Dissertation zur Erlangung der Doktorwürde

an der

Gesamtfakultät für Mathematik,
Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

Thema:

Dissecting cleavage mechanisms in tensed collagen fibrils using reactive molecular dynamics simulations

Vorgelegt von:

Jannik Michael Buhr

Gutachtende:

Prof. Dr. Frauke Gräter
Prof. Dr. Robert Bruce Russell

Dissecting cleavage mechanisms in tensed collagen fibrils using reactive molecular dynamics simulations

Jannik Michael Buhr

Date of the defense: 2026-02-12

Referees:

Prof. Dr. Frauke Gräter
Prof. Dr. Robert Bruce Russell

Table of contents

Preamble 8
Zusammenfassung 9
Abstract 10
Acknowledgments 11
AI statement 12

1 Introduction 13
1.1 Concerning collagen 13
1.1.1 A journey on the HMS Salisbury 13
1.1.2 The biosynthesis of collagen 14
1.2 A rather radical discovery: mechanoradicals in collagen 16
1.3 Aims of this thesis 17

2 Methods and theory 19
2.1 Molecular dynamics 19
2.2 Molecular mechanics 20
2.2.1 Energy terms 20
2.2.2 Force fields and topologies 22
2.3 Periodic boundary conditions 26
2.4 Constraint algorithms 26
2.5 Initial conditions and ensembles 27
2.6 Thermostats and barostats 27
2.7 Long-range interactions 28
2.8 External biasing potentials 29
2.9 Umbrella sampling and the weighted histogram analysis method 29
2.10 Free energy perturbation by slow-growth 30
2.11 Quantum mechanics and density functional theory 31
2.11.1 The Schrödinger equation 31
2.11.2 The electronic Schrödinger equation 32
2.11.3 Hartree product 32
2.11.4 Slater determinant 33
2.11.5 The Hartree-Fock method 33
2.11.6 Hohenberg-Kohn theorems 33
2.11.7 Kohn-Sham density functional theory 34
2.11.8 Basis sets 34
2.12 Quantum mechanics/molecular mechanics 35
2.13 Computational molecular dynamics engines 36
2.14 Kinetic Monte Carlo 36
2.15 Arrhenius equation and transition state theory 37

3 Collagen hydrolysis by quantum mechanics/molecular mechanics 38
3.1 Previous work on base-catalyzed hydrolysis 38
3.2 Goals of the calculations 41
3.3 System setup and simulation pipeline 41
3.3.1 Initial structure preparation 42

3.3.2 Equilibration protocol 42
3.3.3 Quantum mechanics subsystem selection and setup 43
3.4 Results 47
3.4.1 Simulations reveal mechanistic details of the tetrahedral intermediate formation 47
3.4.2 The triple helix marginally impedes hydroxide attack thermodynamically 51
3.4.3 Proton mobility in quantum mechanically described solvent molecules allows breaking of the peptide bond 54
3.5 Discussion and outlook of the hybrid simulations 56

4 Kinetic Monte Carlo Molecular Dynamics: KIMMDY 58
4.1 Running KIMMDY: implementation and usage 60
4.1.1 Installation 60
4.1.2 Configuration and input/output 61
4.1.3 Ensemble generation 63
4.1.4 Reactions and recipes 63
4.1.5 Kinetic Monte Carlo algorithms 64
4.1.6 Topology module 64
4.1.7 Parametrization interface 66
4.1.8 Slow-growth 68
4.1.9 Checkpoints and restarting 69
4.1.10 High performance computing interface 69
4.1.11 Analysis 70
4.1.12 Code quality and testing 70
4.2 Applying KIMMDY to collagen hydrolysis and homolysis 71
4.2.1 The hydrolysis and homolysis plugins 71
4.2.2 System setup 74
4.3 Results 75
4.3.1 Physiologically expected pH changes have only a minor contribution to hydrolysis rates 75
4.3.2 Solvent accessibility of peptide bonds in the fibril is heterogeneous 75
4.3.3 Prolines in collagen may offer protection against hydrolysis 78
4.3.4 Homolysis becomes competitive with hydrolysis by force concentration 80
4.3.5 Discussion and outlook of applying KIMMDY to hydrolysis and homolysis 82
4.4 Beyond hydrolysis: KIMMDY can be applied to a variety of systems and reactions 84
4.4.1 KIMMDY finds a new radical-scavenging candidate in collagen 84

5 Concluding thoughts and next steps 86

Appendix 88
Extended data figures 88

Extended data listings 92
Additional software and resources 95
Knowledge resources 95
Software 95
List of abbreviations and constants 96

List of Figures 98

List of Listings 100

References 101

Preamble

You are reading the print version of this thesis. If you don't like page breaks, you can also read the web version at jmbuhr.de/phd-thesis. It also provides interactive molecule displays, videos, and handy features like preview windows for hovered figure, equation, or citation references.

Zusammenfassung

Kollagen ist unser wichtigstes strukturelles Protein. Als Hauptkomponente von Sehnen, Bändern und anderem Bindegewebe muss es extremen Kräften standhalten und weist eine interessante Chemie unter externer Last auf. Frühere Studien fanden Mechanoradikale in gespannten Kollagenfasern mittels Elektronenspinresonanz. Wie Homolyse, die Reaktion, die eben diese Radikale produziert, mit anderen, nicht-radikalischen und im wässrigen Medium erwartbaren Reaktionen in Konkurrenz steht, war bislang unklar. Vor kurzem wurde gezeigt, dass basenkatalysierte Hydrolyse selbst durch kleine Kräfte beschleunigt wird. In dieser Arbeit bringe ich diese augenscheinlich widersprüchlichen Beobachtungen mit zwei sich ergänzenden, computerbasierten Methoden in Einklang. Hybride quantenmechanische/molekularmechanische Simulationen enthüllen den komplexen Mechanismus der Hydrolysereaktion und zeigen, wie die chemische Umgebung des Lösungsmittels essentiell ist und dass die einzigartige Struktur von Kollagen die Reaktionsbarriere der basenkatalysierten Hydrolyse erhöht. Dann präsentiere ich eine neue Methode, die in Zusammenarbeit mit zwei Kollegen entstanden ist: KIMMDY (Kinetic Monte Carlo Molecular Dynamics) ermöglicht Reaktionen in molekulardynamischen Simulationen auf eine Weise, die Zeitskalen überspannt und ist ein flexibles Framework, das einfach um neue Chemie erweiterbar ist. Mittels KIMMDY zeige ich, wie Kraftkonzentration in komplexen, quervernetzten Kollagenfibrillen Kräfte in Regime verschieben kann, in denen Homolyseraten diejenigen der Hydrolyse übertreffen.

Abstract

Collagen is our most important structural protein. As the chief component of tendons, ligaments, and other connective tissue, it experiences extreme forces and exhibits interesting chemistry under external stress. Earlier investigations found mechanoradicals in tensed collagen fibers through the use of electron paramagnetic resonance. However, it was still unclear how homolytic cleavage, the reaction that produces these radicals, competes with other, non-radical reactions that would be more expected in an aqueous medium such as the body. Base-catalyzed hydrolysis was recently shown to be greatly accelerated even by small forces, which would favor it in the competition. In this thesis, I resolve these seemingly contradictory findings with two complementary computational methods. Hybrid quantum mechanics/molecular mechanics simulations unveil the complexity of the hydrolysis reaction and show how the chemical environment of the solvent is crucial and that the unique collagen structure increases the reaction barrier of base-catalyzed hydrolysis. Then, I showcase a new method I developed in collaboration with two colleagues to tackle the comparison at scale: KIMMDY (Kinetic Monte Carlo Molecular Dynamics) allows simulating reactions in molecular dynamics simulations in a way that bridges timescales and is a flexible framework that can easily be extended to novel chemistries. With KIMMDY, I reveal how force concentration in complex cross-linked collagen fibrils can push forces to regimes where homolysis outcompetes hydrolysis.

Acknowledgments

Firstly, I want to thank Frauke Gräter and the past and present members of our research group that were part of my journey. I couldn't have wished for a better environment for my PhD, and am grateful for the perfect mix of advice, support, and freedom in my research. Special thanks go to my colleagues Eric Hartmann and Kai Riedmiller for being a great team when developing KIMMDY and many insightful scientific discussions.

I would like to thank the additional members of my Thesis Advisory Committee, Ada Cavalcanti-Adam and Dirk-Peter Herten, for helping to keep me on track, and especially Fabio Lolicato for jumping in as a last-minute member. I would like to thank the members of my defense committee, Eva Blasco, Robert Russell, and Ursula Kummer, for giving their time and expertise to evaluate my thesis, and especially Rob for agreeing to be my second referee. I am grateful to the Heidelberg Institute for Theoretical Studies (HITS) and the Max Planck Institute for Polymer Research (MPI-P) for providing the resources and environment to conduct my research. I would like to thank Dmitry Morozov and Gerrit Groenhof for hosting me in Jyväskylä and teaching me the ins and outs of QM/MM simulations in GROMACS.

Finally, I wouldn't be close to where I am today without the support of my family and friends, and the greatest thanks goes to them. To my friends, who provide the much-needed work-life balance with board games (both on- and offline), bouldering, or acrobatics. To my parents and my brother, who put me into the position where I never even had to question whether something is possible, because I know we would always have each other's backs. *"Das macht man so in einer Herde."* And to my partner Franzi for accepting me into her pack and making me so inconceivably happy by just existing.

AI statement

While there exist many valid use cases for machine learning (ML) and large language models (LLMs) in particular, creating genuine human communication and, in doing so, respecting you, the reader, by crafting a text that is worth reading, are not among them. This is why, even though machine learning techniques are featured in my research and even play a central role in the software that came out of it, the text of this thesis was not generated by an LLM. The only exceptions to this are the list of abbreviations in the appendix, which was completed by a script written by an LLM, as well as spell-checking. So, dear reader, I thank you for your time in advance and sincerely hope you will not regret investing it.

1 Introduction

1.1 Concerning collagen

I would like to open with a small historical story.

Figure 1.1: The Capture of Chandernagore, March 1757, by Dominic Serres the Elder in 1771, oil on canvas. The HMS Salisbury can be seen on the left.

1.1.1 A journey on the HMS Salisbury

It is the 20th of May 1747. You are a sailor in the British Royal Navy. As you stand on the deck of the HMS Salisbury (Figure 1.1), you overlook the English Channel you have been patrolling for months now. The creaking of the rigging on the three wooden masts mixes with the moaning of your crewmates who have fallen sick with that all too common ailment: scurvy. At first, they had felt lethargic, short of breath and couldn't keep up with the hustle and bustle aboard the vessel. They felt a pain in their bones, bruised easily, and the usual smaller injuries sustained from working a physically demanding job just would not heal. As you bring fresh water down into their cabin, your crewmates muster a faint smile, exposing their reddened gums. Some are already losing teeth and you know they are not long for this world. They would eventually die from internal bleeding. This was not unusual. During this time it was common to assume that up to 50% of sailors would die from scurvy during a major voyage [1].

But something was different that day. You observe your ship surgeon James Lind take 12 of the scurvy-ridden sailors and divide them into groups of two. He then administers each group a different remedy that had been suggested as a cure. The first got cider, the second vitriol, the third vinegar. The fourth got citrus fruits such as oranges and lemons, and again the next just a pint of seawater, while the last group got a paste made of balsam of Peru, garlic, myrrh, mustard seed, and radish root [2]. As the HMS Salisbury makes its way back to Plymouth at the end of May, the citrus fruits have run out, but the two sailors that had been assigned them as a cure have already recovered.

James Lind was neither the first to use citrus fruits as a cure nor the first to conduct such an experiment [3]; however, he is widely regarded as the first to include control groups. This arguably makes it the first clinical trial in history. I leave the details for historians to debate.

At that time, nobody knew the true cause of scurvy and even Lind himself had various incorrect theories as to the cause and why specifically citrus fruits helped. But they did help, and their ability to treat and prevent scurvy has even made it into popular culture to the extent that there are video game characters that include citrus fruits as part of their appearance to signify them being seafaring. Some of those even include a reference to the disease as well. The League of Legends champion "Gangplank", a pirate character, comes with an ability to "remove scurvy", whereby he eats an orange to be healed of negative status effects (Figure 1.2) and the "Brazen Buccaneer" (Figure 1.3) in the card game "Riftbound", also by Riot Games, is seen in his artwork handing out oranges.

Figure 1.2: The ability icon for "REMOVE SCURVY", © Riot Games.

6
5
UNIT
BILGEWATER
PIRATE
Brazen Buccaneer
As you play me, you may discard a card as an additional cost. If you do, reduce my cost by 2.
"They called my crew 'scurvy dogs,' and I took that personally."
OGN • 002/298
Six More Vodka • ©2025RGI

Figure 1.3: The Brazen Buccaneer card from Riftbound, the League of Legends trading card game, © Riot Games.

As we now know, scurvy is a result of vitamin C deficiency, which in turn leads to issues with the biosynthesis of collagen. This thesis is not about scurvy directly, but it is about collagen.

Collagen is the most abundant protein in mammals[1] [4]. As our primary structural protein, it makes up connective tissue, cartilage, skin, and tendons. Fascinatingly, it is both a component in structures that need stability, such as bones, as well as in those that need flexibility, such as blood vessels. The prevalence and necessity of properly formed collagen explains many of the symptoms of scurvy and also why I chose this obscure 18th-century ailment as a small introductory story from history. But to understand the relationship at a molecular level, we need to look into how collagen is made.

1.1.2 The biosynthesis of collagen

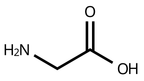


Figure 1.4: Glycine.

The biosynthesis of collagen starts with the pre-pro-peptide[2] at ribosomes in the cytosol. Collagen is characterized by the rather unusual motif Glycine-X-Y (Gly-X-Y), where X and Y are frequently proline (Pro) or 4-hydroxyproline (Hyp) (see Figure 1.4, Figure 1.5, Figure 1.6) and have a slightly higher propensity to be alanine (Ala) [5]. At this stage, however, what will become hydroxyproline is still proline. The pre-pro-collagen peptide makes its way to the endoplasmic reticulum (ER) through an N-terminal signal sequence, which gets cleaved off at its destination. We now have the collagen pro-peptide (pro-collagen).

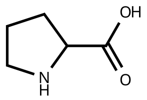


Figure 1.5: Proline.

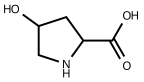


Figure 1.6: Hydroxyproline.

At the ER, two enzymes, prolyl and lysyl hydroxylase, do their work and turn some of the lysines (Figure 1.7) and prolines into hydroxylysines (Figure 1.8) and hydroxyprolines, respectively. Both of these enzymes need vitamin C as a cofactor, so remember to eat your oranges. Pro-collagen consists of a triple helix made up of the individual processed pro-peptides, where two strands share the same composition (alpha 1) and a third has a different composition (alpha 2). Pro-collagen is processed once more in the Golgi apparatus with the addition of oligosaccharides, before finally being secreted to the outside of the cell.

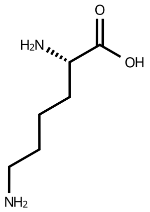


Figure 1.7: Lysine.

Due to the high content of glycine (Figure 1.4), proline (Figure 1.5) and hydroxyproline (Figure 1.6), collagen peptides, and by extension also the triple helices, can be wound very tightly. This is because glycine is the smallest amino acid and proline brings unusual stiffness due to its Φ dihedral angle being locked at -65° by the peptide bond nitrogen (N) being part of a ring structure. This makes collagen an excellent candidate as a structural protein.

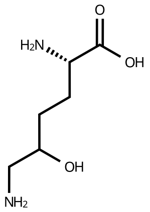


Figure 1.8: Hydroxylysine.

At this point, it is worth noting that there are currently 28 known members of the collagen family. All of which contain at least one such triple helix, but what happens after varies. For this work, the most important members are of the fibrillar type (type I to VI, excluding IV), where understandably the early

[1] going by mass, that is; serum albumin fans may disagree on the counting method

[2] Yes, the naming scheme got a little out of hand.

numbers also constitute the most common forms[3]. More specifically, we will be looking at collagen type I, which is the chief component of tendons, bones, skin, and blood vessels.

Collagen peptidases (collagenases) on the outer membrane of the cell cleave off the ends, leaving us with tropo-collagen. Many tropo-collagen triple helices assemble into fibrillar structures, in which the triple helices are cross-linked to various degrees depending on type and age. The enzyme lysyl oxidase is responsible for creating many of those cross-links. It turns lysine into allysine (its aldehyde), which can then react in an aldol condensation with the hydroxylysines created earlier by lysyl hydroxylase, netting pyridinoline cross-links (PYD). This is where the connection to scurvy lies. A lack of vitamin C leads to a lack of hydroxylated residues, which leads to a lack of collagen cross-links. Thus, scurvy serves to show how crucial the cross-linked, fibrillar structure of collagen is by painting a grim picture of what happens when it is improperly formed.

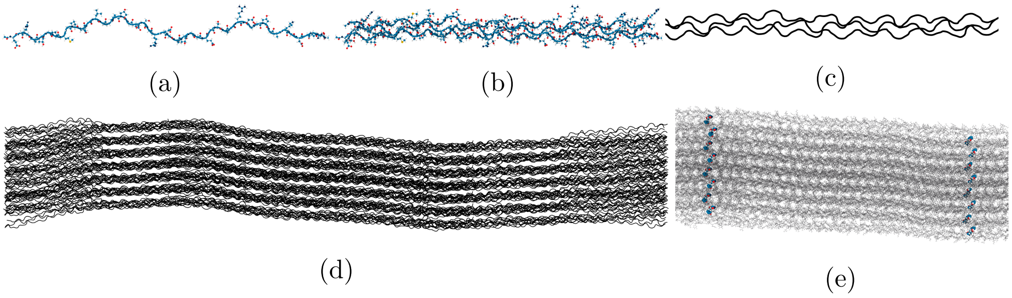


Figure 1.9: Schematic of the structure of collagen type I. Three collagen alpha peptides (a) form a triple helix (b, c). Many triple helices are packed into a collagen fibril (d), in which the triple helices can be cross-linked (e). Only excerpts of the complete chains and fibril are shown.

On a structural level, as seen in Figure 1.9, we then end up with: Three left-handed helical alpha peptides (pre-pro-collagen, Figure 1.9a), wound in a right-handed triple helix 300 nm in length and 1.5 nm in diameter (pro-collagen, tropo-collagen, Figure 1.9b), which is again supercoiled (right-handed) and interwoven and cross-linked with other supercoils to form a collagen fibril (Figure 1.9d, Figure 1.9e). The tropo-collagen molecules in a fibril are not all aligned, but instead exhibit a staggered formation with an offset of around 67 nm (also called the D-Band). This creates a microfibrillar assembly where one part, the overlap region, contains five helices and another, the gap region, contains only 4. This propagates up to the larger fibrils and is thus visible in electron microscopy (Figure 1.10).

[3] Because numbering often happens in the order of discovery and the more common something is the more likely it is to be discovered and characterized.

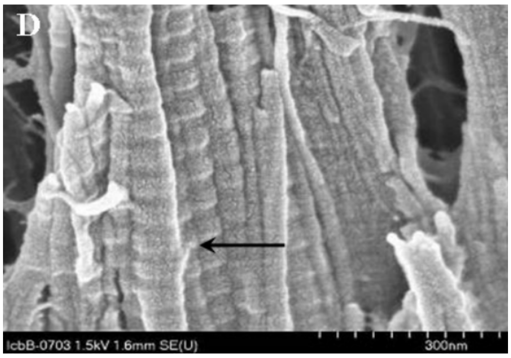


Figure 1.10: Scanning Electron Microscopy (SEM) image of adult mouse articular cartilage collagen by Hughes et al. from Figure 5d of [6].

1.2 A rather radical discovery: mechanoradicals in collagen

As a structural protein, collagen can be under a lot of stress, especially in tendons [7] (Figure 1.11). Take, for example, the semimembranosus muscle, the medial of the three hamstring muscles in the thigh. Its namesake tendon has a cross-sectional area (CSA) of around 100 mm^2 [8] and has to transfer forces up to 1 kN [9], which leads to a stress of 10 MPa. More demanding or rapid exercises can lead to even higher forces, such as 4 kN on the Achilles tendon during vertical jumping [10]. In this case, it also has a higher CSA of around 450 mm^2 [11], so nature is coming prepared, but this still leads to stresses of around 9 MPa. Or take the patellar tendon, which in countermovement jumps has to withstand peak forces of 4 kN [12], but at a CSA of only around 100 mm^2 [13], leading to peak stresses of around 40 MPa. These stresses can become even higher when you take into account the effective CSA, that is the collagen content of the tendon (the part that actually transduces forces), excluding the water content of around 60% [14]. One just has to wonder: What do these forces do to the molecular structure of collagen?

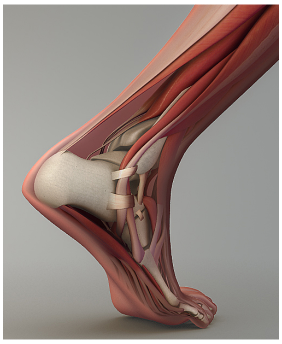


Figure 1.11: Simplified view of the bones, tendons, and ligaments of the foot. The Achilles tendon can be seen attaching the calf muscles to the heel bone. Image Source: cropped version of "Magnified view of a tendon" by Scientfic Animations, https://scientificanimations.com/wiki-images/, CC BY-SA 4.0.

It has long been known that mechanical forces can lead to homolytic bond scission and thus so-called mechanoradicals in polymers [15] and the field of mechanochemistry is well established by now [16]. For decades this thinking had only been applied to synthetic polymers. Biopolymers such as proteins seemed off the table until mechanoradicals in tensed collagen were finally measured through the use of electron paramagnetic resonance (EPR) [17]. Those radicals can then react further into reactive oxygen species (ROS) [18], which can act as signaling molecules in cells or be detrimental at higher levels.

Overlooking biopolymers as candidates for mechanoradicals seems natural still in hindsight. After all, they tend to be solvated in aqueous medium, where other reactions should be more prevalent. One such reaction is hydrolysis, specifically base-catalyzed hydrolysis of the peptide bond of proteins. It was first assumed that such a reaction would be insensitive to external force due to the rate-limiting step being the attack of the hydroxide [19]. This made intuitive sense as an explanation as to why we are able to measure radicals in collagen under force, when a protein could also hydrolyze instead. However, a later study looked deeper into peptide hydrolysis and was able to experimentally show its force-dependence through the use of atomic force microscopy (AFM), supplemented with quantum mechanical (QM) calculations of their own [20].

How can we reconcile seeing radicals in tensed collagen with a peptide bond hydrolysis reaction that is also highly force-dependent? How do the two reactions (Figure 1.12) compete under varying levels of external force?

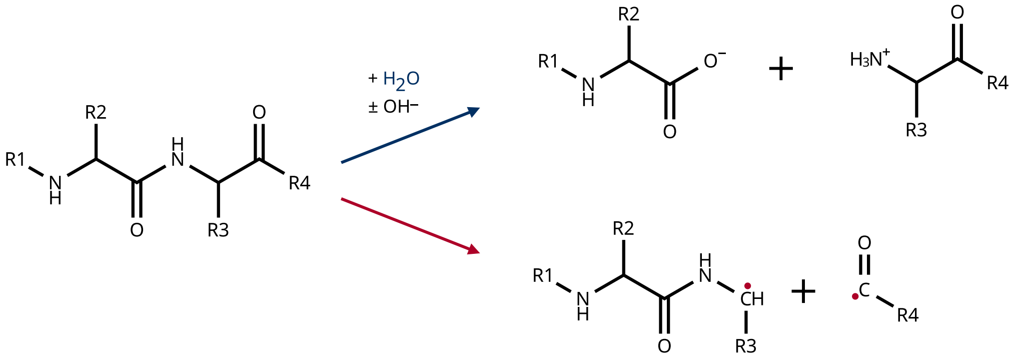


Figure 1.12: Reaction schema of the two competing reactions, hydrolysis (upper pathway) and homolysis (lower pathway).

During base-catalyzed hydrolysis, a hydroxide ion (OH^-) acting as a nucleophile attacks the electrophilic carbon center of the peptide bond. The reaction goes through a tetrahedral intermediate and eventually leads to the rupture of the (C—N) peptide bond, resulting in carboxy- and amino-termini. It is explained in greater detail in Chapter 3. Force-induced homolysis, on the other hand, requires no additional molecules. Bonds along the vector of the applied force are weakened by the energy put into them by the external force, which puts them closer to their bond dissociation energy (BDE). This increases the probability of rupture, whereby the electrons of the scissile bond are split equally between the two binding partners, resulting in unpaired electrons, i.e., radicals. The reaction probability for a particular bond depends on the BDE and the distribution of force through the bonds. How one emulates both reactions with a physical model is explained in more detail in Section 4.2.1.

1.3 Aims of this thesis

While homolysis on the one hand leaves behind radicals as relatively straightforward to detect products, hydrolysis on the other hand just leaves us with unreactive peptides. Furthermore, the wet conditions required for peptide bond hydrolysis are not conducive to measuring radicals in the same sample. EPR measurements require the sample to be dried, because residual water would otherwise be heated up by the microwaves used during measurement and evaporate the sample. I wanted to look deeply at the two competing reactions at a molecular level and thus turned to molecular simulations, the types of which are explained in detail in the Methods and theory section.

The goals of this thesis are twofold. Firstly, there is the biological question at hand. How do the reaction rates of homolysis and hydrolysis compare to each other in collagen fibrils? What factors influence their competition? Secondly, there is the more general question of how chemical reactions can be simulated with computational models, even if the timescales of the reactions are disparate from the timescales that can be achieved with molecular simulations.

The Collagen hydrolysis by quantum mechanics/molecular mechanics chapter, Chapter 3, explores the base-catalyzed hydrolysis reaction in detail. Instead of static calculations, it aims to provide a view into the dynamic nature of the reaction and its complex energy landscape. This is used to directly see how the unique structure of collagen, its tightly wound triple helix, influences the energy barrier of hydrolysis.

But computationally expensive methods such as quantum mechanics/molecular mechanics simulations are inherently limited in the scope of their sampling. The Kinetic Monte Carlo Molecular Dynamics: KIMMDY chapter, Chapter 4, tackles the question of reactive simulations at scale. I will introduce a new method that is the result of a joint project with my colleagues Eric Hartmann and Kai Riedmiller. KIMMDY provides a generalized framework for bridging timescales in reactive molecular dynamics (MD) simulations. As such, the goal of this thesis is not just to answer this one particular question, but also to provide a new tool to the scientific community. In the second half of the chapter, our new tool will be deployed to directly compare the two competing reactions at the scale of a complete collagen fibril.

Lastly, there is one additional goal that can only be read between the lines. I hope for this thesis to be more than a mere enumeration of achievements and instead also provide some joy to the reader. This is why the historical anecdote about scurvy is included and why the upcoming Methods and theory chapter takes on a more educational and, at times, conversational tone.

2 Methods and theory

In the following I will introduce the relevant methods and underlying theory used in the simulations performed for this thesis. The goal is to introduce each topic at least to the extent that the choices made for setting up the simulations and implementing solutions make sense in the context of the theory.

2.1 Molecular dynamics

So how do we simulate molecular systems, especially large ones relevant for biomolecular research? Let's start with a simple small molecule such as this propionaldehyde in Figure 2.1. In the chemical formula, implicit hydrogens are often omitted, so let us also look at a classic ball-and-stick render of the same molecule in Figure 2.2. Ultimately, this representation combines the nuclei and electrons of the atoms involved in the molecule into one representation, but we might also want to show a representation of the electron density like in Figure 2.3. The challenge is now to find a representation in which it becomes feasible to describe the motion of atoms through time in order to build a system that we can observe and then use this system calculate properties by sampling.

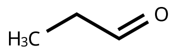


Figure 2.1: Chemical structure formula of propionaldehyde.

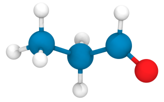


Figure 2.2: Propionaldehyde as a classic ball-and-stick render.

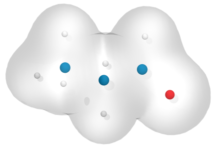


Figure 2.3: Propionaldehyde represented with balls for atomic nuclei and the electron density around them.

As the phrase, usually attributed to statistician George Box, goes: "All models are wrong, but some are useful." So let us start with one of the most simplified ones that will allow us to simulate large systems. Afterward, we will add more detail and then finally combine the simplified with the detailed method.

We can describe our system as a collection of all the atomic positions $\vec{r}$ in 3-dimensional space and their momenta $\vec{p}$, where both vectors are of length 3N with N being the number of atoms. If we can propagate this system through time, we get a path through the 6N-dimensional phase space that we call a trajectory. The overall method is called molecular dynamics (MD).

Let us imagine we already have a way to describe the total energy (kinetic plus potential energy) of the system and its derivative. We can use either Hamiltonian mechanics Equation 1

$$\begin{aligned} \frac{dr_i}{dt} &= \frac{\delta H}{\delta p_i} \\ \frac{dp_i}{dt} &= -\frac{\delta H}{\delta r_i}, \end{aligned} \tag{1}$$

where the Hamiltonian operator H is the total energy operator and r_i is one value of $\vec{r}$, i.e., the position of one atom along one axis and likewise for p_i.

Or we can use Newton's equations of motion Equation 2

$$\begin{aligned} F &= ma \\ -\frac{dV}{dr} &= m\frac{d^2r}{dt^2}, \end{aligned} \tag{2}$$

where V is the potential, whose derivative with respect to the positions corresponds to the force F and the acceleration a is the second derivative of the positions with respect to time t. The mass m is assumed constant for each particle.

Both of these differential equations can be integrated over time to trace out a trajectory given initial conditions for the positions and momenta $\vec{r}_0$ and $\vec{p}_0$. Most implementations use Newton's equations for computations. The integration is performed stepwise over small intervals Δt (or dt). From the initial state, energies and forces are computed for each atom. Under the assumption of those being constant for a small timestep dt, typically in the range of femtoseconds, the new positions are calculated from the forces acting on the atoms. The *leapfrog* algorithm [21] achieves this by interleaving position and velocity updates as shown in Equation 3. This gets repeated until sufficient sampling of phase space is achieved or computational resources, limited by time and money, run out.

$$\begin{aligned}\mathbf{v}(t+\frac{1}{2}\Delta t) &= \mathbf{v}(t-\frac{1}{2}\Delta t)+\frac{\Delta t}{m}\mathbf{F}(t)\\ \mathbf{r}(t+\Delta t) &= \mathbf{r}(t)+\Delta t\mathbf{v}(t+\frac{1}{2}\Delta t)\end{aligned} \tag{3}$$

Because different initial conditions lead to different trajectories, one often runs multiple such computational runs in parallel. For ergodic properties, i.e., those for which the ergodic hypothesis holds, stating that the time average is equal to the ensemble average [22], this can drastically improve sampling.

2.2 Molecular mechanics

How do we obtain a feasible equation for the total energy of our system that scales well to a large number of atoms? This is where we think back to our simplified stick representation in Figure 2.2 and arrive at molecular mechanics (MM), as in classical Newtonian mechanics. Let us assume our system is made up of different kinds of balls connected with springs.

