, J. L. & Garman, E. F. The potential benefits of using higher X-ray energies for macromolecular crystallography. *Journal of Synchrotron Radiation* **26,** 922–930 (2019).

31. Jones, G. D., Lea, J. S., Symons, M. C. & Taiwo, F. A. Structure and mobility of electron gain and loss centres in proteins. *Nature* **330,** 772–773 (1987).

32. Symons, M. C. Electron movement through proteins and DNA. *Free Radical Biology and Medicine* **22,** 1271–1276 (1997).

33. Teng, T. y. & Moffat, K. Primary radiation damage of protein crystals by an intense synchrotron X-ray beam. *Journal of synchrotron radiation* **7,** 313–317 (2000).

34. Henderson, R. Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **241,** 6–8 (1990).

35. Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences* **103,** 4912–4917 (2006).

36. Beitlich, T., Kühnel, K., Schulze-Briese, C., Shoeman, R. L. & Schlichting, I. Cryoradiolytic reduction of crystalline heme proteins: analysis by UV-Vis spectroscopy and X-ray crystallography. *Journal of synchrotron radiation* **14,** 11–23 (2007).

37. Yano, J. *et al.* X-ray damage to the Mn4Ca complex in single crystals of photosystem II: a case study for metalloprotein crystallography. *Proceedings of the National Academy of Sciences* **102,** 12047–12052 (2005).

38. Weik, M. *et al.* Evidence for the formation of disulfide radicals in protein crystals upon X-ray irradiation. *Journal of Synchrotron Radiation* **9,** 342–346 (2002).

39. Meents, A., Gutmann, S., Wagner, A. & Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proceedings of the National Academy of Sciences* **107,** 1094–1099 (2010).
40. Garman, E. F. & Weik, M. Radiation damage in macromolecular crystallography. *Protein Crystallography,* 467–489 (2017).
41. De la Mora, E., Carmichael, I. & Garman, E. F. Effective scavenging at cryotemperatures: further increasing the dose tolerance of protein crystals. *Journal of synchrotron radiation* **18,** 346–357 (2011).
42. Garman, E. F. & Schneider, T. R. Macromolecular cryocrystallography. *Journal of Applied Crystallography* **30,** 211–237 (1997).
43. Wolff, A. M. *et al.* Comparing serial X-ray crystallography and microcrystal electron diffraction (MicroED) as methods for routine structure determination from small macromolecular crystals. *IUCrJ* **7,** 306–323 (2020).
44. Kwon, H. *et al.* XFEL crystal structures of peroxidase compound II. *Angewandte Chemie International Edition* **60,** 14578–14585 (2021).
45. Hirata, K. *et al.* Determination of damage-free crystal structure of an X-ray–sensitive protein using an XFEL. *Nature methods* **11,** 734–736 (2014).
46. Zaloga, G. & Sarma, R. New method for extending the diffraction pattern from protein crystals and preventing their radiation damage. *Nature* **251,** 551–552 (1974).
47. Roots, R. & Okada, S. Estimation of life times and diffusion distances of radicals involved in X-ray-induced DNA strand breaks or killing of mammalian cells. *Radiation research* **64,** 306–320 (1975).
48. Holton, J. M. A beginner's guide to radiation damage. *Journal of synchrotron radiation* **16,** 133–142 (2009).

49. Murray, J. & Garman, E. Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography. *Journal of synchrotron radiation* **9,** 347–354 (2002).

50. Kauffmann, B., Weiss, M. S., Lamzin, V. S. & Schmidt, A. How to avoid premature decay of your macromolecular crystal: a quick soak for long life. *Structure* **14,** 1099–1105 (2006).

51. Barker, A. I., Southworth-Davies, R. J., Paithankar, K. S., Carmichael, I. & Garman, E. F. Room-temperature scavengers for macromolecular crystallography: increased lifetimes and modified dose dependence of the intensity decay. *Journal of synchrotron radiation* **16,** 205–216 (2009).

52. Liebschner, D., Rosenbaum, G., Dauter, M. & Dauter, Z. Radiation decay of thaumatin crystals at three X-ray energies. *Acta Crystallographica Section D: Biological Crystallography* **71,** 772–778 (2015).

53. Yabukarski, F., Doukov, T., Mokhtari, D. A., Du, S. & Herschlag, D. Evaluating the impact of X-ray damage on conformational heterogeneity in room-temperature (277 K) and cryo-cooled protein crystals. *Acta Crystallographica Section D: Structural Biology* **78** (2022).

54. Adman, E. T., Godden, J. & Turley, S. The structure of copper-nitrite reductase from achromobacter cycloclastes at five pH values, with NO-2 bound and with type II copper depleted. *Journal of Biological Chemistry* **270,** 27458–27474 (1995).

55. Sen, K. *et al.* Active-site protein dynamics and solvent accessibility in native Achromobacter cycloclastes copper nitrite reductase. *IUCrJ* **4,** 495–505 (2017).

56. Constable, E. *Comprehensive coordination chemistry II: from biology to nanotechnology* (Newnes, 2003).

57. Schafmeister, C. E., LaPorte, S. L., Miercke, L. J. & Stroud, R. M. A designed four helix bundle protein with native-like structure. *Nature structural biology* **4,** 1039–1046 (1997).

58. Miyamura, Y. *et al.* Regulation of human skin pigmentation and responses to ultraviolet radiation. *Pigment Cell Research* **20,** 2–13 (2007).

59. Radiative Relaxation Quantum Yields for Synthetic Eumelanin¶.

60. Reiter, R. J. *et al.* Melatonin as an antioxidant: under promises but over delivers. *Journal of pineal research* **61,** 253–278 (2016).

61. Mattimore, V. & Battista, J. R. Radioresistance of Deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *Journal of bacteriology* **178,** 633–637 (1996).

62. Musilova, M., Wright, G., Ward, J. M. & Dartnell, L. R. Isolation of radiation-resistant bacteria from Mars analog Antarctic Dry Valleys by preselection, and the correlation between radiation and desiccation resistance. *Astrobiology* **15,** 1076–1090 (2015).

63. Anderson, A. Studies on a radio-resistant micrococcus. I. Isolation, morphology, cultural characteristics, and resistance to gamma radiation. *Food Technol* **10,** 575–578 (1956).

64. Gerard, E., Jolivet, E., Prieur, D. & Forterre, P. DNA protection mechanisms are not involved in the radioresistance of the hyperthermophilic archaea Pyrococcus abyssi and P. furiosus. *Molecular Genetics and Genomics* **266,** 72–78 (2001).

65. Battista, J. R. Against all odds: the survival strategies of Deinococcus radiodurans. *Annual review of microbiology* **51,** 203–224 (1997).

66. Lu, H. *et al.* Deinococcus radiodurans PprI switches on DNA damage response and cellular survival networks after radiation damage. *Molecular & Cellular Proteomics* **8,** 481–494 (2009).

67. Blasius, M., Hübscher, U. & Sommer, S. Deinococcus radiodurans: what belongs to the survival kit? *Critical reviews in biochemistry and molecular biology* **43,** 221–238 (2008).

68. Krisko, A. & Radman, M. Protein damage and death by radiation in Escherichia coli and Deinococcus radiodurans. *Proceedings of the National Academy of Sciences* **107,** 14373–14377 (2010).
69. Krug, M., Weiss, M. S., Heinemann, U. & Mueller, U. XDSAPP: a graphical user interface for the convenient processing of diffraction data using XDS. *Journal of Applied Crystallography* **45,** 568–572 (2012).
70. Potterton, L. *et al.* CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallographica Section D: Structural Biology* **74,** 68–84 (2018).
71. Vagin, A. & Teplyakov, A. MOLREP: an automated program for molecular replacement. *Journal of applied crystallography* **30,** 1022–1025 (1997).
72. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography* **66,** 486–501 (2010).
73. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography* **67,** 355–367 (2011).
74. *Movies* `https://bl831.als.lbl.gov/~jamesh/movies/`.
75. Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47,** D506–D515 (2019).
76. Ginn, H. M. Pre-clustering data sets using cluster4x improves the signal-to-noise ratio of high-throughput crystallography drug-screening analysis. *Acta Crystallographica Section D: Structural Biology* **76,** 1134–1144 (2020).
77. Knuth, K. H. Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing* **95,** 102581 (2019).

78. Pavlopoulou, A. *et al.* Unraveling the mechanisms of extreme radioresistance in prokaryotes: lessons from nature. *Mutation Research/Reviews in Mutation Research* **767,** 92–107 (2016).

79. Chan-Yeung, M. & Xu, R.-H. SARS: epidemiology. *Respirology* **8,** S9–S14 (2003).

80. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet* **395,** 565–574 (2020).

81. Al-Omari, A., Rabaan, A. A., Salih, S., Al-Tawfiq, J. A. & Memish, Z. A. MERS coronavirus outbreak: Implications for emerging viral infections. *Diagnostic microbiology and infectious disease* **93,** 265–285 (2019).

82. Mohammed, I. *et al.* The efficacy and effectiveness of the COVID-19 vaccines in reducing infection, severity, hospitalization, and mortality: A systematic review. *Human vaccines & immunotherapeutics* **18,** 2027160 (2022).

83. *COVID-19 Vaccine* `https://www.pei.de/EN/medicinal-products/vaccines-human/covid-19/covid-19-node.html?cms_gts=251098_list%253DdateOfIssue_dt%252Basc`.

84. Sharma, O., Sultan, A. A., Ding, H. & Triggle, C. R. A Review of the Progress and Challenges of Developing a Vaccine for COVID-19. *Frontiers in immunology* **11,** 585354 (2020).

85. Moghadas, S. M. *et al.* The impact of vaccination on coronavirus disease 2019 (COVID-19) outbreaks in the United States. *Clinical Infectious Diseases* **73,** 2257–2264 (2021).

86. Pritchard, E. *et al.* Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom. *Nature medicine* **27,** 1370–1378 (2021).

87. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582,** 289–293 (2020).

88. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372,** 642–646 (2021).
89. Owen, D. R. *et al.* An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science* **374,** 1586–1593 (2021).
90. Joshi, R. S. *et al.* Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. *Journal of Biomolecular Structure and Dynamics* **39,** 3099–3114 (2021).
91. Banerjee, R., Perera, L. & Tillekeratne, L. V. Potential SARS-CoV-2 main protease inhibitors. *Drug Discovery Today* **26,** 804–816 (2021).
92. Rut, W. *et al.* Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: A framework for anti–COVID-19 drug design. *Science advances* **6,** eabd4596 (2020).
93. Osipiuk, J. *et al.* Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nature communications* **12,** 743 (2021).
94. Zhao, Y. *et al.* High-throughput screening identifies established drugs as SARS-CoV-2 PLpro inhibitors. *Protein & cell* **12,** 877–888 (2021).
95. Ma, C. *et al.* Discovery of SARS-CoV-2 papain-like protease inhibitors through a combination of high-throughput screening and a FlipGFP-based reporter assay. *ACS central science* **7,** 1245–1260 (2021).
96. Macip, G. *et al.* Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Medicinal Research Reviews* **42,** 744–769 (2022).
97. Glasziou, P. P., Sanders, S. & Hoffmann, T. *Waste in covid-19 research* 2020.
98. Palayew, A. *et al.* Pandemic publishing poses a new COVID-19 challenge. *Nature Human Behaviour* **4,** 666–669 (2020).
99. Odone, A., Galea, S., Stuckler, D., Signorelli, C. & literature monitoring working group Amerio Andrea MD PhD Bellini Lorenzo MD Bucci

Daria Capraro Michele MD Gaetti Giovanni MD Salvati Stefano MD, U. V.-S. S. R. C.-1. The first 10 000 COVID-19 papers in perspective: are we publishing what we should be publishing? *European journal of public health* **30,** 849–850 (2020).

100. Kneller, D. W. *et al.* Room-temperature X-ray crystallography reveals the oxidation and reactivity of cysteine residues in SARS-CoV-2 3CL Mpro: insights into enzyme mechanism and drug design. *IUCrJ* **7** (2020).
101. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *The EMBO journal* **21,** 3213–3224 (2002).
102. Wensien, M. *et al.* A lysine–cysteine redox switch with an NOS bridge regulates enzyme function. *Nature* **593,** 460–464 (2021).
103. Ravanfar, R. *et al.* Surface cysteines could protect the SARS-CoV-2 main protease from oxidative damage. *Journal of Inorganic Biochemistry* **234,** 111886 (2022).
104. Drenth, J., Kalk, K. & Swen, H. Binding of chloromethyl ketone substrate analogs to crystalline papain. *Biochemistry* **15,** 3731–3738 (1976).
105. Ozono, S. *et al.* SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nature communications* **12,** 848 (2021).
106. *Biorender* `https://www.biorender.com/`.
107. Shulla, A. *et al.* A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry. *Journal of virology* **85,** 873–882 (2011).
108. Matsuyama, S. *et al.* Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. *Journal of virology* **84,** 12658–12664 (2010).

109. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nature reviews Molecular cell biology* **23,** 3–20 (2022).

110. Michel, C. J., Mayer, C., Poch, O. & Thompson, J. D. Characterization of accessory genes in coronavirus genomes. *Virology journal* **17,** 1–13 (2020).

111. Prydz, K. & Saraste, J. The life cycle and enigmatic egress of coronaviruses. *Molecular microbiology* **117,** 1308–1316 (2022).

112. Cortese, M. *et al.* Integrative imaging reveals SARS-CoV-2-induced reshaping of subcellular morphologies. *Cell host & microbe* **28,** 853–866 (2020).

113. Eymieux, S. *et al.* Ultrastructural modifications induced by SARS-CoV-2 in Vero cells: a kinetic analysis of viral factory formation, viral particle morphogenesis and virion release. *Cellular and Molecular Life Sciences* **78,** 3565–3576 (2021).

114. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300,** 1763–1767 (2003).

115. Tomasello, G., Armenia, I. & Molla, G. The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36,** 2909–2911 (2020).

116. Lee, J. T. *et al.* Genetic surveillance of SARS-CoV-2 Mpro reveals high sequence and structural conservation prior to the introduction of protease inhibitor Paxlovid. *Mbio* **13,** e00869–22 (2022).

117. Cho, E. *et al.* Dynamic profiling of β-coronavirus 3CL Mpro protease ligand-binding sites. *Journal of Chemical Information and Modeling* **61,** 3058–3073 (2021).

118. Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences* **100,** 13190–13195 (2003).

119. Shi, J. & Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *The FEBS journal* **273,** 1035–1045 (2006).

120. Shi, J., Sivaraman, J. & Song, J. Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease. *Journal of virology* **82,** 4620–4629 (2008).

121. Chou, C.-Y. *et al.* Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. *Biochemistry* **43,** 14958–14970 (2004).

122. Chen, S. *et al.* Severe acute respiratory syndrome coronavirus 3C-like proteinase N terminus is indispensable for proteolytic activity but not for enzyme dimerization: biochemical and thermodynamic investigation in conjunction with molecular dynamics simulations. *Journal of Biological Chemistry* **280,** 164–173 (2005).

123. Xue, X. *et al.* Production of authentic SARS-CoV Mpro with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. *Journal of molecular biology* **366,** 965–975 (2007).

124. Swiderek, K. & Moliner, V. Revealing the molecular mechanisms of proteolysis of SARS-CoV-2 M pro by QM/MM computational methods. *Chemical Science* **11,** 10626–10630 (2020).

125. Ramos-Guzmán, C. A., Ruiz-Pernía, J. J. & Tuñón, I. Unraveling the SARS-CoV-2 main protease mechanism using multiscale methods. *ACS catalysis* **10,** 12544–12554 (2020).

126. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorganic & medicinal chemistry letters* **30,** 127377 (2020).

127. Inizan, T. J. *et al.* High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chemical Science* **12,** 4889–4907 (2021).

128. Zhou, X. *et al.* Structure of SARS-CoV-2 main protease in the apo state. *Science China Life Sciences* **64,** 656–659 (2021).

129. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368,** 409–412 (2020).

130. Singh, J., Petter, R. C., Baillie, T. A. & Whitty, A. The resurgence of covalent drugs. *Nature reviews Drug discovery* **10,** 307–317 (2011).

131. Steuten, K. *et al.* Challenges for targeting SARS-CoV-2 proteases as a therapeutic strategy for COVID-19. *ACS infectious diseases* **7,** 1457–1468 (2021).

132. Roe, M. K., Junod, N. A., Young, A. R., Beachboard, D. C. & Stobart, C. C. Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19. *The Journal of general virology* **102** (2021).

133. Chen, H. *et al.* Only one protomer is active in the dimer of SARS 3C-like proteinase. *Journal of Biological Chemistry* **281,** 13894–13898 (2006).

134. Chen, S. *et al.* Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: crystal structure with molecular dynamics simulations. *Journal of Biological Chemistry* **283,** 554–564 (2008).

135. Davis, D. A. *et al.* Regulation of the dimerization and activity of SARS-CoV-2 main protease through reversible glutathionylation of cysteine 300. *Mbio* **12,** e02094–21 (2021).

136. Donaldson, E. F., Graham, R. L., Sims, A. C., Denison, M. R. & Baric, R. S. Analysis of murine hepatitis virus strain A59 temperature-sensitive mutant TS-LA6 suggests that nsp10 plays a critical role in polyprotein processing. *Journal of virology* **81,** 7086–7098 (2007).

137. Paulsen, C. E. & Carroll, K. S. Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chemical reviews* **113,** 4633–4679 (2013).

138. Funk, L.-M. *et al.* Redox regulation of the SARS-CoV-2 main protease provides new opportunities for drug design (preprint) (2022).

139. Rabe von Pappenheim, F. *et al.* Widespread occurrence of covalent lysine–cysteine redox switches in proteins. *Nature Chemical Biology* **18,** 368–375 (2022).

140. Yang, K. S. *et al.* A Novel Y-Shaped, S–O–N–O–S-Bridged Cross-Link between Three Residues C22, C44, and K61 Is Frequently Observed in the SARS-CoV-2 Main Protease. *ACS Chemical Biology* (2023).

141. Reddie, K. G. & Carroll, K. S. Expanding the functional diversity of proteins through cysteine oxidation. *Current opinion in chemical biology* **12,** 746–754 (2008).

142. Salsbury Jr, F. R., Knutson, S. T., Poole, L. B. & Fetrow, J. S. Functional site profiling and electrostatic analysis of cysteines modifiable to cysteine sulfenic acid. *Protein Science* **17,** 299–312 (2008).

143. Claiborne, A. *et al.* Protein-sulfenic acids: diverse roles for an unlikely player in enzyme catalysis and redox regulation. *Biochemistry* **38,** 15407–15416 (1999).

144. Allison, W. S. Formation and reactions of sulfenic acids in proteins. *Accounts of chemical research* **9,** 293–299 (1976).

145. Poole, L. B., Karplus, P. A. & Claiborne, A. Protein sulfenic acids in redox signaling. *Annu. Rev. Pharmacol. Toxicol.* **44,** 325–347 (2004).

146. Cho, S.-H. *et al.* Redox regulation of PTEN and protein tyrosine phosphatases in H2O2-mediated cell signaling. *FEBS letters* **560,** 7–13 (2004).

147. Tonks, N. K. Redox redux: revisiting PTPs and the control of cell signaling. *Cell* **121,** 667–670 (2005).

148. Hamann, M., Zhang, T., Hendrich, S. & Thomas, J. A. in *Methods in enzymology* 146–156 (Elsevier, 2002).

149. Stadtman, E. R. & Levine, R. L. Protein oxidation. *Annals of the New York Academy of Sciences* **899,** 191–208 (2000).

150. Tickle, I. *et al. Staraniso (Global Phasing Ltd, Cambridge, UK)* 2016.

151. Hazes, B. *SFTOOLS (CCP4: Supported Program)*

152. Dodson, E. *CAD (CCP4: Supported Program)*

153. Huang, C., Wei, P., Fan, K., Liu, Y. & Lai, L. 3C-like proteinase from SARS coronavirus catalyzes substrate hydrolysis by a general base mechanism. *Biochemistry* **43,** 4568–4574 (2004).

154. Solowiej, J. *et al.* Steady-state and pre-steady-state kinetic evaluation of severe acute respiratory syndrome coronavirus (SARS-CoV) 3CL-pro cysteine protease: development of an ion-pair model for catalysis. *Biochemistry* **47,** 2617–2630 (2008).

155. Ebrahim, A. *et al.* The temperature-dependent conformational ensemble of SARS-CoV-2 main protease (Mpro). *IUCrJ* **9** (2022).

156. Yang, K. S. *et al.* A quick route to multiple highly potent SARS-CoV-2 main protease inhibitors. *ChemMedChem* **16,** 942–948 (2021).

157. *A Database of Protein Post Translational Modifications* `https://www.abrf.org/delta-mass`.

158. Ray Jr, W. J. & Puvathingal, J. M. A simple procedure for removing contaminating aldehydes and peroxides from aqueous solutions of polyethylene glycols and of nonionic detergents that are based on the polyoxyethylene linkage. *Analytical biochemistry* **146,** 307–312 (1985).

159. Rut, W. *et al.* SARS-CoV-2 M pro inhibitors and activity-based probes for patient-sample imaging. *Nature chemical biology* **17,** 222–228 (2021).

Appendix A

Appendix

A.1 Data collection and refinement statistics.

Table A.1: Data collection and refinement statistics for **anaerobically** purified M^{pro}. Values for the high resolution shell are shown in parentheses.

Anaerobic SARS-CoV-2 M^{pro}	
Data collection	
Space group	3
Cell dimensions	
a,b,c (Å)	44.7, 53.8, 114.7
α, β, γ (°)	90, 101 , 90
Nominal Diffraction range (Å)	48.55 - 1.65
R_{merge} (all $I+$ & $I-$)	0.521 (2.023)
R_{pim}	0.157 (0.72)
I/σ_I	4.4 (1.4)
Completeness spherical (%)	76 (16.3)
Completeness ellipsoidal (%)	93.8 (60.8)

Continued on next page

Table A.1 – continued from previous page

Anaerobic SARS-CoV-2 M$^{\mathrm{pro}}$	
Redundancy	11.7 (8.8)
Anisotropic refinement	
Resolution (Å)	48.55 - 1.65
No. reflections	49469
Rwork / Rfree	0.195/0.256
No. atoms	
Protein	9519
Ligand/ion	24
Water	295
B-factors (Å^2)	
Protein	19.25
Ligand/ion	36.1
Water	30.1
R.m.s. deviations	
Bond lengths (Å)	0.0142
Bond angles (°)	2.139

Principal axes of the ellipsoid fitted to the diffraction cut-off surface
as direction cosines in the orthogonal basis (standard PDB convention),
in terms of reciprocal unit-cell vectors and corresponding diffraction limits (Å):

(0.8913 0.0000 0.4533)	$0.784a^* + 0.621c^*$	1.617
(0.0000 1.0000 0.0000)	b^*	1.81
(-0.4533 0.000 0.8913)	$-0.181a^* + 0.984c^*$	1.884

Eigenvectors of the overall anisotropy tensor as direction cosines in the

Continued on next page

Table A.1 – continued from previous page

Anaerobic SARS-CoV-2 M$^{\text{pro}}$		
orthogonal basis (standard PDB convention), in terms of reciprocal unit-cell vectors and corresponding Eigenvalues of overall anisotropy tensor on |F|s (Å2):		
(0.9671 0.0000 0.2545)	$0.985a^{*} + 0.171c^{*}$	16.98
(0.0000 1.0000 0.0000)	b^{*}	25.21
(-0.2545 0.0000 0.9671)	$-0.099a^{*} + 0.993c^{*}$	23.24

Table A.2: Data collection and refinement statistics for **aerobically** purified M$^{\text{pro}}$. Values for the high resolution shell are shown in parentheses.

Aerobic SARS-CoV-2 M$^{\text{pro}}$	
Data collection	
Space group	3
Cell dimensions	
a,b,c (Å)	45, 54.2, 114.9
α, β, γ (°)	90, 100.8 , 90
Nominal Diffraction range (Å)	39.092 1.165
R_{merge} (all $I+$ & $I-$)	0.245 (2.816)
R_{pim}	0.046 (0.619)
I/σ_I	10.3 (1.5)
Completeness spherical (%)	81.1 (20.5)
Completeness ellipsoidal (%)	94.6 (63.2)
Redundancy	29.9 (21.5)
Refinement	
Resolution (Å)	39.12 - 1.38
No. reflections	91583

Continued on next page

Table A.2 – continued from previous page

Aerobic SARS-CoV-2 M^{pro}	
Rwork / Rfree	0.175/0.224
No. atoms	
Protein	9795
Ligand/ion	111
Water	406
B-factors ($Å^2$)	
Protein	22.83
Ligand/ion	34.36
Water	36.7
R.m.s. deviations	
Bond lengths (Å)	0.0117
Bond angles (°)	1.531

Anisotropic refinement

Principal axes of the ellipsoid fitted to the diffraction cut-off surface as direction cosines in the orthogonal basis (standard PDB convention), in terms of reciprocal unit-cell vectors and corresponding diffraction limits (Å):

(0.9171 0.0000 0.3986)	$0.852a^* - 0.523c^*$	1.355
(0.0000 1.0000 0.0000)	b^*	1.483
(-0.3986 0.0000 0.9171)	$-0.158a^* + 0.987c^*$	1.513

Eigenvectors of the overall anisotropy tensor as direction cosines in the orthogonal basis (standard PDB convention), in terms of reciprocal unit-cell vectors and corresponding Eigenvalues of overall anisotropy tensor on |F|s (Å2):

(0.8656 0.0000 0.5008)	$0.716a^* - 0.698c^*$	16.19

Continued on next page

Table A.2 – continued from previous page

Aerobic SARS-CoV-2 M^{pro}		
(0.0000 1.0000 0.0000)	b^*	18.50
(-0.5008 0.0000 0.8656)	$0.203a^* + 0.979c^*$	20.15

Table A.3: Data collection and refinement statistics for **aerobically (w/o reducing agents)** purified M^{pro}. Values for the high resolution shell are shown in parentheses.

