Dissertation
zur Erlangung der Doktorwürde
an der

Gesamtfakultät für Mathematik,
Ingenieur- und Naturwissenschaften
der
Ruprecht-Karls-Universität Heidelberg

Thema:

Developing a Reactive Molecular Simulation Method for Protein Mechanochemistry

vorgelegt von:

Eric Hans Hartmann

Gutachter: Professor Dr. Frauke Gräter
Professor Dr. Carl Herrmann

Zusammenfassung

Kollagen bildet bei mechanischer Beanspruchung Radikale, gefolgt von einer Reduktion der DOPA-Radikale durch Wasserstoffperoxid. Es ist jedoch nach wie vor unklar, wie die Mechanoradikale zu DOPA gelangen. In der komplexen Umgebung der Kollagenfibrille ist eine große Anzahl verschiedener Reaktionen möglich, darunter auch der Wasserstoffatomtransfer (HAT).

In dieser Arbeit wird eine kombinierte Molekulardynamik- (MD) und kinetische Monte-Carlo-Methode (KMC) als adaptiven KMC Ansatz implementiert. Die entwickelte Software, KIMMDY, ist in der Lage, Trajektorien reaktiver Systeme in kondensierter Phase mit langer Zeitskala zu simulieren. Mehr als 600 HAT-Reaktionen werden in einem Kollagenfibrillenmodell mit bis zu 20 aufeinanderfolgenden Reaktionen simuliert.

Um MD-Simulationen von Aminosäureradikalen zu ermöglichen, wird ein klassisches Kraftfeld mit Hilfe der Grappa-Methode auf QM-Energien und -Kräfte trainiert. Darüber hinaus wird ein neuronales Graphen-Netzwerk angepasst, um HAT-Raten für Ensembles vorherzusagen, welche aus MD-Simulationen in einem neuartigen *transition path sampling* Ansatz erzeugt wurden.

Diese drei methodischen Fortschritte ermöglichen die Anwendung reaktiver Simulationen in der Kollagenfibrille zur Beobachtung der HAT *pathways* vom Ort der Homolyse zum posttranslationalen Oxidationsprodukt von Phenylalanin und Tyrosin, DOPA. Das DOPA-Radikal wird in Simulationen Reaktionsprodukt und die kinetischen Eigenschaften bestätigen die Rolle als *radical scavenger*. Ein weiterer *radical scavenger*, Pyridinolin (PYD), wird identifiziert und seine mechanochemischen Eigenschaften charakterisiert. KIMMDY bietet eine neue Perspektive auf radikalische Reaktionen in Kollagen und ist so konzipiert, dass es auf neuartige Moleküle und Reaktionen angewendet werden kann.

Abstract

Collagen has recently been found to form radicals when subjected to mechanical stress, followed by detoxification of DOPA radicals via hydrogen peroxide. However, it remains unclear how mechanoradicals reach DOPA. Within the complex environment of the collagen fibril, numerous different reactions are possible, including hydrogen atom transfer (HAT).

In this work a combined molecular dynamics (MD) and kinetic Monte Carlo (KMC) method is implemented within an adaptive KMC framework. The developed software, KIMMDY, is capable of simulating long timescale trajectories of reactive condensed phase systems. More than 600 HAT reactions are simulated in a collagen fibril model with up to 20 consecutive reactions.

To make MD simulations of amino acid radicals possible, a highly accurate classical force field is trained on QM energies and forces using the Grappa method. Furthermore, a graph neural network is adapted to predict HAT rates for ensembles generated from MD simulations in a novel approach to transition path sampling.

These three methodological advances facilitate the application of reactive simulations in the collagen fibril to observe HAT pathways from the homolysis site to the post-translational oxidation product of phenylalanine and tyrosine, DOPA. The DOPA radical can be observed in simulations and kinetic properties confirm the radical scavenger role. Another radical scavenger, pyridinoline (PYD) is proposed and its mechanochemical properties characterised. KIMMDY provides a new perspective on radical reactions in collagen and is designed to be applied to novel molecules and reactions.