2.2.1 Energy terms

The total potential energy of our system is described as the sum of bonded and non-bonded terms Equation 4, which we can further decompose with Equation 5 and Equation 6.

$$V_{\text{total}} = V_{\text{bonded}} + V_{\text{nonbonded}} \tag{4}$$

$$V_{\text{bonded}} = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} \tag{5}$$

$$V_{\text{non-bonded}} = V_{\text{electrostatic}} + V_{\text{van der Waals}} \tag{6}$$

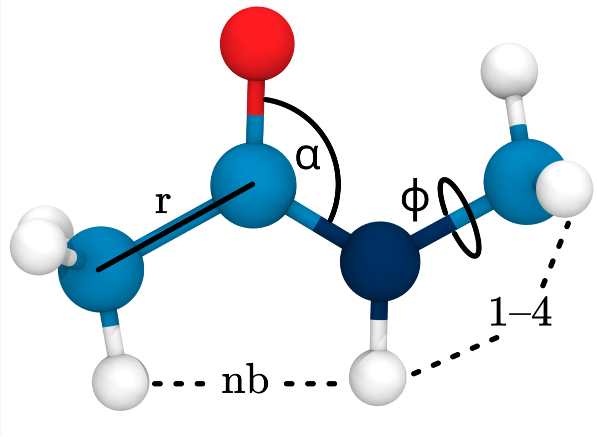


Figure 2.4: Excerpt from a glycine dipeptide with representative MM bonded terms, a bond r, an angle α, a dihedral as well as non-bonded terms nb and 1–4.

The bonded and non-bonded terms are illustrated by examples in Figure 2.4. For each of these terms we find approximations that are easy to compute on modern hardware, which typically means that exponentials are avoided in favor of polynomials obtained through Taylor expansion [23]. For example, while the potential energy in a bond between two atoms can be described with a Morse potential [24] (Equation 7, Figure 2.5):

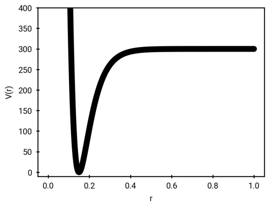


Figure 2.5: A Morse potential.

$$V_{\text{Morse}}(r) = D\left(1 - e^{-\beta(r-r_0)}\right)^2, \tag{7}$$

with the well depth D, the equilibrium bond distance r_0 and β relating to the force constant k via $\beta = \sqrt{\frac{k}{2D}}$, in practice this is often simplified to a harmonic potential (Equation 8, Figure 2.6):

$$V_{\text{harmonic}}(r) = \frac{1}{2}k(r - r_0)^2 \tag{8}$$

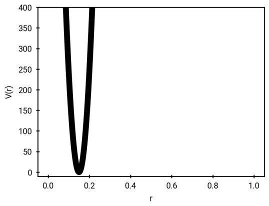


Figure 2.6: A harmonic potential.

This is a good approximation in the low-energy regime and has the added bonus that our molecule is unlikely to fall apart due to the infinitely scaling potential instead of the plateau in the Morse potential. Likewise, angles between three atoms can also be approximated with harmonic potentials. Finally, bonded connections between four atoms give rise to dihedral angles that are periodic in nature and are thus represented with cosines (Equation 9) as the result of a Fourier series [25].

$$V_{\text{cosine}}(\phi) = k(1 + \cos(n\phi - \phi_0)) \tag{9}$$

Sometimes additional bonded terms labelled "improper dihedrals" are used to, e.g., enforce planarity of certain molecules by adding potential energy for out-of-plane bending.

In addition to those bonded terms, we have the non-bonded terms split into electrostatic interactions and van der Waals interactions.

Van der Waals forces (or consequently the potential energy from which they arise) are the result of a mixture of repulsion through overlapping electron density at very close distances and attraction at intermediate distances due to spontaneous and induced dipole-dipole interactions between the electron clouds of two atoms. While the long-range attraction scales with $1/R^6$, the short-range repulsion features an exponential term. This is why in practice the Lennard-Jones potential [26] is used instead of a more accurate formulation, simplifying computations (Equation 10, Figure 2.7).

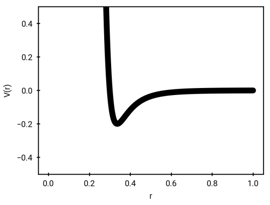


Figure 2.7: A Lennard-Jones potential.

$$V_{\text{vdw}}(r_{ab}) = 4\epsilon_{ij}\left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}\right], \tag{10}$$

where r is the distance between two atoms i and j and σ and ϵ are parameters that depend on the pair of atoms.

The electrostatic or Coulomb [27] term (Equation 11 and Figure 2.8) is also summed for all pairs of atoms.

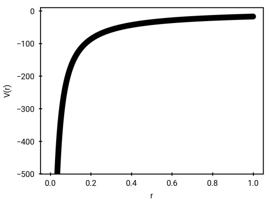


Figure 2.8: A Coulomb potential.

$$V_{\text{Coulomb}}(r_{ij}) = \frac{Q_a Q_b}{\epsilon r^2}, \tag{11}$$

with the effective dielectric constant ϵ.

Here, the charge Q of an atom is not just the result of ionization. Instead, it includes so-called partial charges, where the total electron density of a molecule has been (artificially) assigned beforehand to individual atoms based on quantum mechanical calculation procedures like the Restrained Electrostatic Potential (RESP) [28].

Those non-bonded interactions are sometimes scaled down when atoms are, albeit not directly, connected by being part of the same dihedral, which is then called a 1–4 interaction, and are usually omitted when atoms are direct or second neighbors. The contributions are then already integrated into the bonded terms.

2.2.2 Force fields and topologies

Now that we have equations for the energy terms of our system that are also easily derivable, we need to feed some values into them. How do we get, store, and recall the parameters for all those equations? The answer comes in the form of force fields.

Many PhD students before me and other clever people have performed more detailed quantum mechanical *ab initio* calculations to figure out what those parameters should be, like, for example, the force constant k (in Equation 8) for a bond between two particular atoms. This will vary by molecule and chemical environment; not every carbon (C) to carbon bond is the same, but we can group similar atoms into "atomtypes". In protein simulations, a popular choice are the Amber force fields [29]. Since the early version developed in 1994, Amber94 [30] has been continuously refined. Amber99 improved upon it with RESP partial charges [31] and provided a basis for further modifications. The parameters are not only fit to *ab initio* QM calculations. They are also adjusted to match observables from simulations to experimental values, such as the dihedral angles of a peptide in a Ramachandran diagram [32]. The version used for the simulations in this thesis is `amber99sb*ildnp`, where `sb` stands for Stony Brook University, `*` stands for protein backbone modifications, `ildn` improved side-chain torsion potentials, and the `p` stands for improved proline and hydroxyproline parameters [33–35]. Other well-known force fields include CHARMM [36], GROMOS [37] and OPLS-AA [38]. The choice of a force field is one of the many crucial choices when setting up a molecular dynamics simulation, as different force fields have been optimized for different types of molecules or observables. Amber was chosen as it is specifically geared towards protein simulations. We will be going through examples from this force field, because how a force field is structured will be relevant to Chapter 4 in particular. Specifically, the lines shown in the upcoming listings (Listing 1, Listing 2) are from a version of the files [39] ported to the molecular dynamics engine GROMACS [40] that I have been using during the course of my PhD.

Just like the total energy function Equation 4, the main entry point for the force field (`forcefield.itp`, Listing 1) includes references to the non-bonded and bonded parameters. In addition to that, it also holds more general parameters, such as the combination rule used to compute Lennard-Jones (LJ) parameters for a pair of atoms and the fudge parameters, which are factors for scaling down LJ and Coulomb (QQ) parameters for 1–4 interactions. `itp` stands for "include topology" and we will talk more about the topology concept later.

Listing 1 Excerpt from forcefield.itp

```
[ defaults ]
; nbfunc  comb-rule  gen-pairs  fudgeLJ  fudgeQQ
1         2          yes        0.5      0.8333

#include "ffnonbonded.itp"
#include "ffbonded.itp"
```

For any given pair of atoms that have been assigned atom types and are not bonded, we can fill in the parameters of the LJ equation (Equation 10). The combination rule as mentioned in the general parameters decides how the parameters ϵ and σ from both atoms are combined into one parameter each. Type 1 would indicate a geometric average, while type 2, as in this example, used the Lorentz-Berthelot [41, 42] rules, where σ used an arithmetic average and ϵ uses a geometric average (Equation 12).

Listing 2 Select lines from ffnonbonded.itp

```
[ atomtypes ]
; name  at.num   mass     charge ptype  sigma       epsilon
C        6       12.01    0.0000  A    3.39967e-01  3.59824e-01
CA       6       12.01    0.0000  A    3.39967e-01  3.59824e-01
H        1        1.008   0.0000  A    1.06908e-01  6.56888e-02
HC       1        1.008   0.0000  A    2.64953e-01  6.56888e-02
```

$$\begin{aligned} \sigma_{ij} &= \left(\sigma_{ii}\,\sigma_{jj}\right)^{1/2} \\ \epsilon_{ij} &= \left(\epsilon_{ii}\,\epsilon_{jj}\right)^{1/2} \end{aligned} \tag{12}$$

Likewise, the MD engine can use the charges from Listing 2 to calculate the electrostatic potential. However, you might have noticed that all the charges in the default `atomtypes` are `0.0`. This is correct, as, for example, a carbon when saturated is not charged by default, but based on its chemical environment it should have a partial charge. How large should the partial charge be? This is defined in the `topology`.

The topology of a system lists all molecules and atoms of a system and how they are connected. So while a force field lists all potential interactions between types of atoms that it knows about, a topology is a concrete instance of the interactions and their parameters. The topology files shown are in the format of my MD engine of choice, GROMACS. A topology file (Listing 3) may include the force field and then list a number of "moleculetypes" that can then be used to describe the system with the number of each type listed in the molecules section. Note though that a topology is only concerned with connectivity and types, but not with coordinates. Coordinates are specified in a separate file such as a `.gro` or `.pdb` [43] file, excerpts of which can be seen in Listing 8 and Listing 9. When GROMACS creates a topology from coordinates, it uses information such as the "residuetype" to fill out and assign interactions. For example, in Listing 3 the atom with identifier (ID) 1 is a carbon of type CT that belongs to the acyl cap of a glycine dipeptide, so it has residuetype ACE. `.rtp` files in the force fields contain specialized residue topologies. In this case, given that we are dealing with a peptide, GROMACS will look in `aminoacids.rtp` (Listing 5), find the partial charges for the atoms of the residue, and write their values directly into the topology as part of the `atoms` section. Additionally, improper dihedrals are added based on the `impropers` section.

Not all values have to be written explicitly. Take the first bond in Listing 3 between atoms with IDs 1 and 2. The topology only specifies the functional form of the interaction, where type 1 is a harmonic potential [44]. The values for the parameters are then taken from the `bondtypes` section based on the atomtypes of the atoms involved in the bond, where in this case the `bondtypes` section was included from the force field file as seen in Listing 4. It is also possible to specify the parameters directly, which will be especially relevant for our method KIMMDY, Chapter 4. The same is done for higher-order interactions such as angles and dihedrals. A special case is the `pairs` section, which indicates pairs of atoms that have scaled LJ and Coulomb terms, usually due to them being part of a 1–4 interaction. This section also allows one to specify the Coulomb and van der Waals parameters directly.

Listing 3 Selected, simplified lines from a .top file of a capped glycine dipeptide

```
#include "./amber99sb-star-ildnp.ff/forcefield.itp"

[ moleculetype ]
; Name            nrexcl
Protein             3

[ atoms ]
; nr   type resnr residue  atom  cgnr   charge  mass   typeB ..
  1    CT     1    ACE     CH3    1  -0.3662   12.01
  2    HC     1    ACE    HH31    2   0.1123   1.008

[ bonds ]
;  ai    aj funct    c0    c1      c2    c3
    1     2     1
    1     3     1

[ angles ]
;  ai    aj    ak funct    c0   c1   c2    c3
    2     1     3     1

[ molecules ]
; Compound        #mols
Protein             1
SOL         39427
NA                 110
CL                 110
```

Listing 4 Selected, simplified lines from ffbonded.itp

```
[ bondtypes ]
; i   j  func    b0         kb
  CT  HC   1     0.10900   284512.0
```

Listing 5 Selected lines from aminoacids.rtp

```
[ ACE ]
 [ atoms ]
  HH31    HC           0.11230    1
   CH3    CT          -0.36620    2
  HH32    HC           0.11230    3
  HH33    HC           0.11230    4
     C    C            0.59720    5
     O    O           -0.56790    6
 [ bonds ]
  HH31   CH3
   CH3  HH32
   CH3  HH33
   CH3     C
     C     O
 [ impropers ]
   CH3    +N     C     O
```

2.3 Periodic boundary conditions

When simulating systems such as those used in biology, like a protein solvated in water, we can extend our system at little cost by applying periodic boundary conditions (PBC) (Figure 2.9). A central simulation cell is assumed to infinitely repeat in all directions. A particle exiting the cell enters at the opposite side. This allows one to mimic a large homogeneous solvation environment of the protein, though one has to be careful that the simulation box is sufficiently large to prevent the protein from interacting with its own periodic image.

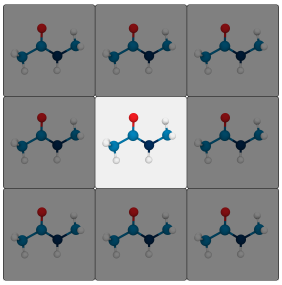


Figure 2.9: A unit cell surrounded by periodic images.

2.4 Constraint algorithms

The higher we can set our timestep Δt, the longer the total time we can simulate with our computational resources. However, higher timesteps also lead to more unstable simulations. Imagine a harmonic bond in a molecule between an oxygen and a hydrogen. The harmonic potential (Equation 8) has a force constant k of 462750 kJ/mol/nm^2. The masses of the atoms are 1.008 u and 16 u where 1 u = 1.66×10^{-27} kg. Together this is a reduced mass μ of 1.574×10^{-27} kg.

The period T of the harmonic oscillator is $T = 2\pi\sqrt{\frac{\mu}{k}} \approx 1.5$ fs. The fastest oscillation in a system sets an upper bound for the timestep, as taking a stride in time during which a bond should complete a full oscillation naturally leads to instabilities. As such we have an interest in constraining those high-frequency bonds to allow for larger timesteps.

The SHAKE algorithm [45] applies an update to the coordinates after the unconstrained position update to fulfill a list of distance constraints by applying constraint forces. The SETTLE algorithm [46] is a dedicated implementation for rigid water molecules, which provides a significant speedup (after all, a large part of most molecular dynamics simulations is the box of water molecules). The LINCS algorithm [47] is used in practice for its significant speed. It is non-iterative and sets bond lengths to their correct lengths after the unconstrained update. This allows for timesteps upwards of 1 fs.

2.5 Initial conditions and ensembles

The initial conditions for molecular dynamics simulations typically come from a structure determined experimentally through x-ray crystallography or cryo-electron microscopy (cryo-EM), though with the rise of AlphaFold [48], sometimes an amino acid sequence can be a good-enough starting point for protein simulations. The initial structure is then minimized, solvated, and equilibrated, where typically restraints are employed to ease the protein into the ensemble. Here, we differentiate between the microcanonical (NVE), canonical (NVT), and isothermal-isobaric (NPT) ensemble, where N stands for the fixed number of particles, V is a fixed volume, E is a fixed total energy, T is a fixed temperature, and P is a fixed pressure. This is to make sure that the system is stable and to reduce the bias due to the initial conditions.

2.6 Thermostats and barostats

In order to achieve the various ensembles, it is necessary to control the respective constants. Keeping the number of particles the same is trivial, the other observables require a bit more math.

From the equipartition theorem [49] it follows that the temperature T of an ideal gas system with N particles can be described by Equation 13 as a function of the individual particle velocities.

$$T = \frac{E_{\text{kin}}}{\frac{3}{2}Nk_{\text{B}}} = \frac{\langle \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 \rangle}{\frac{3}{2}Nk_{\text{B}}}, \tag{13}$$

where m_i is the mass of particle i and v_i is its velocity. Accordingly, the temperature of a system can be controlled by adjusting the velocities of the atoms in the system. Given a virtual external heat bath of temperature T_0, the Berendsen algorithm [50] couples the temperature of the system to it by Equation 14.

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{14}$$

The time constant τ of the exponentially decaying temperature deviation adjusts the strength of the coupling. (Stochastic) velocity-rescaling [51] is a newer method for temperature coupling based on the Berendsen thermostat that ensures correct kinetic energy distribution. It keeps the advantages of the Berendsen thermostat, the fast first-order decay of temperature deviations and the absence of oscillations, while producing a correct canonical ensemble.

Likewise, the Berendsen algorithm can couple the pressure to an external "pressure bath" by scaling the volume by a factor λ via Equation 15.

$$\lambda = 1 - \beta \frac{\Delta t}{\tau P} \left(P_0 - P \right), \tag{15}$$

with the isothermal compressibility β, the simulation timestep Δt and the reference pressure P_0. The Berendsen barostat does not result in the exact NPT ensemble and does not compute the correct fluctuations in pressure or volume, but it is fast for relaxation without oscillations. As such, it is typically only used briefly at the start of a simulation during equilibration, while for longer simulations the Parrinello-Rahman barostat [52] is used.

It is worth noting that pressure coupling does not play well with QM/MM simulations (Section 2.12), because the MD engine does not account for the contribution of QM/MM forces to the system virial. As such, those simulations will use the canonical ensemble.

2.7 Long-range interactions

Explicitly computing all non-bonded interactions between all particles becomes exceedingly costly as the number of particles increases. It gets further complicated by the use of periodic boundary conditions. Coulomb interactions decrease with a rate of $1/r^2$ (see Equation 11), while the Lennard-Jones potential decreases with $1/r^6$ (see Equation 10). As such, the non-bonded van der Waals interactions can readily be treated with a cutoff, whereby no forces are calculated for interactions beyond a certain distance (typically around 1 nm), as their contribution would be negligible. The interactions for the slowly decaying electrostatic interactions, on the other hand, require more finesse. Particle Mesh Ewald (PME) [53] is based on Ewald summation [54], but operates on a grid. Charges are assigned to the nearest grid point and forces on grid points as the result of Coulomb interactions are redistributed to the atoms around the grid point. Having a grid and periodic boundary conditions allows the PME method to solve the problem in frequency space after a Fourier transform, simplifying the computation down to a single sum over the grid. With this, PME scales with $\mathcal{O}(N \log N)$.

If we had to compute which particle is within the cutoff distance of which other particle for all particles at every step, we would not have gained much from using a cutoff in the first place. To alleviate this, a buffered neighbor list is used. Particles keep track of their neighbors within a region that is slightly larger than the cutoff distance (the Verlet buffer [55]). The neighbor list can then be updated less frequently than at every timestep and use clusters of particles [56].

2.8 External biasing potentials

Since we are investigating the influence of external forces on collagen, we also need a way of adding those in our molecular dynamics simulations. Force-probe MD simulations allow biasing the energy of a system by external forces [57]. The force can be added to individual atoms or the center of mass (COM) of groups of atoms. To this end, a virtual spring is attached and moved in the direction of a defined pulling direction. This can either be defined by absolute coordinates or through the distance between pull groups. The spring can either add a constant force (constant force pulling), increase the force over time (force-ramp) or the virtual attachment point of the spring itself can travel at a constant speed along the pulling coordinate (constant velocity pulling). In the latter case the force can vary over time and will depend on the pulling speed and the force constant k of the harmonic potential of the virtual spring.

2.9 Umbrella sampling and the weighted histogram analysis method

By using a harmonic potential, sometimes also called umbrella potential, to trace out a reaction coordinate, we get an initial trajectory. From this initial trajectory, snapshots (umbrella windows) are extracted along the reaction coordinate. The reaction coordinate is a collective variable that can be made up from a combination of measurements such as positions, distances, and angles. From those snapshots, more simulations are started, with the umbrella potential this time applied to keep the collective variable around the same value as initially in the snapshot. Together, these simulations can be used to estimate the free energy profile along the reaction coordinate. This free energy profile is sometimes also called the potential of mean force (PMF), although this is not technically correct due to the inclusion of entropic effects in the sampling [58].

Figure 2.10: Because the harmonic potential holding each umbrella window in place along the reaction coordinate ξ is often also called an umbrella potential, I was initially quite confused: “That’s not how you hold an umbrella!” It starts to make more sense when referring to the histograms of the umbrella windows (lots of small umbrellas, Figure 3.9d), but I nonetheless made this squirrel holding an upside-down umbrella for a seminar using a combination of manual painting and StableDiffusion [59] running locally via InvokeAI [60].

The weighted histogram analysis method (WHAM) [61–63] is then used to analyze the resulting forces and collective variable values. With a reaction coordinate ξ, the goal is to find the free energy profile $A(\xi)$. Given the probability $P(\xi)$ for visiting a point in phase space along the reaction coordinate, the free energy is given by Equation 16.

$$A(\xi) = -k_{\mathrm{B}} T \ln P(\xi), \tag{16}$$

with the temperature T and the Boltzmann constant k_{B}. In the limit of infinite sampling, we could just run our simulations forever to obtain the true probability distribution P, but with the umbrella windows along the reaction coordinate we can speed things up. Each umbrella window i from 1 to n produces a biased probability distribution $P_i^{bias}(\xi)$ with the biasing umbrella potential $V_i(\xi)$ and the free-energy contribution from the biasing potential F_i. Together, the free energy of one window is described in terms of the biased distributions by Equation 17.

$$A_i(\xi) = -k_{\mathrm{B}}T \ln P_i^{\mathrm{bias}}(\xi) - V_i(\xi) + F_i \tag{17}$$

Next, the weights w_i of Equation 18 are adjusted to minimize statistical error.

$$P(\xi) = \sum_{i=1}^{n} w_i P_i^{\mathrm{bias}}(\xi) \tag{18}$$

This makes it another self-consistent method, aided by the fact the umbrella windows, or rather their histograms along the reaction coordinate, ξ overlap. The more the umbrella windows agree on the resulting unbiased probability distribution in a region $\Delta\xi$, the higher the convergence. The result is a free energy profile and the histograms of the umbrella windows, which can be used to visualize overlap. Bootstrapping allows for an estimate of the uncertainty of the free energy profile by comparing the predictions with varying holdouts [63].

2.10 Free energy perturbation by slow-growth

Slow-growth is another method of calculating free energy differences, although you will find that in this thesis it is used for a different purpose. The underlying idea is that the Hamiltonian is being made dependent on an additional term λ next to the positions $\vec{r}$ and momenta $\vec{p}$ (Equation 19):

$$\lambda : H = H(\vec{r}, \vec{p}, \lambda), \tag{19}$$

where the coupling parameter λ is used to transition from the Hamiltonian in state A at $\lambda = 0$ to the Hamiltonian in state B at $\lambda = 1$ in Equation 20.

$$\begin{array}{rcl} H(\vec{r}, \vec{p}, 0) & = & H^{\mathrm{A}}(\vec{r}, \vec{p}) \\ H(\vec{r}, \vec{p}, 1) & = & H^{\mathrm{B}}(\vec{r}, \vec{p}) \end{array} \tag{20}$$

The computational engine provides a way of specifying different parameters for state A and state B for the energy terms. For example, a harmonic bond may start with a spring constant k_{A} and transition to a spring constant[4] k_{B}.

[4] though at this point it is a bit ironic to call it a spring *constant*.

2.11 Quantum mechanics and density functional theory

A key limitation of MM is that bonds can not form or break, i.e., chemistry can not happen, because all connectivity and parameters thereof have to be defined ahead of the simulation. So, let us take a look at a more detailed but also computationally expensive technique. In quantum mechanical (QM) simulations, we are no longer treating atoms as indivisible units and instead also simulate the electrons around the nuclei to some extent. To what extent depends on the concrete method and implementation. We will be looking at density functional theory (DFT) in particular, but first we need to lay some groundwork. The overarching idea of MD remains, but the form of our energy function will change quite a bit.

2.11.1 The Schrödinger equation

The general equation for the energy of a molecule (or system) is the Schrödinger equation (Equation 21) [64].

$$\hat{H}\Psi = E\Psi \tag{21}$$

This is an eigenvalue equation where the eigenvalue of the Hamilton operator $\hat{H}$ acting on the wavefunction Ψ (a set of eigenfunctions of the Hamiltonian) is the total energy E. For each eigenfunction Ψ_n we have an associated energy E_n. The wavefunction describes the matter wave of each particle, where, according to de Broglie, the wavelength λ of the wave is the Planck constant h divided by the momentum p (Equation 22) [65].

$$\lambda = \frac{h}{p} \tag{22}$$

Furthermore, the Born rule [66] tells us that the wavefunction times its own complex conjugate results in a probability density. So the probability of finding a particle in a volume slice dr is equal to the integral over dr, normalized such that Equation 23 holds:

$$\int_{-\infty}^{\infty} \Psi^*(r)\Psi(r)\,dr = P = 1 \tag{23}$$

The Schrödinger equation is still incredibly hard to solve for, but we can get a step closer by invoking the Born-Oppenheimer approximation [67]: Due to the electrons being so much lighter and faster than the nuclei, we can solve for both systems independently of each other. As far as the electrons are concerned, the nuclei are pretty much stationary.

2.11.2 The electronic Schrödinger equation

This leads us to the electronic Schrödinger equation (Equation 24):

$$\hat{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R})\Psi_{\mathrm{el}}(\mathbf{r},\mathbf{R}) = E_{\mathrm{el}}\Psi_{\mathrm{el}}(\mathbf{r},\mathbf{R}) \tag{24}$$

In the following, Ψ will refer to Ψ_{el} and we can expand the equation to Equation 25

$$\begin{aligned} \hat{H}\Psi &= E\Psi \\ &= \left[\hat{T} + \hat{V} + \hat{U}\right]\Psi \\ &= \left[\sum_{i=1}^{N}\left(-\frac{\hbar^2}{2m_i}\nabla_i^2\right) + \sum_{i=1}^{N} V(\mathbf{r}_i) + \sum_{i<j}^{N} U\left(\mathbf{r}_i, \mathbf{r}_j\right)\right]\Psi, \end{aligned} \tag{25}$$

where T is the kinetic energy of the electrons (or respectively $\hat{T}$ is the kinetic energy operator), V is the potential energy of the electrons due to the interaction with the (now assumed static) nuclei and U is the potential energy of electron-electron repulsion. N is the number of electrons.

This equation, although looking deceptively simple, is still near impossible to solve for any real-world system. The next step is to find easier-to-solve-for formulations of the wavefunction.

2.11.3 Hartree product

The Hartree product [68] describes the electronic wavefunction as a product of individual electron wavefunctions (Equation 26):

$$\Psi(x_1, x_2, x_3, \dots, x_n) = \psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\dots\psi_n(x_n), \tag{26}$$

where x is the state of an electron given by its position r and spin ω. This is not quite sufficient, because, as fermions, electrons are meant to be indistinguishable, and their wavefunction is meant to be antisymmetric, meaning it should change sign upon exchange of any two coordinates. We can enforce this by writing the wavefunction a bit differently. For a two-particle system this looks like Equation 27:

$$\Psi(x_1, x_2) = \frac{1}{\sqrt{2}}\left[\psi_1(x_1)\psi_2(x_2) - \psi_1(x_2)\psi_2(x_1)\right], \tag{27}$$

which satisfies antisymmetry (Equation 28).

$$\Psi(x_1, x_2) = -\Psi(x_2, x_1) \tag{28}$$

2.11.4 Slater determinant

This generalizes to more particles in the form of a determinant, the Slater determinant [69] (Equation 29):

$$\Psi(x_1, x_2)\frac{1}{\sqrt{2}} = \begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) \end{vmatrix}, \tag{29}$$

where taking a 2-electron system to an N-electron system means replacing the 2×2 matrix with an $N \times N$ matrix and 2 turning into $N!$. We now have a way of constructing wavefunctions or molecular orbitals (MOs), in which electrons can live. But how do we find the correct functions?

2.11.5 The Hartree-Fock method

The Hartree-Fock method [70] now assumes that the electronic energy eigenfunctions can be described as a single Slater determinant made up of individual one-electron wavefunctions. Each such one-electron wavefunction (a molecular orbital) is built up from a finite set of basis functions and is thus a linear combination of atomic orbitals (LCAO). The orbitals are then varied such that the Hartree-Fock energy is minimized under the assumption that each electron only interacts with the average charge density of all other electrons (variational method and mean-field approximation). This is also called the self-consistent field method.

2.11.6 Hohenberg-Kohn theorems

With the Hohenberg-Kohn theorems [71]:

1. The ground state energy is a unique functional of the electron density:
   $E = E_0[\eta_0(r)]$
2. The electron density function that minimizes the energy is the true electron density:
   $\forall \eta : E[\eta(r)] \geq E_0[\eta_0(r)]$,

we can apply the basic principles to larger systems, because we now “only” have to compute the electron density function as opposed to all individual electron wavefunctions.

2.11.7 Kohn-Sham density functional theory

In Kohn-Sham Density Functional Theory (KS DFT) [72] we compute the energy of electrons by assuming that the density consists of a single Slater determinant (non-interacting limit) (Equation 30), which we can then treat like in the Hartree-Fock method (i.e., with a variational, self-consistent field approach).

$$E[\rho] = T[\rho] + \int d\mathbf{r}\; v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) + E_{\text{Coulomb}}[\rho] + E_{\text{xc}}[\rho], \tag{30}$$

with electron density ρ, kinetic energy T, external potential v, position in space $\mathbf{r}$, and the errors from electron-electron exchange and correlation lumped into an exchange correlation energy E_{xc}. The type of exchange correlation functional (XC) used has a large impact on computational cost and accuracy of the calculations. XC functionals can be ordered in terms of their accuracy. The local density approximation (LDA), where the functional only depends on the local electron density at a given point, is exceeded by the generalized gradient approximation (GGA), where the functional also depends on the local gradient of the density (e.g., LYP after Lee, Yang, and Parr [73]). Adding a second derivative leads to meta-GGA and hybrid DFT mixes in exact exchange (e.g., B3LYP after Becke 3-parameter and LYP [73, 74] or PBE0 after Perdew–Burke–Ernzerhof [75, 76]). Lastly, double-hybrid methods such as MP2 [77] add Møller–Plesset perturbation theory [78] electron correlation as well.

2.11.8 Basis sets

A collection of basis functions that we can use in DFT is called a basis set. A linear combination of basis functions makes up a molecular orbital (MO) for one electron, which is then combined with other MOs in one Slater determinant for the whole system. For the basis functions, we have the choice between Slater-type orbitals (STO), which are the typical atomic orbitals $Nx^a y^b z^c e^{-\zeta r}$ (where N is a normalization constant, a, b, c control the angular momentum, and ζ controls the width) and Gaussian-type orbitals (GTO) $Nx^a y^b z^c e^{-\zeta r^2}$. Gaussians have the added benefit that the product of two Gaussians is another Gaussian, which makes them mathematically favored to work with, even though they are less physical. Contracting multiple GTOs via a linear combination (CGTO) is used to mimic a single STO at lower computational cost.