Aerobic (w/o reducing agents) SARS-CoV-2 M^{pro}	
Data collection	
Space group	3
Cell dimensions	
a,b,c (Å)	44.9, 53.8, 112.7
α, β, γ (°)	90, 100.6 , 90
Nominal Diffraction range (Å)	48.427 1.390
R_{merge} (all $I+$ & $I-$)	0.674 (4.442)
R_{pim}	0.210 (1.681)
I/σ_I	3.8 (1.6)
Completeness spherical (%)	70.4 (14.7)
Completeness ellipsoidal (%)	91.6 (52.6)
Redundancy	11.2 (7.9)
Refinement	
Resolution (Å)	38.68 - 2.2
No. reflections	26923
Rwork / Rfree	0.299/0.373
No. atoms	
Protein	9368
Ligand/ion	36

Continued on next page

Table A.3 – continued from previous page

Aerobic (w/o reducing agents) SARS-CoV-2 M$^{\text{pro}}$		
Water	36	
B-factors (Å^2)		
Protein	60.727	
Ligand/ion	66.6	
Water	49.8	
R.m.s. deviations		
Bond lengths (Å)	0.005	
Bond angles (°)	1.347	
Anisotropic refinement		
Principal axes of the ellipsoid fitted to the diffraction cut-off surface as direction cosines in the orthogonal basis (standard PDB convention), in terms of reciprocal unit-cell vectors and corresponding diffraction limits (Å):		
(0.9964 0.0000 -0.0846)	$0.831a^* - 0.557c^*$	2.014
(0.000 1.000 0.000)	b^*	2.175
(0.0846 0.0000 0.9964)	$0.035a^* + 0.999c^*$	2.445
Eigenvectors of the overall anisotropy tensor as direction cosines in the orthogonal basis (standard PDB convention), in terms of reciprocal unit-cell vectors and corresponding Eigenvalues of overall anisotropy tensor on |F|s (Å2):		
(0.9970 0.0000 -0.0774)	$0.838a^* - 0.546c^*$	40.86
(0.0000 1.0000 0.0000)	b^*	51.67
(0.0774 0.0000 0.9970)	$0.032a^* + 0.999c^*$	67.70

A.2 Supporting Figures

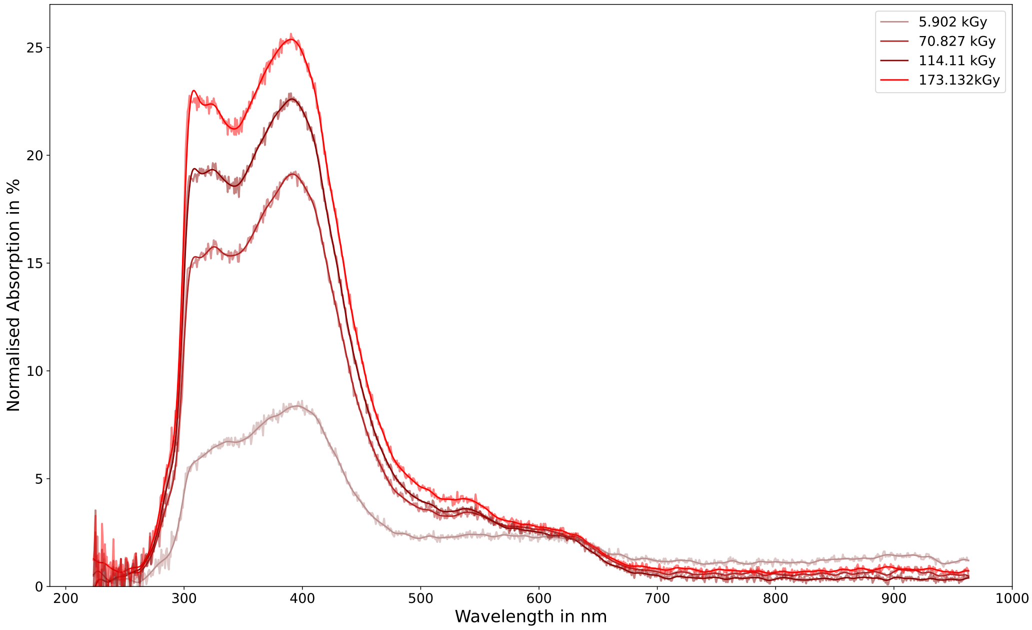


Figure A.1: UV/Vis difference spectrum showing the increase in signal during irradiation at 100 K. Multiple time points (increasing doses) are shown for an apo lysozyme crystal. The Y-axis shows the percent increase in signal relative to the starting reference spectrum at t=0 (d = 0 Gy).

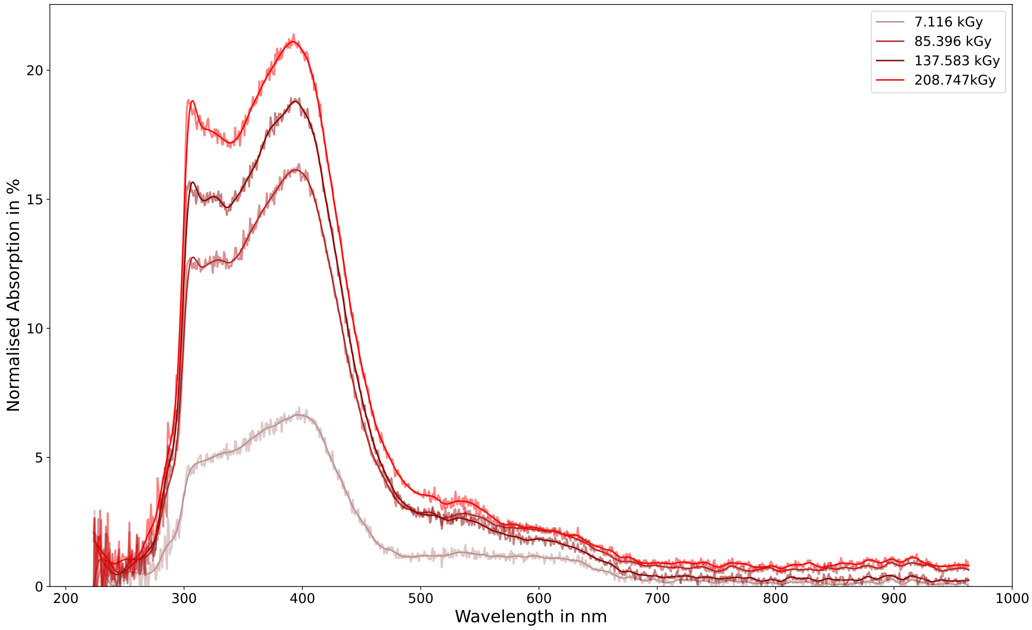


Figure A.2: UV/Vis difference spectrum showing the increase in signal during irradiation at 100 K. Multiple time points (increasing doses) are shown for a lysozyme crystal soaked with 100 mM tryptophan. The Y-axis shows the percent increase in signal relative to the starting reference spectrum at t=0 (d = 0 Gy).

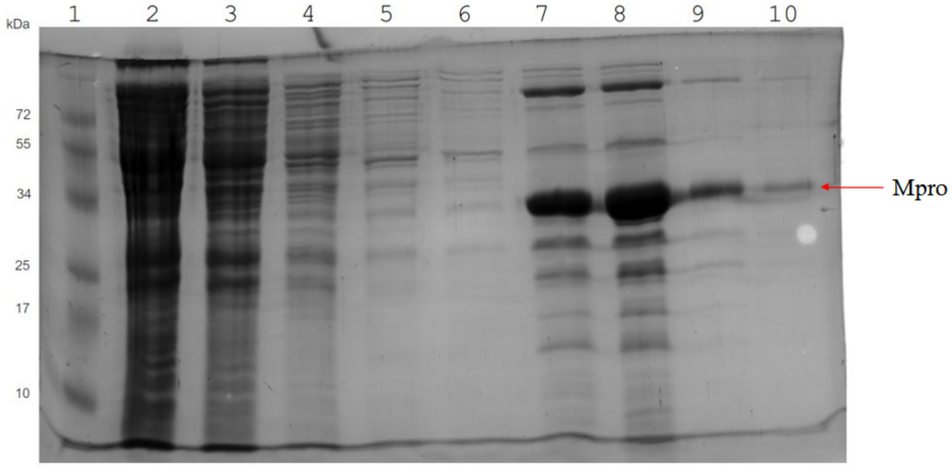


Figure A.3: 12 % SDS Tris-Tricine PAGE gel of anaerobically purified M^{pro} after Ni-affinity chromatography. Lane 1: protein MW marker prestained, lane 2: flow through, lane 3: wash 1, lane 4: wash 2, lane 5: wash 3, lane 6: wash 4, lane 7: elution 1, lane 8: elution 2, lane 9: elution 3, lane 10: elution 4.

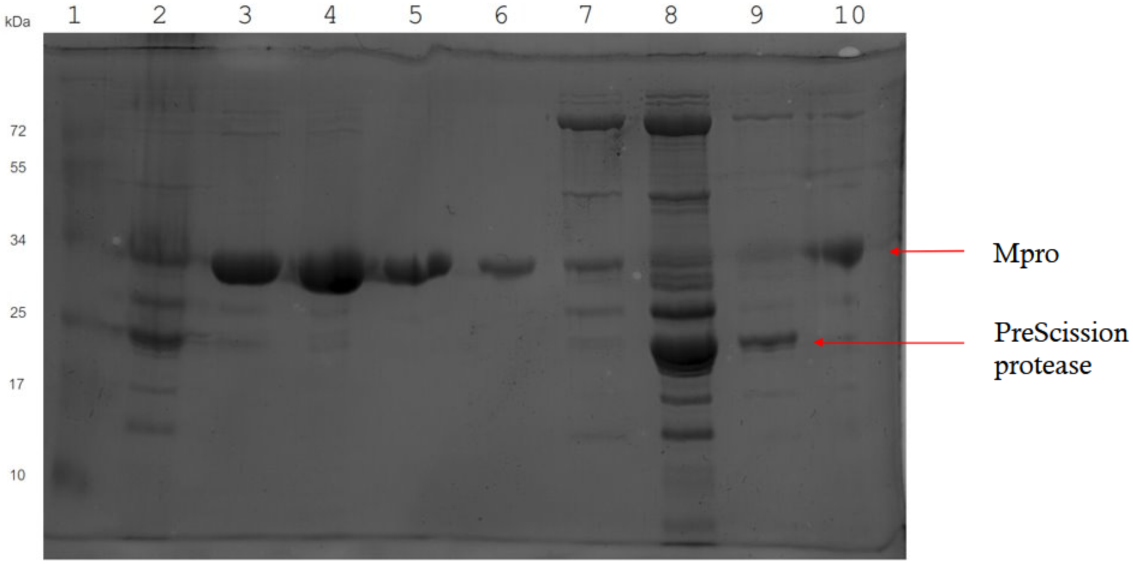


Figure A.4: 12 % SDS Tris-Tricine PAGE gel of anaerobically purified M^{pro} after reverse Ni-affinity chromatography. Lane 1: protein MW marker prestained, lane 2: post-cleavage, lane 3: flow trough, lane 4: wash 1, lane 5: wash 2, lane 6: wash 3, lane 7: elution 1, lane 8: elution 2, lane 9: elution 3, lane 10: elution 2 from previous gel.