Acknowledgements

I would like to thank Frauke for her guidance and the opportunity to develop my own ideas and follow through with them during the work on this dissertation.

I would like to thank Jannik and Kai with whom I spent many hours working on KIMMDY and learning about software development.

I would like to thank Leif and Eddie, who were my students and helped to realise my ideas while making their projects their own.

I would like to thank Benedikt for setting the foundation of my work and introducing me to collagen and KMC/MD simulations.

I would like to thank Evgeni, Denis, Johanna and Daniel for bringing new perspectives to the KIMMDY project and pushing it forward.

I would like to thank the whole research group for providing a helpful and open environment.

I would like to thank the people who make HITS a place I look forward to visit every day.

I would like to thank my thesis committee, reviewers, people who organise conferences, workshops and summer schools who contribute to the special character of the scientific community.

I would like to thank my family and friends for their love and support.

Contents

List of Figures

List of Abbreviations and Acronyms

AM1-BCC	AM1 Bond Charge Correction, method for partial charge derivation based on the semiempirical method AM1
CHARMM	Chemistry at HARvard Macromolecular Mechanics, simulation software and force field family
DFT	Density Functional Theory
DOPA	Dihydroxyphenylalanine
Grappa	Graph Attentional Protein Parametrisation, method to predict MM parameters
GROMACS	Groningen Machine for Chemical Simulations, MD simulation software
HAT	Hydrogen Atom Transfer
HLKNL	Hydroxylysino-Ketonorleucine
KIMMDY	Kinetic Monte Carlo Molecular Dynamics, combined KMC-MD simulation method
KMC	Kinetic Monte Carlo
LY2, LY3, L4Y, L5Y, LYX	Cross-link amino acid segments in force field definition
MAE	Mean Average Error
MSE	Mean Squared Error
MD	Molecular Dynamics
MM	Molecular Mechanics
ML	Machine Learning
MLFF	Machine Learning Force Field
MLP	Multilayer Perceptron
GNN	Graph Neural Network
PES	Potential Energy Surface
PYD	Hydroxylysyl Pyridinoline
QM	Quantum Mechanics
RESP	Restrained Electrostatic Potential, method for partial charge derivation
RMSE	Root Mean Squared Error
TST	Transition State Theory

Chapter 1

Introduction

1.1 Molecular simulations may aid in understanding protein ageing

Proteins are frequently subjected to chemical modifications after biosynthesis.[1] Some modifications are performed at specifically targeted sites by enzymes, contributing to diverse protein functions[2, 3], while non-enzymatic modifications are of a more stochastic nature[4, 5].

Because these unspecific modifications are difficult to control, they have solely been thought of as damage[4] but there is a delicate balance between damage and regulatory signal[6]. However, for long-lived proteins like lens crystallin or collagen, these modifications can accumulate over time and impair protein function.[7, 8, 9, 10] Some long-lived proteins are load-bearing, which is known to lead to polymer ageing[11], an insight that recently led to the investigation of the structural protein collagen in this regard[12, 13, 14]

Due to the stochastic nature of reactions in protein ageing, diverse reaction products are formed, which are hard to investigate by bulk methods. Molecular simulations are "computational microscopes"[15] and could be an ideal tool to study reactive pathways in single molecules. However, this is complicated by the fact that force fields used for biomolecular simulations assume molecules to be chemically inert to improve simulation speed in order to reach biologically relevant time scales.[16] This shortcoming can be circumvented, paving the way for molecular simulations of protein ageing processes.

1.2 Several chemical modifications occur in collagen

Collagen is a prime candidate for the study of protein ageing because it is subjected to mechanical stress and long-lived. A high stress experiment is a well defined scenario with a high potential for ageing events to occur and can be modelled with constant force molecular dynamics simulations. Furthermore, tissue such as cartilage, tendons and ligaments has a high concentration of collagen, which facilitates experimental validation of computational results.