A minimal basis set encompasses one basis function for each atomic orbital (AO). Using two basis functions per AO is called double-zeta (and so forth). This allows one to more accurately capture interactions between atoms. Doing so only for the valence AOs and keeping the core AOs at single-zeta is called a split valence basis set. Allowing AOs to be polarized by interactions with other atoms requires mixing each AO with an orbital of a higher angular momentum (l+1), e.g., mixing s orbitals with p orbitals. This is then called a polarized basis set.

2.12 Quantum mechanics/molecular mechanics

While it is feasible to simulate millions of atoms with molecular mechanics, DFT can only reach orders of thousands of atoms at great computational cost. We need a method that not only can simulate a large enough portion of a collagen fibril with sufficient sampling time but can also model the chemistry of hydrolysis. Quantum mechanics/molecular mechanics (QM/MM) combines both methods. A large system is simulated by molecular mechanics, while a small subsystem is treated quantum mechanically.

There are different ways of combining the two systems. The method used in this thesis is called Gaussian expansion of the electrostatic potential (GEEP) [79] and is a form of electrostatic embedding. While in a purely mechanical embedding, the electrostatic interactions between MM and QM atoms would only be treated by molecular mechanics terms, in an electrostatic (or electronic) embedding, the QM subsystem also sees the charges of the MM atoms and can be polarized by those. One step up from that would be a polarized embedding, where the MM subsystem can also be polarized by the QM region. The QM/MM simulations in this thesis use the electrostatic embedding (Figure 2.11).

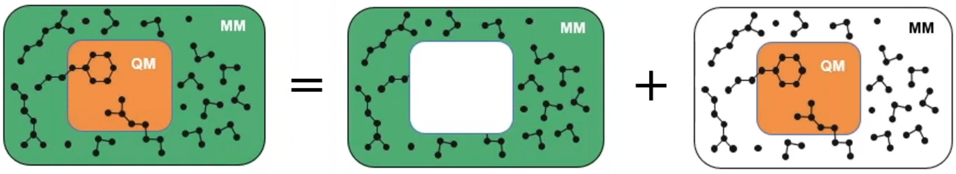


Figure 2.11: Electrostatic embedding scheme for QM/MM. Figure by Dmitry Morozov [80].

The new Hamiltonian of the QM subsystem consists of the original Hamiltonian with the addition of the electrostatic interactions between the electrons of the QM region and the MM atoms and between the nuclei of the QM region and the MM atoms (Equation 31, from the GROMACS documentation [81]):

$$H^{\mathrm{QM/MM}} = H_{\mathrm{el}}^{QM} - \sum_{i}^{n}\sum_{J}^{M}\frac{e^2 Q_J}{4\pi\epsilon_0 r_{iJ}} + \sum_{A}^{N}\sum_{J}^{M}\frac{e^2 Z_A Q_J}{e\pi\epsilon_0 R_{AJ}}, \tag{31}$$

with the number of electrons n and nuclei N in the QM region and the number of charged MM atoms M. Here, “charged” also includes partial charges, not just ions.

H_{el}^{QM} is the original electronic Hamiltonian of an isolated QM system. The first double sum (n,M) is the electrostatic interaction between the QM electrons and the MM atoms. The second double sum (N,M) is the electrostatic interaction of the QM nuclei with the MM atoms. Lennard-Jones and Coulomb interactions between QM atoms are excluded in the MM Hamiltonian to avoid double counting, but Lennard-Jones interactions between QM and MM atoms are calculated.

Bonded interactions at the QM-MM interface that contain at most one QM atom are described by MM. Bonds crossing the QM-MM interface are capped with a so-called link atom, a hydrogen that only exists in the QM region. The force on this atom is distributed over the two atoms of the bond.

2.13 Computational molecular dynamics engines

The molecular dynamics engine of choice for this thesis is GROMACS [40] with documentation available at manual.gromacs.org [57]. Our newly developed software method, KIMMDY, was tested with GROMACS versions ranging from v2021 to v2025.3. Early QM/MM simulations to test feasibility used the developer build of GROMACS v2022. The production QM/MM simulations in this thesis use GROMACS v2023.2. The QM engine used and interfaced with GROMACS is CP2K [82] v2023.2 available from cp2k.org.

2.14 Kinetic Monte Carlo

Given that our method KIMMDY, which will be introduced in Chapter 4, stands for Kinetic Monte Carlo Molecular Dynamics, it makes sense to also explain what kinetic Monte Carlo (kMC) is. Monte Carlo methods in general are methods that use random sampling to solve a numerical problem. In kinetic Monte Carlo, we are modeling the time progression through a collection of states connected by processes with known transition rates. The simplest example is rejection-free kMC (rf-kMC) [83, 84]:

We start at time $t = 0$ in a state S_k and collect the rates r_{ki} for all processes leading to new states S_l. The list of rates has length N_k and is indexed by i. Stack the rates on top of each other and normalize to a total rate of 1, which results in the normalized cumulative function $R'_k(i)$ in Equation 32 with a rate sum of Q_k (Figure 2.12):

$$\begin{aligned} R_k(i) &= \sum_{j=1}^{i} r_{ki} \\ Q_k &= R_k(N_k) \\ R'_k(i) &= \frac{R_k(i)}{Q_k} \end{aligned} \tag{32}$$

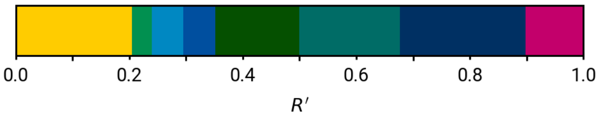


Figure 2.12: Stacked rates to sample from during kMC.

Then, draw a random number $u \in (0, 1]$ from a uniform distribution. From the normalized cumulative function, find the event i for which $R'_k(i-1) < u \leq R'_k(i)$. Execute the process tied to event i. Draw a second uniform random number $u' \in (0, 1]$. Determine $\Delta t = \frac{\ln(1/u')}{Q_k}$ and set the new time $t = t + \Delta t$. Repeat all steps until sufficient sampling of the states has been achieved.

The event list of a state and the list of possible states are not always known beforehand. Adaptive kMC [85] characterizes a method in which the event list is calculated on the fly.

2.15 Arrhenius equation and transition state theory

In order to apply kinetic Monte Carlo, we need reaction rates.

The Arrhenius equation Equation 33 describes an empirical relationship between the reaction rate k and the activation energy E_a of a reaction [86].

$$k = A\,\mathrm{e}^{\left(\frac{-E_\mathrm{a}}{RT}\right)}, \tag{33}$$

with the pre-exponential factor A and the gas constant R as well as the temperature T. If E_a is given per molecule and not per mole, R is replaced by the Boltzmann constant k_B.

The Eyring equation [87] contains a way of calculating the pre-exponential factor, or attempt frequency, via Equation 34.

$$A = \kappa \frac{k_\mathrm{B} T}{h}, \tag{34}$$

with the transmission coefficient κ, which represents the fraction of reaction paths through the transition state that end in the product state without passing the transition state again. A transition state is a saddle point in the potential energy surface between the educt and product state. κ is often assumed to be 1, which equals accepting the fundamental assumption of transition state theory, in which the educt state and the transition state are in equilibrium, but once the transition state has been passed, the reaction proceeds to the product state.

When experimental data for reaction rates at different temperatures is available, A can be determined for the specific reaction based on an Arrhenius plot. Here, the logarithm of the reaction rate ($\ln(k)$) is plotted against the reciprocal temperature $1/T$. With the Arrhenius equation translated to the form in Equation 35:

$$\ln(k) = \ln(A) - \frac{E_\mathrm{a}}{R}\left(\frac{1}{T}\right), \tag{35}$$

the y-intercept of a log-log-linear fit corresponds to $\ln(A)$ and the slope of the fit equals $-E_\mathrm{a}/R$.

3 Collagen hydrolysis by quantum mechanics/molecular mechanics

As introduced earlier, collagen under mechanical stress can undergo two competing reactions: peptide bond hydrolysis (breaking the bond through reaction with water) and homolytic bond scission (direct mechanical rupture producing radicals). While experiments have detected radicals in tensed collagen [17], and other studies showed that hydrolysis is accelerated by force [20], it remained unclear how these two pathways compete under physiologically relevant conditions.

Simulating an entire collagen fibril at the quantum mechanical level of detail would be computationally infeasible. This is why I reduced the core question to a comparison between a single peptide, as was the case in the single-molecule pulling experiments by Pill et al. [20], and a collagen triple helix. In graphics and schematics these systems will be referred to as `single` and `triple`. This chapter addresses the hydrolysis pathway using hybrid quantum mechanics/molecular mechanics (QM/MM) simulations, the method of choice for reactive subsystems for large biomolecular systems (see Section 2.12). The effect of the fibrillar assembly will come into play again in Chapter 4, but this chapter will only be concerned with *how* the hydrolysis reaction happens.

It is worth pointing out that while it makes sense to tell the story in this order, QM/MM followed by KIMMDY (Chapter 4), many discoveries are the culmination of things happening in parallel. Thus, a key reason as to why it makes sense to focus the QM/MM simulations on just a triple helix as opposed to a complete fibril is only explained later in Section 4.3.2, where I talk about the solvent accessibility of the peptide bond.

Originally I had planned to simulate both homolytic and hydrolytic bond cleavage in the same QM/MM setup to get a direct comparison. However, it is a fundamental limitation of my chosen QM methodology, density functional theory (DFT), that a system has to be representable by a single Slater determinant – meaning the total spin of the system cannot change during the simulation. This renders a direct sampling approach to homolytic (and thus radical) cleavage reactions futile, as bond homolysis creates two unpaired electrons. As such, this thesis will not contain QM/MM simulations of homolysis, and I will assume instead, as is commonly done in the field [88–90], that the transition state energy of the cleavage reaction is very close to the dissociation energy of the respective bond, wherever a direct comparison is required. In the following I will thus focus on base-catalyzed hydrolysis and my extensive sampling of the reaction in a QM/MM scheme.

3.1 Previous work on base-catalyzed hydrolysis

Base-catalyzed hydrolysis is a multi-step reaction that breaks a peptide bond through nucleophilic attack by a hydroxide ion (OH^-) (Figure 3.1). The mechanism proceeds as follows: First, the catalytic hydroxide (acting as a nucleophile) attacks the electrophilic carbon atom of the peptide bond's carbonyl group. This attack goes through a transition state TS1, overcoming the associated energy barrier, to form a tetrahedral intermediate (TI), where the carbon is temporarily bonded to four substituents. The TI is stabilized by hydrogen bonds with surrounding solvent molecules or potentially with protons from the protein backbone – a stabilization mechanism worth investigating. Next, the protonation of the peptide bond nitrogen leads to a zwitterionic intermediate (ZI), which can then rupture at the peptide bond (passing transition state TS2), leaving separate C- and N-termini that can, depending on the pH value, also be protonated or deprotonated. Ultimately, since the proton to form the ZI comes from water, a hydroxide ion is regenerated, making it truly catalytic.

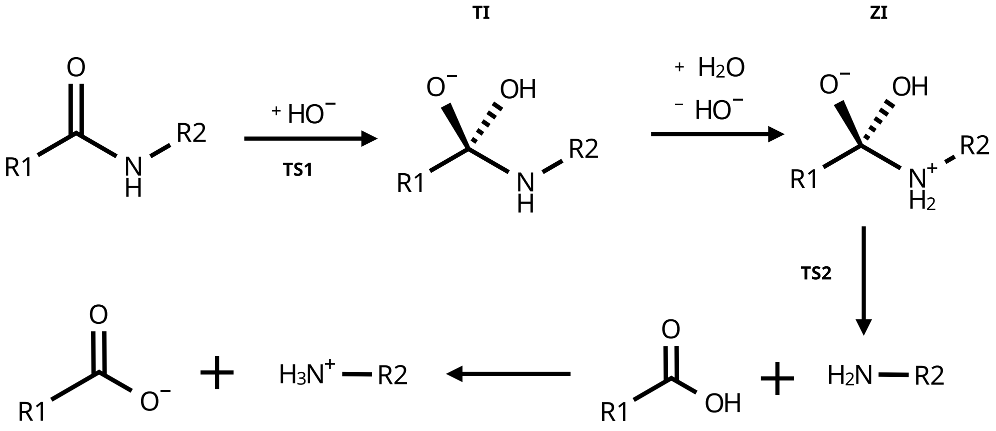


Figure 3.1: Reaction schema of base-catalyzed peptide bond hydrolysis. TS is a transition state, TI is the tetrahedral intermediate and ZI is the zwitterionic state.

To put the results of this chapter into context, I compiled previous results by Pill et al. [20] into a single figure (Figure 3.3b). It is worth going into some detail about how these results were obtained to make the comparison to my results clearer. Broadly, they obtained reaction rates and energies by two distinct methods, designed to investigate how external force affects base-catalyzed hydrolysis:

Experimental approach: A single-molecule pulling setup with an atomic force microscopy (AFM) cantilever measured bond lifetimes under varying applied forces and temperatures. These lifetimes can be translated to activation energies and attempt frequencies (pre-exponential factors) via an Arrhenius plot (see Section 2.15).

Computational approach: They calculated *ab initio* single-point energies at the MP2/TZVP level of theory for B3LYP-optimized structures (using the 6-31+G(d) basis set) under external force (Figure 3.2) for each of the states shown in Figure 3.1. As shown in Figure 3.3a, those energies are comprised of the energies of three auxiliary systems that are minimized independently of each other

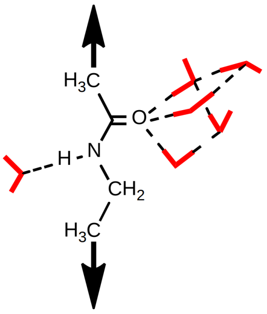
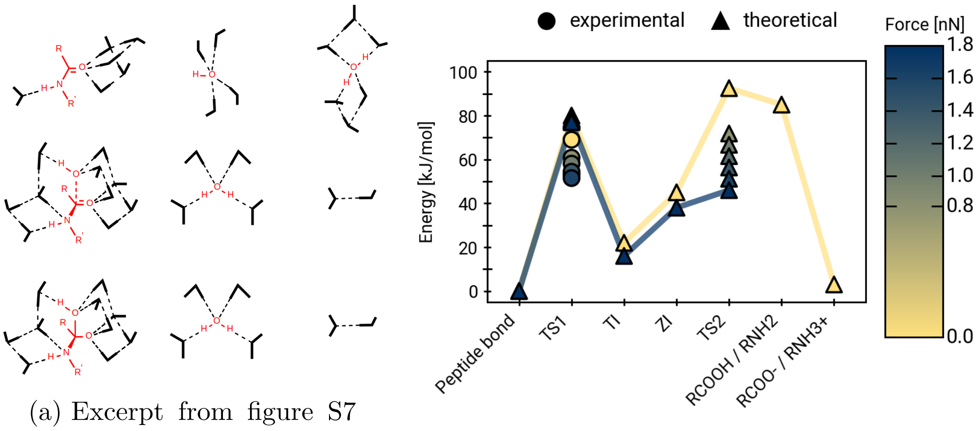


Figure 3.2: Illustration of the pulling setup from the SI of [20].

and chosen in such a way that the total number of water molecules, hydrogen bridges, and number of strained HOH bond angles stays the same. The first subsystem contains the actual peptide bond as a part of N-ethylethanamide; the second and third contain the initial hydroxide ion and more water molecules to balance out the quantities mentioned above that are to stay constant.



(a) Excerpt from figure S7 of the SI of [20] showing the auxiliary subsystems for the first 3 reaction steps. From top to bottom: educt, TS1, TI.

(b) (Force-dependent) energies of the reaction steps based on AFM (experimental) and *ab initio* QM (theoretical) by [20]. The energy of the educt state is subtracted as a baseline.

Figure 3.3: Visual summary of the setup and data points obtained by [20] that are relevant for this thesis.

Using this setup, they found that the first energy barrier (TS1) is barely lowered by adding force, while the second energy barrier (TS2) is dramatically lowered even by adding small external forces. This makes intuitive sense: the attack of the hydroxide occurs perpendicular to the axis of applied force, while the breaking apart of the ZI happens along the pulling axis. Notably, the energy barriers for TS1 calculated from experimental bond lifetimes are significantly lower than the barriers from the *ab initio* QM calculations. For comparison, the reaction rate obtained from experiments at an external force of 1 nN was k=6.89 ± 0.85 s^{-1} with an energy barrier of 58.1 ± 2.8 kJ/mol and an attempt frequency of $1.3\times10^{11\pm0.5}$ s^{-1} (or an overall average attempt frequency A of 10^{11} s^{-1}. The total rate constant from the QM calculations can be derived from the individual energies of TS1 and TS2 with Equation 36

$$\begin{aligned} k1 &= A1e^{-E_{TS1}/k_{\mathrm{B}}T} \\ k2 &= A1e^{-E_{TS2}/k_{\mathrm{B}}T} \\ k &= \frac{k_1k_2[OH^-]}{k_1+k_2}, \end{aligned} \tag{36}$$

with the concentration of hydroxide $[OH^-]$ coming in as 10^-(14 - 7.4)^ at the experimental pH of 7.4 and a temperature of 21 °C (294.15 K). The auxiliary attempt frequency $A1$ was assumed fixed such that $A1 * [OH^-] = A$ matches the average attempt frequency A gathered from the AFM experiments. Together this results in a theoretical reaction rate of 10^{-3} s^{-1}.

These results serve as a reference for my calculations and on top of that they will also play a role in Chapter 4.

3.2 Goals of the calculations

The goal of my research into hydrolysis was to go beyond highly optimized and reduced model systems and explore how hydrolysis behaves in a complex, dynamical environment. Rather than using small, pre-optimized structures, the following will employ large-scale QM/MM simulations of peptides where the QM region encompasses two whole residues (and their connecting peptide bond) plus surrounding water molecules. This approach directly samples the reaction in a more realistic, fluctuating environment, providing the unique opportunity to observe reaction dynamics and the sequence of events to complement single-point energies.

Two caveats apply when comparing energies between different approaches. Firstly, my pipeline will measure free energy differences, which incorporate entropic contributions, as opposed to the pure electronic energies shown in Figure 3.3b. Secondly, the most favorable conformation, as is found by energy minimization, might not be covered by phase space sampling. Exploring the accessibility of said conformation is in itself a feature of the sampling approach, but it does also mean that values for energy barriers will differ between approaches. This is why the focus will be on trends between systems and not on the absolute values of the energies and energy barriers.

The overarching hypothesis is that the triple helical structure of collagen peptides in particular should hinder the approach of the hydroxide and energetically disfavor the tetrahedral intermediate.

3.3 System setup and simulation pipeline

In my early tests I found that a combination of 14 QM water molecules and the TZV2P-MOLOPT-GTH basis set [91] with the PBE functional [92] provides the most stable simulations. Due to its good polarizability and diffuse characteristics, this basis set can stabilize the negative charge in the QM region (originating from the hydroxide ion) and allows charge delocalization via proton transfer reactions. The dynamic nature of the protons in the QM region (meaning they can hop between water molecules and the reactants) is a key feature of this setup and will be explored further in the results section.

3.3.1 Initial structure preparation

The triple helix initial structure is a cut-out from a *Homo sapiens* collagen type I fibril without cross-links that was obtained from ColBuilder [93]. It is composed of three chains (A, B, and C) with individual lengths of 45, 43, and 45 amino acids, respectively (Figure 3.5). The N-terminus (displayed on the left of each molecular render, Figure 3.5, Figure 3.6, Figure 3.8) was acetyl-capped (ACE), while the C-terminus was capped with an N-methyl group (NME) in PyMOL [94] to account for the fact that the actual collagen protein is longer than the cut-out simulated here. From this point onward, everything is handled by my custom automated simulation and analysis pipeline. The code for this is available at github.com/graeter-group/hydrolysis-qmmm-workflow. This version does not include the outputs of the simulations, as those would be way too large. They are instead archived locally at the Heidelberg Institute for Theoretical Studies (HITS) and available upon request. The workflow is organized in Python modules and controlled from a central notebook, `index.qmd`, loading the modules. Simulations are computed locally or dispatched to the in-house Slurm [95] high performance computing (HPC) cluster. Interactions with GROMACS are solely based on the command line interface via files and system calls to shell commands from Python. A `settings.py` file creates a permutation matrix of input conditions and other parameters that were varied and iterated on during development of the pipeline. Templating allows sharing assets such as `mdp` files across systems and parameters. A file-based caching mechanism makes sure that work doesn't have to be repeated unless explicitly required. Iterations of the code were tracked with git, but the public repository is a freshly instanced one to avoid uploading large files that were only tracked during development to provide synchronization between a remote and local workstation. In addition to the simulation pipeline, there is an analysis pipeline based on a directed acyclic graph (DAG) built with Apache Hamilton. The cache of the DAG can also be synchronized between workstations with a custom script, which allows local analysis of otherwise unwieldy data.

Throughout my work I embrace reproducible research and open access and a lot of work went into crafting a robust and extensible workflow. But I like talking about workflows way too much, which is why I will stop myself here and refer to the code linked above or questions directed at future me instead. For now, what is more important than how the workflow is built, is what it does and the results thereof.

3.3.2 Equilibration protocol

Each system (the triple helix and each of the chains individually) starts as a PDB file. It is converted to GROMACS coordinates and a topology with `gmx pdb2gmx`. All simulations use the `amber99sb-star-ildnp` force field and partial charges for the hydroxide ion in its MM representation were ported from CHARMM36 [36]. A box is created with a minimum distance to the protein of 2 nm in each dimension, then scaled by 130%. This ensures sufficient space, especially in the direction of pulling. The protein is then solvated in TIP3P water and neutralized with 0.15 M NaCl to mimic physiological salt concentration. After energy minimization, the system is equilibrated first in the NVT ensemble (constant number of particles, volume, and temperature) and then in the NPT ensemble (constant number of particles, pressure, and temperature) while applying position restraints to the protein atoms for 100 ps each. This staged equilibration allows the solvent to relax around the protein structure before applying external forces. Each system is then equilibrated under an external force to simulate mechanical stress. In theory, this external force is an input parameter, but in practice the sampling ended up so computationally expensive that only one value for it will be explored here. A force of 1 nN per chain (603 kJ/mol/nm in GROMACS units) is applied to the center of mass (COM) of the α carbons of the capping residues in opposing directions for 100 ps (50000 steps at dt=0.002 ps). The triple helix configuration gains additional rotational restraints to prevent unwinding. This mimics it being packed inside a fibril and restrained by the surrounding triple helices. The length of the equilibration was found to be sufficient by observing the distances of the pull groups (Figure 3.4).

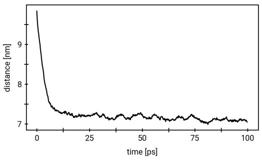
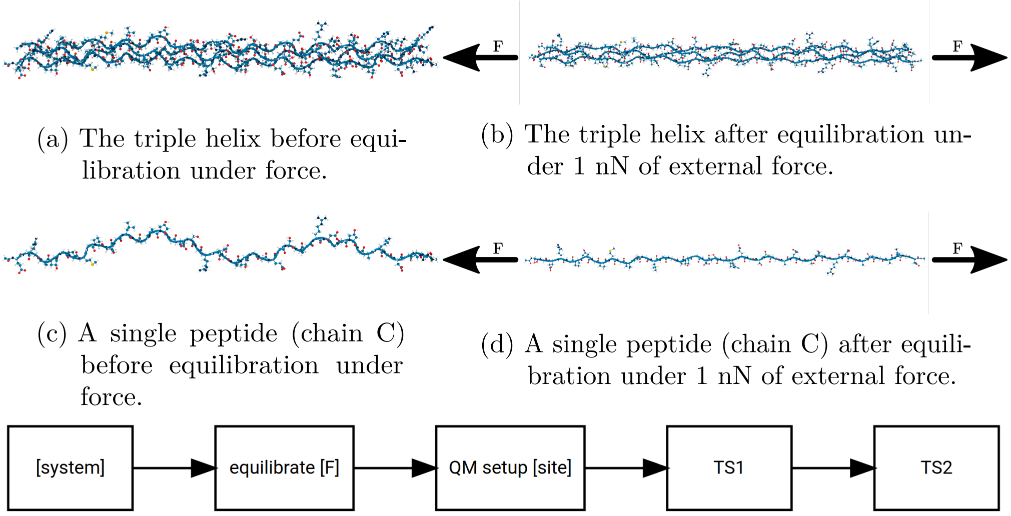


Figure 3.4: The distance between the pull groups for the external force as a function of time for the equilibration of the triple helix under 1 nN.



(a) The triple helix before equilibration under force.

(b) The triple helix after equilibration under 1 nN of external force.

(c) A single peptide (chain C) before equilibration under force.

(d) A single peptide (chain C) after equilibration under 1 nN of external force.

(e) QM/MM pipeline with choices from the input parameter space highlighted by square brackets [].

Figure 3.5: The collagen triple helix and single peptide before and after equilibration under external force.

3.3.3 Quantum mechanics subsystem selection and setup

From the force equilibration trajectory, the last 10 frames are extracted to provide multiple starting points for the actual hydrolysis reaction. This is done because QM/MM simulations can be unstable and finding an initial stable trajectory suitable for umbrella sampling is not always guaranteed. Additionally, using multiple starting configurations increases the sampling of phase space and improves statistical convergence.

In the triple helix, 8 peptide bonds were chosen at random but close to the center to avoid anisotropies in the external force, which attaches at the ends of the peptides. Each site is defined by the indices of its carbon and nitrogen atom, where the C is contributed by the residue on the N-terminal side and the N is contributed by the residue from the C-terminal side. From each starting point and chosen peptide bond, the workflow now looks for water molecules that can execute the first step of the reaction: the attack of the hydroxide (TS1 in Figure 3.1).

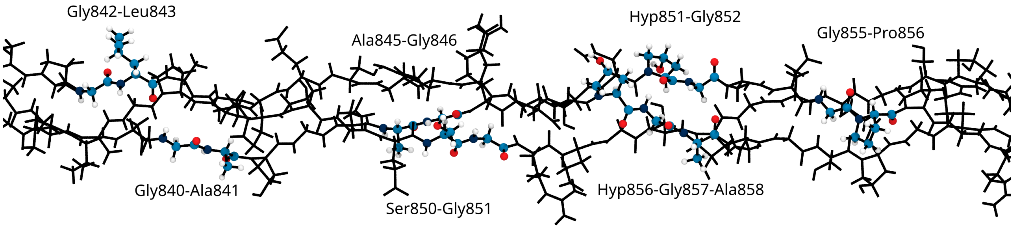


Figure 3.6: A close-up of the labeled sites that were sampled for hydrolysis. Each site exists as part of the triple helix and just in its own single chain. The QM region of each site, that is, all atoms of both residues that are connected with the peptide bond that is to be hydrolyzed, is rendered with the ball-and-stick representation in color, while the rest of the protein is shown with black lines.

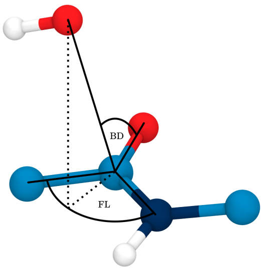


Figure 3.7: Schema of the Bürgi-Dunitz and Flippin-Lodge angles.

As visualized in Figure 3.7, the Bürgi-Dunitz angle (BD) is the angle between the approach vector of a nucleophile (OH^-) and an electrophilic center (the carbonyl carbon of the peptide bond), where the optimal angle is 107° [96]. Similarly, the Flippin-Lodge angle [97, 98] characterizes the displacement of the nucleophile at its elevation towards or away from the substituents attached to the electrophilic center. For the purpose of finding the best candidate, the optimal Flippin-Lodge angle is assumed to be the one where the distance to both substituents is maximized and defined as 0° at the angle bisector. This leads to the following penalty function Equation 37:

$$p = \sqrt{(BD - 107)^2 + (FL - 0)^2}, \tag{37}$$

where BD is the Bürgi-Dunitz angle and FL is the angle between the projection of the OH^- approach vector onto the electrophile plane and the angle bisector of the angle between the substituents and the electrophilic center.

Thus, first the 15 [5] solvent molecules closest to the carbonyl C are chosen, and out of those, the one with the lowest penalty is designated as the attack hydroxide. The remaining 14 solvent molecules are also designated as QM atoms in the QM/MM scheme. All atoms of both residues that make up the target peptide bond are marked as QM atoms. Figure 3.8 shows a close-up render of the

[5] technically this is also an input parameter.

QM/MM setup for the site Gly842-Leu843 (i.e., the peptide bond between glycine with residue ID 842 and leucine with ID 843) in the single peptide configuration. Black atoms are MM atoms and non-QM solvent molecules are not shown. Added forces are signified with the symbol F. It is worth noting that in addition to the constant external force applied to the ends of the peptide and the force from the harmonic potential that will pull the hydroxide onto the carbonyl C, there is an additional potential on the COM of the QM water molecules to keep them centered around the target peptide bond. This is to prevent them from being exchanged with MM water and I will later show in Figure 3.13 that this is indeed working as intended. Note that the hydroxide is also part of this pull group, as otherwise it would artificially increase the pressure in the QM region by being pulled towards the carbonyl carbon.