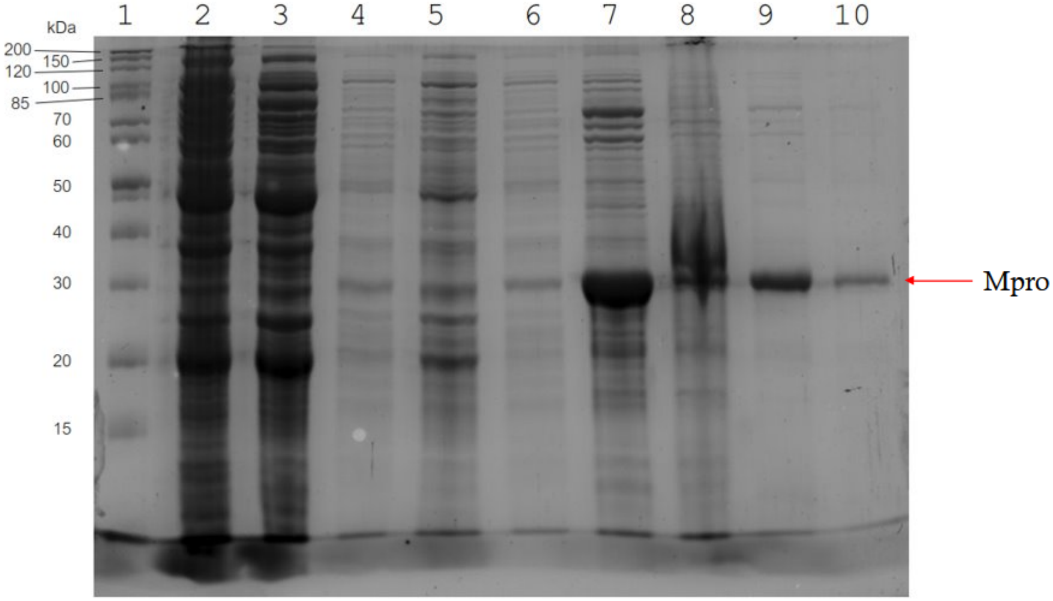


Figure A.5: 12 % SDS Tris-Tricine PAGE gel of aerobically purified M^{pro} (no reducing agents) after Ni-affinity chromatography. Lane 1: PageRuler unstained Protein Ladder, lane 2: flow through, lane 3: wash 1, lane 4: wash 2, lane 5: wash 3, lane 6: wash 4, lane 7: elution 1, lane 8: elution 2, lane 9: elution 3, lane 10: elution 4.

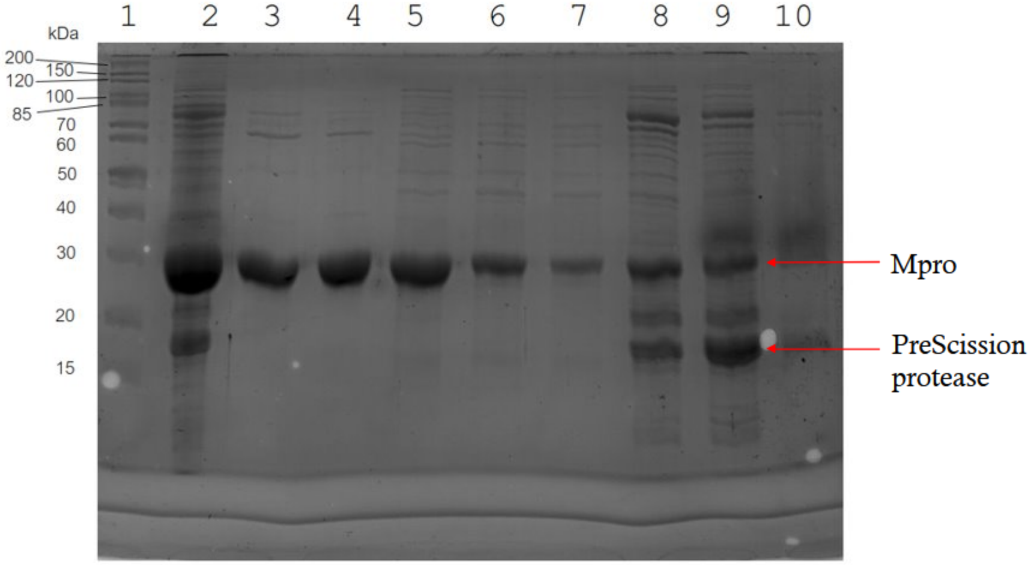


Figure A.6: 12 % SDS Tris-Tricine PAGE gel of aerobically purified M^{pro} (no reducing agents) after reverse Ni-affinity chromatography. Lane 1: PageRuler unstained Protein Ladder, lane 2: post-cleavage, lane 3: flow trough, lane 4: wash 1, lane 5: wash 2, lane 6: wash 3, lane 7: wash 4, lane 8: elution 1, lane 9: elution 2, lane 10: elution 3.

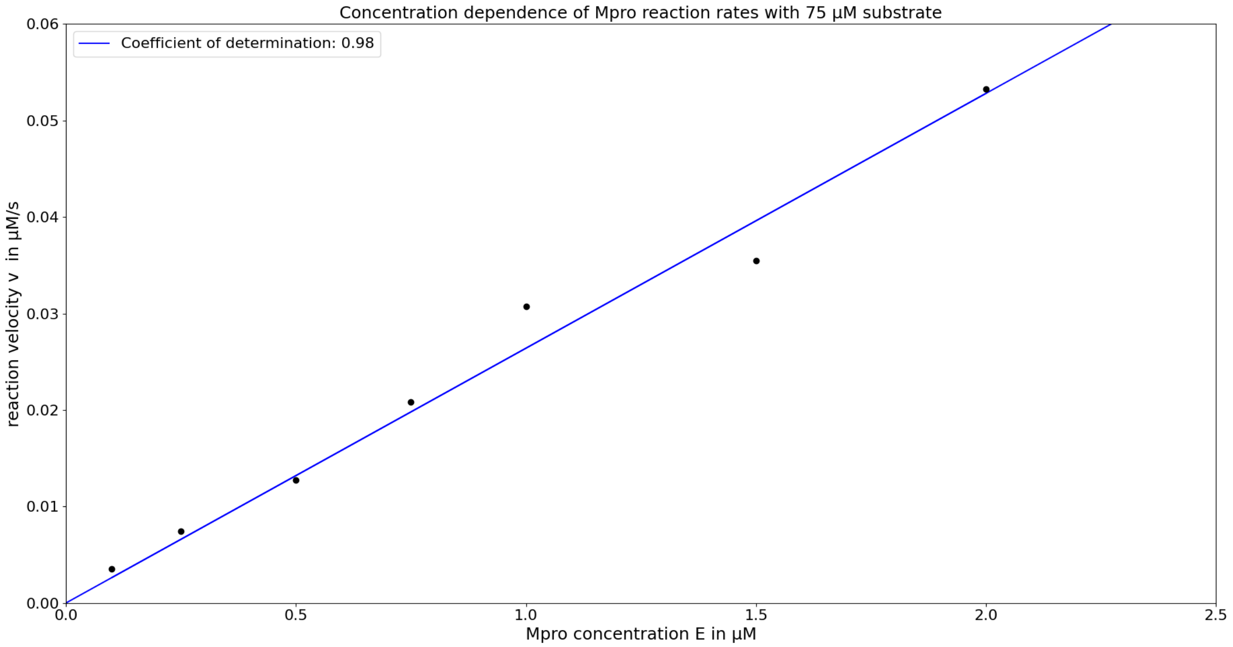


Figure A.7: Reaction velocity v for different M^{pro} concentrations [E] at a constant substrate concentration of 75 µM. Measurements were done in duplicates. The data is fitted with a least squares fit shown in blue.