The hierarchical structure of collagen spans several orders of magnitude.

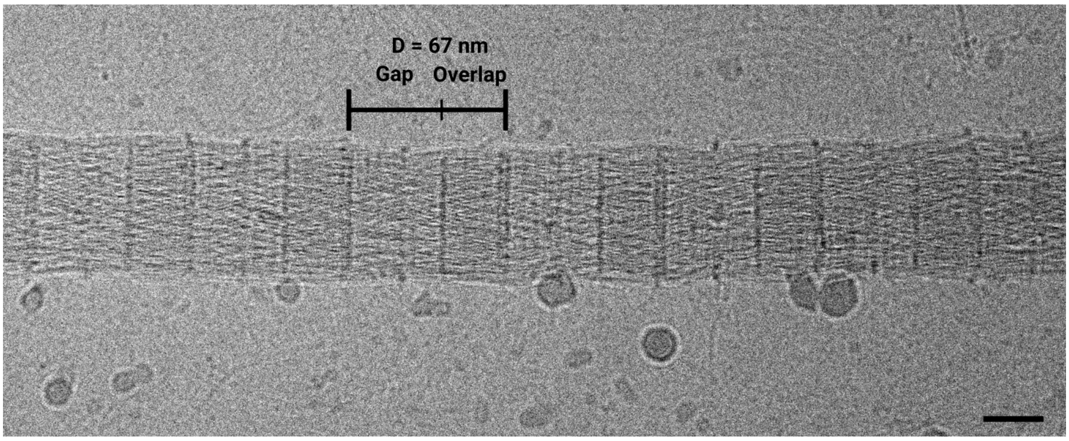


Figure 1.1: **Collagen fibrils have a periodic pattern.** Cryo EM micrograph of a *rattus norvegicus* tendon sample. The scale bar has a length of 25 nm. Figure provided by Aysecan Ünal.

[17] Collagen-rich tissue is made up of collagen fibres with a length in the millimetre range. Proteoglycans act as interfibrillar connections between fibrils on the μm scale.[18] Fibrils (Fig. 1.1) are built out of cross-linked triple helices and also stabilised by non-covalent interactions. On the smallest scale, the triple helix is a unique quaternary structure of individual collagen strands with a hydrogen bonding pattern that is facilitated by a repeating Gly-Pro-Xaa or Gly-Xaa-Hyp sequence.[19] Here, Xaa refers to any amino acid.

Hydroxyproline (Hyp) is a post-translational modification that stabilises triple helices because of a preference for a certain proline ring conformation.[19]. In the collagen model studied here, it is the fourth most common amino acid (Fig. 1.2).