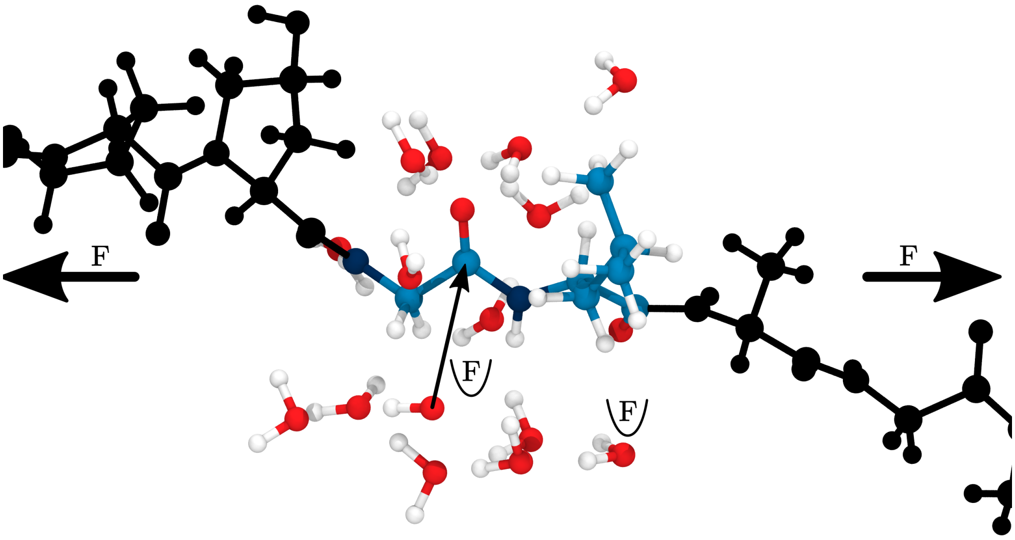
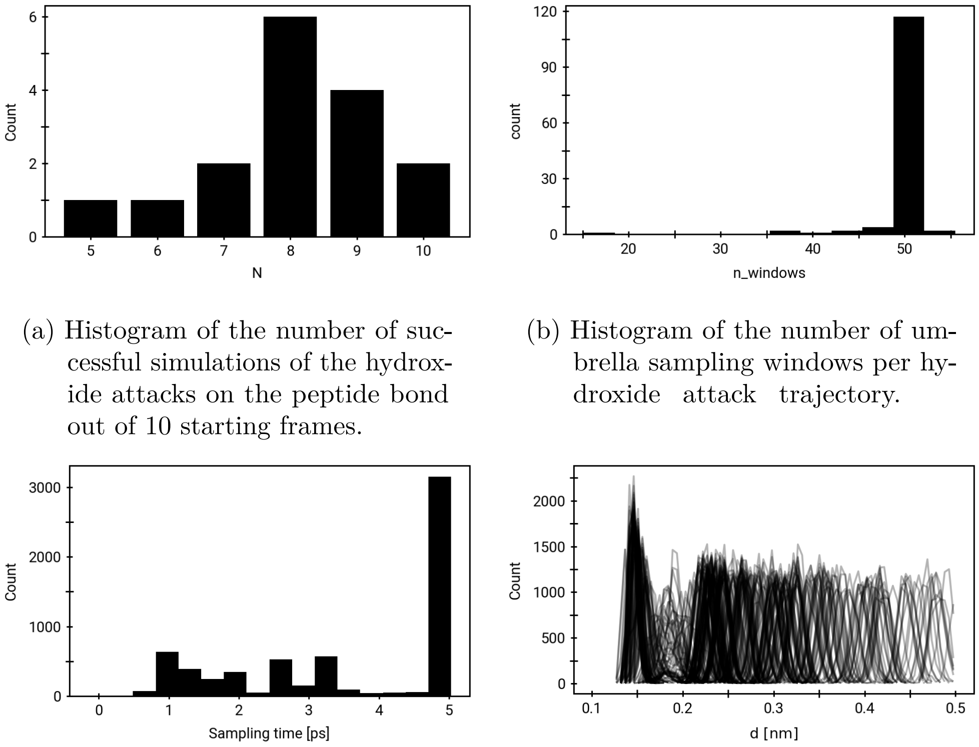


Figure 3.8: The QM subsystem (in color) embedded in the MM system (in black) with arrows for the forces involved at the start of the attack of the hydroxide for the single peptide system at a glycine-leucine site. MM water molecules are not shown.

Once the QM atoms and the hydroxide have been designated by the workflow, the QM/MM system is set up. To this end, it creates the necessary index files that GROMACS uses to label groups of atoms as, e.g., QM atoms or pull groups and fills templates for other input files such as `mdp` for GROMACS and `inp` for CP2K A short 100-step QM/MM warm-up simulation with a small delta t of just 0.2 fs increases the stability of the system before the actual approach trajectory is run. During the warm-up the hydroxide is kept at its original distance by an umbrella potential. The workflow then checks for successful warm-ups and continues with the approach simulation. The approach is simulated at a timestep of 1 fs and typically takes around 2 ps. The force constant of the harmonic potential pulling the hydroxide closer to the carbonyl C is 40000 kJ/mol/nm^2. From this initial trajectory, 50 snapshots are extracted such that there are 30 equidistant snapshots (with regard to the $OH^{-}\cdots C$ distance) below a distance of 0.25 nm and 20 snapshots above the cutoff distance. This is done to improve sampling in the critical transition state (TS) region. From these snapshots, also called umbrella sampling windows, additional simulations are started, where the distance between the hydroxide and peptide bond is kept near the initial distance of the snapshot. These simulations use a rather steep potential with a force constant of 50000 $kJ/mol/nm^2$ to ensure the peak of the free energy surface (the transition state) can actually be sampled. This results in sufficient coverage of the reaction coordinate and adequate overlap of the umbrella windows (Figure 3.9d). The umbrella sampling simulations are analyzed with WHAM (see Section 2.9) to obtain the free energy profile of the reaction step. After the umbrella sampling of TS1, the pipeline moves on to the second reaction step. Because the system contains enough QM solvent molecules that facilitate proton mobility, it is not required to artificially separate establishing the zwitterionic state (ZI) from the breaking of the peptide bond (TS2). Instead, from each of the stable TIs, a snapshot is used as the starting point of a breaking simulation, whereby an additional umbrella potential pulls apart the C and N atoms. This is then used as the initial trajectory for umbrella sampling of TS2.

Out of the 8 initially chosen hydrolysis sites, all were able to get at least 5 successful hydroxide attack trajectories (starting from the last 10 frames of the equilibration as detailed above). Many got 8 successful approaches and 2 had all 10 simulations succeed (Figure 3.9a). The criterion for success in this case is a simulation that ran without crashing at least until the distance between the carbonyl C and the hydroxide O is below 0.145 nN, which would allow for sampling of the tetrahedral intermediate (whose optimal C−O distance for the free energy minimum will later be found to be around 0.15 nN). Out of these trajectories, most managed to start the target 50 umbrella sampling windows, although a few did not establish the sampling simulation due to instabilities during the QM/MM warm-up (Figure 3.9b). Most sampling simulations managed to run their entire 5 ps, but some terminated earlier (Figure 3.9c).



(a) Histogram of the number of successful simulations of the hydroxide attacks on the peptide bond out of 10 starting frames.

(b) Histogram of the number of umbrella sampling windows per hydroxide attack trajectory.

(c) Histogram of the achieved simulation lengths of the umbrella window simulations.

(d) Histograms of the coverage of the umbrella windows over the reaction coordinate of TS1.

Figure 3.9: Statistics of the umbrella sampling simulations of TS1. All windows combined total a sampling time of 22.72 ns.

3.4 Results

3.4.1 Simulations reveal mechanistic details of the tetrahedral intermediate formation

With the simulation setup described above, initial reactive trajectories are obtained. Within the QM subsystem defined around one of the sites (see Figure 3.6), the hydroxide ion, surrounded by water, is pulled towards the peptide bond and in most cases stably forms the tetrahedral intermediate shown in Figure 3.10. The tetrahedral intermediate (also see Figure 3.11) and how it is formed warrant further investigation. For this we can look not only at the initial approach trajectory, but also at the sampling windows spawned from each initial trajectory.

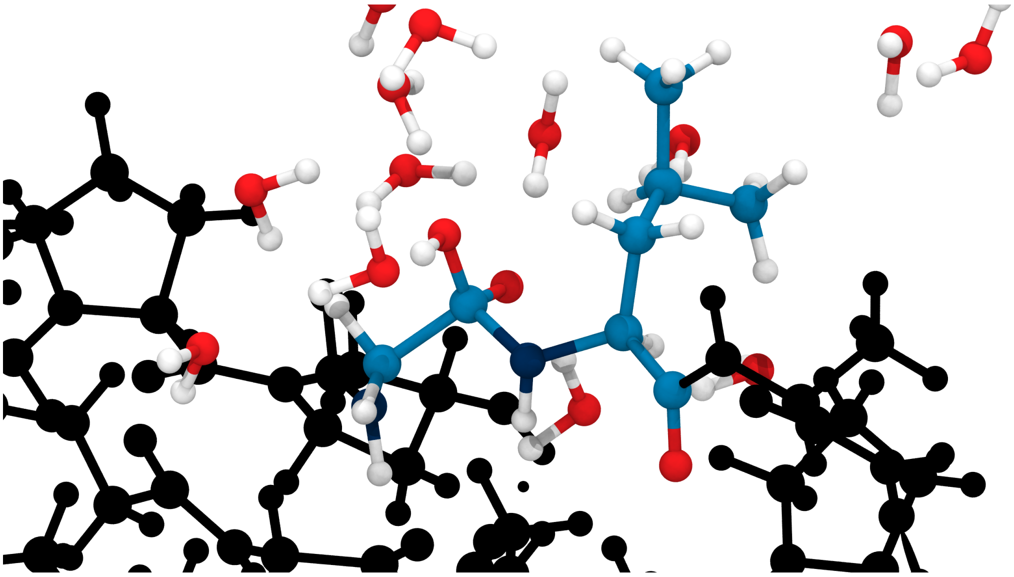


Figure 3.10: The tetrahedral intermediate (TI) state stably formed in a QM/MM simulation of the hydroxide attack in the triple helix system at a glycine-leucine site. The QM subsystem is shown in color embedded in the MM system in black. MM water molecules are not shown.

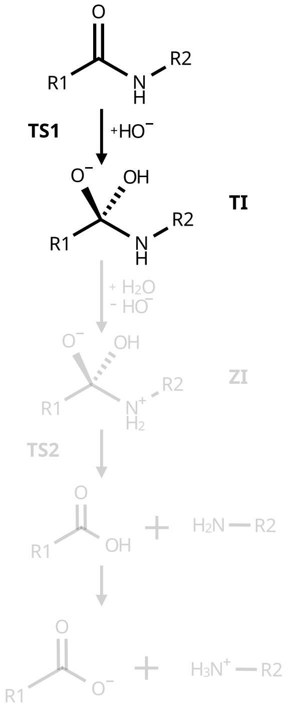


Figure 3.11: Reminder of where we are in the hydrolysis reaction.

Very few constraints are imposed on the system, so many different states with regard to protonation states are sampled. The attacking hydroxide is not forced to remain a hydroxide as it approaches the peptide bond in the initial trajectory and during the umbrella window simulations. It can be protonated by the surrounding QM water, delocalizing the negative charge over the QM region. Likewise, the carbonyl oxygen can be stabilized by hydrogen bridges or even fully protonated when the TI is formed. This led to an unexpected progression of the reaction, where instead of the addition reaction, forming the TI, a substitution reaction can take place whereby the original carbonyl O is displaced by the attacking hydroxide. This effectively results in no net reaction taking place. However, this was a rare event and only 62 out of 6359 umbrella windows had to be discarded for sampling the wrong reaction. The decision to discard a window was made based on a distance criterion as shown in Figure 5.1a. Windows where the average distance of the carbonyl C to the carbonyl O exceeds 0.155 nm were deemed accidental substitution reactions and those where the sum of the average distances of C to the carbonyl O and C to the hydroxide O was lower than 0.277 nm were deemed as collapsing TIs and also excluded. The effect of this filtering, by example of the protonation plots (Figure 3.15), is shown in extended Figure 5.2.

Looking at Figure 3.13 we can see how the TI is stabilized. Note that Figure 3.12 provides a visual representation of the legend. As the hydroxide approaches, which is read from right to left as the distance decreases, the hydroxide reaches a critical point around 0.2 nm where its negative charge influences the carbonyl O. This carbonyl consequently accumulates more charge density and is stabilized more by the surrounding QM water, which can be seen in the form of the distance to the closest proton of any QM water molecule to the carbonyl O decreasing. The QM water consistently contributes the closest proton, indicating that no MM water molecule has snuck into the QM region to replace it. Interestingly, protein H atoms rarely contribute the closest H. This is mostly due to the geometry of the peptide bonds in the sampled sites. In the triple helix formation, the carbonyl O of the peptide bond can be oriented in such a way that it points away from the other two chains, like in the exemplary TI in Figure 3.10. On average this is not more likely than the bond pointing the other way, into the helix, but the sampled sites presented in this thesis have a higher propensity of pointing outwards or being parallel to the center axis of the helix (Figure 5.1b). At the time of writing, computations for additional sites that feature the inverted orientation are running, but the results will be for another publication.

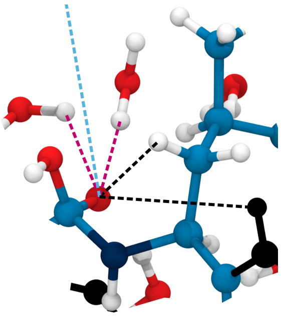


Figure 3.12: Visual explanation of the distance type color code in Figure 3.13.

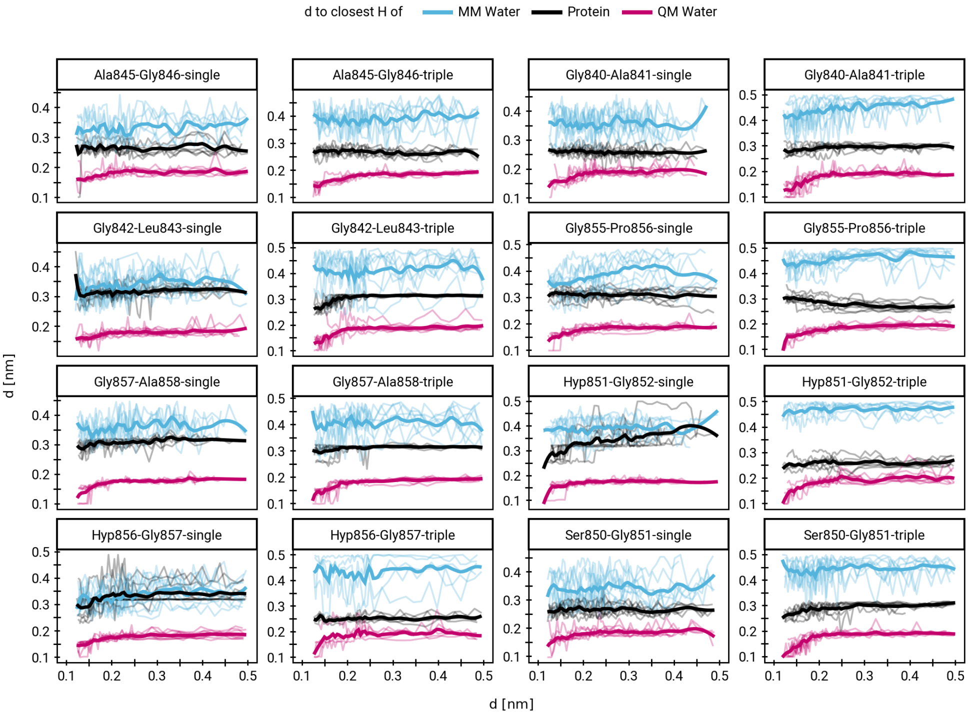


Figure 3.13: Small multiples plot of the approach of the hydroxide in terms of distances of the carbonyl O of the TI to the nearest proton of either an MM solvent molecule (cyan), the protein (black), or the QM water (magenta) across all umbrella windows. Distances are average within each window and the windows belonging to the same approach simulation are connected. The thick lines are a LOESS curve across all windows. Figure 3.12 provides a visual aid for the color coding.

Going over all trajectories and assigning each proton of a QM water or the initial QM hydroxide to the closest QM oxygen out of the QM water molecules, the carbonyl O and the initial hydroxide, results in Figure 3.15 (with annotations in Figure 3.14). The number of protons assigned to the totality of QM water molecules is not shown, as this is implicit in the sum of protons assigned to the attacking hydroxide and the carbonyl O. As the hydroxide approaches the carbonyl group, it is embedded in QM water, so it can either be in its deprotonated form or be protonated to a simple water molecule. This is why it fluctuates between having two protons and just one. At higher OH^--distances the carbonyl O very rarely gets assigned a proton, but as the hydroxide approaches and the TI is formed, the higher negative charge density attracts more protons on average. For windows in which the TI is formed, the original carbonyl O and the original hydroxide O are indistinguishable, so the hydroxide can also lose its original proton and the carbonyl O can gain a proton. All this is facilitated by the proton mobility in the QM water. Interestingly, in the triple helix systems, the carbonyl O gets protonated more in the TI than in the single chain system (Figure 3.16).

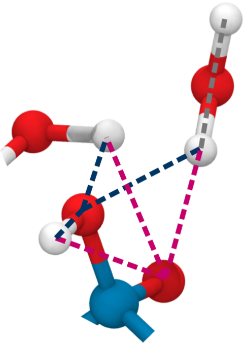


Figure 3.14: Visual explanation of the proton counting color code in Figure 3.15.

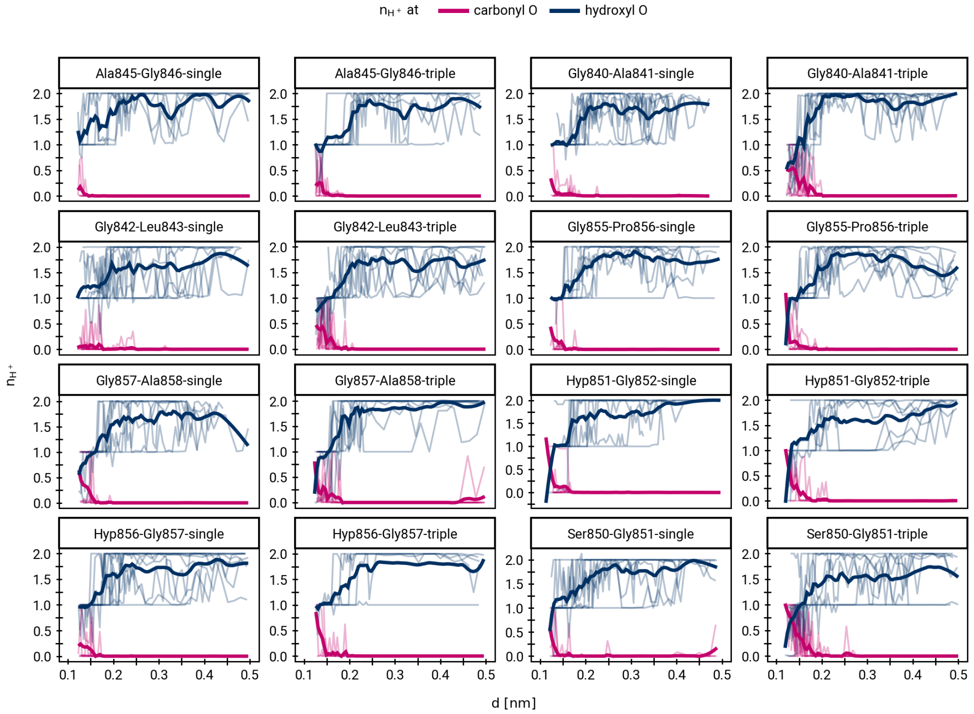


Figure 3.15: Protonation states of the carbonyl O of the TI and approaching hydroxide as the TI is formed. QM water and hydroxide protons were assigned uniquely to one oxygen (carbonyl O, hydroxide O or QM water O) for each time point within each umbrella window. Proton counts are average within each window. The windows belonging to the same approach simulation are connected and the thick lines are LOESS curves across all windows Figure 3.14 provides a visual aid for the color coding.

In the region that will be used to determine the free energy barrier of the reaction, TS1, shaded cyan in Figure 3.16, a number of data points can be seen, where the hydroxyl O initiating the reaction still has two protons assigned to it, meaning it is still a water molecule at this point. In fact, the average number of protons in the middle of the TS1 region is around 1.5. The availability of a proton acceptor around the attacking hydroxide to turn it into an electrophile should be crucial for overcoming the energy barrier towards the TI.

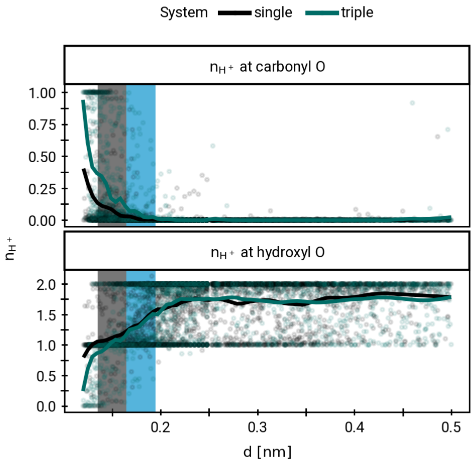


Figure 3.16: Further average of Figure 3.15. The protonation states are shown as the average number of protons assigned to either the hydroxyl O or the carbonyl O with all sampling windows averaged for each system. The region for detection of TS1 is shaded cyan, while the region for detection of the TI is shaded gray.

3.4.2 The triple helix marginally impedes hydroxide attack thermodynamically

Umbrella sampling (see Section 2.9) allows calculating free energies along a reaction coordinate, which in this case is the distance between the attacking hydroxide and the carbonyl center. To provide better sampling coverage, all sampling windows belonging to the same site (Figure 3.6) are pooled regardless of the initial approach trajectory (of which there are 5 to 10 per site, see Figure 3.9a)

Applying WHAM to the windows, after filtering out the unwanted side reactions, results in free energy profiles of the reaction coordinate (Figure 3.17). From these, the free energies for TS1 and the TI are computed. All energies (ΔG) are reported as differences from the baseline of the educt state, where the hydroxide ion is far away and solvated. This baseline energy is defined as the mean free energy at the reaction coordinate of 0.45 ± 0.02 nm for each profile. After subtraction of the baseline, the TS1 free energy barrier is the highest point in the cyan shaded region, while the TI energy is the lowest point in the region shaded in gray.

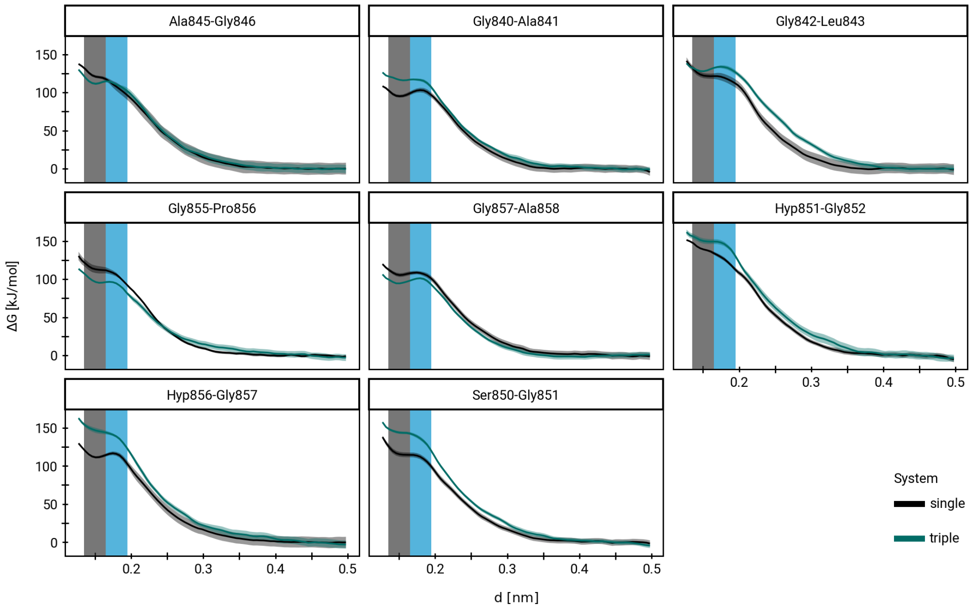


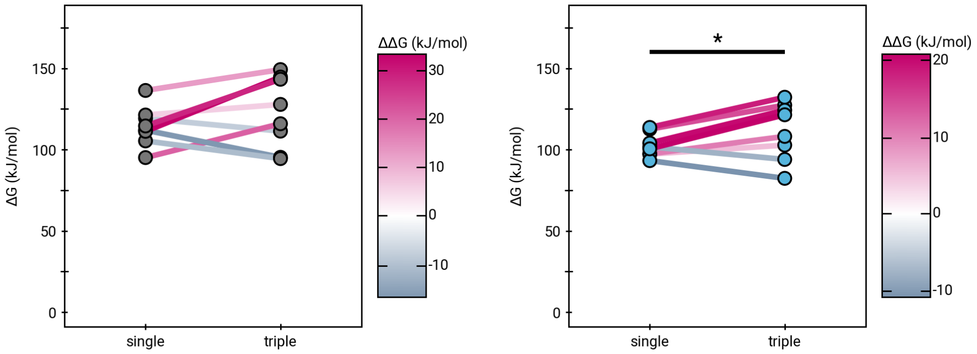
Figure 3.17: Free energy profiles obtained via WHAM of the umbrella sampling of the attack of the hydroxide (via TS1). The region for detection of TS1 is shaded cyan, while the region for detection of the TI is shaded gray. The baseline energy of the educt state has been subtracted. The shaded region around the lines represents the error estimate from bootstrapping.

The profiles are converged and exhibit low bootstrapping errors. As the hydroxide approaches (from right to left), the free energy increases. While all profiles at least level off around the distance that is characteristic for the TI, not all show a clear dip for the TI. This indicates that in order to form a stable intermediate, the TI needs a high degree of coordination with the solvent molecules that can be entropically unfavored in a complex dynamical system.

Notably, no amino-acid-specific effects can be observed in the free energy profiles and the free energy barriers have a considerable spread. The standard deviation of the TS1 barrier is 7.3 kJ/mol for the single peptide and 17.7 kJ/mol for the triple helix. According to the observations from Figure 3.13 and Figure 3.15, the stabilization of the TI is mostly provided by the QM water molecules and not hydrogen bridges with the rest of the protein (be it QM or MM atoms). This indicates that while the local chemical environment around the sampled site has a strong influence on the energetics of the reaction, this influence is mostly governed by the solvent arrangement around the reaction center. In turn, how well the highly dynamic solvent system can organize around the peptide bond is subject to fluctuations and entropic contributions that take a lot of computational time to sample.

A pairwise comparison between the single peptide and triple helix energies for each site shows no significant difference for the TI (Figure 3.18a) and a significant difference for TS1 (Figure 3.18b). The average reaction free energy of TS1 for the single chain system across all sites was 102.5 ± 2.6 kJ/mol (±SEM) and 111.6 ± 6.6 kJ/mol for the triple helix. However, it should be noted that the most appropriate significance test here is a paired *t*-test, because each site is represented in its single chain and triple helix configuration and the residuals are roughly normally distributed, and the alternative hypothesis is the original hypothesis that the triple helix would lead to higher energies (protecting collagen from being hydrolyzed). All these assumptions in place result in a *p*-value of 0.04. This also means that due to the large spread, dropping any of the requirements to perform the significance test in the manner described would result in a failure to reject the null hypothesis.

The average ΔΔG going from a single peptide to the same site in the triple helix was 9.02 ±4.41 kJ/mol (±SEM). While this may not seem like a large difference in free energy, it should be noted that since the reaction barrier enters exponentially into reaction rate calculations (see Equation 33), this difference still corresponds to a decrease in the reaction rate by a factor of 30 to 40.



(a) Comparison of the reaction free energies of the TI after subtraction of the educt state energy as a baseline for the single peptide and triple helix systems. The same site in both systems is connected with a line, color coded by the difference.

(b) Comparison of the reaction free energy barriers of TS1 (baseline is the educt state) for the single peptide and triple helix systems. The same site in both systems is connected with a line, color coded by the difference. The star indicates statistical significance at α=0.05. The *p*-value of a paired *t*-test with the alternative hypothesis $\Delta\Delta G > 0$ is 0.04.

Figure 3.18: Results of the umbrella sampling of the attack of the hydroxide, reaching the TI via TS1.

3.4.3 Proton mobility in quantum mechanically described solvent molecules allows breaking of the peptide bond

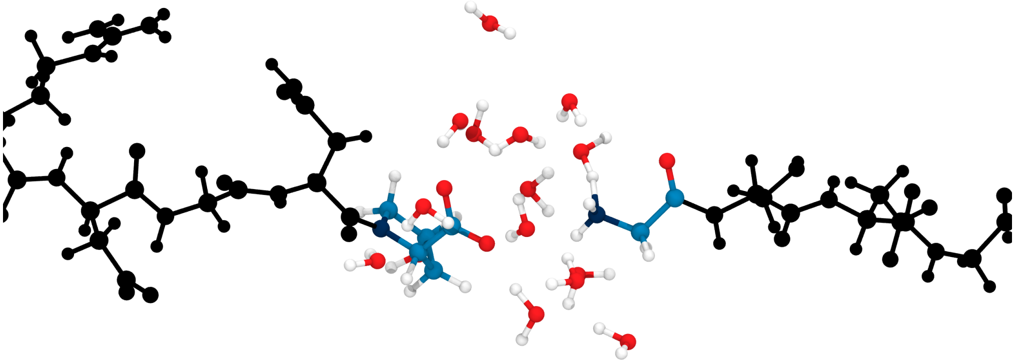


Figure 3.19: The QM subsystem (in color) and the MM subsystem (in black) of a single chain system during the rupture of the peptide bond (C—N).

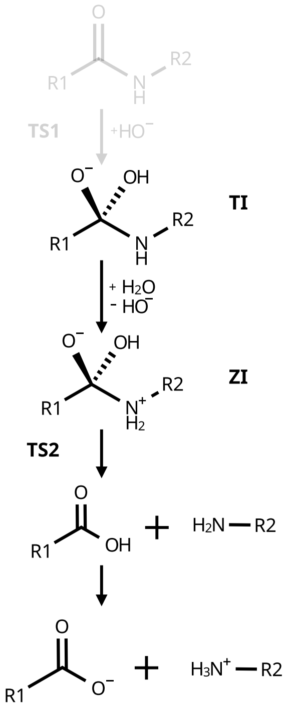
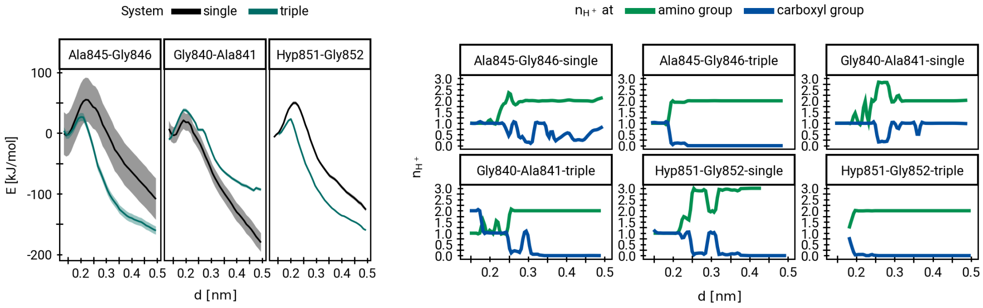


Figure 3.20: Reminder of where we are in the hydrolysis reaction.

After the TI is formed, in a second step of this multi-step hydrolysis reaction, the C—N bond along the protein backbone is cleaved, after first being protonated at the N to form a zwitterionic intermediate (ZI) according to the reaction scheme (Figure 3.20). The scission is again followed by protonation state changes to form a negatively charged C-terminus (carboxyl group) and a positively charged N-terminus (amino group). This was simulated as a single reaction coordinate, the elongation of the C—N bond. Again, initial reactive trajectories were obtained (e.g., Figure 3.19) by pulling with a harmonic potential, from which 50 equally spaced snapshots along the reaction coordinate were extracted. Those snapshots then served as the starting points for umbrella sampling simulations during which an umbrella potential kept the pulling distance near the starting value of each snapshot. While TS1 and the TI are the focus of this study, comparisons of the free energy profile of the breaking of the peptide bond obtained via WHAM are shown in Figure 3.21a for three pairs of single chain and triple helix systems. As they only have limited sampling, they don't allow conclusions about differences between the single chain and triple helix for breaking of the peptide bond, but they do show that it is possible to run end-to-end simulations of the complete base-catalyzed peptide bond hydrolysis reaction with minimal restraints.