A.3 Scripts

A.3.1 Reflection decay of a dose series

```
# This script evaluates a dose series and plots the normalised
# total intensity after each data set against the dose.
# Two files types are necessary:
# 1. A dose file "output-Summary.txt" from Raddose 3D which
# specifies the Dose after each DS
# 2. ASCII.HKL files from XDS for each data set

import matplotlib.pyplot as plt
import matplotlib.pylab as pylab
params = {'legend.fontsize': 'x-large',
#'figure.figsize': (15, 5),
'axes.labelsize': 'x-large',
'axes.titlesize':'x-large',
'xtick.labelsize':'x-large',
'ytick.labelsize':'x-large'}
pylab.rcParams.update(params)
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

# directory = '/home/gieselh/PycharmProjects/Crystalfuneral/input_files/'

# print('What is the name of the crystal? (E.g. Lys_122_ctrl)')
# crystal_name = input()


print('How many images per data set?')
```

```
Images_per_dataset = input()

# extract doses for each data set from file

mylines_dose = []
Dose_file = open(’/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/Programs/
RADDOSE-3D-2.1.0/ID30A_3_8_12_18/4HB1/2hr_block/’ + ’output-Summary.txt’,’rt’)
for line in Dose_file:
mylines_dose.append(line[44:53])


# Dose_after_dataset1 = float(mylines_dose[13])
# Dose_after_dataset2 = float(mylines_dose[47])
# Dose_after_dataset3 = float(mylines_dose[81])
# Dose_after_dataset4 = float(mylines_dose[115])
# Dose_after_dataset5 = float(mylines_dose[149])
# Dose_after_dataset6 = float(mylines_dose[183])
# Dose_after_dataset7 = float(mylines_dose[217])
# Dose_after_dataset8 = float(mylines_dose[251])
# Dose_after_dataset9 = float(mylines_dose[285])
# Dose_after_dataset10 = float(mylines_dose[319])
Dose_after_dataset1 = float(mylines_dose[13])
Dose_after_dataset2 = float(mylines_dose[46])
Dose_after_dataset3 = float(mylines_dose[79])
Dose_after_dataset4 = float(mylines_dose[112])
Dose_after_dataset5 = float(mylines_dose[145])
Dose_after_dataset6 = float(mylines_dose[178])
Dose_after_dataset7 = float(mylines_dose[211])
Dose_after_dataset8 = float(mylines_dose[244])
Dose_after_dataset9 = float(mylines_dose[277])
```

```
Dose_after_dataset10 = float(mylines_dose[310])



print(Dose_after_dataset1, Dose_after_dataset2, Dose_after_dataset3,
Dose_after_dataset4, Dose_after_dataset5, Dose_after_dataset6,
Dose_after_dataset7, Dose_after_dataset8, Dose_after_dataset9,
Dose_after_dataset10)

Dose_per_image_in_dataset_1 = float(Dose_after_dataset1) /
int(Images_per_dataset)
Dose_per_image_in_dataset_2 = float(Dose_after_dataset2 -
Dose_after_dataset1) / int(Images_per_dataset)
Dose_per_image_in_dataset_3 = float(Dose_after_dataset3 -
Dose_after_dataset2) / int(Images_per_dataset)
Dose_per_image_in_dataset_4 = float(Dose_after_dataset4 -
Dose_after_dataset3) / int(Images_per_dataset)
Dose_per_image_in_dataset_5 = float(Dose_after_dataset5 -
Dose_after_dataset4) / int(Images_per_dataset)
Dose_per_image_in_dataset_6 = float(Dose_after_dataset6 -
Dose_after_dataset5) / int(Images_per_dataset)
Dose_per_image_in_dataset_7 = float(Dose_after_dataset7 -
Dose_after_dataset6) / int(Images_per_dataset)
Dose_per_image_in_dataset_8 = float(Dose_after_dataset8 -
Dose_after_dataset7) / int(Images_per_dataset)
Dose_per_image_in_dataset_9 = float(Dose_after_dataset9 -
Dose_after_dataset8) / int(Images_per_dataset)
Dose_per_image_in_dataset_10 = float(Dose_after_dataset10 -
Dose_after_dataset9) / int(Images_per_dataset)
```

```
# creating an empty list for every variable and observable
# to put it in a data frame.
h_list = []
k_list = []
l_list = []
I_list = []
logI_list = []
sigI_list = []
dose_list = []

# creating empty data frames for each dataset
DS1 = pd.DataFrame()
DS2 = pd.DataFrame()
DS3 = pd.DataFrame()
DS4 = pd.DataFrame()
DS5 = pd.DataFrame()
DS6 = pd.DataFrame()
DS7 = pd.DataFrame()
DS8 = pd.DataFrame()
DS9 = pd.DataFrame()
DS10 = pd.DataFrame()


# extract Intensities from files
for i in range(1, 11):
dataset_file = open(’/media/hgiesele/Volume/work/Data/ID30A_3_07_12_18/
mx2083_4HB1_block_25mM_trp_4_w1_2hr/data/xds’ + str(i) +
’/XDS_ASCII.HKL’).readlines()

for line in dataset_file:
```

```
if not line.startswith('!'):
hx = int(line[3:6].strip())
kx = int(line[9:12].strip())
lx = int(line[15:18].strip())

Ix = float(line[19:29])
sigIx = float(line[31:40])
ZDx = float(line[60:65])

h_list.append(hx)
k_list.append(kx)
l_list.append(lx)
I_list.append(Ix)
sigI_list.append(sigIx)
if Ix > 0:  # or F_reference <= 0:
logI_list.append(np.log10(Ix))
else:
logI_list.append(0)

if i == 1:
Dose = ZDx * Dose_per_image_in_dataset_1
dose_list.append(Dose)
elif i == 2:
Dose = ZDx * Dose_per_image_in_dataset_2 + Dose_after_dataset1
dose_list.append(Dose)
elif i == 3:
Dose = ZDx * Dose_per_image_in_dataset_3 + Dose_after_dataset2
dose_list.append(Dose)
elif i == 4:
Dose = ZDx * Dose_per_image_in_dataset_4 + Dose_after_dataset3
```

```
dose_list.append(Dose)
elif i == 5:
Dose = ZDx * Dose_per_image_in_dataset_5 + Dose_after_dataset4
dose_list.append(Dose)
elif i == 6:
Dose = ZDx * Dose_per_image_in_dataset_6 + Dose_after_dataset5
dose_list.append(Dose)
elif i == 7:
Dose = ZDx * Dose_per_image_in_dataset_7 + Dose_after_dataset6
dose_list.append(Dose)
elif i == 8:
Dose = ZDx * Dose_per_image_in_dataset_8 + Dose_after_dataset7
dose_list.append(Dose)
elif i == 9:
Dose = ZDx * Dose_per_image_in_dataset_9 + Dose_after_dataset8
dose_list.append(Dose)
elif i == 10:
Dose = ZDx * Dose_per_image_in_dataset_10 + Dose_after_dataset9
dose_list.append(Dose)
else:
Dose = 0

if i == 1:
DS1['h'] = pd.Series(h_list)
DS1['k'] = pd.Series(k_list)
DS1['l'] = pd.Series(l_list)
DS1['Dose'] = pd.Series(dose_list)
DS1['I'] = pd.Series(I_list)
DS1['sigI'] = pd.Series(sigI_list)
DS1['logI'] = pd.Series(logI_list)
```

```
elif i == 2:
DS2['h'] = pd.Series(h_list)
DS2['k'] = pd.Series(k_list)
DS2['l'] = pd.Series(l_list)
DS2['Dose'] = pd.Series(dose_list)
DS2['I'] = pd.Series(I_list)
DS2['sigI'] = pd.Series(sigI_list)
DS2['logI'] = pd.Series(logI_list)

elif i == 3:
DS3['h'] = pd.Series(h_list)
DS3['k'] = pd.Series(k_list)
DS3['l'] = pd.Series(l_list)
DS3['Dose'] = pd.Series(dose_list)
DS3['I'] = pd.Series(I_list)
DS3['sigI'] = pd.Series(sigI_list)
DS3['logI'] = pd.Series(logI_list)

elif i == 4:
DS4['h'] = pd.Series(h_list)
DS4['k'] = pd.Series(k_list)
DS4['l'] = pd.Series(l_list)
DS4['Dose'] = pd.Series(dose_list)
DS4['I'] = pd.Series(I_list)
DS4['sigI'] = pd.Series(sigI_list)
DS4['logI'] = pd.Series(logI_list)

elif i == 5:
DS5['h'] = pd.Series(h_list)
```

```
DS5['k'] = pd.Series(k_list)
DS5['l'] = pd.Series(l_list)
DS5['Dose'] = pd.Series(dose_list)
DS5['I'] = pd.Series(I_list)
DS5['sigI'] = pd.Series(sigI_list)
DS5['logI'] = pd.Series(logI_list)

elif i == 6:
DS6['h'] = pd.Series(h_list)
DS6['k'] = pd.Series(k_list)
DS6['l'] = pd.Series(l_list)
DS6['Dose'] = pd.Series(dose_list)
DS6['I'] = pd.Series(I_list)
DS6['sigI'] = pd.Series(sigI_list)
DS6['logI'] = pd.Series(logI_list)

elif i == 7:
DS7['h'] = pd.Series(h_list)
DS7['k'] = pd.Series(k_list)
DS7['l'] = pd.Series(l_list)
DS7['Dose'] = pd.Series(dose_list)
DS7['I'] = pd.Series(I_list)
DS7['sigI'] = pd.Series(sigI_list)
DS7['logI'] = pd.Series(logI_list)

elif i == 8:
DS8['h'] = pd.Series(h_list)
DS8['k'] = pd.Series(k_list)
DS8['l'] = pd.Series(l_list)
DS8['Dose'] = pd.Series(dose_list)
```

```
DS8[’I’] = pd.Series(I_list)
DS8[’sigI’] = pd.Series(sigI_list)
DS8[’logI’] = pd.Series(logI_list)

elif i == 9:
DS9[’h’] = pd.Series(h_list)
DS9[’k’] = pd.Series(k_list)
DS9[’l’] = pd.Series(l_list)
DS9[’Dose’] = pd.Series(dose_list)
DS9[’I’] = pd.Series(I_list)
DS9[’sigI’] = pd.Series(sigI_list)
DS9[’logI’] = pd.Series(logI_list)

elif i == 10:
DS10[’h’] = pd.Series(h_list)
DS10[’k’] = pd.Series(k_list)
DS10[’l’] = pd.Series(l_list)
DS10[’Dose’] = pd.Series(dose_list)
DS10[’I’] = pd.Series(I_list)
DS10[’sigI’] = pd.Series(sigI_list)
DS10[’logI’] = pd.Series(logI_list)



h_list.clear()
k_list.clear()
l_list.clear()
I_list.clear()
logI_list.clear()
```

```
sigI_list.clear()
dose_list.clear()


# If increasing exposure times/ flux settings are used throughout
# the dose series, they can be corrected for by adjusting the * 1
# to the multiple compared to the first DS.

normalised_total_I_DS1 = DS1[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS2 = DS2[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS3 = DS3[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS4 = DS4[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS5 = DS5[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS6 = DS6[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS7 = DS7[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS8 = DS8[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS9 = DS9[’I’].sum() / (DS1[’I’].sum() * 1)
normalised_total_I_DS10 = DS10[’I’].sum() / (DS1[’I’].sum() * 1)


normalised_total_logI_DS1 = np.log(DS1[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS2 = np.log(DS2[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS3 = np.log(DS3[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS4 = np.log(DS4[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS5 = np.log(DS5[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS6 = np.log(DS6[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS7 = np.log(DS7[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS8 = np.log(DS8[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS9 = np.log(DS9[’I’].sum() / (DS1[’I’].sum() * 1))
normalised_total_logI_DS10 = np.log(DS10[’I’].sum() / (DS1[’I’].sum() * 1))
```

```
total_I = {’Dose’: [Dose_after_dataset1, Dose_after_dataset2,
Dose_after_dataset3, Dose_after_dataset4,
Dose_after_dataset5, Dose_after_dataset6,
Dose_after_dataset7, Dose_after_dataset8,
Dose_after_dataset9, Dose_after_dataset10],
’TotalI’: [normalised_total_I_DS1, normalised_total_I_DS2,
normalised_total_I_DS3, normalised_total_I_DS4,
normalised_total_I_DS5, normalised_total_I_DS6,
normalised_total_I_DS7, normalised_total_I_DS8, normalised_total_I_DS9,
normalised_total_I_DS10],
’TotallogI’: [normalised_total_logI_DS1, normalised_total_logI_DS2,
normalised_total_logI_DS3, normalised_total_logI_DS4,
normalised_total_logI_DS5, normalised_total_logI_DS6,
normalised_total_logI_DS7, normalised_total_logI_DS8,
normalised_total_logI_DS9, normalised_total_logI_DS10]}

total_I_df = pd.DataFrame(total_I, columns=[’Dose’, ’TotalI’, ’TotallogI’])

# calculate linear regression
X = np.array([Dose_after_dataset1, Dose_after_dataset2, Dose_after_dataset3,
Dose_after_dataset4, Dose_after_dataset5, Dose_after_dataset6,
Dose_after_dataset7, Dose_after_dataset8, Dose_after_dataset9,
Dose_after_dataset10]).reshape(-1, 1)
Y = np.array([normalised_total_logI_DS1, normalised_total_logI_DS2,
normalised_total_logI_DS3, normalised_total_logI_DS4,
normalised_total_logI_DS5, normalised_total_logI_DS6,
normalised_total_logI_DS7, normalised_total_logI_DS8,
normalised_total_logI_DS9, normalised_total_logI_DS10]).reshape(-1, 1)
```

```
linear_regressor = LinearRegression().fit(X, Y)
Y_prediction = linear_regressor.predict(X)

v = float(linear_regressor.coef_[0])
b = float(linear_regressor.intercept_)
# fit least squares
Y_exp = np.exp(b) * np.exp(X * v)
# fit exponetial decay
Y_decay = 1 * np.exp(X * v)

half_life = np.log(0.5) / v
print(half_life)
print(v, b)
print('RÂ² score: {}'.format(linear_regressor.score(X, Y)))
print(total_I_df)

plt.figure(figsize=(2, 2))
plt.subplot(122)
plt.title("linear plot", fontsize = 16)
plt.scatter(total_I_df.Dose, total_I_df.TotalI, linewidth=1,
marker='o', color='#2c7bb6', label="data set")
plt.plot(total_I_df.Dose, Y_exp, linewidth=1, color='black',
    label="exponential fit")
plt.ylabel('Normalised Total Intensity I', fontsize=18)
plt.xlabel('Diffraction weighted Dose MGy', fontsize=18)
plt.text(8, .6, "Half life %.2f MGy" % (round(half_life, 2)),
fontsize = 16)
plt.legend()
plt.ylim(0)
plt.xlim(0)

plt.subplot(121)
plt.title("semilogarithmic plot", fontsize = 16)
plt.scatter(total_I_df.Dose, total_I_df.TotalI, linewidth=1,
marker=’o’, color=’#2c7bb6’, label="data set")
plt.plot(total_I_df.Dose, Y_exp, linewidth=1, color=’black’,
label="linear fit")
plt.yscale(’log’)
plt.ylabel(’Normalised Total Intensity I’, fontsize=18)
plt.xlabel(’Diffraction weighted Dose MGy’, fontsize=18)
plt.text(6, .9, ’RÂ² score: {}’.format(round
(linear_regressor.score(X, Y), 3)), fontsize = 16)
plt.legend()
plt.xlim(0)
#plt.suptitle(’Diffraction Decay of Lysozyme 311_20mM_trp’)
plt.show()
```