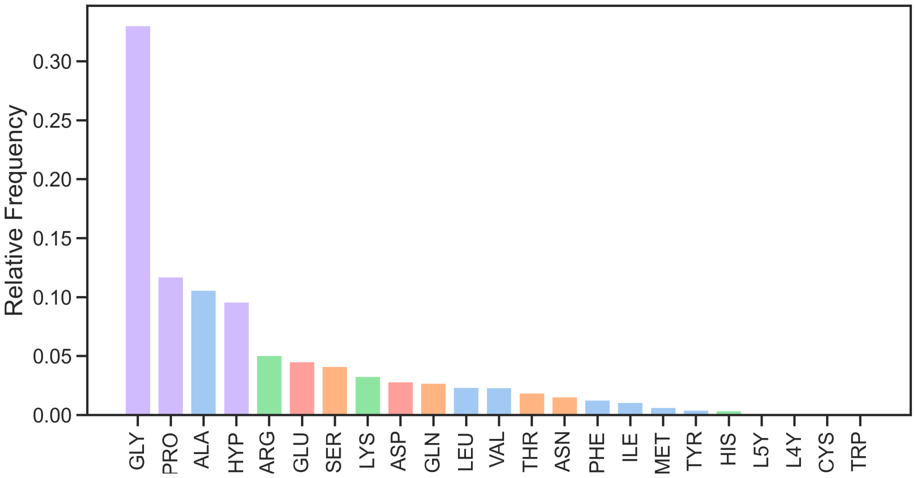


Figure 1.2: **Frequency of amino acids in collagen.** The amino acids frequency is analysed for a triple helix in collagen type I of *rattus norvegicus*. The color scheme is chosen according to physico-chemical properties of the amino acids. L5Y and L4Y are cross-link amino acids and constitute HLKNL.

Another set post-translational modifications of collagen are the aforementioned cross-links. Most cross-links are derived from lysine but cross-links involving other amino acids and non-protein components exist as well.[20, 21, 22] Among the lysine derived cross-links, a main distinction is drawn between divalent and "mature" trivalent cross-links. Interestingly, lysine is enzymatically modified in the endoplasmatic reticulum by the lysyl hydroxylase and in the extracellular matrix by the lysyl oxidase but these reactions are only performed to create the reactants for the cross-linking reaction.[23] Cross-linking occurs spontaneously with a range of possible mechanisms, especially for trivalent cross-link formation (Fig. 1.3a,b). [24]