(a)

(b)

Figure 3.21: Results of the umbrella sampling of the breaking of the peptide bond. (a) Free energy profiles obtained via WHAM of the umbrella sampling of the breaking of the peptide bond (TS2). (b) Protonation states of the carboxyl group (COO^-) and the amino group ($NH_3{}^+$) as the C—N bond is pulled apart. QM protons H^+ were assigned uniquely to one oxygen or nitrogen (carboxyl O, QM water O or amino group N) for each time point within each umbrella window. Proton counts are average within each window. The lines are LOESS curves across all sampling windows. Protons assigned to the QM water molecules are not shown, as they can be inferred from the sum of protons assigned to either of the groups. Figure 3.22 provides a visual aid for the color coding.

Furthermore, looking into the movement of the protons in the QM subsystem during all sampling windows reveals different possible sequences of events for the final steps of the reaction (Figure 3.21b). As chemists, we like to simplify reactions into easy-to-manage reaction steps. While this is an incredibly useful model, reality is messier than that [6]. We see that the reaction can occur in the sequence shown in the reaction schema, like in panel Gly840-Ala841-single in the top right or Ala854-Gly846-single in the top left, where the ZI is formed by first protonating the N, making it go from 1 to 2 protons, before the carboxyl group is deprotonated. But it can also happen in a concerted manner, whereby the amino group gets protonated at the same time as the carboxyl group gets deprotonated, as seen in the top and bottom middle panels Ala845-Gly846-triple and Hyp851-Gly852-single. As of now there is not enough data to make inferences about the energy landscape of both pathways.

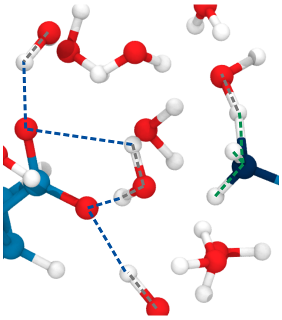


Figure 3.22: Visual explanation of the proton counting color code in Figure 3.21b

[6] Though in this case it is also just a different kind of model, a computational model, a simulation.

3.5 Discussion and outlook of the hybrid simulations

It should be noted that the absolute energies and activation energies of this study cannot directly be compared to the energies obtained by [20] shown in Figure 3.3. The former correspond to free energies (G), which include entropic contributions, while the latter represent the electronic energy of minimized structures. This is why it is not unusual that the average energy of the TI in this study is around 125 kJ/mol over the baseline of the educt state, while the average electronic TI energy for the minimized structure is only around 20 KJ/mol. Comparisons within each reference frame are still valid, but a direct comparison between the two approaches remains elusive.

The unconstrained nature of the chosen approach is a double-edged sword. It enabled sampling of interesting regions of phase space and direct observation of physically relevant behavior, such as proton mobility in the QM region during the reaction. This also showed how the base-catalyzed hydrolysis reaction can proceed in a varying sequence of steps, with the ZI forming dynamically as the peptide bond is rupturing. However, the broad sampling makes it harder to attribute the results to any specific structural effect without even higher computational costs.

While a difference in free energy was found for the activation energy of the hydroxide attack between single chain and triple helix systems, the small effect size does not fully explain the prevalence of the competing reaction (homolytic backbone cleavage) that is observed experimentally in tissue samples.

Furthermore, there is a risk of survivorship bias. Finding the set of parameters that allows simulating the reaction without crashing and the initial selection of the hydrolysis sites is likely to favor easier-to-access and easier-to-simulate systems. This is also apparent in Figure 5.1b, which shows that the current selection of peptide bonds tends to have the carbonyl oxygen point away from the helix, making the tetrahedral intermediate easier-to-access by the solvent molecules. It also highlights sites of future simulations to mitigate this bias. The trajectories that are harder to simulate are those with a rougher free energy landscape and those with higher energy states. The new simulations include more of those, such as a peptide bond between a proline and a hydroxyproline residue. Interestingly, because the sites that are harder to stabilize in the triple helix due to their orientation were underrepresented in the initial dataset, intuition would predict that the true effect of the triple helix might be even greater than what was shown here already.

The fundamental limitation is computational cost: simulating a single umbrella window on one HPC node using 20 cores at 2.1 GHz for 5 ps took on average 60 hours. With 50 windows per approach trajectory and 8 starting configurations on average, this sums up to 1000 CPU-days per site.

Going to even larger systems (full fibrils) is computationally unfeasible with the current QM/MM approach. Further sampling at additional sites would shed more light on the details of the reaction and eliminate potential bias in the selected attack sites, but it would not overcome the fundamental scaling limit. While the compute time could be reduced through further optimizations and targeted sampling of the most critical regions, comparing two competing reactions at scale in realistic system sizes with sufficient sampling requires a fundamentally different approach.

This limitation motivates the development of KIMMDY (Chapter 4), which then incorporates insights from these QM/MM simulations into an efficient emulator of hydrolysis that can simulate many reactions even in large collagen fibrils.

4 Kinetic Monte Carlo Molecular Dynamics: KIMMDY

Figure 4.1: The logo of KIMMDY, designed by Denis Kieswetter and fine-tuned by me.

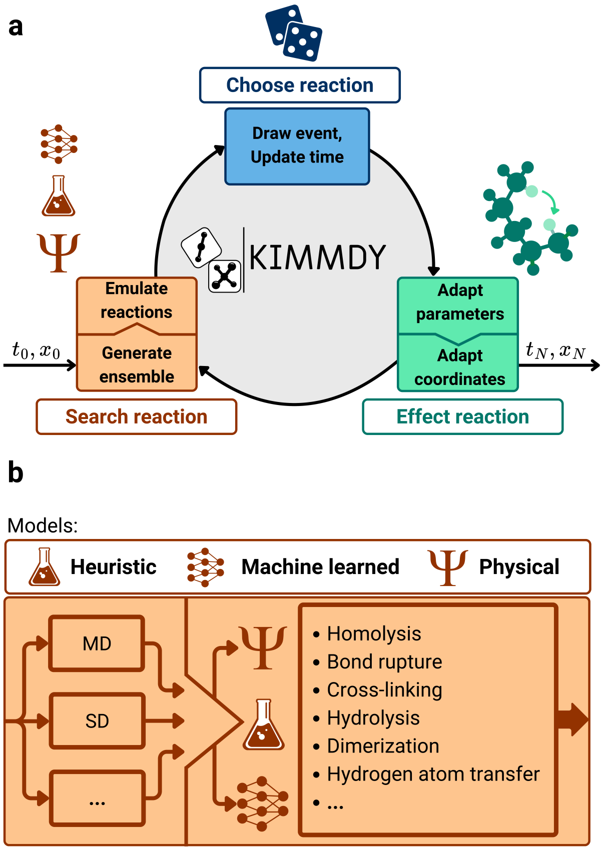


Figure 4.2: The KIMMDY schema that Denis Kieswetter made for [99]. Given an initial state at t_0, KIMMDY generates an ensemble, typically via MD. Reaction plugins then provide rates for their potential reactions. KIMMDY then chooses a reaction and updates the global time. Next, it effects the reaction, adapting the topology and coordinates of the system to the new state after the reaction. A new round of sampling can begin.

KIMMDY stands for Kinetic Monte Carlo Molecular Dynamics, a concept that was first devised by Benedikt Rennekamp to tackle the question of homolytic rupture in large molecular dynamics simulations of collagen fibrils [100]. The core idea is this: Molecular dynamics simulations, in force field based methods, allow simulating and sampling large systems over relatively long periods of time. However, they lack the key feature of life, chemistry. Bonds can not break or form in these simulations, so no reactions can occur. Furthermore, even new reactive simulation methods like reactive force fields [101, 102], or machine-learned potentials [103–108], reach a fundamental limit when it comes to reactions whose dwell time exceeds the time that can reasonably be simulated by orders of magnitude. This was previously solved by my colleague Benedikt for force-dependent homolysis, by viewing a molecular dynamics simulation as a sampling step in a Monte Carlo scheme, during which the bond lengths are recorded and used in a physical model of homolysis that converts forces acting on bonds into activation energies for the homolysis. Those activation energies are then converted to reaction rates with the Arrhenius equation (Equation 33). A set of Python scripts then realized the scheme to sample with an MD simulation, wait for the simulation to complete, and calculate the predicted rupture rates for all backbone bonds. A reaction was then chosen with the Monte Carlo scheme (see Section 2.14) and the topology of the system was modified to accommodate the change. This then allows for the next sampling MD and looking for follow-up reactions.

When my colleagues Kai Riedmiller, Eric Hartmann, and I heard of this, we decided that such a scheme should be applicable not just to homolysis, but to all possible chemical reactions. The name KIMMDY was kept, but it was now realized as a user-friendly, extensible, and powerful framework written in Python and accessible via the command line and configuration files. Some of the figures I will show in this chapter have already been published in our manuscript, which is available on bioRxiv [99] and at the time of writing under review at Nature Communications. As not only a joint first author paper, but also a collaborative software project with tens of thousands of lines of code added, removed, and changed, it is not straightforward to pin any given feature on a single one of us. For this thesis, I will do my best to highlight especially the parts on which I had a particular focus, but nonetheless it remains the product of an incredibly fruitful collaboration.

Figure 4.2 shows an overview of how KIMMDY works. KIMMDY is based on a plugin architecture. Possible reactions, such as bond rupture, hydrolysis, or hydrogen atom transfer (HAT), are implemented as plugins that report back reaction rates. KIMMDY then takes care of executing the chosen reaction, stabilizing the system, and running the sampling simulations. In practice those are MD simulations, though KIMMDY is agnostic with regard to the type of simulation that generates the ensembles. The reaction plugins are also free to use any method necessary to obtain reaction rates based on the state of the system. This could be a heuristic fit to experimental data, a physical model, or a machine-learned one. All are present in the currently available plugins, such as the machine-learned model predicting HAT by Kai, a plugin based on Benedikt's original physical model for homolysis, or my heuristic model for peptide bond hydrolysis. But most importantly, we provide extensive documentation to enable any user to write their own reaction plugin and get started with KIMMDY: graeter-group.github.io/kimmdy/.

The structure of our documentation is based on the Diátaxis framework developed by Daniele Procida, who explains it further on diataxis.fr [109]. Briefly, each piece of information and writing is categorized by whether its primary purpose is user-action- or user-cognition oriented and whether the goal is to acquire knowledge or to solve an immediate application problem (Figure 4.3). This creates distinct types of documentation. Tutorials are oriented towards learning

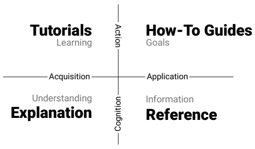


Figure 4.3: Schema of the diataxis documentation framework, based on a figure by Daniele Procida [109].

and require active participation by the user. How-To Guides are directed at a specific task, which is actively performed by the user. Explanations are for understanding and are consumed passively. A Reference serves to provide task-specific documentation.

In this chapter, I will first go over how KIMMDY works and use the typical flow of a user setting up a KIMMDY run as the guiding thread to explain how they interact with KIMMDY and what goes on under the hood. Then I will apply KIMMDY to the biological question posed in the very beginning, integrating insights from the QM/MM simulations into an emulator for peptide bond hydrolysis to directly compare reaction rates of homolysis and hydrolysis in a large collagen fibril simulation.

4.1 Running KIMMDY: implementation and usage

Note: the implementation details of KIMMDY have already been described by me and edited by my co-authors in the supplementary materials of [99]. As such, sections of text will be similar and may only have minor adaptations for the context of this thesis.

4.1.1 Installation

Let us picture a user who has a new type of reaction that they want to simulate within MD. They are already familiar with their MD engine GROMACS and have the files to simulate their system without any reactions. After going to our documentation website, they install KIMMDY. KIMMDY can just be installed from the Python package index (PyPI) with `pip install`, but to speed up installing dependencies, especially those of our machine learning plugins, we recommend using `uv`, e.g., with `uv tool install -p 3.11 kimmdy`. Now the `kimmdy` command is available on the command line. Typing `kimmdy --help` brings up more options:

```
usage: kimmdy [-h] [--input INPUT] [--restart] [--loglevel LOGLEVEL]
              [--logfile LOGFILE] [--show-plugins] [--generate-
jobscript]
              [--version] [--debug] [--callgraph]

Welcome to KIMMDY. `kimmdy` runs KIMMDY. Additional tools are available as
`kimmdy-...` commands. These are `-analysis`, `-modify-top` and `-
build-
examples`. Access their help with `kimmdy-... -h.`Visit the documentation
online at <https://graeter-group.github.io/kimmdy/>

options:
  -h, --help            show this help message and exit
  --input INPUT, -i INPUT
                        Kimmdy input file. Defaults to `kimmdy.yml`. See
                        <https://graeter-
                        group.github.io/kimmdy/guide/references/input.html>
                        for all options. CLI flags (e.g. --
restart or
                        --loglevel) have precedence over their counterparts in
                        the input file.
  --restart, -r         Restart or continue from a previous run instead of
                        incrementing the run number for the output directory.
                        If the output directory does not exist, it will be
                        like a regular fresh run.
  --loglevel LOGLEVEL, -l LOGLEVEL
                        Logging level (CRITICAL, ERROR, WARNING, INFO, DEBUG)
  --logfile LOGFILE, -f LOGFILE
                        Logfile
  --show-plugins        List available plugins
  --generate-jobscript  Instead of running KIMMDY directly, generate the
                        output directory and a jobscript `jobscript.sh` for
                        slurm HPC clusters. You can then run this jobscript
                        with sbatch jobscript.sh.
  --version             Show version and exit.
  --debug               On error, drop into debugger
  --callgraph           Generate a visualization of function calls for
                        debugging and documentation.
```

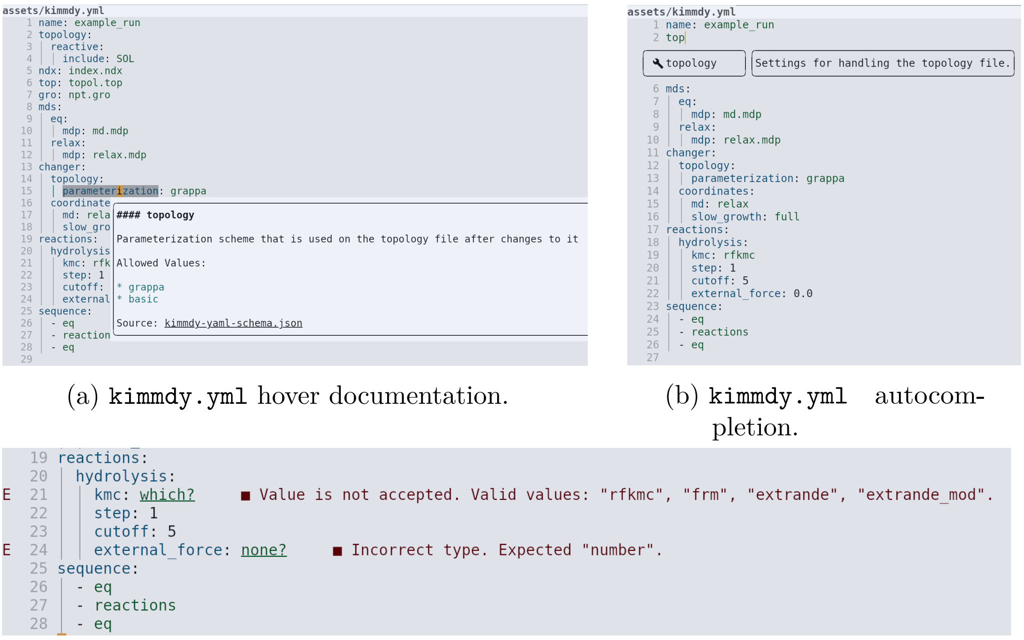
Additional provided tools include `kimmdy-analysis` for post-simulation analysis, and `kimmdy-modify-top` for modifying or parametrizing a system outside of a KIMMDY run.

4.1.2 Configuration and input/output

KIMMDY is controlled via the command-line interface (CLI) and a YAML-formatted configuration file. In there, every task KIMMDY performs has a corresponding configuration section. The sequence of tasks KIMMDY executes is also defined in this configuration, making KIMMDY highly flexible. The configuration file contains input files such as the topology and initial coordinates, high performance computing (HPC) options, logging and working directory options, and general settings for KIMMDY. For the complete documentation of all options, see the automatically updated list at graeter-group.github.io/kimmdy/guide/references/input.html.

When the `yaml` file, typically named `kimmdy.yml`, is opened in a properly configured code editor supporting the language server protocol, such as the popular Visual Studio Code or Neovim, the user will automatically get features such as autocompletion for keywords and options (Figure 4.4b), hover documentation

(Figure 4.4a). They will also see warnings for missing, unknown keys or incompatible values (Figure 4.4c). This reduces the likelihood of disappointing application crashes, because the validity of the input can be checked before KIMMDY is even started. Not everything can be checked before runtime, such as the availability of some plugin dependencies or the existence of input files, but further consistency checks are front-loaded at the start of KIMMDY, such that it stops earlier rather than later in a run. KIMMDY checks that all input files are present, which avoids running a whole MD simulation only to find out that the mdp file for the next step is missing.



```
assets/kimmdy.yml
name: example_run
topology:
  reactive:
    include: SOL
ndx: index.ndx
top: topol.top
gro: npt.gro
mds:
  eq:
    mdp: md.mdp
  relax:
    mdp: relax.mdp
changer:
  topology:
    parameterization: grappa
  coordinate
    md: rela
    slow_gro
reactions:
  hydrolysis
    kmc: rfk
    step: 1
    cutoff:
    external
sequence:
  - eq
  - reaction
  - eq
```

```
#### topology

Parameterization scheme that is used on the topology file after changes to it

Allowed Values:

* grappa
* basic

Source: kimmdy-yaml-schema.json
```

(a) `kimmdy.yml` hover documentation.

```
assets/kimmdy.yml
name: example_run
top
```

```
mds:
  eq:
    mdp: md.mdp
  relax:
    mdp: relax.mdp
changer:
  topology:
    parameterization: grappa
  coordinates:
    md: relax
    slow_growth: full
reactions:
  hydrolysis:
    kmc: rfkmc
    step: 1
    cutoff: 5
    external_force: 0.0
sequence:
  - eq
  - reactions
  - eq
```

(b) `kimmdy.yml` autocompletion.

```
reactions:
  hydrolysis:
E   kmc: which?     ■ Value is not accepted. Valid values: "rfkmc", "frm", "extrande", "extrande_mod".
    step: 1
    cutoff: 5
E   external_force: none?     ■ Incorrect type. Expected "number".
sequence:
  - eq
  - reactions
  - eq
```

(c) `kimmdy.yml` warnings.

Figure 4.4: Various features of modern text editors working with a `kimmdy.yml` file.

The user then sets up the configuration file (the config) and decides on a `sequence`, a series of tasks that KIMMDY will run. This could be for example different MD simulations, such as an equilibration and a production run, followed by sampling for reactions, and a slow-growth simulation (see below) to equilibrate after each reaction. The sequence can include nested tasks and tasks with a multiplicity, i.e., tasks that are run multiple times. Each task gets its own output subdirectory in the KIMMDY run output directory and KIMMDY tracks the latest version of each type of file. This means that if one task runs an MD trajectory, the next task will continue after the trajectory (or in the middle of it, if a reaction should occur at a certain time point) and if a task modifies the topology, the next task will use the modified topology.

In the example shown in Figure 4.4, the first task is an MD simulation.

4.1.3 Ensemble generation

In theory, KIMMDY can work with any ensemble generation method that results in atomistic trajectories. This includes *ab initio* MD, machine-learned potential (MLP) simulations, classical MD, or other ensemble generation methods. In practice, the KIMMDY architecture is oriented towards GROMACS [40, 110] simulations, limiting the immediate uses to unbiased MD, biased MD, and in the near future MLP simulations.

KIMMDY interfaces with GROMACS via its command-line interface, which makes KIMMDY robust to GROMACS version changes. In the KIMMDY config, the GROMACS binary name and flags for commands such as `gmx grompp` and `gmx mdrun` are exposed as options. The `.mdp` files for requested MD tasks and all files defining the molecular structure are taken as user input. It should be noted that albeit MD simulations are conducted, no continuous reactive trajectories are obtained. Rather, separate trajectories are generated for every state and the slow-growth simulations (see below) in between. The concatenated trajectories that KIMMDY analysis tools can create from all states are only for visualization and inspection purposes.

The next task in the example is checking for reactions.

4.1.4 Reactions and recipes

KIMMDY manages reactions via a plugin system, providing users with an easy method to expand KIMMDY's existing functionalities. A plugin generates reaction rates alongside reaction recipes based on the latest conformational ensemble.

How the reaction rates are determined is up to the plugin. They can be calculated based on simple or complex physical models, predicted by a machine-learned model or be determined based on experimental heuristics. Plugins consume simulation outputs in various ways, but popular choices include using MDAnalysis [111, 112] to read trajectories or PLUMED [113] to monitor distances.

The reaction recipes describe all modifications to the topology and structure required to perform the reaction corresponding to the calculated rate. A recipe is composed of recipe steps, which can break a bond, create a bond, place an atom at some coordinates, modify the topology, or request a specific MD simulation. Importantly, plugins only have to think about those elemental modifications, while KIMMDY takes care of the complex implications of them (see Section 4.1.6). Additionally, plugins can defer returning the actual steps of the recipe until a reaction has been chosen by KIMMDY to reduce computational cost. For example, the hydrolysis plugin returns rates for all peptide bonds in the system, but only chooses a specific water molecule to execute the attack once a bond has been chosen as the target. Reaction rates can be valid for the whole period of the sampling simulation or vary over time for each possible reaction, which is also reported back to KIMMDY.

The example only asks the hydrolysis plugin for rates, which are then passed on to the chosen kMC algorithm.

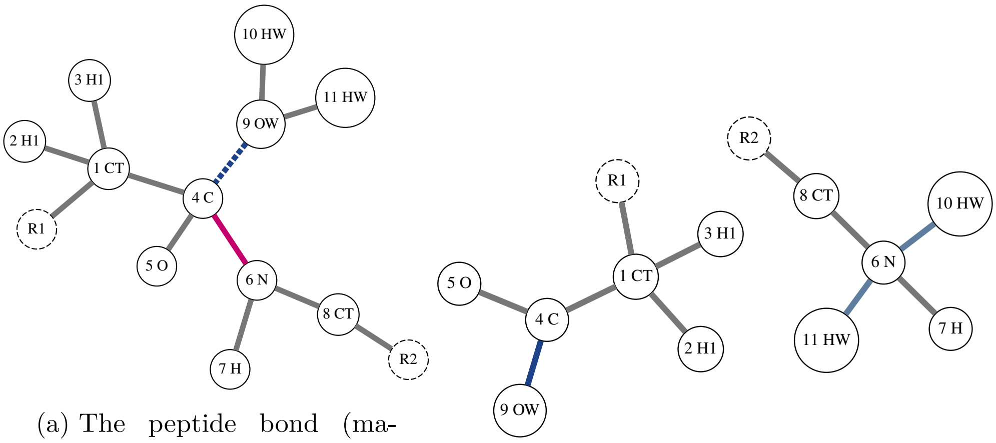
4.1.5 Kinetic Monte Carlo algorithms

KIMMDY makes use of Adaptive kMC. It collects all reaction rates from the plugins and uses a user-selected choice of kMC algorithm to sample a reaction from the distribution. For a more detailed theoretical background of kMC than Section 2.14, please also refer to the SI of [99]. Choices for kMC algorithms currently include rejection-free kMC ("rfkmc"), "extrande" [114], a modified version of extrande ("extrande_mod"), rejection-free kMC with multiple choices at the same time ("multi_rfkmc"), as well as a debugging option that always chooses the first (or last) reaction. Individual reactions can declare their preferred algorithm in addition to the global choice of kMC algorithm. In this thesis only rejection-free kMC is used.

4.1.6 Topology module

From the initially user-supplied topology, an internal `Topology` object is build. The force field, as well as topology *includes*, are explicitly resolved, meaning every interaction has individual parameters. This enables the use of GRAPPA (see Section 4.1.7) for parametrizing on-the-fly, but is still compatible with established force fields, like Amber [29] or Charmm [36]. The `Topology` supports breaking bonds, creating bonds, and deleting hydrogen atoms. For all of these actions, the interaction terms, like bonds, angles, dihedrals, pair exclusions, etc., are updated accordingly. Furthermore, it identifies, and keeps track of possible radicals and whether it needs parametrization due to some modifications. Creating bonds, deleting bonds, and moving atoms do not change the index of the atoms stored in the topology. This ensures compatibility with standard analysis techniques, but residues are not necessarily continuous in the resulting topology afterward. If continuous residues are needed, the `Topology` provides a function to re-index itself. After an atom is deleted, the topology is re-indexed by default. To use the `Topology` in an MD simulation, a new `top` file is generated from the internal topology object.

The example focuses on peptide bond hydrolysis. Assuming a bond has been chosen to be hydrolyzed, KIMMDY then gets the recipe to be executed. For simplicity, only a small portion of the complete molecule is shown around the peptide bond in Figure 4.6 and Figure 4.5. The plugin only had to specify bonds that are to be deleted and the new ones to create; KIMMDY takes care of higher-order interactions. For example, deleting the peptide bond $4\ \mathrm{C} - 6\ \mathrm{N}$ will also remove the angle $6\ \mathrm{N} - 4\ \mathrm{C} - 5\ \mathrm{O}$ and the dihedral $7\ \mathrm{H} - 6\ \mathrm{N} - 4\ \mathrm{C} - 5\ \mathrm{O}$ and so forth. Similarly, the atoms that were previously excluded due to being bound or part of a 1–3 or 1–4 interaction will now have non-bonded interactions. Newly formed bonds such as $9\ \mathrm{OW} - 4\ \mathrm{C}$ will create the corresponding angles, dihedrals and pairs for 1–4 interactions. Even improper dihedrals that are defined in the residuetypes, such as those for peptide bonds between amino acids, are re-created based on the residuetype recorded in the topology. If the basic operations are not sufficient or non-standard changes are required, a `RecipeStep` can also bring a custom function that is free to modify the topology in any way necessary.



(a) The peptide bond (magenta) between 4 C and 6 N is to be hydrolyzed. 9 OW (O Water) will be bound to 4 C and the two HW atoms will be bound to 6 N.

(b) The newly created bonds are colored in blue.

Figure 4.5: Simplified graph representation of a peptide bond undergoing hydrolysis.

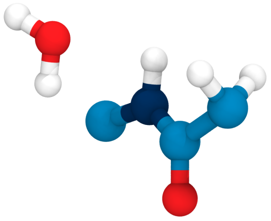


Figure 4.6: A water molecule next to a peptide bond, simplified.

All modifications to the topology are fast and don't scale with system size through the use of hash maps to store interactions. Thus, if the key for an interaction is known, the lookup scales with O(1). Each atom in the topology atom list also keeps a list of atoms that it is connected to, which drastically reduces the search space for interactions. For example, if one wants to know if a bond exists between two particular atoms, one look into the `bound_to` list of either one suffices. And if angles or dihedrals are to be checked, a small local topology can be built up from the `bound_to` lists without having to iterate over all bonds, angles, or dihedrals of a large topology. This is how KIMMDY stays performant even for multi-million atom systems.

By default, only the first moleculetype of a system is available for reactions, as many reactions don't involve the solvent molecules. However, one can easily configure includes and excludes for the `Reactive` moleculetype (see Figure 4.4). When multiple molecules of the same moleculetype are included, such as 100 water molecules in addition to the protein (Listing 4.7), KIMMDY creates new atoms, bonds etc. for each instance of the moleculetype with increasing atom- and residue-numbers. In doing so, the atom ids in the topology file (`.top`) and the line numbers of the coordinate file (`.gro`) always line up. This is important, as the atom indices written in a `gro` file may be wrong due to the fixed-width nature of the format (they wrap around at 1 million); only their order (i.e., line number) is guaranteed to be unique.

Figure 4.7: A topology molecules section.

```
[ molecules ]
Protein 1
SOL 100
```

Topology and force field objects are easy to inspect and modify when working on a reaction plugin or interactively working with KIMMDY code in a notebook. Even the topology visualizations in Figure 4.5 were originally created with KIMMDY code. The order of operations when describing a reaction does not matter, because KIMMDY keeps a consistent state at each step. For example, Listing 6 shows how Figure 4.5b is created from Figure 4.5a and what the output of inspecting the topology object looks like (Listing 7).

Listing 6 Creating a topology and modifying it.

```
top = Topology.from_path(
  "./assets/gly-hoh.top",
  # [...]
)
top.break_bond(('4', '6'))
top.break_bond(('9', '10'))
top.break_bond(('9', '11'))
top.bind_bond(('4', '9'))
top.bind_bond(('10', '6'))
top.bind_bond(('11', '6'))
```

Listing 7 Inspecting the topology object.

```
Topology with the following
molecules:
Reactive: 1

With Moleculetype Reactive with:
11 atoms,
9 bonds,
13 angles,
6 pairs,
[...]

ForceField parameters with
66 atomtypes,
98 bondtypes,
257 angletypes,
134 proper dihedraltypes
58 improper dihedraltypes
153 residuetypes
```

Due to their strongly typed nature, KIMMDY objects are exceptionally easy to work with. Expressive type annotations and documentation strings (also see Listing 10 and Listing 11) provide a fast and efficient development experience and allow for interactive exploration with modern editor tooling.

4.1.7 Parametrization interface

After each reaction, the product state needs to be parametrized. Hydrolyzing a peptide bond effectively removes one water molecule and creates two terminal amino acids. Terminal amino acids are nothing unusual for most force fields, so KIMMDY can adjust the topology parameters for the state after the reaction based on the force field. This is called the basic parametrizer. Usage of the basic parametrizer assumes that the product molecule parameters are contained in the force field included in the topology file or that additional required changes are done explicitly with the custom topology modification `RecipeStep`.

Other reactions are not as straightforward, but KIMMDY comes equipped to handle all of them. Grappa [115] is a Graph Attentional Protein Parametrization model developed by Leif Seute and Eric that can generate new MM parameters for unseen structures, even for radicals such as those as the result of bond homolysis or HAT. Grappa parametrizes around the atoms involved in the reaction. Additionally, the whole `Reactive` moleculetype can be parametrized with Grappa either before a KIMMDY run or at the start, producing consistent parameters for any reaction.

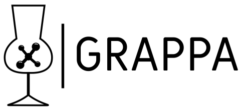


Figure 4.8: The logo of GRAPPA, designed by Denis Kieswetter.

Applying correct parameters is important not only to maintain stable simulations, but also to ensure that physically accurate states are sampled. For example, the radical center in Figure 4.9 should have its substituents in the same plane as the carbon atom, as it is now best represented as being sp^2 instead of sp^3 hybridized. Please note that while you will see an interactive element to explore the molecule via a plugin I wrote for our documentation system in the web version of this thesis, the print version will have just a screenshot. The plugin is available at github.com/jmbuhr/quarto-molstar.