A.3.2 Create difference spectra at multiple time points during a UV/Vis measurement

```
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab
import scipy.signal
import pandas as pd
params = {’legend.fontsize’: ’x-large’,
#’figure.figsize’: (15, 5),
’axes.labelsize’: ’x-large’,
’axes.titlesize’:’x-large’,
'xtick.labelsize':'x-large',
'ytick.labelsize':'x-large'}
pylab.rcParams.update(params)


# Specify crystal identifier
Sample_name ="acnir_trp_3"
# Specify path to directory where spectral data is located
Path = "/media/hgiesele/PEARSON_BMBF1/oldpchenry/Beamtime_data/
25_06_18_FIP/AcNiR_trp_3/"
# Specify path to directory where the dose file for
# the particular crystal is located
Dose_file = open("/afs/physnet.uni-hamburg.de/users/inf_bio/
hgiesele/Programs/RADDOSE-3D-2.1.0/FIP_25_06_18/AcNiR_50mM_3/
output-Summary.txt")

def Read_spectrum_file(file):
# reads a spectrum file and extracts wavelength and
# corresponding absorption values which are then saved
# in two lists. The function returns the lists.

counter = 0
wavelength = []
absorption = []
for line in file:

if counter > 16 and counter < 1061:
wavelength.append(float(line.split()[0]))
absorption.append(float(line.split()[-1]))
counter += 1
```

```
# Crop noisy part of the spectrum
del wavelength[0:30]
del wavelength[970:]
del absorption[0:30]
del absorption[970:]

return wavelength, absorption

def create_DF (Spectrum_list):
# Takes a list of spectrum files and generates one datafarme
# containing wavelength and corresponding absorption values
# from each spectrum file. The dataframe is returned.
DS = pd.DataFrame()
counter = 1
for i in Spectrum_list:
wavelength, absorption = Read_spectrum_file(i)

DS[’wavelength’] = pd.Series(wavelength)
DS[’Absorption’ + str(counter)] = pd.Series(absorption)
counter += 1
return DS

def calculate_difference_Spectrum(Dataframe):
# Uses a dataframe to calculate a difference spectrum for each
# subsequent spectrum (timepoint) with regard to the the first
# spectrum. The difference spectrum is normalised to show the
# relative change in signal in percent compared to the first
# spectrum. The difference spectra are saved to a new column
# within the dataframe and the dataframe is returned.
```

```
Dataframe["Difference Absorption0"] = ((Dataframe[’Absorption1’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption75"] = ((Dataframe[’Absorption2’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption150"] = ((Dataframe[’Absorption3’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption225"] = ((Dataframe[’Absorption4’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption300"] = ((Dataframe[’Absorption5’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption375"] = ((Dataframe[’Absorption6’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption450"] = ((Dataframe[’Absorption7’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption525"] = ((Dataframe[’Absorption8’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption600"] = ((Dataframe[’Absorption9’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption750"] = ((Dataframe[’Absorption10’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption900"] = ((Dataframe[’Absorption11’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption1050"] = ((Dataframe[’Absorption12’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption1200"] = ((Dataframe[’Absorption13’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption1350"] = ((Dataframe[’Absorption14’] /
Dataframe[’Absorption1’])-1)*100
Dataframe["Difference Absorption1500"] = ((Dataframe[’Absorption15’] /
Dataframe[’Absorption1’])-1)*100
```

```
return Dataframe

def low_pass_filter(Y_values):
# creates a low pass filter to smooth the signal
b, a = scipy.signal.butter(3, 0.1, "lowpass")
data = Y_values
filtered = scipy.signal.filtfilt(b, a, data)
return filtered

def plot_DifferenceSpectrum(Dataframe):
# extracts the "average Dose" value for each timepoint in kGy

mylines_dose = []
for line in Dose_file:
mylines_dose.append(line[44:53])
Dose_for_spectrum0 = 0
Dose_for_spectrum75 = round(float(mylines_dose[17]) * 1000, 3)
Dose_for_spectrum150 = round(float(mylines_dose[50]) * 1000, 3)
Dose_for_spectrum225 = round(float(mylines_dose[83]) * 1000, 3)
Dose_for_spectrum300 = round(float(mylines_dose[116]) * 1000, 3)
Dose_for_spectrum375 = round(float(mylines_dose[149]) * 1000, 3)
Dose_for_spectrum450 = round(float(mylines_dose[182]) * 1000, 3)
Dose_for_spectrum525 = round(float(mylines_dose[215]) * 1000, 3)
Dose_for_spectrum600 = round(float(mylines_dose[248]) * 1000, 3)
Dose_for_spectrum750 = round(float(mylines_dose[281]) * 1000, 3)
Dose_for_spectrum900 = round(float(mylines_dose[314]) * 1000, 3)
Dose_for_spectrum1050 = round(float(mylines_dose[347]) * 1000, 3)
Dose_for_spectrum1200 = round(float(mylines_dose[380]) * 1000, 3)
Dose_for_spectrum1350 = round(float(mylines_dose[413]) * 1000, 3)
```

```
Dose_for_spectrum1500 = round(float(mylines_dose[446]) * 1000, 3)


# generate a list of X-values to plot
X_Dose = [Dose_for_spectrum0, Dose_for_spectrum75,
Dose_for_spectrum150, Dose_for_spectrum225, Dose_for_spectrum300,
Dose_for_spectrum375, Dose_for_spectrum450, Dose_for_spectrum525,
Dose_for_spectrum600, Dose_for_spectrum750, Dose_for_spectrum900,
Dose_for_spectrum1050, Dose_for_spectrum1200, Dose_for_spectrum1350,
Dose_for_spectrum1500]

# generate a dataframe containing X and Y values for plotting
DF = pd.DataFrame(list(zip(X_Dose, Dataframe.loc[288,
'Difference Absorption0':'Difference Absorption1500'])),
columns=["Dose", "Absorption at 400 nm " + Sample_name])
# save the new dataframe
DF.to_csv(
"/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/PycharmProjects/
Crystalfuneral/venv/400nm/DF_400nm_" + Sample_name + "_25_06_18.csv")
# and the input dataframe
Dataframe.to_csv(
"/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/PycharmProjects/
Crystalfuneral/venv/Spectra/DF_spectrum_" + Sample_name + ".csv")

# plot the difference spectrum of 4 selected time points
plt.plot(Dataframe["wavelength"], low_pass_filter(
Dataframe["Difference Absorption150"]), color="rosybrown",
linewidth=1.5,label=str(Dose_for_spectrum150) + " kGy")
plt.plot(Dataframe["wavelength"], low_pass_filter(
Dataframe["Difference Absorption600"]) , color="firebrick",
linewidth=1.5,label=str(Dose_for_spectrum600) + " kGy")
plt.plot(Dataframe["wavelength"], low_pass_filter(
Dataframe["Difference Absorption900"]), color="darkred",
linewidth=1.5, label=str(Dose_for_spectrum900) + " kGy")
plt.plot(Dataframe["wavelength"], low_pass_filter(
Dataframe["Difference Absorption1200"]) ,color="red",
linewidth=1.5,label=str(Dose_for_spectrum1200) + "kGy")
plt.plot(Dataframe["wavelength"], Dataframe["Difference Absorption150"],
color = "rosybrown", linewidth=2, alpha=0.5)
plt.plot(Dataframe["wavelength"], Dataframe["Difference Absorption600"],
color ="firebrick", linewidth=2, alpha=0.5)
plt.plot(Dataframe["wavelength"], Dataframe["Difference Absorption900"],
color = "darkred", linewidth=2,alpha=0.5)
plt.plot(Dataframe["wavelength"], Dataframe["Difference Absorption1200"],
color="red", linewidth=2, alpha=0.5)

#plt.title("Difference UV/Vis Spectrum", fontsize=16)
plt.ylabel(’Normalised Absorption in %’, fontsize=18)
plt.xlabel(’Wavelength in nm’, fontsize=18)
#plt.ylim(0)
plt.legend()

plt.show()



# open a number of spectrum files from different
# time points in the measurement
Spectrum_file0 =
open(Path +  Sample_name +"_00000.txt", "rt")
Spectrum_file75 =
open(Path +  Sample_name +"_00075.txt", "rt")
Spectrum_file150 =
open(Path + Sample_name +"_00" + "150" + ".txt", "rt")
Spectrum_file225 =
open(Path +  Sample_name +"_00" + "225" + ".txt", "rt")
Spectrum_file300 =
open(Path +  Sample_name +"_00" + "300" + ".txt", "rt")
Spectrum_file375 =
open(Path +  Sample_name +"_00" + "375" + ".txt", "rt")
Spectrum_file450 =
open(Path +  Sample_name +"_00" + "450" + ".txt", "rt")
Spectrum_file525 =
open(Path +  Sample_name +"_00" + "525" + ".txt", "rt")
Spectrum_file600 =
open(Path +  Sample_name +"_00" + "600" + ".txt", "rt")
Spectrum_file750 =
open(Path +  Sample_name +"_00" + "750" + ".txt", "rt")
Spectrum_file900 =
open(Path +  Sample_name +"_00" + "900" +".txt", "rt")
Spectrum_file1050 =
open(Path +  Sample_name +"_0" + "1050" + ".txt", "rt")
Spectrum_file1200 =
open(Path +  Sample_name +"_0" + "1200" +".txt", "rt")
Spectrum_file1350 =
open(Path +  Sample_name +"_0" + "1350" + ".txt", "rt")
Spectrum_file1500 =
open(Path +  Sample_name +"_0" + "1500" + ".txt", "rt")


# create a list containing these files as input
# for the "create_DF" function
Spectrum_list = [Spectrum_file0,Spectrum_file75,Spectrum_file150,
Spectrum_file225,Spectrum_file300,Spectrum_file375,Spectrum_file450,
Spectrum_file525, Spectrum_file600,Spectrum_file750,
Spectrum_file900,Spectrum_file1050, Spectrum_file1200,
Spectrum_file1350,Spectrum_file1500]

Dataframe = create_DF(Spectrum_list)
plot_DifferenceSpectrum(calculate_difference_Spectrum(Dataframe))
```