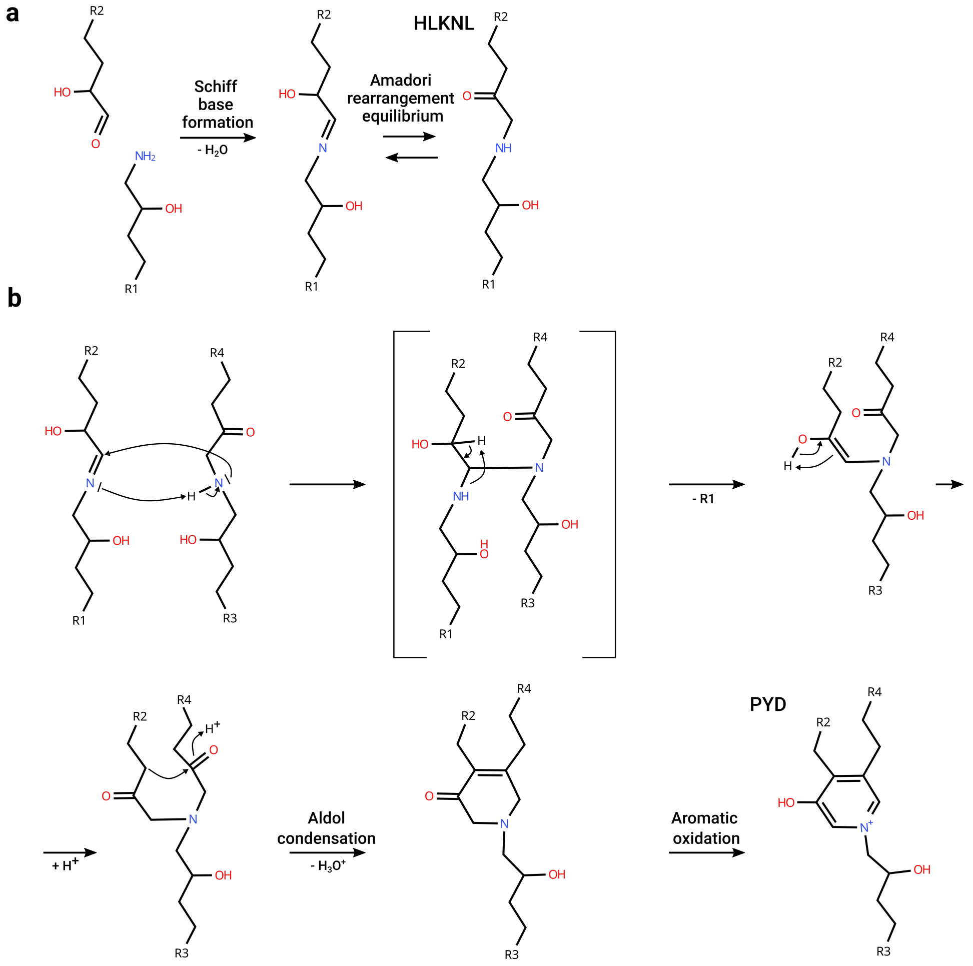


Figure 1.3: **Cross-link formation is spontaneous.** **a**, Mechanism of divalent cross-link formation (here HLKNL) from allysine. **b**, Mechanism of trivalent cross-link formation from two divalent cross-links. Multiple reaction mechanisms for trivalent cross-link formation have been proposed, one of which is this variation.[24]

DOPA is an oxidation product of tyrosine and phenylalanine in collagen (Fig. 1.4). This oxidation can occur spontaneously by radical reactions or enzymatically in the free or protein-bound tyrosine and phenylalanine.[25, 26] A free L-DOPA can be incorporated into proteins during biosynthesis. In recent studies, DOPA has been proposed as radical scavenger for the detoxification of mechanoradicals.[12, 13, 14]. The role of DOPA in radical mechanisms seems to be a more general pattern as it has also been observed in a ribonucleotide reductase.[27]

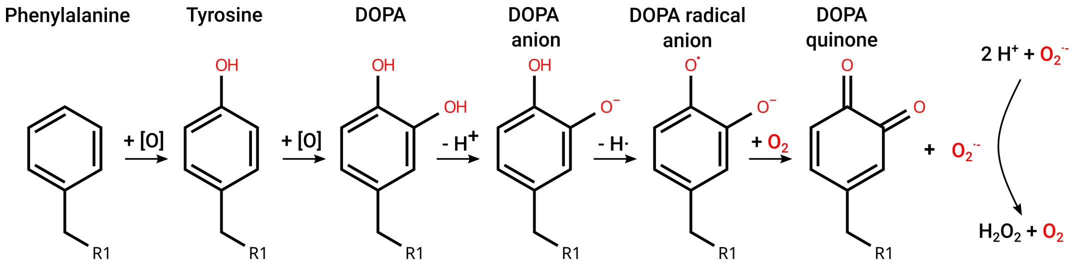


Figure 1.4: **Oxidation of aromatic amino acids.** Oxidation of phenylalanine and tyrosine leads to DOPA formation. Further reactions can include deprotonation and HAT to a DOPA radical anion that reacts with molecular oxygen to superoxide with an eventual hydrogen peroxide follow-up reaction.

Taken together, the family of collagen proteins has a diverse post-translational chemistry. Several unusual modifications are known to occur in collagen, some of which accumulate over the long lifespan of collagens.

1.3 Collagen radicals are scavenged by DOPA

Many cellular processes involve redox reactions, requiring a balance between oxidising and reducing agents. Both ends of the redox spectrum are detrimental with a physiological state termed ”redox eustress” in-between.[28] To regulate the redox status of a cell, feedback loops are used.[28] Diverse sources of oxidative stress, including endogenous stress like aerobic cell metabolism or exogenous stress such as UV irradiation, can produce reactive species. These species function as redox signals that are sensed, often including reversible thiol redox switches. As a consequence, gene expression is modulated and cell stress proteins expressed.