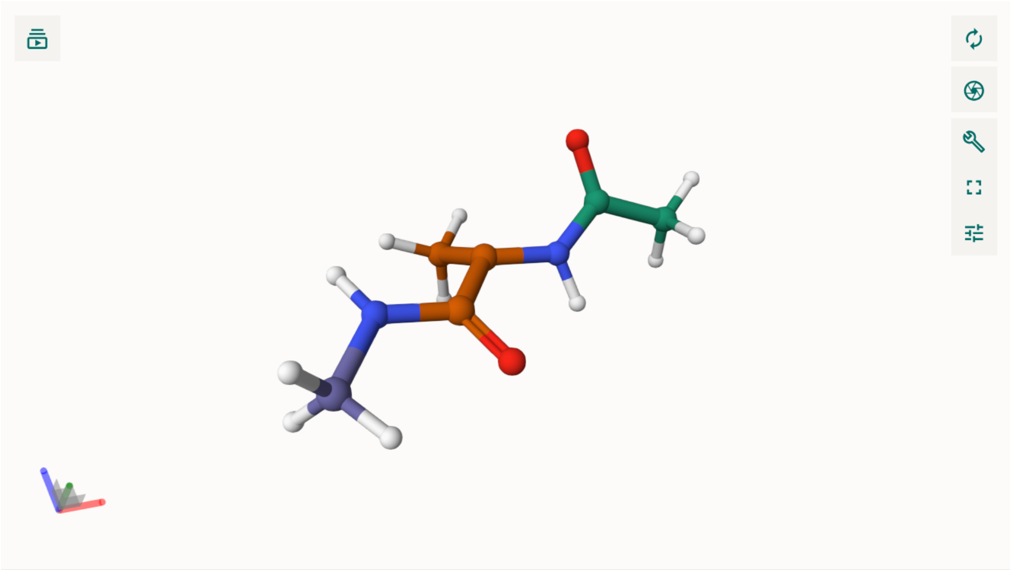


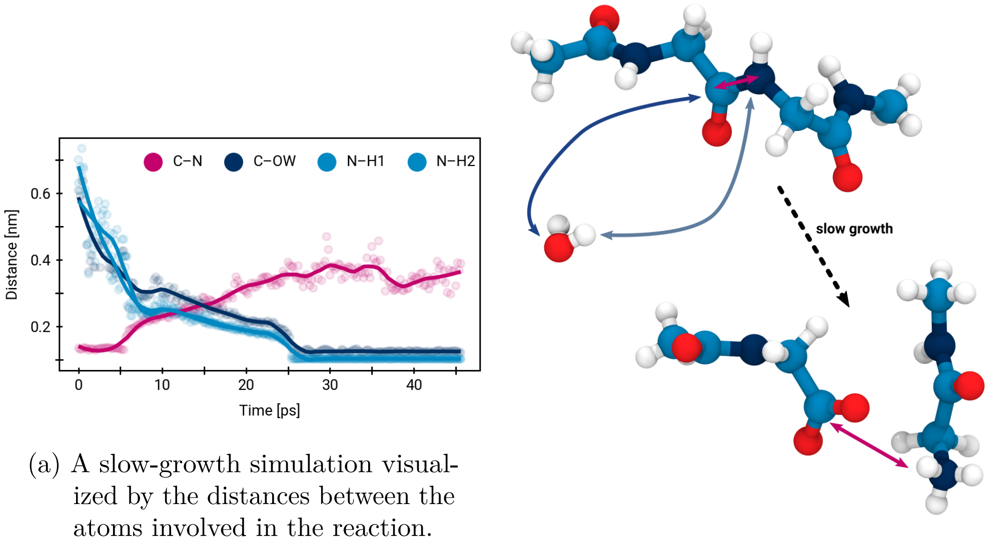
Figure 4.9: Screenshot of the interactive molecule display of an alanine dipeptide with a radical center.

Now KIMMDY has a trajectory with a chosen reaction time and knows the reaction that should happen. It also has the topology as it should be before and after the reaction, including all required parameters. Swapping out the topology immediately would likely explode the simulation system, as the coordinates may be very far away from a minimum energy state with the new parameters. This is where slow-growth comes in.

4.1.8 Slow-growth

For creating an initial structure of the reaction product state, minimizing and equilibrating the molecular systems is viable. However, this may take a long time for large systems, resulting in inefficient simulations. We use the slow-growth method implemented in the GROMACS free-energy module to interpolate between product and reactant state for continuous parameter and coordinate changes. Subsequently, a short equilibration ensures the states are Markovian, i.e., sampling happens from the Boltzmann distribution of the new state.

To this end, KIMMDY generates a merged topology that transitions between state A (educt) and state B (product). KIMMDY supports two modes of slow-growth. In “morse-only” mode, breaking or forming bonds are modeled entirely with Morse potentials, where the unbound is mimicked by a Morse potential with parameters such that the shape matches the van der Waals potential of the two now non-bonded atoms. In “full” slow-growth mode all parameters around the affected atoms are transitioned, where non-bonded parameters are modeled using pair potentials. Slow-growth of non-bonded parameters, especially for van der Waals interactions, can quickly lead to high energies and forces due to the exponential nature of the potential. As such the best choice of KIMMDY options and mdp parameters to relax the structure between reactions and sampling depends on the system. An example trajectory of a reaction including slow-growth can be seen at https://youtu.be/tgS5uleN288 (or directly in the web version of my thesis at jmbuhr.de/phd-thesis) and is visualized in Figure 4.10.



(a) A slow-growth simulation visualized by the distances between the atoms involved in the reaction.

(b) Visual explanation of the distances plotted in Figure 4.10a. Coloured double-arrows correspond to the distances in the plot.

Figure 4.10: KIMMDY’s slow-growth applied to hydrolysis of a peptide bond.

Currently, exact continuations of MD simulations in GROMACS are only possible from checkpoint files, which do not allow for topology modifications. We use coordinates and velocities to continue MD simulations, which leads to a short period where thermostat and barostat equilibrate again. Switching between "free-energy" on and off in GROMACS also currently has some artifacts, albeit non-breaking, which may be improved in the future.

After the slow-growth task the next set of MD simulations and sampling for reactions can begin. But what if a single KIMMDY run includes long-running simulations or expensive models to compute?

4.1.9 Checkpoints and restarting

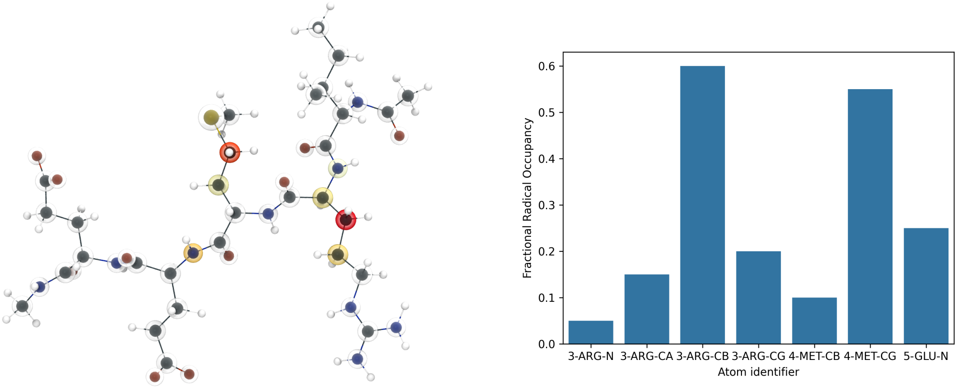
When KIMMDY is started with an already existing output directory, it can either create a new output directory with incrementing run numbers or attempt to restart an existing run in the output directory. The behavior can be set by the `--restart` flag of the CLI or the `restart:` keyword in the `kimmdy.yml` file. When restarting, KIMMDY first checks if the existing run has already completed, in which case no additional work is done. It then scans the existing output subdirectories of the already executed tasks in the sequence for incomplete tasks and for tasks during or after which a continuation is possible. Conveniently, GROMACS simulations can be restarted from checkpoints and the end of a simulation the current system state is written out explicitly. Thus, KIMMDY can restart after GROMACS simulations or even continue halted simulations from the latest checkpoint. Files previously written by KIMMDY are parsed to faithfully reproduce the current state of the KIMMDY simulation. All tasks started after the last MD task cannot be used and are discarded. This is not typically an issue, because the simulations are the computationally expensive part, not the reaction rate queries. However, reaction plugins can implement their own caching mechanisms to speed up restarting after a known simulation. Examples for this are the hydrolysis and homolysis plugins, which both need the query bond distances and average them over the simulation. Because of this, they use a shared cache file of those average distances, which not only allows fast restarting after a completed simulation but also allows for a speed-up when both reactions are in a competition setting, in which case only one of the plugins has to compute the average distances and the other can use the same shared cache.

4.1.10 High performance computing interface

KIMMDY simulations are easily parallelizable and due to the MD or machine-learning reaction plugins they require extensive resources, typically available on high performance computing (HPC) systems. We added utilities to run KIMMDY on HPC systems using Slurm [95] and expose prefixes to the GROMACS commands necessary for running it on a system using a message passing interface (MPI) such as OpenMPI. Restarting as described above is used in conjunction with automated resubmission of Slurm jobs for HPC clusters that have a time limit on job runtime.

4.1.11 Analysis

KIMMDY also includes analysis tools via `kimmdy-analysis`. Currently this involves `trjcat` to concatenate the individual trajectories into a continuous one for visualization, `energy` to plot various energy terms during the simulations, `rates` to visualize reaction rates for all reactions, and tools specially designed for radical migration via HAT (see Figure 4.11).



(a)

(b)

Figure 4.11: KIMMDY analysis tools visualizing the radical occupancy after many successive HAT reactions. (a) Radical occupancy is written to a pdb file as beta values and then visualized with VMD. (b) Radical occupancy as a bar plot.

4.1.12 Code quality and testing

Academic software development teams often change composition rapidly due to the nature of short-term projects, such as Master's and doctoral degrees. Hence, writing robust and maintainable code was made a top priority in the development of KIMMDY. To minimize friction when working with multiple people on the same code base, we adhered to best practices of research software development and established clear conventions, outlined in our online documentation. KIMMDY code is strongly typed and extensively tested. Further, KIMMDY has infrastructure in place to automate running unit- and integration tests before a change is merged. New KIMMDY versions are automatically created and released based on the Conventional Commits specification [116] to clearly identify new features and label backwards-incompatible changes. The version used for the results in this thesis is `v8.0.0`.

4.2 Applying KIMMDY to collagen hydrolysis and homolysis

Next, KIMMDY will be applied to directly compare peptide hydrolysis to homolysis of the backbone, to answer why we see radicals in experiments on collagen [17], even though peptide hydrolysis should be significantly sped up under force [20], which would mean the backbone would hydrolyse before it had a chance to break homolytically. To compare two reactions in KIMMDY, reaction rates are required.

4.2.1 The hydrolysis and homolysis plugins

Rates for homolysis use the physical model developed by Benedikt Rennekamp [100], which follows a Bell-Evans model [117, 118]: Bonds are modeled using a Morse potential (Equation 7 and Figure 2.5) in the MD simulation. The potential is shifted by the work already performed on the bond by the external force, which is calculated as the bond elongation multiplied with the force on the bond at the average bond distance. This assumes that the system is well equilibrated under the external force. If this is the case, the force in a bond is the derivative of the potential at the average bond distance $\overline{r}$ and computed by Equation 38, which is also shown in Figure 4.12.

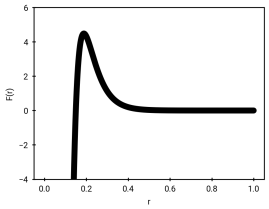


Figure 4.12: The derivative of the Morse potential, i.e., the force as a function of the bond distance r.

$$F(\overline{r}) = 2\beta D^{-\beta(\overline{r}-r_0)} \left(1 - e^{-\beta(\overline{r}-r_0)}\right), \tag{38}$$

with $\beta = \sqrt{\frac{k}{2D}}$,the spring constant k, and the dissociation energy D.

The effective potential (Equation 39)

$$V_{eff}(r) = V_{\text{Morse}}(r) - F(\overline{r})\ r, \tag{39}$$

has a minimum and a maximum given by Equation 40

$$r_{\text{min/max}} = r_0 - \frac{1}{\beta}\left(\beta\ D \pm \sqrt{\frac{\beta^2 D^2 - 2D\beta F(\overline{r})}{2\beta D}}\right), \tag{40}$$

and consequently provides the energy barrier ΔE for force-dependent homolysis via Equation 41:

$$\Delta E = V_{eff}(r_{\text{max}}) - V_{eff}(r_{\text{min}}), \tag{41}$$

which is then used in the Arrhenius equation (Equation 33) to obtain a reaction rate (Equation 42).

$$k_{\text{homolysis}} = A\,\mathrm{e}^{\left(\frac{-\Delta E}{RT}\right)} \tag{42}$$

Rates for hydrolysis are also force-dependent. The experiments by Pill et al. [20] already allow fitting a model to match external forces applied to a single peptide bond with bond lifetimes and thus reaction rates (Figure 4.13). Because they observed a bi-exponential decay at 0.7 nN and thus fit two Bell models, one for the region below 0.7 and one for above 0.7, in order to obtain continuous rates to use in KIMMDY, the two models were interpolated around 0.7 nN in a log-linear fashion. The rates as one would obtain them if the theoretical energy barriers from their *ab initio* QM calculations were used, are also shown. In the KIMMDY plugin for hydrolysis, the choice between experimental and theoretical rates is a configuration option, but here I will use just the experimental rates from AFM to be as close to the experimental setting as possible.

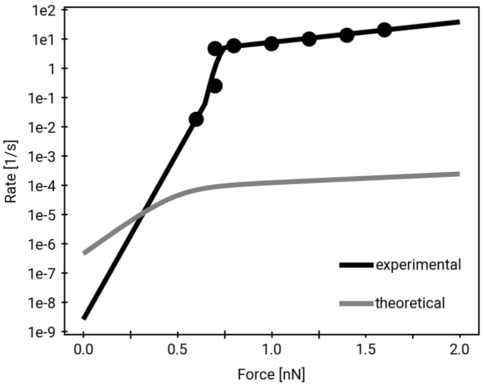


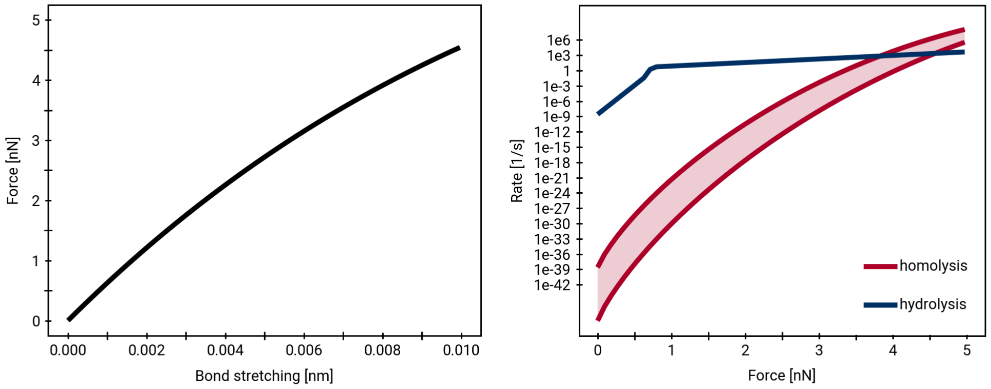
Figure 4.13: Experimental (AFM) and theoretical (QM) reaction rates used as an input to the hydrolysis plugin as a function of force. Points represent the averages of the AFM experiments [20].

Crucially, the hydrolysis plugin cannot simply use the difference of the bond extension to the equilibrium distance defined in the force field to relate to externally applied forces. When forces are applied just like in Figure 3.5, an externally applied force of 0 nN should relate to a force of 0 nN in the model from the AFM experiments. However, when no force is applied, the average equilibrium bond distances of all bonds are not equal to the equilibrium distance of the Morse potential from the force field. This is due to the entropic contributions in the simulation and the laws of statistical mechanics: The probability P_i of finding a system in a state i is given by the Boltzmann distribution as a function of the states' energy E_i by Equation 43:

$$P_i \propto e^{(-\frac{E_i}{k_\mathrm{B}T})} \tag{43}$$

For an asymmetric potential energy function like the Morse potential, this leads to the aforementioned effect. The solution is straightforward. The internal bond forces calculated from bond distances of each peptide bond are recorded during the equilibrium simulations at 0 applied force and used as a baseline for all other forces. This way, the force that is nominally applied to the simulation matches the force that was nominally applied during the AFM experiment.

The force responses for both reactions can now be directly compared (Figure 4.13). As the external force acts on the system, the average bond distances increase, which leads to an apparent force acting directly on the bond. The values for one such bond are plotted in Figure 4.14a and the reaction rates for both reactions, homolysis and hydrolysis, as a function of those forces are shown in Figure 4.14b.



(a) (b)

Figure 4.14: Force response of the homolysis and hydrolysis reaction rates. (a) The force that has performed work on a bond as a function of bond stretching, i.e., $r-r_0$. (b) Reaction rates for hydrolysis (experimental model) and homolysis (physical model) as a function of force. For homolysis, the rates depends on the BDE, so two exemplary lines are shown. The lower line for a C-C bond in a protein backbone with a BDE of 341 kJ/mol and the upper line for a C-C bond in a Pyridinoline (PYD) cross-link with a BDE of 282 kJ/mol.

A few more modifications to the hydrolysis reaction rate apply. Firstly, the experiment used a single, freely accessible peptide bond. The hydrolysis plugin captures that a peptide bond may be less accessible by taking into account the solvent accessible surface are (SASA). SASA is calculated with the Lee-Richards algorithm [119] via a small module within the hydrolysis plugin that I coined `minisasa` by interfacing with the freesasa C-library [120]. SASA acts as a scaling factor for the attempt frequency (A) in the form of the ratio of the SASA of the bond to the SASA of a freely accessible peptide bond in a glycine dipeptide.

Secondly, the AFM experiments happened at a fixed pH of 7.4, while the hydrolysis plugin can take this into account again as a scaling factor to the attempt frequency. The concentration of hydroxide ions, which are required catalytically for the reaction, is given by $c_{\mathrm{OH}^-} = 10^{-(14-\mathrm{pH})}$ and consequently

$c_{\mathrm{OH}^-_{\mathrm{exp}}} = 10^{-(14-\mathrm{pH}_{\mathrm{exp}})}$. Since the concentration of OH^- is proportional to the probability that any given solvent molecule is a catalytically active OH^-, the scaling factor is the ratio of concentrations, which leads to the final equation for the reaction rate of a peptide bond in a sampled state, Equation 44.

$$\mathrm{Rate}(bond, t) = \frac{\mathrm{SASA}(bond, t)}{\mathrm{SASA}_{\mathrm{max}}} \cdot \frac{c_{\mathrm{OH}^-}}{c_{\mathrm{OH}^-_{\mathrm{exp}}}} \cdot \mathrm{rate}_{\mathrm{exp}}(F_{bond,t}) \tag{44}$$

Both the force and the SASA can be time-dependent or averaged over the entire state. Once a peptide bond has been chosen to react by the kMC step, the hydrolysis plugin calculates the optimal solvent molecule to execute the attack based on distance and angles (Bürgi-Dunitz and Flippin-Lodge, see Figure 3.7).

The code for the hydrolysis plugin is available at github.com/graeter-group/kimmdy-hydrolysis and the homolysis plugin is bundled with other built-in reactions at github.com/graeter-group/kimmdy-reactions.

4.2.2 System setup

In addition to the single chain and triple helix systems as introduced for Chapter 3 (just without the QM/MM), a complete collagen fibril is introduced. It is an all-atom model of a *Rattus norvegicus* collagen fibril (PDB ID: 3HR2) comprised of 41 triple helices spanning one central overlap and one gap region for a total of roughly 320 000 protein atoms. ColBuilder [93] was used to generate a model with N- and C-terminal Pyridinoline (PYD) cross-links with a connectivity of 9.C-5.B-944.B and 1047.C-1047.A-98.B, respectively. Furthermore, all Phe and Tyr residues were randomly mutated to either dihydroxyphenylalanine (DOPA) deprotonated at the $C\epsilon$ (DO1) or $C\zeta$ (DO2) hydroxy group.

Figure 4.15: A collagen fibril.

The fibril was solvated and neutralized with 0.15 M NaCl. After an energy minimization using the steepest descent method, a 10 ns NVT equilibration and then a 10 ns NPT equilibration were performed. It was then equilibrated under the different external forces ranging from 0 to 3 nN per chain for 2 ns and the convergence of the pulling simulation was verified based on the distances between the pull groups. The equilibration was followed by 2 ns of sampling simulations, during which bond distances where recorded with PLUMED [113]. Likewise, a single triple helix out of a collagen fibril and a single peptide out of the same triple helix where solvated in TIP3 water, neutralized with 0.15 M NaCl, energy minimized and equilibrated under NVT and NPT ensembles at a temperature of 300 K. The same temperature is used in the rate calculations for all reactions. The external forces applied range from no external force to 3 nN. In the case of the complete collagen fibril, an additional, more physiological, sheared pulling was introduced by sampling the force on each triple helix from a Gaussian distribution with a standard deviation of 30 % of the mean force centered around the nominal external force. This mimics the effect of unequal loading under strain in a real tendon or ligament. The code for the entire setup and analysis is available at github.com/graeter-group/kimmdy-hydrolysis-examples.

4.3 Results

Please note that some of the results of this section have also already been discussed in [99] and thus will have a high text similarity.

4.3.1 Physiologically expected pH changes have only a minor contribution to hydrolysis rates

With KIMMDY it is now possible to rapidly sample variations of various parameters, one of which is the pH value. This is visualized in extended data Figure 5.3a, which explores the impact of external force and the pH value on the two systems already introduced for QM/MM in Chapter 3, the single peptide chain and the triple helix. In the current model, the pH value only enters the rate calculation as a global input parameter and is not affected by the outcome of the molecular dynamics sampling step. So by just looking at Equation 44, we can calculate that a drop in pH from 7.4 to pH 7.0, as one would expect to see physiologically, e.g., in muscle tissue during exercise [121], is expected to reduce the rate by a factor of 2.5. This is because it only contributes linearly to the pre-exponential factor.

4.3.2 Solvent accessibility of peptide bonds in the fibril is heterogeneous

The solvent accessible surface area (SASA) warrants a deeper look.

Figure 4.16 explores the SASA of the peptide bonds in the three systems under no force and 1 nN of pulling (with shear pulling for the fibril). The SASA encompasses seven orders of magnitude, but plenty of peptide bonds remain relatively accessible even in the fibril system compared to the reference SASA of the peptide bond in a Glycine-Glycine dipeptide. This is especially true under force, as the backbones of the chains get stretched out, making the peptide bonds even more accessible. Also note how the shape of the distribution for the triple helix matches the shape for the fibril.

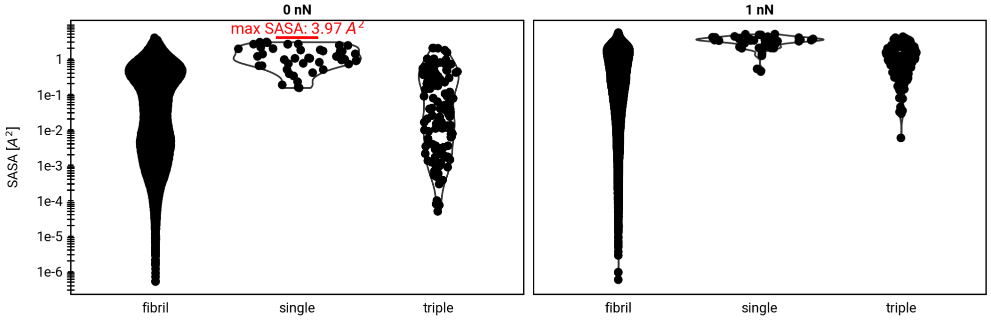


Figure 4.16: Average Solvent Accessible Surface Area (SASA, log10-scale) per peptide bond in the collagen fibril, the single chain and the triple helix system while no or 1 nN of force is applied. The 1 nN case for the fibril represents the shear pulling setup. The SASA of a freely accessible peptide bond in a Glycine-Glycine dipeptide is shown as a reference in red.

Because of this I wondered specifically about the effect of the fibril. Is a triple helix in the middle of the fibril less accessible than a fibril near the edge of the packing? This would lead to boundary effects that should be excluded. However, looking at slices through the fibril system (Figure 4.17 and Figure 4.18) reveals that even the inner triple helices in the overlap region contain bonds with high SASA.

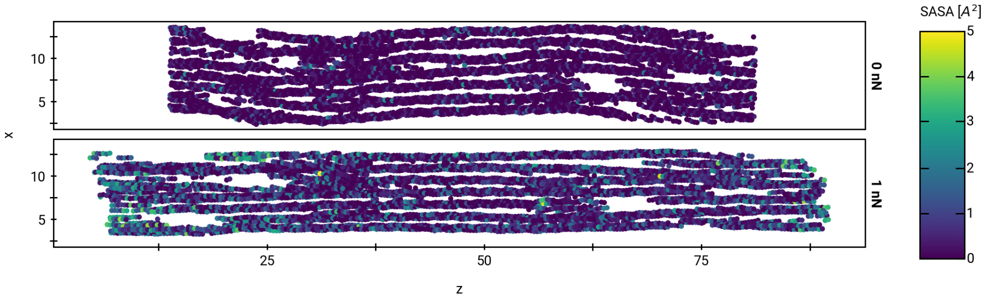


Figure 4.17: A slice through the collagen fibril in the x-z-plane with y-coordinates filtered to be between 7.9 and 9.6 nm. Each peptide bond is colored by SASA for the 0 nN and 1 nN shear pulling setup.

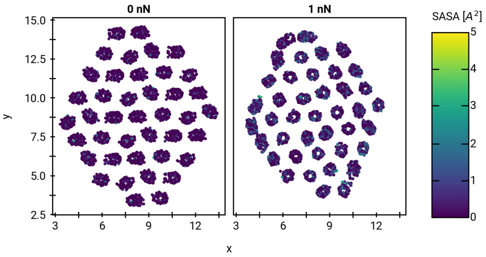


Figure 4.18: A slice through the collagen fibril in the x-y-plane with the z-coordinates filtered to be between 40 and 50 nm. Each peptide bond is colored by SASA for the 0 nN and 1 nN shear pulling setup.

More quantitatively, plotting the SASA against the radial distance from the center in the x-y-plane for peptide bonds in the overlap region reveals no significant increase of the SASA for the outer triple helices (Figure 4.19).

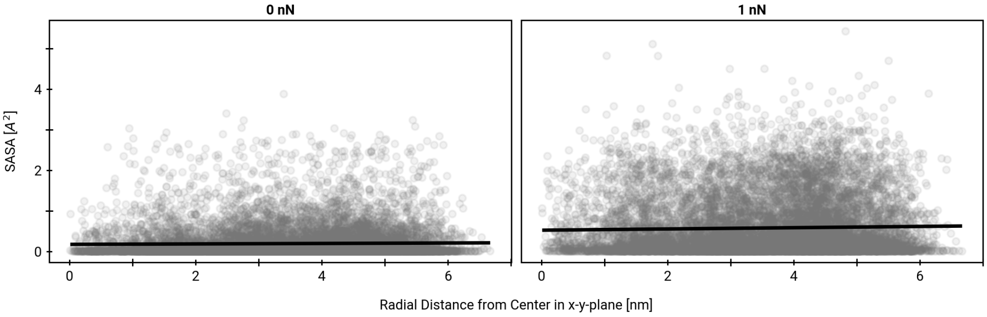


Figure 4.19: SASA of the peptide bonds in the overlap region of the collagen fibril (z-coordinate between 30 and 60 nm) plotted against their radial distance from the center in the x-y plane. Black lines are linear models. 1 nN refers to the 1 nN shear pulling setup.

I discarded the hypothesis that pulling would “squeeze out” water from the fibril, as instead the extension of the chains creates more surface area. Thus, the SASA remains high and the influence of nearby triple helices can be assumed negligible for a large fraction of the peptide bonds. Still it can be explored in more detail with regard to the unique properties of collagen in the next section, Section 4.3.3.

4.3.3 Prolines in collagen may offer protection against hydrolysis

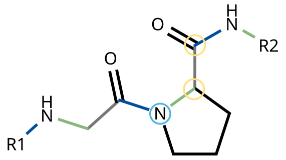
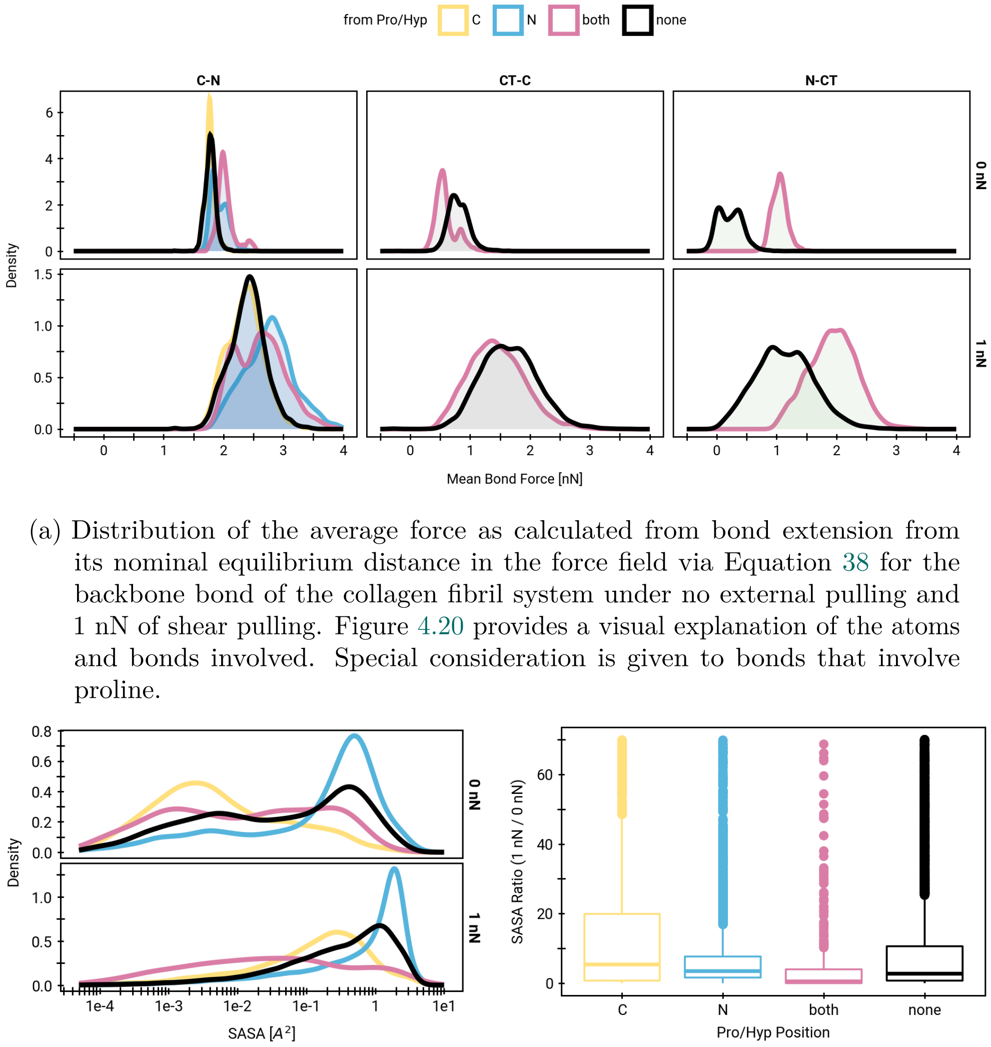


Figure 4.20: A Glycine-Proline peptide highlighting the three bond types of the protein backbone. Blue: amide C – N (the peptide bond with atom types C-N); Green: N – C (with atom types N-CT); Gray: C – C (with atom types CT-C).