A.4 plot 400nm spectrum

```
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab
import pandas as pd
params = {’legend.fontsize’: ’x-large’,
#’figure.figsize’: (15, 5),
’axes.labelsize’: ’x-large’,
’axes.titlesize’:’x-large’,
’xtick.labelsize’:’x-large’,
’ytick.labelsize’:’x-large’}
pylab.rcParams.update(params)

Path = "/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/
PycharmProjects/Crystalfuneral/venv/400nm/"
def Average_plot_acnir ():
df_ctrl  = pd.read_csv(Path + "acnir_ctrl_average_25_06_18.csv")
df_trp50 = pd.read_csv(Path + "acnir_50mM_average_25_06_18.csv")
```

```
plt.errorbar(df_ctrl["Dose_average"],df_ctrl["Absorption_average"],
xerr=df_ctrl["Dose_stdev"], yerr=df_ctrl["Absorption_stdev"],
fmt=’o’, color=’red’,ecolor=’lightgray’, elinewidth=2, capsize=4,
label = "20.04.18 AcNiR ctrl n = 3")

plt.errorbar(df_trp50["Dose_average"],df_trp50["Absorption_average"],
xerr=df_trp50["Dose_stdev"], yerr=df_trp50["Absorption_stdev"],
fmt=’o’, color=’blue’,ecolor=’lightgray’, elinewidth=2, capsize=4,
 label = " 20.04.18 AcNiR 50 mM Trp n = 3")

plt.xlabel("Average Dose (exposed region) in kGy", fontsize = 18)
plt.ylabel("Normalised Absorption at 450 nm", fontsize = 18)
plt.xlim(0)
plt.legend(loc=’lower right’)
plt.show()
def Average_plot_thau ():
df_ctrl  = pd.read_csv(Path + "thau_ctrl_average_20_04_18.csv")
df_trp50 = pd.read_csv(Path + "thau_50mM_average_20_04_18.csv")


plt.errorbar(df_ctrl["Dose_average"],df_ctrl["Absorption_average"],
xerr=df_ctrl["Dose_stdev"], yerr=df_ctrl["Absorption_stdev"],
fmt=’o’, color=’red’,ecolor=’lightgray’, elinewidth=2, capsize=4,
 label = "20.04.18 thau ctrl n = 3")

plt.errorbar(df_trp50["Dose_average"],df_trp50["Absorption_average"],
xerr=df_trp50["Dose_stdev"], yerr=df_trp50["Absorption_stdev"],
fmt=’o’, color=’blue’,ecolor=’lightgray’, elinewidth=2, capsize=4,
 label = " 20.04.18 thau 50 mM Trp n = 3")

plt.xlabel("Average Dose (exposed region) in kGy", fontsize = 18)
plt.ylabel("Normalised Absorption at 400 nm", fontsize = 18)
plt.xlim(0)
plt.legend(loc=’lower right’)
plt.show()
def Average_plot_lys ():
df_ctrl50  = pd.read_csv(Path + "lys_ctrl_average_20_04_18.csv")
df_trp50 = pd.read_csv(Path + "lys_50mM_average_20_04_18.csv")
df_ctrl100  = pd.read_csv(Path + "lys_ctrl_average_25_06_18.csv")
df_trp100 = pd.read_csv(Path + "lys_100mM_average_25_06_18.csv")

plt.errorbar(df_ctrl50["Dose_average"],df_ctrl50["Absorption_average"],
xerr=df_ctrl50["Dose_stdev"], yerr=df_ctrl50["Absorption_stdev"],
fmt=’o’, color=’red’,ecolor=’lightgray’, elinewidth=2, capsize=4,
label = "20.04.18 lys ctrl n = 3")

plt.errorbar(df_trp50["Dose_average"],df_trp50["Absorption_average"],
xerr=df_trp50["Dose_stdev"], yerr=df_trp50["Absorption_stdev"],
fmt=’o’, color=’blue’,ecolor=’lightgray’, elinewidth=2, capsize=4,
label = " 20.04.18 lys 50 mM Trp n = 3")
plt.errorbar(df_ctrl100["Dose_average"],df_ctrl100["Absorption_average"],
xerr=df_ctrl100["Dose_stdev"],  yerr = df_ctrl100["Absorption_stdev"],
fmt=’o’, color=’darkred’,ecolor=’lightgray’, elinewidth=2, capsize=4,
label = " 25.06.18 lys ctrl n = 4")
plt.errorbar(df_trp100["Dose_average"],df_trp100["Absorption_average"],
xerr=df_trp100["Dose_stdev"], yerr = df_trp100["Absorption_stdev"],
fmt=’o’, color=’darkblue’,ecolor=’lightgray’, elinewidth=2, capsize=4,
 label = "25.06.18 lys 100 mM Trp n = 4")
```

```
plt.xlabel("Average Dose (exposed region) in kGy", fontsize = 18)
plt.ylabel("Normalised Absorption at 400 nm", fontsize = 18)
plt.legend(loc='lower right')
plt.xlim(0)
plt.show()

def indivudal_plot():
# lys controls
df = pd.read_csv(Path + "DF_400nm_lys_control_1_20_04_18.csv")
df1 = pd.read_csv(Path + "DF_400nm_lys_control_3_20_04_18.csv")
df2 = pd.read_csv(Path + "DF_400nm_lys_control_4_20_04_18.csv")
df3 = pd.read_csv(Path + "DF_400nm_lys_control_1_25_06_18.csv")
df4 = pd.read_csv(Path + "DF_400nm_lys_control_2_25_06_18.csv")
df5 = pd.read_csv(Path + "DF_400nm_lys_control_3_25_06_18.csv")
df6 = pd.read_csv(Path + "DF_400nm_lys_control_4_25_06_18.csv")
# lys trp soaks
df7 = pd.read_csv(Path + "DF_400nm_lys_trp_50mm_1_20_04_18.csv")
df8 = pd.read_csv(Path + "DF_400nm_lys_trp_50mm_2_20_04_18.csv")
df9 = pd.read_csv(Path + "DF_400nm_lys_trp_50mm_3_20_04_18.csv")
df10 = pd.read_csv(Path + "DF_400nm_lys_trp_1_25_06_18.csv")
df11 = pd.read_csv(Path + "DF_400nm_lys_trp_2_25_06_18.csv")
df12 = pd.read_csv(Path + "DF_400nm_lys_trp_3_25_06_18.csv")
df13 = pd.read_csv(Path + "DF_400nm_lys_trp_4_25_06_18.csv")


df["normalised"] = df["Absorption at 400 nm lys_control_1"] /
df.iloc[12]["Absorption at 400 nm lys_control_1"]
df1["normalised"] = df1["Absorption at 400 nm lys_control_3"] /
df1.iloc[14]["Absorption at 400 nm lys_control_3"]
```

```
df2["normalised"] = df2["Absorption at 400 nm lys_control_4"] /
df2.iloc[14]["Absorption at 400 nm lys_control_4"]
df3["normalised"] = df3["Absorption at 400 nm lys_control_1"] /
df3.iloc[14]["Absorption at 400 nm lys_control_1"]
df4["normalised"] = df4["Absorption at 400 nm lys_control_2"] /
df4.iloc[14]["Absorption at 400 nm lys_control_2"]
df5["normalised"] = df5["Absorption at 400 nm Lys_control"] /
df5.iloc[14]["Absorption at 400 nm Lys_control"]
df6["normalised"] = df6["Absorption at 400 nm lys_control_4"] /
df6.iloc[14]["Absorption at 400 nm lys_control_4"]

df7["normalised"] = df7["Absorption at 400 nm lys_trp_50mm_1"] /
df7.iloc[14]["Absorption at 400 nm lys_trp_50mm_1"]
df8["normalised"] = df8["Absorption at 400 nm lys_50mm_trp_2"] /
df8.iloc[14]["Absorption at 400 nm lys_50mm_trp_2"]
df9["normalised"] = df9["Absorption at 400 nm lys_trp_50mm_3"] /
df9.iloc[14]["Absorption at 400 nm lys_trp_50mm_3"]
df10["normalised"] = df10["Absorption at 400 nm lys_trp_1"] /
df10.iloc[14]["Absorption at 400 nm lys_trp_1"]
df11["normalised"] = df11["Absorption at 400 nm Lys_trp_2"] /
df11.iloc[14]["Absorption at 400 nm Lys_trp_2"]
df12["normalised"] = df12["Absorption at 400 nm Lys_trp_3"] /
df12.iloc[14]["Absorption at 400 nm Lys_trp_3"]
df13["normalised"] = df13["Absorption at 400 nm lys_trp_4"] /
df13.iloc[14]["Absorption at 400 nm lys_trp_4"]


plt.plot(df["Dose"], df["normalised"],
label ="lys ctrl 1 20.04.18", color = "darkred")
plt.plot(df1["Dose"], df1["normalised"],
label ="lys ctrl 3 20.04.18", color = "darkred")
plt.plot(df2["Dose"], df2["normalised"],
label ="lys ctrl 4 20.04.18", color = "darkred")
plt.plot(df3["Dose"], df3["normalised"],
label ="lys ctrl 1 25.06.18", color = "red")
plt.plot(df4["Dose"], df4["normalised"],
label ="lys ctrl 2 25.06.18", color = "red")
plt.plot(df5["Dose"], df5["normalised"],
label ="lys ctrl 3 25.06.18", color = "red")
plt.plot(df6["Dose"], df6["normalised"],
label ="lys ctrl 4 25.06.18", color = "red")

plt.plot(df7["Dose"], df7["normalised"],
label ="lys 50 mM trp 1 20.04.18", color = "darkblue")
plt.plot(df8["Dose"], df8["normalised"],
label ="lys 50mM trp 2 20.04.18", color = "darkblue")
plt.plot(df9["Dose"], df9["normalised"],
label ="lys 50mM trp 3 20.04.18", color = "darkblue")
plt.plot(df10["Dose"], df10["normalised"],
label ="lys 100mM trp 1 25.06.18", color = "blue")
plt.plot(df11["Dose"], df11["normalised"],
label ="lys 100mM trp 2 25.06.18", color = "blue")
plt.plot(df12["Dose"], df12["normalised"],
label ="lys 100mM trp 3 25.06.18", color = "blue")
plt.plot(df13["Dose"], df13["normalised"],
label ="lys 100mM trp 4 25.06.18", color = "blue")

plt.xlabel("Average Dose in kGy", fontsize = 18)
plt.ylabel("in %", fontsize = 18)
plt.legend()
plt.show()

print(df)

Average_plot_acnir()
Average_plot_thau()
Average_plot_lys()
```

A.4.1 Download and prepare proteome data

```
import glob
import pandas as pd
import requests



def generate_identifier_list(API_link):
# creates a list of identifiers for proteome data that
# complies with the search criteria defined in the API_link
url = API_link
all_identifiers = requests.get(url).text.split("\n")
identifierlist = []
f_prot = open("bacteria_list.txt", "w")
for i in all_identifiers:
new_identifier = "3AUP" + i.lstrip("UP")
url_fasta = "https://rest.uniprot.org/uniprotkb/stream?format=
fasta&query=%28proteome%" + str(new_identifier) + "%29"
data = requests.get(url_fasta).text
organism_name = data.partition("OS=")[2].partition("OX=")[0] + i
f_prot.write(organism_name + ",")
```

```
print(organism_name)
identifierlist.append(organism_name)
f_prot.close()
return identifierlist


def download_files(API_link):
# downloads fasta files that follow certain search criteria
# specified in the API_link

# "https://rest.uniprot.org/proteomes/stream?format=list&query=
%28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%20%28upid%3AUP000002215%
29%20AND%20%28proteome_type%3A1%29&sort=protein_count%20asc"
url = API_link
# a list of all proteome fasta files agreeing with these criteria
# is produced
all_identifiers = requests.get(url).text.split("\n")
# each fasta file is then separately downloaded and saved
for i in all_identifiers:
new_identifier = "3AUP" + i.lstrip("UP")
url_fasta = "https://rest.uniprot.org/uniprotkb/stream?format=
fasta&query=%28proteome%" + str(new_identifier) + "%29"
data = requests.get(url_fasta).text
# strips the full proteome identifier for better readability
organism_name = data.partition("OS=")[2].partition("OX=")[0] + i
print(organism_name)
output_path = "/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/
Documents/4HB1 general documents/proteome_comparison/
reverse search/" + str(organism_name.replace("/", " ")) + ".fasta"
```

```
# writes the proteome data with a simplified name to the output directory
with open(output_path, 'w') as f:
f.write(data)


def get_list_of_all_fasta_files(path_to_files):
list_with_filenames = glob.glob(path_to_files)
print(len(list_with_filenames))
return list_with_filenames


def sum_of_all_amino_acids(g, a, l, m, f, w, k, q, e, s, p, v, i,
c, y, h, r, n, d, t):
sum_aa = g + a + l + m + f + w + k + q + e + s + p + v + i + c +
y + h + r + n + d + t
return sum_aa


def calculate_percentage(x, total):
if x == 0:
return 0
else:
percentage_of_x = round(x * 100 / total, 2)
return percentage_of_x


def calculate_aa_comp(fasta_file_list):
list_of_names = []
list_of_G = []
list_of_A = []
```

```
list_of_L = []
list_of_M = []
list_of_F = []
list_of_W = []
list_of_K = []
list_of_Q = []
list_of_E = []
list_of_S = []
list_of_P = []
list_of_V = []
list_of_I = []
list_of_C = []
list_of_Y = []
list_of_H = []
list_of_R = []
list_of_N = []
list_of_D = []
list_of_T = []

for fasta in fasta_file_list:

fasta_name = fasta.partition("reverse search/")[2].partition(".fasta")[0]
list_of_names.append(fasta_name)

with open(fasta, ’r’) as f:
data = f.readlines()
number_of_g = 0
number_of_a = 0
number_of_l = 0
number_of_m = 0
```

```
number_of_f = 0
number_of_w = 0
number_of_k = 0
number_of_q = 0
number_of_e = 0
number_of_s = 0
number_of_p = 0
number_of_v = 0
number_of_i = 0
number_of_c = 0
number_of_y = 0
number_of_h = 0
number_of_r = 0
number_of_n = 0
number_of_d = 0
number_of_t = 0

for line in data:
if not line.startswith(">"):
number_of_g += line.count(’G’)
number_of_a += line.count(’A’)
number_of_l += line.count(’L’)
number_of_m += line.count(’M’)
number_of_f += line.count(’F’)
number_of_w += line.count(’W’)
number_of_k += line.count(’K’)
number_of_q += line.count(’Q’)
number_of_e += line.count(’E’)
number_of_s += line.count(’S’)
number_of_p += line.count(’P’)
```

```
number_of_v += line.count('V')
number_of_i += line.count('I')
number_of_c += line.count('C')
number_of_y += line.count('Y')
number_of_h += line.count('H')
number_of_r += line.count('R')
number_of_n += line.count('N')
number_of_d += line.count('D')
number_of_t += line.count('T')

sum_of_all_aa = sum_of_all_amino_acids(number_of_g, number_of_a,
number_of_l, number_of_m, number_of_f, number_of_w, number_of_k,
number_of_q, number_of_e, number_of_s, number_of_p, number_of_v,
number_of_i, number_of_c, number_of_y, number_of_h, number_of_r,
number_of_n, number_of_d, number_of_t)

list_of_G.append(calculate_percentage(number_of_g, sum_of_all_aa))
list_of_A.append(calculate_percentage(number_of_a, sum_of_all_aa))
list_of_L.append(calculate_percentage(number_of_l, sum_of_all_aa))
list_of_M.append(calculate_percentage(number_of_m, sum_of_all_aa))
list_of_F.append(calculate_percentage(number_of_f, sum_of_all_aa))
list_of_W.append(calculate_percentage(number_of_w, sum_of_all_aa))
list_of_K.append(calculate_percentage(number_of_k, sum_of_all_aa))
list_of_Q.append(calculate_percentage(number_of_q, sum_of_all_aa))
list_of_E.append(calculate_percentage(number_of_e, sum_of_all_aa))
list_of_S.append(calculate_percentage(number_of_s, sum_of_all_aa))
list_of_P.append(calculate_percentage(number_of_p, sum_of_all_aa))
list_of_V.append(calculate_percentage(number_of_v, sum_of_all_aa))
list_of_I.append(calculate_percentage(number_of_i, sum_of_all_aa))
list_of_C.append(calculate_percentage(number_of_c, sum_of_all_aa))
```

```
list_of_Y.append(calculate_percentage(number_of_y, sum_of_all_aa))
list_of_H.append(calculate_percentage(number_of_h, sum_of_all_aa))
list_of_R.append(calculate_percentage(number_of_r, sum_of_all_aa))
list_of_N.append(calculate_percentage(number_of_n, sum_of_all_aa))
list_of_D.append(calculate_percentage(number_of_d, sum_of_all_aa))
list_of_T.append(calculate_percentage(number_of_t, sum_of_all_aa))

dictio = {"Proteome": list_of_names, ’Ala’: list_of_A, ’Gly’: list_of_G,
’Val’: list_of_V, ’Ile’: list_of_I,’Leu’: list_of_L, ’Pro’: list_of_P,
’Met’: list_of_M, ’Cys’: list_of_C, ’Phe’: list_of_F,’Tyr’: list_of_Y,
’Trp’: list_of_W, ’Arg’: list_of_R, ’His’: list_of_H, ’Lys’: list_of_K,
’Asp’: list_of_D, ’Glu’: list_of_E, ’Ser’: list_of_S, ’Thr’: list_of_T,
’Asn’: list_of_N, ’Gln’: list_of_Q}
df = pd.DataFrame(dictio)
print(df)
return df


def generate_kingdom_dataframes():
# Eukaryota = generate_identifier_list(
# "https://rest.uniprot.org/proteomes/stream?format=list&query=
# %28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%20%28upid%3AUP000002215%
# 29%20AND%20%28proteome_type%3A1%29%20AND%20%28superkingdom%
# 3AEukaryota%29")

Archea = generate_identifier_list(
"https://rest.uniprot.org/proteomes/stream?format=list&query=
%28%28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%20%28upid%3AUP000002215%
29%20AND%20%28proteome_type%3A1%29%29%20AND%20%28superkingdom%
3AArchaea%29")
```

```
# Bacteria = generate_identifier_list(
# "https://rest.uniprot.org/proteomes/stream?format=list&query=
# %28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%20%28upid%3AUP000002215%
# 29%20AND%20%28proteome_type%3A1%29%20AND%20%28superkingdom%
# 3ABacteria%29")
all_res = []

for identifier in Archea:
res = proteome_dataframe[proteome_dataframe[’Proteome’] == identifier]
all_res.append(res)

df_archea = pd.concat(all_res)
df_archea.to_csv(
"/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/Documents/
4HB1 general documents/proteome_comparison/reverse search/
archeae_proteome.csv")

# for identifier in Bacteria:
#     res = proteome_dataframe[proteome_dataframe[’Proteome’] == identifier]
#     all_res.append(res)
#
# df_bacteria = pd.concat(all_res)
# df_bacteria.to_csv("/afs/physnet.uni-hamburg.de/users/
# inf_bio/hgiesele/Documents/ 4HB1 general documents/
# proteome_comparison/reverse search/bacteria_proteome.csv")

# for identifier in Eukaryota:
#     res = proteome_dataframe[proteome_dataframe[’Proteome’] == identifier]
#     all_res.append(res)
```

```
#
# df_eukaryota = pd.concat(all_res)
# df_eukaryota.to_csv("/afs/physnet.uni-hamburg.de/users/
# inf_bio/hgiesele/Documents/4HB1 general documents/
# proteome_comparison/reverse search/eukaryota_proteome.csv")


path = "/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/Documents/
4HB1 general documents/proteome_comparison/reverse search/*.fasta"
# download_files("https://rest.uniprot.org/proteomes/stream?
# format=list&query=%28%28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%
# 20%28upid%3AUP000002215%29%29%20AND%20%28proteome_type%3A1%29")
# dataframe = calculate_aa_comp(get_list_of_all_fasta_files(path))
# dataframe.to_csv("/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele
# /Documents/4HB1 general documents/proteome_comparison/reverse search/"
#  + "proteome_dataframe.csv", sep=’\t’)

proteome_dataframe = pd.read_csv("/afs/physnet.uni-hamburg.de/users/
inf_bio/hgiesele/Documents/4HB1 general documents/proteome_comparison/
reverse search/proteome_dataframe_no_numerator.csv",delimiter=",")

list_of_column_identifier = ["Ala", ’Gly’, ’Val’, ’Ile’, ’Leu’, ’Pro’,
’Met’, ’Cys’, ’Phe’, ’Tyr’, ’Trp’, ’Arg’, ’His’,’Lys’, ’Asp’, ’Glu’,
’Ser’, ’Thr’, ’Asn’, ’Gln’]


generate_kingdom_dataframes()
# f_bac = open("bacteria_list.txt", "w")
# f_bac.write(generate_identifier_list("https://rest.uniprot.org/
# proteomes/stream?format=list&query=%28%28busco%3A%5B50%20TO%20%2A%
# 5D%29%20NOT%20%28upid%3AUP000002215%29%20AND%20%28proteome_type%
# 3A1%29%29%20AND%20%28superkingdom%3ABacteria%29"))
# f_bac.close()

# f_euk = open("eukaryota_list.txt", "w")
# f_euk.write(generate_identifier_list("https://rest.uniprot.org/
# proteomes/stream?format=list&query=%28%28busco%3A%5B50%20TO%20%2A%
# 5D%29%20NOT%20%28upid%3AUP000002215%29%29%20AND%20%28proteome_type%
# 3A1%29%20AND%20%28superkingdom%3AEukaryota%29"))
# f_euk.close()


#generate_identifier_list("https://rest.uniprot.org/proteomes/stream?
#format=list&query=%28%28busco%3A%5B50%20TO%20%2A%5D%29%20NOT%20%28upid%
#3AUP000002215%29%20AND%20%28proteome_type%3A1%29%29%20AND%20%
#28superkingdom%3ABacteria%29")
```