Interestingly, there are no cysteines in mammalian tropocollagen, which precludes the existence of any thiol-based redox sensors.[29] This absence is likely due to an incompatibility of disulfide bonds with the triple helix structure, as cysteine mutations are associated with collagen diseases.[30, 31]

Mechanoradicals have recently been discovered in collagen under tension through DOPA radical measurements,[12] but whether the quantities released under these conditions are physiologically relevant remained unclear. The existence of a non-thiole radical sensing system in collagen would point to a certain significance from an evolutionary standpoint. Because collagen is a long-lived protein, even slow accumulation of damage would be detrimental and could justify a cellular response. In a follow-up study, Rennekamp *et al.*[14] identified specific homolysis sites, especially the Cα-Cβ bond of the short arm of the PYD cross-link (R2 in fig. 1.3). The radical scavenging activity of DOPA in collagen was also confirmed and the occurrence of DOPA shown to be at phenylalanine and tyrosine positions.[13] Detoxification of DOPA radicals can occur via hydrogen peroxide (Fig. 1.4).

At this point, the radical initiation step has been established to be homolytic cleavage at cross-link bonds and the termination, at least within the context of collagen, occurs at DOPA. However, the propagation steps that lead to radical migration from the homolysis site to DOPA are not clear. Protein radicals can undergo a range of reactions, including hydrogen atom transfer (HAT)[32], dimerisation, peroxidation, decarboxylation, β-scission or deprotonation.[33, 34]. Out of these, only HATs could lead to a DOPA radical. While dimerisation and peroxidation reactions generally seem more favourable than HAT[34], a limited solvent accessibility and low radical species concentration in collagen makes HAT a plausible alternative. HATs have been observed in free amino acids with rates in the order of $10^6\,\mathrm{s}^{-1}$[33] but steric effects in a folded protein may influence the reaction rate. Hence, a mechanism that includes HATs to DOPA after homolytic cleavage at a cross-link seems plausible (Fig. 1.5).

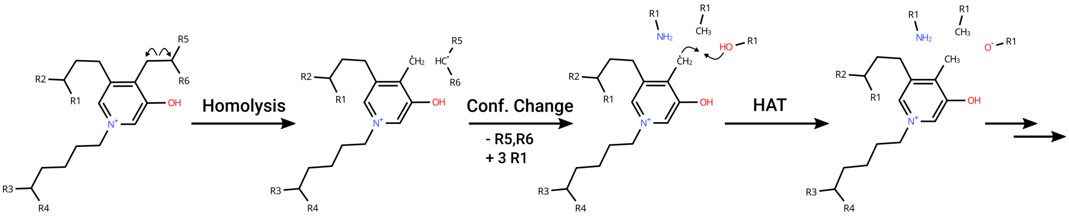


Figure 1.5: **Homolysis is followed by HATs.** Homolytic cleavage occurs at the Cα-Cβ bond of the short arm of a PYD cross-link. Conformational change, in part due to the cleavage, leads to a different local environment of the radicals. HAT reactions lead to different radical species, potentially including tyrosine or DOPA. This mechanism was first proposed by Zapp *et al.*[12]

Without nearby thioles, DOPA seems to be the best radical scavenger available[35] in collagen. It has also been noted that aromatic residues are enriched in the vicinity of cross-links[12], which follows the principle of co-localisation of reactive species sources and targets[6]. A remaining question is whether DOPA is not only thermodynamically the most stable radical in collagen but also kinetically accessible after homolysis. An indication for this would be if a direct HAT from the homolysis site to DOPA is possible. Also, for determining the role of tyrosine as alternative redox sensor, a tyrosine reducing system would also need to be identified, as irreversible modifications are more indicative of molecular damage than of a sensing system.[28]

Apart from the lack of thioles, collagen also produces radicals in the extracellular matrix, which is an untypical location.[6] However, this is not problematic for radical sensing and a response to oxidative stress because the signalling molecule hydrogen peroxide can diffuse from the extracellular matrix into cells via aquaporins or interact with receptors[28] and there are extracellular cell stress proteins [36].

1.4 Reactive simulation methods have a diverse scope

Using computational methods, the radical propagation mechanism can be further investigated. A suitable model should satisfy two conditions: First, the protein environment of radicals should be taken into consideration. HAT rates from free amino acids cannot be expected to apply to protein radicals because the conformational flexibility in proteins is limited. Second, protein HATs occur roughly in the μs range[33], which must be attainable by the computational method.