Given the unusually high content of proline in collagen, this special amino acid deserves a closer look.



(a) Distribution of the average force as calculated from bond extension from its nominal equilibrium distance in the force field via Equation 38 for the backbone bond of the collagen fibril system under no external pulling and 1 nN of shear pulling. Figure 4.20 provides a visual explanation of the atoms and bonds involved. Special consideration is given to bonds that involve proline.

(b) Distribution of the SASA of the peptide bonds by proline participation.

(c) Ratio of SASA with and without pulling split by proline participation.

Figure 4.21: The effect of pulling on bond strain and solvent accessibility in the collagen fibril.

Figure 4.21 explores the impact of shear pulling on the internal bond forces as calculated via Equation 38 and the solvent accessibility of the peptide bond. The average force is a function of how much the average bond distance in the simulation differs from the equilibrium bond distance in the force field. As such, it is expected to have positive values due to entropic contributions in the simulation and the asymmetric nature of the Morse potential (Figure 2.5, also see Section 4.2.1). In Figure 4.21a, the distributions of the average force are classified by the type of bond and whether one or both of the participating atoms are part of a proline residue. The bond types and atoms involved are visually explained in Figure 4.20.

In equilibrium (0 nN external pulling), the peptide bond (bond type C-N) tends to have a higher extension compared to its equilibrium distance (r_0) when it is between two prolines compared to the extension if no prolines are involved. The same is true if the N of the peptide bond is contributed by a proline and the C by some other amino acid (cyan curve) for one part of the bimodal distribution. If just the C-terminal end is from a proline (yellow curve), the situation is the same as if no proline is part of the bond (black curve). Under external force (1 nN), the peptide bonds with two prolines or the proline N are under the highest strain. The C $-$ C bond (with atom types CT-C) is under more strain in non-proline residues, but the N $-$ C bond (with atom types N-CT) is more strained in prolines due to the ring structure.

Notably, despite the higher strain, peptide bonds between proline residues have a lower solvent accessibility than those between other residues (Figure 4.21b) and the accessibility also increases less under external pulling (Figure 4.21c). The effect is attenuated when just the N-terminal side of the peptide bond is a proline, because these cases then have a high chance of the smallest amino acid Glycine on the C-terminal side. In nature, the metallopeptidases that cleave collagen peptides (collagenase) are also less likely to cleave between proline residues according to the MEROPS database [122] (ebi.ac.uk/merops/cgi-bin/pepsum?id=M10.001, see ID M10.001 and up)

However, because the SASA ratio only enters into the hydrolysis rate calculation linearly in the current model, its influence on the final rate is small compared to the influence of the bond stretching by external force (Figure 4.22).

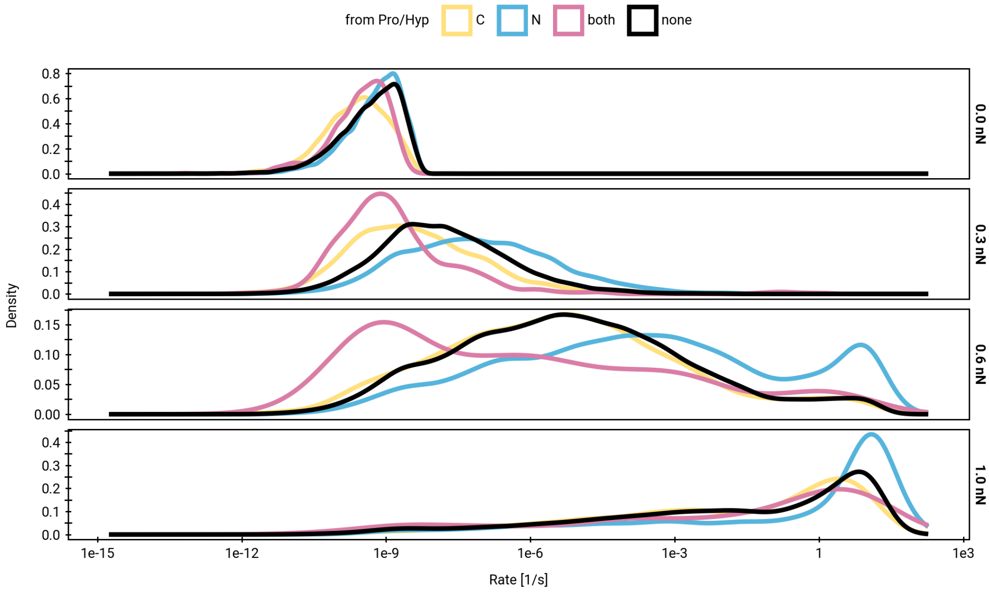


Figure 4.22: Distribution of hydrolysis rates by proline participation for the fibril setup with increasing shear pulling forces.

4.3.4 Homolysis becomes competitive with hydrolysis by force concentration

The externally applied force and in turn the stretching of the bonds enters exponentially into the rate calculations for both reactions (see Figure 4.14b). Furthermore, a small divergence in bond length from the equilibrium bond length has a large influence on the measured force in the bond (see Figure 4.14a).

It follows from Figure 4.23 that in the absence of an external pulling force, both hydrolysis and homolysis yield low rates. This is unsurprising, given that spontaneous fragmentation of peptides is highly unlikely in ambient conditions. When the single chain protein systems are pulled at 1 nN, hydrolysis outcompetes homolysis for a single peptide chain (Figure 4.23, right). However, for a densely packed and cross-linked collagen system pulled with shear stress and the same average external force of 1 nN, with all else being equal, homolysis is accelerated drastically (Figure 4.23, left). It reaches rates of similar orders of magnitude as hydrolysis and outcompetes it when comparing the highest rates.

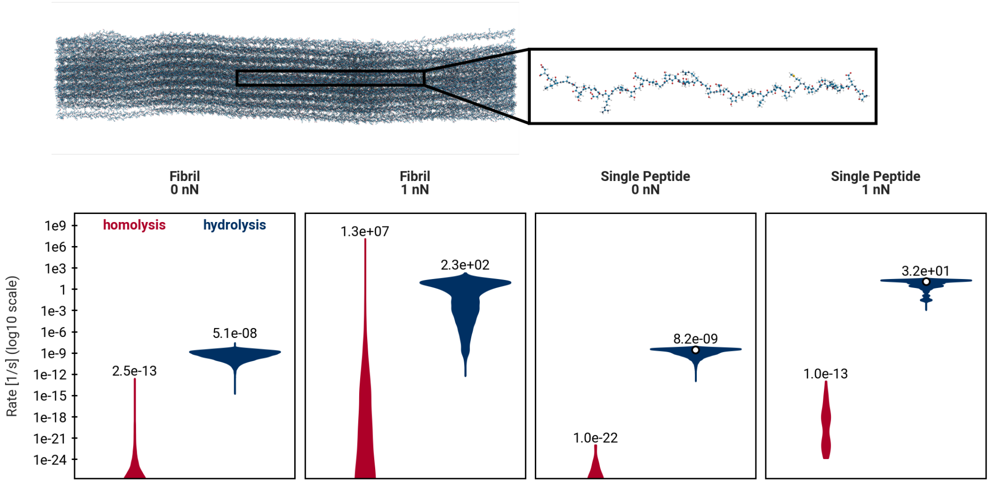


Figure 4.23: Reaction rates (log10-scale) for homolysis and hydrolysis in the collagen fibril system (left) and the single chain (right), with the systems visualized above the plot. The 1 nN case for the fibril system represents the shear pulling setup. The value of the highest rate for each setup is shown above the respective distribution. The white points in the hydrolysis rate distributions of the single peptide denote the experimental reference value from AFM experiments [20] for the respective force. Rates lower than 10^{-25} 1/s are not shown.

This is due to areas of high stress concentration in the large collagen fibril system that push the reaction rates into a region of the force response curve where hydrolysis rates already taper off and homolysis rates still continue to rise with force (see Figure 4.14 and [100]). This force concentration is the result of the shear pulling in combination with the cross-links. Ultimately, these high rates are what we are concerned with the most, because once a reaction has happened during instantaneous loading, the stress should subside. This puts more focus on the fastest rates. The competition would be even more in favor of homolysis if the hydrolysis model used the theoretical reaction rates based on the *ab initio* activation energy calculations by [20] as they can be seen in the extended data Figure 5.5. Extended data Figure 5.6 shows a more fine-grained force profile between 0 and 1 nN of shear pulling.

4.3.5 Discussion and outlook of applying KIMMDY to hydrolysis and homolysis

We were able to simulate direct competition of the two reactions and provide qualitative answers as to how hydrolysis and homolysis compete. But directly connecting quantitative values from experiments to simulated reaction rates remains a challenge. KIMMDY bridges the gap between AFM experiments on single peptides and large scale biomolecular simulations of collagen fibrils, but we are still missing a concrete link to experimental data of real tissues. Zapp et al. pulled with stresses of 2-40 MPa on tissue samples of tendons and saw signals of radicals in EPR measurements [17], but it is unclear how the number of radicals could be quantified in a way that would map down to reaction rates at the molecular level. The external pulling force (with shear pulling) at which we see the first homolysis reaction rates at the second timescale in the KIMMDY simulations is greater than 0.6 nN per chain (see Figure 5.6) and hydrolysis would be predicted to occur before that, according to the model based on AFM experiments. This force equates to 1.8 nN per triple helix, which has a cross-sectional area of 1.77 nm^2 (1.5 nm diameter) and thus results in a stress of 1.02 GPa, which is two to three orders of magnitude higher than what is used in experiments and what is measured physiologically (Section 1.2), even with taking into account that only 60% of the tendon weight would be collagen and subtracting the water content to get an effective CSA. Together this suggests that the forces within complex biological systems can momentarily and locally confined exceed computationally determined thresholds due to irregularities as they can only be modeled to a limited extent even with the large collagen fibril simulation system. Thus, a model for the missing scale between complete tendons and sections of fibrils would be highly interesting for studying force propagation.

The current model for hydrolysis only takes the pH value into account in the form of a global scaling factor. Combining KIMMDY with advances in constant pH simulations in GROMACS [123] could allow for a more fine-grained influence and lead to more accurate predictions of the hydrolysis rate.

Likewise, the influence of the solvent accessibility of the peptide bond is highly likely to be underestimated in the current model for multiple reasons. Firstly, as shown in Chapter 3, the mobility of protons, the formation of hydrogen bonds, and in general the arrangement of water and hydroxide molecules around the peptide bond is crucial for the hydrolysis reaction. So is the angle of attack of the hydroxide towards the electrophilic center. It is a current limitation of the simple model that SASA only influences the rate linearly, which becomes clear especially when looking at the range of possible values in the collagen fibril system. The SASA only entering the rate calculation linearly is a very conservative assumption. It is expected that a more accurate model that takes into account the likelihood of water molecules arranging favorably around the reaction site will produce lower reaction rates, which would further swing the reaction rates in favor of homolysis. Secondly, the collagen fibril model, and the way external pulling is applied, best describes a slice through an infinite collagen fiber. But nature is not infinite. Thus, even with Section 4.3.2 showing that large portions of peptide bonds remain accessible to solvent even near the center of the fibril, it remains to be seen if this is actually the case in even larger scale simulations. Those could potentially better take into account the thinning of the system as it is stretched and answer how this affects the water content and the packing of the triple helices in the fibril.

4.4 Beyond hydrolysis: KIMMDY can be applied to a variety of systems and reactions

KIMMDY is designed as a framework that can be easily extended to different chemistries. As such, we included examples for various systems and reactions in our manuscript [99]. I would like to especially shout out my colleagues Eric Hartmann, Kai Riedmiller, Evgeni Ulanov, Boris Schüpp, Denis Kiesewetter, Daniel Sucerquia and Camilo Aponte-Santamaría. The examples I am about to show were set up by them, while I only provided implementation help for the reaction plugins as well as features and specialized fixes on the KIMMDY side.

We showed that KIMMDY can be applied to a very diverse set of systems. Boris Schüpp applied KIMMDY to cross-linking reactions in DNA origami and Camilo Aponte-Santamaría showed that KIMMDY can even work with coarse-grained polymer systems. Both of these systems are interesting in their own right, but don't quite fit into the core story of this thesis. They provide a great way of showing the versatility of KIMMDY, but there is one result I would like to focus on in particular, because it complements the rest of the collagen discoveries very well.

4.4.1 KIMMDY finds a new radical-scavenging candidate in collagen

With the power of KIMMDY we can simulate the time evolution of reactions through large biomolecular systems. First, with KIMMDY we simulated homolytic bond breaks in the collagen fibril system (see Figure 4.23 and Figure 4.15). Next, we used it to simulate what happens to the radicals that form as the result of those bond scissions (Figure 4.24). A radical can move through a molecule by a process known as hydrogen atom transfer (HAT), whereby a nearby hydrogen atom forms a covalent bond with the radical, leaving its original binding partner as the new radical. Kai Riedmiller trained a graph neural network to emulate the reaction energy barrier of HAT with high accuracy [124], which in turn provides reaction rates for KIMMDY in the form of the HAT-plugin. Evgeni Ulanov showed in simulations of small n-alkyl radicals that KIMMDY robustly reproduces the overall reaction dynamics compared to experiments, with only slight deviations for the most prominent shifts relative to each other. Eric Hartmann and Denis Kiesewetter then applied this to the large collagen system.

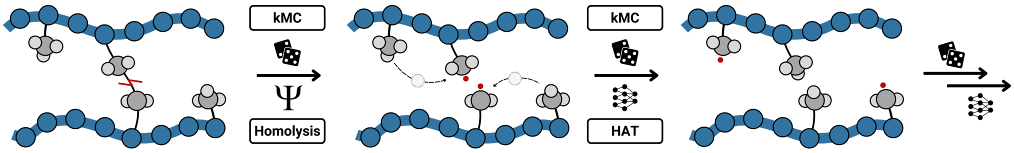
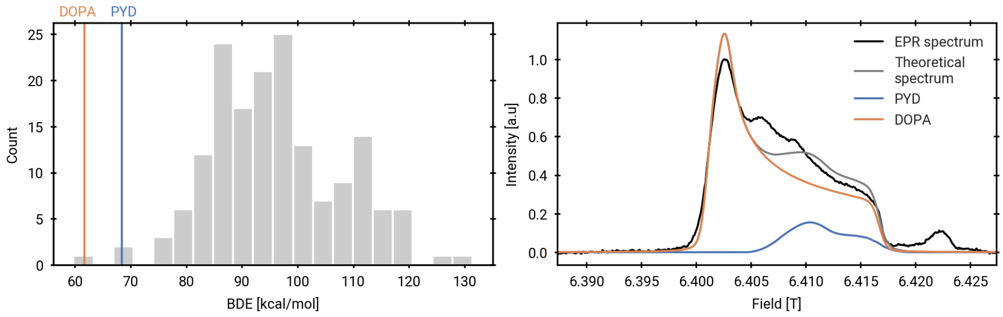


Figure 4.24: Schema of successive HAT reactions after homolytic bond scission simulated in KIMMDY [99].

After many consecutive HAT reactions, the dwell time for the residues was analyzed and two compounds stood out. The first is the familiar dihydroxyphenylalanine or DOPA (Figure 4.26), which forms a stable radical at one of the hydroxyl groups. The second is a new face, the pyridinoline cross-link (PYD, Figure 4.27). It can also form a stable radical at its hydroxide, as BDE calculations then revealed (Figure 4.28a). Furthermore, when Daniel Sucerquia calculated the expected EPR spectrum of the PYD radical, this radical was shown to make up for part of the missing intensity that is seen in experimental spectra compared to just using DOPA (Figure 4.28b). This is an exciting discovery, because it suggests that not only is collagen more likely to break at or near the cross-links [100], nature also came prepared with a radical scavenger in the very vicinity.



(a)

(b)

Figure 4.28: Analysis of the DOPA and PYD radicals. The figures are from our manuscript [99]. (a) Bond dissociation energy BDE of PYD (blue line) and DOPA (orange line) compared to a distribution of BDE of abstractions processes in different amino acids from [125] (gray bars). (b) Experimental EPR spectrum compared to theoretical EPR spectrum for a DOPA and PYD radical.

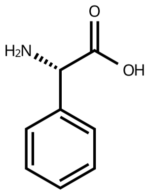


Figure 4.25: Phenylalanine.

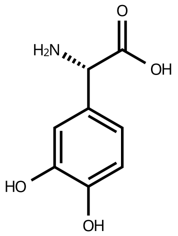


Figure 4.26: Dihydroxyphenylalanine (DOPA).

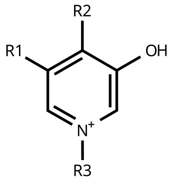


Figure 4.27: Center of a pyridinoline crosslink.

5 Concluding thoughts and next steps

I set out to answer how two reactions, peptide bond hydrolysis and homolytic bond scission, compete under external force in complex biological systems, specifically in our most important structural protein, collagen, with its unique multiscale structure (Figure 1.9). To address this question, I used and developed two complementary methods.

Hybrid quantum mechanics/molecular mechanics simulations provided atomistic insights into the reaction mechanisms of base-catalyzed hydrolysis. They highlighted the dynamic nature of the hydrolysis reaction, showing how proton mobility within the solvent is crucial. The free energy profiles of the reaction for the same sites in both a single peptide and a triple helix revealed that the triple helix does marginally impede peptide hydrolysis thermodynamically. Since the reaction barrier enters exponentially into reaction rate calculations (see Equation 33), this seemingly small difference of 9.02 kJ/mol still corresponds to a decrease in the reaction rate by a factor of 30 to 40. However, the computational costs of QM/MM simulations in combination with the excessive sampling requirements of a reaction as complex as hydrolysis, limit the scope when looking into the two competing reactions at scale. This limitation motivated the development of KIMMDY (Kinetic Monte Carlo Molecular Dynamics), a generalized framework for bridging timescales in reactive molecular dynamics simulations.

Built as a collaborative effort with Eric Hartmann and Kai Riedmiller, KIMMDY combines classical molecular dynamics simulations with a plugin architecture for incorporating diverse reaction types based on physical models, machine learning, or experimental heuristics. KIMMDY allowed us to directly compare competing reactions at scale and revealed the critical role of force concentration in cross-linked fibrils. Localized stress concentrations lead to a regime where homolysis becomes competitive with or even outcompetes hydrolysis. Additional analysis uncovered that solvent accessibility of peptide bonds remains surprisingly high even in the fibril interior, particularly under tension. Proline residues, which are part of the characteristic Gly-X-Y motif in collagen, appear to offer some protection against hydrolysis by reducing solvent accessibility, even though they experience higher mechanical strain. The findings from the QM/MM simulations strongly suggest that the influence of solvent accessibility should be greater than how it is currently modeled, but more data is required to answer this question quantitatively. This provides an opportunity to build a more accurate model for hydrolysis in the future. One approach could be a machine-learned emulator trained on QM/MM data that predicts reaction barriers based on the local chemical environment of each peptide bond, similar to the HAT model by Kai Riedmiller [124]. Due to the complexity of the reaction mechanism, this still poses a considerable challenge, but the plethora of QM/MM simulations of the entire hydrolysis reaction presented here provide a great starting point as training data and reference systems. The homolysis reaction could also be modeled more accurately, by using more information about the chemical environment than just the current bond stretching and the BDE.

Beyond the specific question of hydrolysis versus homolysis, KIMMDY enabled the simulation of radical dynamics through hydrogen atom transfer reactions. This led to the discovery that pyridinoline cross-links can act as radical scavengers, complementing DOPA, which helped explain previously unaccounted features in experimental EPR spectra. The finding that nature has placed radical scavengers in the very vicinity where bonds are most likely to break [126] reveals an elegant protective mechanism.

The forces at which homolysis reactions become prevalent in simulations remain orders of magnitude higher than experimentally measured tissue stresses, suggesting that additional factors beyond what can be captured even in large-scale fibril simulations contribute to force amplification in biological systems. Thus, it would be a great opportunity in the future to build a model that bridges the missing scale between all-atom molecular dynamics simulations and real tissue samples. Coarse-grained force fields such as Martini3 [127] provide a way of simulating larger systems by merging multiple atoms into a combined representation and KIMMDY is already prepared to work with such topologies. This leaves only the reaction plugins to be adapted to the coarse-grained representation. An even more ambitious approach would be to couple KIMMDY with finite element models [128, 129] of entire tissues [130], which would allow for simulating tendons or ligaments, albeit at a lower resolution.

Crossing the experimental to computational divide from the other side with more single-molecule experiments should also be fruitful. While we now have experimental lifetimes of peptide bonds [20], seeing the same for homolysis would be incredibly exciting. This could be achieved with an AFM setup with a medium containing a probe for radicals or the products of their follow-up reactions with the solvent (reactive oxygen species) such as 2',7'-Dichlorodihydrofluorescein (DCFH2, Figure 5.1) [131]. Combining this with the experimental setup of Pill et al. [20], could directly test the predictions made by KIMMDY about the competition of hydrolysis and homolysis under force.

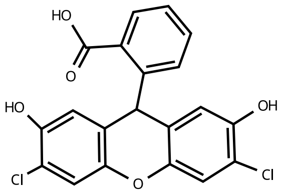


Figure 5.1: 2',7'-Dichlorodihydrofluorescein (DCFH2).

Quite a while ago I was working on my Bachelor's thesis in the lab of Matthias Mayer at the Center for Molecular Biology of Heidelberg University (ZMBH). During a group meeting I presented my findings and came across two seemingly contradictory data points, about which I then stated: "… and now I'm just confused." Matthias immediately jumped in to say: "That's brilliant! Being confused is always the first step." Since then, I cherish those moments of confusion, those drivers of discovery. What I didn't fully grasp at that time is that every next step after the first, simultaneously has the potential to be the first step of something new. I sincerely hope that in this thesis, I provided some answers and built a tool to find more in the future, but more importantly, also opened up many new questions, so that we can all revel in confusion together.

« Being confused is always the first step! »

Appendix

Extended data figures

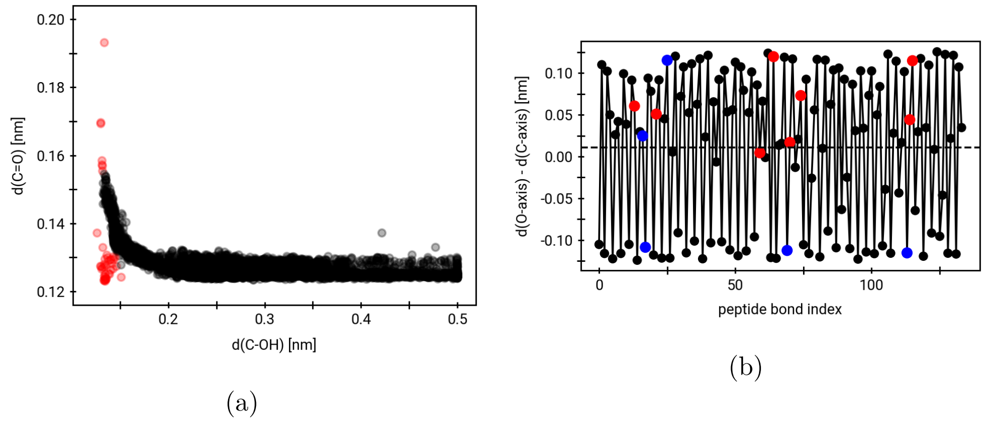
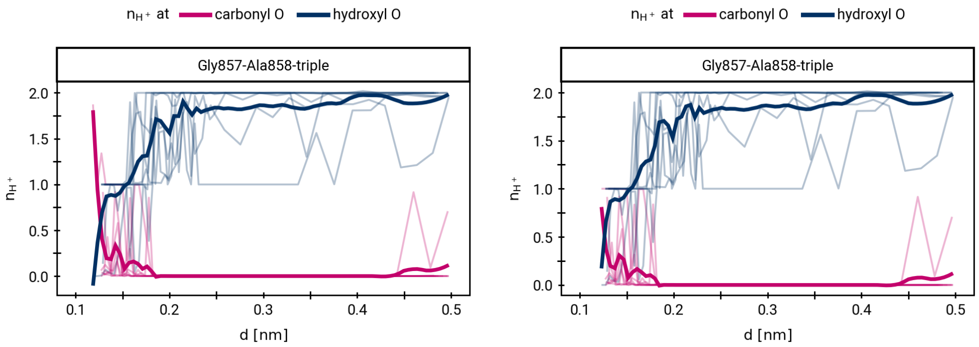


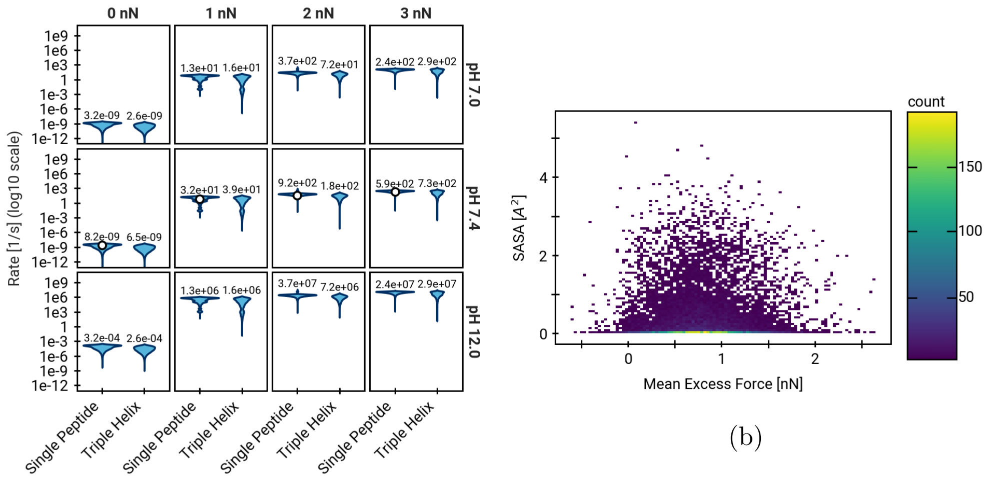
Figure 5.1: Extended QM/MM sampling statistics. (a) Bond length of the original carbonyl C=O bond plotted against the distance of the attacking hydroxide to the carbonyl C. Each point represents the average within one umbrella sampling window. Windows colored red were discarded due to being side reactions.
(b) The distance of the carbonyl O minus the distance of the carbonyl C of the peptide bond from the central axis of the force-equilibrated collagen triple helix structure. This represents the orientation of the carbonyl C=O bond, where a number greater than 0 means that the bond is pointing outwards from the backbone, away from the respective other two chains of the helix and a value smaller than 0 means it is pointing into the helix. The sampled sites are highlighted in red. The blue points are additional sites that are being sampled for the next publication.



(a) Protonation states before filtering.

(b) Protonation states after filtering.

Figure 5.2: The effect of filtering out side reactions for one affected example.



(b)

(a)

Figure 5.3: Extended hydrolysis results from KIMMDY runs. (a) Hydrolysis reaction rates for increasing forces and varying pH values in the single peptide chain and triple helix systems as sampled with KIMMDY. (b) SASA in 1 nN shear pulling fibril simulation plotted against the Mean Excess Force, i.e., the force derived from bond extension in each peptide bond after accounting for average bond forces in the baseline, 0 nN simulations.

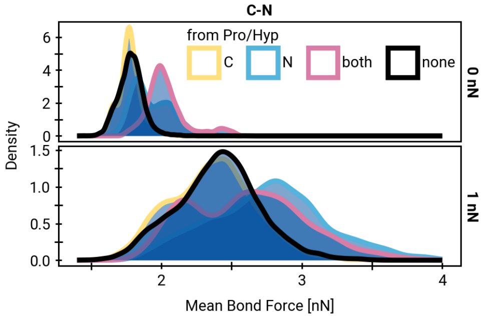


Figure 5.4: A closer look at the mean bond force in the peptide bonds of the collagen fibril system under shear pulling.

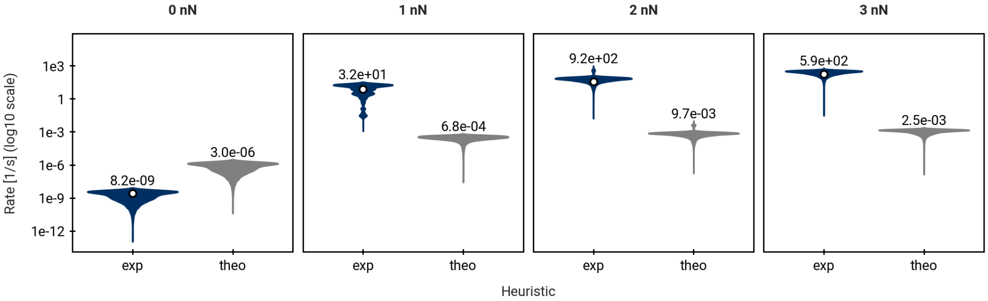


Figure 5.5: Hydrolysis reaction rates (log10-scale) for a single chain under force increasing from 0 to 3 nN using the experimental (exp) model compared to the physical model based on single-point energies. Rates lower than 10^{-13} 1/s are not shown.

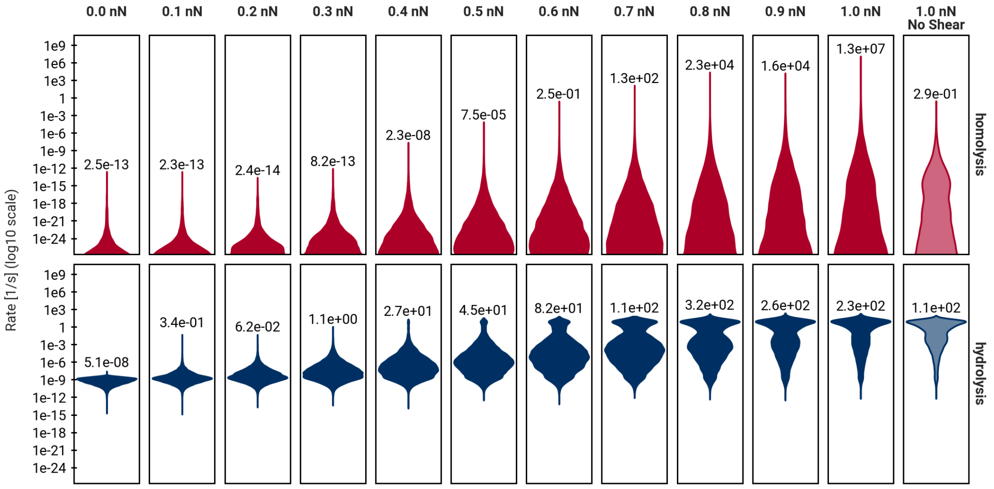


Figure 5.6: Reaction rates (log10-scale) for hydrolysis and homolysis in the collagen fibril system under shear stress for increasing external force with the 1 nN no-shear setup shown for comparison. Rates lower than 10^{-25} 1/s are not shown.