A.4.2 Generating a proteome heatmap

```
import matplotlib.pylab as pylab
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from astropy.stats import histogram
from matplotlib.ticker import PercentFormatter

params = {’legend.fontsize’: ’x-large’,
# ’figure.figsize’: (15, 5),
’axes.labelsize’: ’x-large’,
’axes.titlesize’: ’x-large’,
’xtick.labelsize’: ’x-large’,
’ytick.labelsize’: ’x-large’}
pylab.rcParams.update(params)


def Archeae_heatmap():
# Load data
residues = [’Ala’, ’Gly’, ’Val’, ’Ile’, ’Leu’, ’Pro’, ’Met’,
’Cys’, ’Phe’, ’Tyr’, ’Trp’, ’Arg’, ’His’, ’Lys’, ’Asp’,
’Glu’, ’Ser’, ’Thr’, ’Asn’, ’Gln’]

df = pd.read_csv(
’/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/Documents/
4HB1 general documents/proteome_comparison/reverse search/
archeae_proteome.csv’, delimiter=’,’, index_col=0)

df2 = pd.read_csv(
’/afs/physnet.uni-hamburg.de/users/inf_bio/hgiesele/Documents/
4HB1 general documents/proteome_comparison/reverse search/
proteome_dataframe_no_numerator.csv’, sep=’\t’)

z = df2[df2["Proteome"] == "Halobacterium salinarum
(strain ATCC 700922   JCM 11081   NRC-1) UP000000554"]
y = df2[df2["Proteome"] == "Thermococcus radiotolerans
UP000250085"]
w = df2[df2[
"Proteome"] == "Thermococcus gammatolerans
```

```
(strain DSM 15229 / JCM 11827 / EJ3) (DSM 15229 / JCM 11827 / EJ3)"]
v = df2[df2["Proteome"] == "Thermococcus kodakarensis
(strain ATCC BAA-918   JCM 12380   KOD1) UP000000536"]
# u = df2[df2["Proteome"] == "Chroococcidiopsis thermalis
#(strain PCC 7203) UP000010384"]

# Bin data
binned_data = []
fig1, axes = plt.subplots(nrows=4, ncols=5, sharex=True, sharey=True)
axes = np.ravel(axes)  # flatten the axes array to a simple list
for i, res in enumerate(residues):
frequencies, bin_edges = histogram(df[res], bins="knuth")
bin_width = bin_edges[1] - bin_edges[0]  # bins have fixed width
bin_centers = bin_edges + (bin_width / 2)
bin_centers = bin_centers[:-1]  # Remove last element

# Fitting
#   Here you would do your fitting of the binned data
x_space = np.linspace(0, 17, 1000)  # i= 500 -> x_space[500]

# Interpolate binned data to new x-grid
frequencies_interp = np.interp(x=x_space, xp=bin_centers,
fp=frequencies, left=0, right=0)
# # This will evaluate your frequencies values (y-data) which
# # correspond to the initial bin_centers x-grid, to a new
# # x-grid (x_space). Any missing values on the left and the
# # right are set to 0 (i.e. this takes care of the padding
# # issue we talked about).

# Plot histograms
```

```
ax = axes[i]
ax.bar(bin_centers, frequencies, width=bin_width, alpha=0.3,
label=’Binned’)
ax.plot(x_space, frequencies_interp, c=’red’,
label=’Interpolated’)  # this overlays the interpolated data on top
# of your initial bins
ax.text(0.5, 0.95, res, ha=’center’, va=’top’,
transform=ax.transAxes, fontsize=18)

# Append data to array for saving
no_of_points = len(x_space)
max_freq = np.max(frequencies_interp)
datapoints = np.vstack([[res] * no_of_points, x_space,
frequencies_interp / max_freq]).T

# this will run on the first iteration of the loop only
if isinstance(binned_data, list):
binned_data = datapoints
else:  # this runs every other time
binned_data = np.vstack([binned_data, datapoints])

# Export data
np.savetxt(’proteome_data_binned.csv’, binned_data, delimiter=’,’,
header=’Residue,Occurrence,Frequency’, fmt=’%s’)

# Plot
fig1.supxlabel(’Occurrence in proteomes (%)’, fontsize=18)
fig1.supylabel(’Counts’, fontsize=18)
fig1.suptitle(’Amino acid distribution over 200 proteomes
from archeae’, fontsize=18)
```

```
# fig1.show()

# Load data
df = pd.read_csv('proteome_data_binned.csv', delimiter=',',
header=0, names=['Residue', 'Occurrence (%)', 'Frequency'])

# Create heatmap
heatmap_data = pd.pivot(df, index='Occurrence (%)',
columns='Residue', values='Frequency')
heatmap_data = heatmap_data.reindex(residues, axis=1)
fig2, ax2 = plt.subplots(1, 1)
# save the object as a variable, so you can change various
# properties later
hm = sns.heatmap(heatmap_data, ax=ax2)


# Invert y-axis, so that occurrences are in ascending order
hm.invert_yaxis()

# Set y-axis labels every 1% occurrence
y_tick_labels = [i for i in range(0, 18)]
y_tick_locations = [np.argmin(np.abs(x_space - y_tick))
for y_tick in y_tick_labels]
hm.set_yticks(ticks=y_tick_locations,
labels=y_tick_labels, fontsize=18)

# Reset x-labels rotation
hm.set_xticklabels(hm.get_xticklabels(), rotation=0, fontsize=18)

# Change colorbar labels to percentages
```

```
cbar = hm.collections[0].colorbar
cbar.ax.yaxis.set_major_formatter(PercentFormatter(1, 0))
cbar.set_label(’Normalized frequency’, fontsize=18)

# Set plot title
hm.set_title(’Amino acid distribution over 200 proteomes
from archeae’)

# Overlay line plot on top of the heatmap

list_x = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 11, 12, 13, 14, 15, 16, 17, 18, 19]

trend_x = pd.Series(list_x)
trend_y1 = z[residues].squeeze()
trend_y2 = y[residues].squeeze()
trend_y3 = w[residues].squeeze()
trend_y4 = v[residues].squeeze()

print(trend_y3)
# trend_y5 = u[residues].squeeze()
#
# # # Transform datapoints to Axes coordinates
xs = trend_x + 0.5
ys1 = trend_y1 * len(x_space) / 17
ys2 = trend_y2 * len(x_space) / 17
ys3 = trend_y3 * len(x_space) / 17
ys4 = trend_y4 * len(x_space) / 17
# ys5 = trend_y5 * len(x_space) / 17
#
```

```
ax2.plot(xs, ys1, 'o-', c='black', linewidth=4, markersize=6)
ax2.plot(xs, ys1, '.-', c='white',
label="Halobacterium salinarum (strain ATCC 700922
JCM 11081   NRC-1) UP000000554")
#
ax2.plot(xs, ys2, 'o-', c='black', linewidth=4, markersize=6)
ax2.plot(xs, ys2, '.-', c='blue', label="Thermococcus radiotolerans
(EJ2) UP000250085  ")
# #
ax2.plot(xs, ys3, 'o-', c='black', linewidth=4, markersize=6)
ax2.plot(xs, ys3, '.-', c='green',
label="Thermococcus gammatolerans (strain DSM 15229 / JCM 11827 /
EJ3) UP000001488")
# # #
ax2.plot(xs, ys4, 'o-', c='black', linewidth=4, markersize=6)
ax2.plot(xs, ys4, '.-', c='cyan',
label="Thermococcus kodakarensis (strain ATCC BAA-918
JCM 12380   KOD1) UP000000536")
#
# ax2.plot(xs, ys5, 'o-', c='black', linewidth=4, markersize=6)
# ax2.plot(xs, ys5, '.-', c='lightgreen', label =
# "Chroococcidiopsis thermalis (strain PCC 7203) UP000010384")

#
# # Plot both figures at once
plt.legend()
plt.show()

Archeae_heatmap()
```