The second condition is problematic for most reactive atomistic simulation methods. QM/MM simulations[37] have the additional limitation that only a small region is reactive, ReaxFF[38] would need to be parametrised for the target application and EVB[39] simulations are quite fast but difficult to parametrise and scales dependent on the number of possible reactions in a system. Machine learning force fields (MLFFs) promise to emulate QM calculations at a much higher simulation speed. Recently, molecular systems of a size unattainable for QM calculations have been simulated with MLFFs[40]. However, current MLFFs are still 50 - 100 times slower than classical MM force fields, especially when considering long-range interactions.[41] Classical MM force fields are non-reactive but simulation times in the μs to ms range can be reached.

For simulating reactions, it is not strictly necessary to simulate the movement of atoms. In fact, if the likelihood of reactions within a time interval is known, the time evolution of conformations need not be simulated at all. Kinetic Monte Carlo methods propagate reactive systems over time and time steps depend on the fastest reactions in a given system.[42, 43] Reaction rates can be informed by molecular structures, thus both conditions can be satisfied. This is typically done by QM-based transition state search[44] but can also be emulated by cheaper methods, including graph neural networks (GNNs)[45]. In practice, commonly used KMC implementations require a fixed list of possible events with associated rates.[44, 46] This would mean an exhaustive list of every possible HAT reaction in a collagen model is required before starting a KMC simulation, posing a strong limitation.

In adaptive kinetic Monte Carlo, the event/reaction list is constructed anew for every KMC step.[47] It follows that only reaction rates for reactions starting from populated states need to be calculated, greatly reducing the number of rate calculations for large systems. So far, the event list has been constructed using QM-based transition state search[47, 48], high-temperature MD for non-reactive transitions[49] and heuristic models based on observables of structures generated with MD simulations[50, 51, 52, 53, 54].

The idea to sample collagen dynamics in the μs timescale with classical MM force fields and using a heuristic to determine HAT rates seems promising. Effectively, the current reactive state is sampled and transitions to neighbouring states can be found from the conformational ensemble. In contrast to a simple heuristic model, one with quantitative interpretability would allow for valuable insights from reactive simulations. To that end, a HAT GNN method has recently been developed that emulates DFT calculations.[45]

Chapter 2

Research Aim

The aim of this work is to gain insights on the radical migration from collagen homolysis sites to DOPA sites using reactive molecular simulations. No current simulation method is capable of performing these simulations, so a method is developed that combines KMC and MD simulations with a machine learning (ML) model for HAT rate prediction (Fig. 2.1). Apart from this method, further method development is also needed for MD simulations of protein radicals and HAT rate prediction.