Extended data listings

Listing 8 Select lines from a .gro file of a capped glycine dipeptide

```
   26
    1ACE    CH3    1  -0.351  -0.014  -0.085
    1ACE   HH31    2  -0.365  -0.002   0.013
[...]
   1.11109   0.40658   0.49846
```

Listing 9 Select lines from a .pdb file of a capped glycine dipeptide, shortened.

```
ATOM  1  C  ACE  1  -2.031  -0.018  -1.163  1.00  0.00  C
ATOM  2  O  ACE  1  -1.615  -0.128  -2.324  1.00  0.00  O
```

Listing 10 Excerpts from the code of KIMMDY.

```
class MoleculeType:
    """One moleculetype in the topology

    Attributes
    ----------
    atoms : dict[str, Atom]
    bonds : dict[tuple[str, str], Bond]
    pairs : dict[tuple[str, str], Pair]
    angles : dict[tuple[str, str, str], Angle]
    proper_dihedrals :
      dict[tuple[str, str, str, str], MultipleDihedrals]
    improper_dihedrals :
      dict[tuple[str, str, str, str], Dihedral]
    position_restraints : dict[str, PositionRestraint]
    dihedral_restraints :
      dict[tuple[str, str, str, str], DihedralRestraint]
    radicals : dict[str, Atom]
        dict mapping atom indices to atom objects
        for storing all radical atoms

    """
```

Listing 11 Excerpts from the code of KIMMDY.

```python
@dataclass()
class Atom:
    """Information about one atom

    A class containing atom information as in the atoms section
    of the topology. An atom keeps a list of which atoms it is
    bound to and its radical state.

    From gromacs topology:
    ; nr type resnr residue atom cgnr charge mass typeB chargeB massB
    """

    nr: str
    type: str
    resnr: str
    residue: str
    atom: str
    cgnr: str
    charge: str
    mass: Optional[str] = None
    typeB: Optional[str] = None
    chargeB: Optional[str] = None
    massB: Optional[str] = None
    bound_to_nrs: list[str] = field(
      default_factory=list, compare=False
    )
    is_radical: bool = field(default=False, compare=False)
    comment: Optional[str] = field(default=None, compare=False)

    @classmethod
    def from_top_line(cls, l: list[str]):
        return cls(
            nr=l[0],
            type=l[1],
            resnr=l[2],
            residue=l[3],
            atom=l[4],
            cgnr=l[5],
            charge=l[6],
            mass=field_or_none(l, 7),
            typeB=field_or_none(l, 8),
            chargeB=field_or_none(l, 9),
            massB=field_or_none(l, 10),
        )
```

Additional software and resources

The source code for this thesis and for auxiliary figures is available at github.com/jmbuhr/phd-thesis.

I would like to highlight a couple of additional resources that were instrumental not only during the research for my PhD, but also for the writing of this thesis. This is important to me, because I strongly believe in Open Source and Open Science. Many of these projects are maintained by volunteers and researchers that put in a lot of effort to make them available to the community without expecting anything in return. Especially research software rarely gets the recognition it deserves, despite being a crucial part of modern science.

Knowledge resources

- David Sherrill of Georgia Tech University, who has excellent lectures online for brushing up on QM and MM simulations: youtube.com/@DavidSherrill1

Software

- GROMACS: gromacs.org [40]
- PLUMED: plumed.org [113]
- VMD: ks.uiuc.edu/Research/vmd/ [132]
- R: r-project.org [133]
  - Tidyverse: tidyverse.org [134]
- Python: python.org
  - Pandas: pandas.pydata.org [135, 136]
  - Polars: pola-rs.github.io/polars/
  - Numpy: numpy.org [137]
  - Matplotlib: matplotlib.org/ [138]
  - Plotnine: plotnine.org/ [139]
  - MDAnalysis: mdanalysis.org [111, 112]
  - uv: docs.astral.sh/uv/
- moldraw (ketcher): moldraw.com
- Neovim: neovim.io
- Zotero: zotero.org
- Quarto: quarto.org [140]
- Apache Hamilton: hamilton.apache.org
- git: git-scm.com
- mol*: molstar.org [141]

List of abbreviations and constants

AA: amino acid
ACE: acetyl-capped
AFM: atomic force microscopy
Ala: alanine
AO: atomic orbital
BD: Bürgi-Dunitz angle
BDE: bond dissociation energy
CGTO: contracted Gaussian-type orbitals
CLI: command-line interface
COM: center of mass
cryo-EM: cryo-electron microscopy
CSA: cross-sectional area
DAG: directed acyclic graph
DFT: density functional theory
DOPA: dihydroxyphenylalanine
e.g.: exempli gratia
EM: electron microscopy
EPR: electron paramagnetic resonance
ER: endoplasmic reticulum
GEEP: Gaussian expansion of the electrostatic potential
GGA: generalized gradient approximation
Gly: glycine
GTO: Gaussian-type orbitals
HAT: hydrogen atom transfer
HITS: Heidelberg Institute for Theoretical Studies
HPC: high performance computing
Hyp: hydroxyproline
ID: identifier
i.e.: id est
kMC: kinetic Monte Carlo
KS: Kohn-Sham
LCAO: linear combination of atomic orbitals
LDA: local density approximation
LJ: Lennard Jones
LLMs: large language models
MD: molecular dynamics
ML: machine learning
MLP: machine-learned potential
MM: molecular mechanics
MO: molecular orbital
MPI: message passing interface
MPI-P: Max Planck Institute for Polymer Research
NME: N-methyl group
NPT: isothermal-isobaric ensemble

NVE: microcanonical ensemble
NVT: canonical ensemble
PBC: periodic boundary conditions
PME: particle mesh Ewald
PMF: potential of mean force
Pro: proline
PYD: pyridinoline
PyPI: Python package index
QM/MM: quantum mechanics/molecular mechanics
QM: quantum mechanics
QQ: Coulomb
RESP: restrained electrostatic potential
ROS: reactive oxygen species
SASA: solvent accessible surface area
SEM: scanning electron microscopy (also: standard error of the mean)
Ser: serine
Slurm: simple linux utility for resource management
STO: Slater-type orbitals
TI: tetrahedral intermediate
TS: transition state
WHAM: weighted histogram analysis method
XC: exchange correlation functional
ZI: zwitterionic intermediate
ZMBH: Center for Molecular Biology of Heidelberg University

N_A: Avogadro constant (6.022×10^{23} mol^{-1})
k_B: Boltzmann constant (1.380649×10^{-23} J K^{-1})
h: Planck constant (6.626070×10^{-34} J Hz^{-1})
$\hbar$: reduced Planck constant (1.054572×10^{-34} J s)
R: gas constant ($N_\mathrm{A}k_\mathrm{B}$)

List of Figures

1.1 The Capture of Chandernagore. 13
1.2 Remove scurvy. 13
1.3 Brazen Buccaneer. 13
1.4 Glycine. 14
1.5 Proline. 14
1.6 Hydroxyproline. 14
1.7 Lysine. 14
1.8 Hydroxylysine. 14
1.9 Schematic of the structure of collagen type I. 15
1.10 SEM of cartilage collagen. 16
1.11 Achilles tendon render. 16
1.12 Hydrolysis vs. homolysis reaction schema. 17
2.1 Propionaldehyde, chemical formula. 19
2.2 Propionaldehyde, ball-and-stick render. 19
2.3 Propionaldehyde, nuclei with electron density. 19
2.4 Bonded energy terms. 21
2.5 Morse potential. 21
2.6 Harmonic potential. 21
2.7 Lennard-Jones potential. 22
2.8 Coulomb potential. 22
2.9 Periodic boundary conditions. 26
2.10 Squirrel holding upside-down umbrella. 29
2.11 QM/MM electrostatic embedding. 35
2.12 Stacked rates to sample from during kMC. 36
3.1 Hydrolysis reaction scheme. 39
3.2 Pill et al. pulling setup. 40
3.3 Summary of datapoints from Pill et al. [20]. 40
3.4 Pulling equilibration. 43
3.5 The collagen triple helix and single peptide before and after equilibration under external force. 43
3.6 Sampled hydrolysis sites. 44
3.7 Schema of the Bürgi-Dunitz and Flippin-Lodge angles. 44
3.8 The QM/MM subsystems, initial state. 45
3.9 Statistics of the umbrella sampling simulations of TS1. 47
3.10 The QM/MM subsystems, TI. 48
3.11 Reminder of where we are in the hydrolysis reaction. 48
3.12 Visual explanation of the proton distances. 49
3.13 Distances of the TI to the nearest proton. 49
3.14 Visual explanation of the proton counting. 50
3.15 Protonation states of the carbonyl during hydroxide attack. . . . 50
3.16 Averaged protonation states during hydroxide attack. 51
3.17 Free energy profiles of TS1 and the TI. 52
3.18 Results of the umbrella sampling of the attack of the hydroxide, reaching the TI via TS1. 53

3.19 The QM subsystem during peptide bond rupture. 54
3.20 Reminder of where we are in the hydrolysis reaction. 54
3.21 Results of the umbrella sampling of the breaking of the peptide bond. 55
3.22 Visual explanation of the distances during peptide bond breaking. 55
4.1 KIMMDY logo. 58
4.2 KIMMDY schema. 58
4.3 Schema of the diataxis documentation framework, based on a figure by Daniele Procida [109]. 60
4.4 `kimmdy.yml` features. 62
4.5 Simplified graph representation of a peptide bond undergoing hydrolysis. 65
4.6 Water and a peptide bond. 65
4.7 A topology molecules section. 66
4.8 GRAPPA logo. 67
4.9 Screenshot of the interactive molecule display of an alanine dipeptide with a radical center. 67
4.10 KIMMDY's slow-growth applied to hydrolysis of a peptide bond. 68
4.11 KIMMDY analysis tools. 70
4.12 Derivative of the Morse potential. 71
4.13 Hydrolysis reaction rates fit. 72
4.14 Force response of the homolysis and hydrolysis reaction rates. . . 73
4.15 A collagen fibril. 74
4.16 Average SASA in the collagen fibril. 76
4.17 SASA in a x-z-plane-slice of the collagen fibril. 76
4.18 SASA in a x-y-plane-slice of the collagen fibril. 77
4.19 SASA in the collagen fibril against radial distance from the center. 77
4.20 Glycine-Proline peptide. 78
4.21 The effect of pulling on bond strain and solvent accessibility in the collagen fibril. 78
4.22 Hydrolysis rates by proline participation. 80
4.23 Reaction rates for homolysis and hydrolysis in the collagen fibril system. 81
4.24 Schema of successive HAT reactions after homolytic bond scission simulated in KIMMDY [99]. 84
4.25 Phenylalanine. 85
4.26 Dihydroxyphenylalanine (DOPA). 85
4.27 Center of a pyridinoline crosslink. 85
4.28 Analysis of the DOPA and PYD radicals. 85
5.1 2',7'-Dichlorodihydrofluorescein (DCFH2). 87
5.1 Extended QM/MM sampling statistics. 88
5.2 The effect of filtering out side reactions for one affected example. 89
5.3 Extended hydrolysis results from KIMMDY runs. 89
5.4 Mean bond force in the peptide bonds. 90
5.5 Theoretical vs. experimental hydrolysis reaction rates. 90
5.6 Hydrolysis reaction rates for increased shear stress. 91

List of Listings

1 Excerpt from forcefield.itp 23
2 Select lines from ffnonbonded.itp 24
3 Selected, simplified lines from a .top file of a capped glycine dipeptide 25
4 Selected, simplified lines from ffbonded.itp 25
5 Selected lines from aminoacids.rtp 26
6 Creating a topology and modifying it. 66
7 Inspecting the topology object. 66
8 Select lines from a .gro file of a capped glycine dipeptide 92
9 Select lines from a .pdb file of a capped glycine dipeptide, shortened. 92
10 Excerpts from the code of KIMMDY. 93
11 Excerpts from the code of KIMMDY. 94

References

1. Price, C. The Age of Scurvy. (2017).
2. Lind, J. *A Treatise on the Scurvy: In Three Parts. Containing an Inquiry Into the Nature, Causes, and Cure, of That Disease. Together with a Critical and Chronological View of What Has Been Published on the Subject. S. Crowder.* (S. Crowder, 1772).
3. Baron, J. H. Sailors' scurvy before and after James Lind – a reassessment. *Nutrition Reviews* **67**, 315–332 (2009).
4. Verzár, F. Aging of the Collagen Fiber. in *Elsevier.* vol. 2 243–300 (Elsevier, 1964).
5. Ricard-Blum, S. The Collagen Family. *Cold Spring Harbor Perspectives in Biology* **3**, a004978 (2011).
6. The University of Wales Bioimaging laboratory, Institute of Biological Sciences, The University of Wales, Aberystwyth, Wales, UK, Hughes, L., Archer, C. & Ap Gwynn, I. The ultrastructure of mouse articular cartilage: Collagen orientation and implications for tissue functionality. A polarised light and scanning electron microscope study and review. *European Cells and Materials* **9**, 68–84 (2005).
7. Fang, F. & Lake, S. P. Experimental evaluation of multiscale tendon mechanics. *Journal of Orthopaedic Research* **35**, 1353–1365 (2017).
8. Sahinis, C., Kellis, E., Dafkou, K. & Ellinoudis, A. Reliability of Distal Hamstring Tendon Length and Cross-sectional Area Using 3-D Freehand Ultrasound. *Ultrasound in Medicine and Biology* **47**, 2579–2588 (2021).
9. Besier, T. F., Fredericson, M., Gold, G. E., Beaupré, G. S. & Delp, S. L. Knee muscle forces during walking and running in patellofemoral pain patients and pain-free controls. *Journal of Biomechanics* **42**, 898–905 (2009).
10. Fukashiro, S., Komi, P. V., Järvinen, M. & Miyashita, M. Comparison between the directly measured achilles tendon force and the tendon force calculated from the ankle joint moment during vertical jumps. *Clinical Biomechanics* **8**, 25–30 (1993).
11. Sponbeck, J. K., Perkins, C. L., Berg, M. J. & Rigby, J. H. Achilles Tendon Cross Sectional Area Changes Over a Division I NCAA Cross Country Season. *International Journal of Exercise Science* **10**, 1226–1234 (2017).
12. Finni, T., Komi, P. V. & Lepola, V. In vivo human triceps surae and quadriceps femoris muscle function in a squat jump and counter movement jump. *European Journal of Applied Physiology* **83**, 416–426 (2000).
13. Svensson, R. B., Hansen, P., Hassenkam, T., *et al.* Mechanical properties of human patellar tendon at the hierarchical levels of tendon and fibril. *Journal of Applied Physiology* **112**, 419–426 (2012).
14. Chimich, D., Shrive, N., Frank, C., Marchuk, L. & Bray, R. Water content alters viscoelastic behaviour of the normal adolescent rabbit medial collateral ligament. *Journal of Biomechanics* **25**, 831–837 (1992).

15. Kauzmann, W. & Eyring, H. The Viscous Flow of Large Molecules. *Journal of the American Chemical Society* **62**, 3113–3125 (1940).
16. Beyer, M. K. & Clausen-Schaumann, H. Mechanochemistry: The Mechanical Activation of Covalent Bonds. *Chemical Reviews* **105**, 2921–2948 (2005).
17. Zapp, C., Obarska-Kosinska, A., Rennekamp, B., *et al.* Mechanoradicals in tensed tendon collagen as a source of oxidative stress. *Nature Communications* **11**, 2315 (2020).
18. Fitch, K. R. & Goodwin, A. P. Mechanochemical Reaction Cascade for Sensitive Detection of Covalent Bond Breakage in Hydrogels. *Chemistry of Materials* **26**, 6771–6776 (2014).
19. Xia, F., Bronowska, A. K., Cheng, S. & Gräter, F. Base-Catalyzed Peptide Hydrolysis Is Insensitive to Mechanical Stress. *The Journal of Physical Chemistry B* **115**, 10126–10132 (2011).
20. Pill, M. F., East, A. L. L., Marx, D., Beyer, M. K. & Clausen-Schaumann, H. Mechanical Activation Drastically Accelerates Amide Bond Hydrolysis, Matching Enzyme Activity. *Angewandte Chemie International Edition* **58**, 9787–9790 (2019).
21. Hockney, R. W., Goel, S. P. & Eastwood, J. W. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics* **14**, 148–158 (1974).
22. Boltzmann, L. *Vorlesungen über Gastheorie. J.A. Barth.* (J.A. Barth, Leipzig, 1896).
23. Taylor, B. *Methodus incrementorum directa & inversa. Auctore Brook Taylor ... Londini : Typis Pearsonianis prostant apud Gul. Innys ad Insignia Principis in Coemeterio Paulino, 1715.* (Londini : Typis Pearsonianis prostant apud Gul. Innys ad Insignia Principis in Coemeterio Paulino, 1715, 1715).
24. Morse, P. M. Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Physical Review* **34**, 57–64 (1929).
25. Mémoire sur la propagation de la chaleur dans les corps solides. in *Cambridge University Press.* (eds. Fourier, J. B. J. & Darboux, J. G.) vol. 2 213–222 (Cambridge University Press, Cambridge, 2013 (1890)).
26. Lennard-Jones, J. E. Cohesion. *Proceedings of the Physical Society* **43**, 461 (1931).
27. Coulomb, A. First memoir on electricity and magnetism. *A Source Book in Physics* 408–413 (1785).
28. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *American Chemical Society.* (2002) doi:10.1021/j100142a004.
29. Case, D. A., Aktulga, H. M., Belfon, K., *et al.* AmberTools. *Journal of Chemical Information and Modeling* **63**, 6183–6191 (2023).

30. Cornell, W. D., Cieplak, P., Bayly, C. I., *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179–5197 (1995).

31. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049–1074 (2000).

32. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95–99 (1963).

33. Best, R. B. & Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix—Coil Transition of Polypeptides. *The Journal of Physical Chemistry B* **113**, 9004–9015 (2009).

34. Lindahl, Abraham, Hess & van der Spoel. GROMACS 2020 Source code. *Zenodo.* (2020) doi:10.5281/zenodo.3562495.

35. Aliev, A. E., Kulke, M., Khaneja, H. S., *et al.* Motional timescale predictions by molecular dynamics simulations: Case study using proline and hydroxyproline sidechain dynamics: Proline Force Field Parameters. *Proteins: Structure, Function, and Bioinformatics* **82**, 195–215 (2014).

36. Brooks, B. R., Brooks, C. L., Mackerell, A. D., *et al.* CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **30**, 1545–1614 (2009).

37. Schuler, L. D., Daura, X. & van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry* **22**, 1205–1218 (2001).

38. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation* **11**, 3499–3509 (2015).

39. Sorin, E. J. & Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophysical Journal* **88**, 2472–2493 (2005).

40. Abraham, M. J., Murtola, T., Schulz, R., *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

41. Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. (1881) doi:10.1002/andp.18812480110.

42. Berthelot, D. Sur le mélange des gaz. *Compt. Rendus* **126**, 15 (1898).

43. Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. The Protein Data Bank archive as an open data resource. *Journal of Computer-Aided Molecular Design* **28**, 1009–1014 (2014).

44. File formats - GROMACS 2025.3 documentation.

45. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *Journal of Computational Physics* **23**, 327–341 (1977).

46. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**, 952–962 (1992).

47. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472 (1997).

48. Jumper, J., Evans, R., Pritzel, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

49. Boltzmann, L. E. *Einige Allgemeine Sätze Über Wärmegleichgewicht. K. Akad. der Wissensch.* (K. Akad. der Wissensch., 1871).

50. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).

51. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).

52. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182–7190 (1981).

53. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *Journal of chemical physics* **98**, 10089–10089 (1993).

54. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369**, 253–287 (1921).

55. Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **159**, 98–103 (1967).

56. Páll, S. & Hess, B. A flexible algorithm for calculating pair interactions on SIMD architectures. *Computer Physics Communications* **184**, 2641–2650 (2013).

57. GROMACS 2025.3 documentation.

58. Grossfield, A. WHAM: The weighted histogram analysis method.

59. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv.* (2022) doi:10.48550/arXiv.2112.10752.

60. Invoke-ai/InvokeAI. *InvokeAI.* (2025).

61. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011–1021 (1992).

62. Souaille, M. & Roux, B. Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations. *Computer Physics Communications* **135**, 40–57 (2001).

63. Hub, J. S., de Groot, B. L. & van der Spoel, D. G_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *Journal of Chemical Theory and Computation* **6**, 3713–3720 (2010).

64. Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Physical Review* **28**, 1049–1070 (1926).
65. De Broglie, L. Recherches sur la théorie des quanta. *Migration-université en cours d'affectation.* (Migration-université en cours d'affectation, 1924).
66. Born, M. Zur Quantenmechanik der Stoßvorgänge. *Zeitschrift für Physik* **37**, 863–867 (1926).
67. Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **389**, 457–484 (1927).
68. Hartree, D. R. *The Calculation of Atomic Structures. J. Wiley.* (J. Wiley, 1957).
69. Slater, J. C. The Theory of Complex Spectra. *Physical Review* **34**, 1293–1322 (1929).
70. Fock, V. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik* **61**, 126–148 (1930).
71. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Physical Review* **136**, B864–B871 (1964).
72. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **140**, A1133–A1138 (1965).
73. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **37**, 785–789 (1988).
74. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98**, 5648–5652 (1993).
75. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **105**, 9982–9985 (1996).
76. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **110**, 6158–6170 (1999).
77. Head-Gordon, M., Pople, J. A. & Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical Physics Letters* **153**, 503–506 (1988).
78. Møller, Chr. & Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **46**, 618–622 (1934).
79. Laino, T., Mohamed, F., Laio, A. & Parrinello, M. An Efficient Real Space Multigrid QM/MM Electrostatic Coupling. *Journal of Chemical Theory and Computation* **1**, 1176–1184 (2005).
80. Morozov, D. Webinar: Multiscale QM/MM simulations: Exploring chemical reactions using novel GROMACS/CP2K interface (2020-12-08). (2020).
81. Hybrid Quantum-Classical simulations (QM/MM) with CP2K interface - GROMACS 2025.3 documentation.
82. Kühne, T. D., Iannuzzi, M., Del Ben, M., *et al.* CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **152**, 194103 (2020).

83. Bortz, A. B., Kalos, M. H. & Lebowitz, J. L. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics* **17**, 10–18 (1975).

84. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22**, 403–434 (1976).

85. Henkelman, G. & Jónsson, H. Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table. *The Journal of Chemical Physics* **115**, 9657–9666 (2001).

86. Arrhenius, S. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für Physikalische Chemie* **4U**, 226–248 (1889).

87. Eyring, H. The Activated Complex in Chemical Reactions. *The Journal of Chemical Physics* **3**, 107–115 (1935).

88. Liang, Y., Zheng, W., Xie, H., Zha, X. & Wang, T. A quantum chemistry study on C–H homolytic bond dissociation enthalpies of five-membered and six-membered heterocyclic compounds. *Journal of the Indian Chemical Society* **99**, 100527 (2022).

89. Kosar, N., Ayub, K., Gilani, M. A. & Mahmood, T. Benchmark DFT studies on C–CN homolytic cleavage and screening the substitution effect on bond dissociation energy. *Journal of Molecular Modeling* **25**, 47 (2019).

90. John, P. C. S., Guan, Y., Kim, Y., Kim, S. & Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nature Communications* **11**, 2328 (2020).

91. VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *The Journal of Chemical Physics* **127**, 114105 (2007).

92. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865–3868 (1996).

93. Obarska-Kosinska, A., Rennekamp, B., Ünal, A. & Gräter, F. ColBuilder: A server to build collagen fibril models. *Biophysical journal* (2021) doi:10.1016/j.bpj.2021.07.009.

94. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. (2015).

95. Jette, M. A. & Wickberg, T. Architecture of the Slurm Workload Manager. in *Springer Nature Switzerland.* (eds. Klusáček, D., Corbalán, J. & Rodrigo, G. P.) 3–23 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43943-8_1.

96. Bürgi, H. B., Dunitz, J. D., Lehn, J. M. & Wipff, G. Stereochemistry of reaction paths at carbonyl centres. *Tetrahedron* **30**, 1563–1572 (1974).

97. Lodge, E. P. & Heathcock, C. H. Steric effects, as well as *-orbital energies, are important in diastereoface differentiation in additions to chiral aldehydes. *Steric effects, as well as *-orbital energies, are important in diastereoface differentiation in additions to chiral aldehydes* **109**, 3353–3361 (1987).

98. Flippin, L. A. & Heathcock, C. H. ACYCLIC STEREOSELECTION. XVI: HIGH DIASTEREOFACIAL SELECTIVITY IN LEWIS ACID MEDIATED ADDITIONS OF ENOLSILANES TO CHIRAL ALDEHYDES. *ACYCLIC STEREOSELECTION. XVI: HIGH DIASTEREOFACIAL SELECTIVITY IN LEWIS ACID MEDIATED ADDITIONS OF ENOLSILANES TO CHIRAL ALDEHYDES* (1983).

99. Buhr*, J., Hartmann*, E., Riedmiller*, K., *et al.* KIMMDY: A biomolecular reaction emulator. *bioRxiv.* (2025) doi:10.1101/2025.07.02.662624 *These authors contributed equally to this work.

100. Rennekamp, B., Kutzki, F., Obarska-Kosinska, A., Zapp, C. & Gräter, F. Hybrid Kinetic Monte Carlo/Molecular Dynamics Simulations of Bond Scissions in Proteins. *Journal of Chemical Theory and Computation* **16**, 553–563 (2020).

101. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A* **105**, 9396–9409 (2001).

102. Senftle, T. P., Hong, S., Islam, M. M., *et al.* The ReaxFF reactive force-field: Development, applications and future directions. *npj Computational Materials* **2**, 1–14 (2016).

103. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148**, 241722 (2018).

104. Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **98**, 146401 (2007).

105. Smith, J., Isayev, O. & E. Roitberg, A. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **8**, 3192–3203 (2017).

106. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. in *Curran Associates, Inc.* (eds. Koyejo, S., Mohamed, S., Agarwal, A., et al.) vol. 35 11423–11436 (Curran Associates, Inc., 2022).

107. Musaelian, A., Batzner, S., Johansson, A., *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications* **14**, 579 (2023).

108. Mazitov, A., Bigi, F., Kellner, M., *et al.* PET-MAD, a universal interatomic potential for advanced materials modeling. *arXiv.* (2025) doi:10.48550/arXiv.2503.14118.

109. Procida, D. Diátaxis documentation framework.

110. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, 43–56 (1995).

111. Gowers, R. J., Linke, M., Barnoud, J., *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *scipy* (2016) doi:10.25080/Majora-629e541a-00e.

112. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32**, 2319–2327 (2011).

113. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **185**, 604–613 (2014).

114. Voliotis, M., Thomas, P., Grima, R. & Bowsher, C. G. Stochastic Simulation of Biomolecular Networks in Dynamic Environments. *PLOS Computational Biology* **12**, e1004923 (2016).

115. Seute, L., Hartmann, E., Stühmer, J. & Gräter, F. Grappa – a machine learned molecular mechanics force field. *Chemical Science* **16**, 2907–2930 (2025).

116. Conventional Commits.

117. Bell, G. I. Models for the Specific Adhesion of Cells to Cells. *Science* **200**, 618–627 (1978).

118. Evans, E. & Ritchie, K. Dynamic strength of molecular adhesion bonds. *Biophysical Journal* **72**, 1541–1555 (1997).

119. Lee, B. & Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology* **55**, 379–400 (1971).

120. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research.* (2016) doi:10.12688/f1000research.7931.1.

121. Street, D., Bangsbo, J. & Juel, C. Interstitial pH in human skeletal muscle during and after dynamic graded exercise. *The Journal of Physiology* **537**, 993 (2001).

122. Rawlings, N. D., Barrett, A. J., Thomas, P. D., *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* **46**, D624–D632 (2018).

123. Aho, N., Buslaev, P., Jansen, A., *et al.* Scalable Constant pH Molecular Dynamics in GROMACS. *Journal of Chemical Theory and Computation* **18**, 6148–6160 (2022).

124. Riedmiller, K., Reiser, P., Bobkova, E., *et al.* Substituting density functional theory in reaction barrier calculations for hydrogen atom transfer in proteins. *Chemical Science* **15**, 2518–2527 (2024).

125. Treyde, W., Riedmiller, K. & Gräter, F. Bond dissociation energies of X–H bonds in proteins. *RSC Advances* **12**, 34557–34564 (2022).

126. Rennekamp, B., Karfusehr, C., Kurth, M., *et al.* Collagen breaks at weak sacrificial bonds taming its mechanoradicals. *Nature Communications* **14**, 2075 (2023).

127. Souza, P. C. T., Alessandri, R., Barnoud, J., *et al.* Martini 3: A general purpose force field for coarse-grained molecular dynamics. *Nature Methods* **18**, 382–388 (2021).

128. Hrennikoff, A. Solution of Problems of Elasticity by the Framework Method. *Journal of Applied Mechanics* **8**, A169–A175 (2021).

129. Courant, R. Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society* **49**, 1–23 (1943).

130. Freutel, M., Schmidt, H., Dürselen, L., Ignatius, A. & Galbusera, F. Finite element modeling of soft tissues: Material models, tissue interaction and challenges. *Clinical Biomechanics* **29**, 363–372 (2014).

131. Chen, X., Zhong, Z., Xu, Z., Chen, L. & Wang, Y. 2 ,7 - Dichlorodihydrofluorescein as a fluorescent probe for reactive oxygen species measurement: Forty years of application and controversy. *Free Radical Research* **44**, 587–604 (2010).

132. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

133. R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

134. Wickham, H., Averick, M., Bryan, J., *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

135. team, T. pandas development. Pandas-dev/pandas: Pandas. *Zenodo.* (2025) doi:10.5281/zenodo.17229934.

136. McKinney, W. Data Structures for Statistical Computing in Python. in (eds. van der Walt, S. & Millman, J.) 56–61 (2010). doi:10.25080/Majora-92bf1922-00a.

137. Harris, C. R., Millman, K. J., van der Walt, S. J., *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

138. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

139. team, T. plotnine development. Plotnine: A grammar of graphics for Python. doi:10.5281/zenodo.1325308.

140. Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y. & Dervieux, C. Quarto. (2022) doi:10.5281/zenodo.5960048.

141. Sehnal, D., Bittrich, S., Deshpande, M., *et al.* Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research* **49**, (2021).