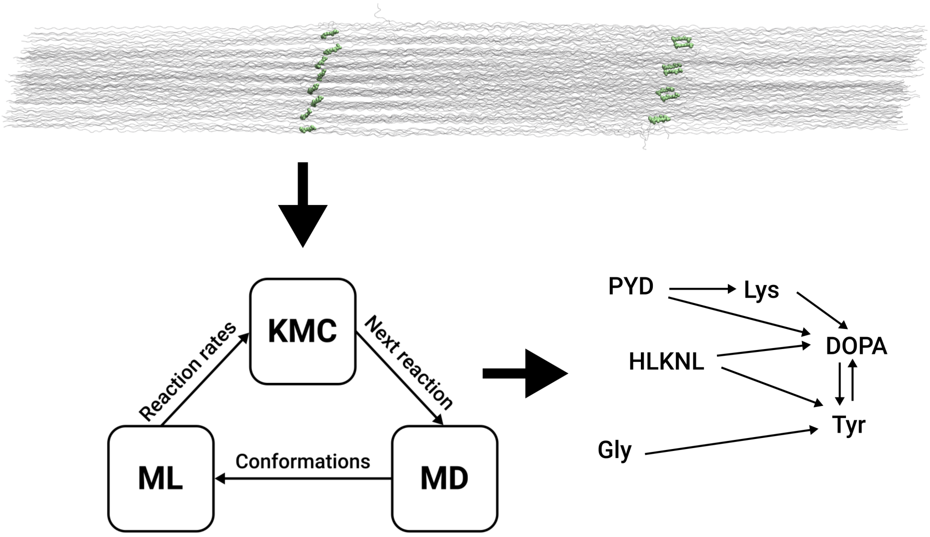


Figure 2.1: **A reactive molecular simulation method.** A collagen fibril model is simulated with a novel method. This method combines kinetic Monte Carlo and molecular dynamics simulations with a machine learning model. MD simulations generate a conformational ensemble, the ML method predicts reaction rates based on the ensemble and a KMC algorithm is used to choose the next reaction. This triad can run in a loop. From this, a radical migration network can be constructed.

To generate an ensemble of collagen structures after homolytic cleavage using MD simulations, a force field for protein radicals is built. A novel ML method for predicting classical MD parameters of arbitrary molecular systems is trained on radical peptides to obtain the required parameters.

The aforementioned HAT GNN has not been applied to predict HAT rates from MD simulations before. For this use case, the HAT GNN method is adapted and conditions for robustly predicting HAT rates from MD simulations are investigated.

The reactive molecular simulation method is implemented and generalised to facilitate the simulation of a range of reactions. The method is named after a previous MD/KMC method[55], KIMMDY, which stands for Kinetic Monte Carlo Molecular DYnamics.

Finally, the newly developed simulation method is applied to simulations of a collagen fibril. The role of DOPA as a radical scavenger is investigated from a kinetics standpoint and further radical scavengers are proposed.

Chapter 3

Theory and Methods

3.1 Molecular dynamics

Algorithm

GROMACS[56] is used as MD engine and algorithmic details follow the GROMACS documentation[57]. MD simulations are used to propagate the position of atoms over time using Newton's equations of motion,

$$\frac{d^2\boldsymbol{r}_i}{dt^2} = \frac{\boldsymbol{F}_i}{m_i}, \tag{3.1}$$

for an atom i with a force F, mass m and position r. Thus, to propagate the system over time, initial coordinates, velocities and forces acting on atoms according to a potential V,

$$\boldsymbol{F}_i = -\frac{\delta V}{\delta \boldsymbol{r}_i}, \tag{3.2}$$

are necessary. Initial coordinates need to be supplied by the user and velocities are generated for every atom along every dimension j according to the Boltzmann distribution at a certain temperature using a standard normally distributed random variable s,

$$v_{i,j} = s\sqrt{\frac{k_bT}{m_i}}. \tag{3.3}$$

The potential acting on the atom is described by the force field.

Using the leap-frog algorithm[58], velocity and position are updated in alternating steps.

$$\begin{aligned} \boldsymbol{v}(t+\frac{1}{2}\Delta t) &= \boldsymbol{v}(t-\frac{1}{2}\Delta t) + \frac{\Delta t}{m}\boldsymbol{F}(t) \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \Delta t\, \boldsymbol{v}(t+\frac{1}{2}\Delta t) \end{aligned} \tag{3.4}$$

The algorithm up to now would sample states in the NVE ensemble. For most applications, the use of a thermostat and barostat is sensible to sample from the NPT ensemble. Several options exist for both. The Berendsen thermostat[59] and related methods[60] couple the molecular system to an external heat bath. Effectively, the kinetic energy of the molecular system is scaled to correct the current temperature in the direction of the given temperature. This leads to a scaling of the velocities and in consequence the atom positions. A barostat scales the box size and the atom coordinates in it to correct the pressure.

The size of a MD time step is determined by the fastest vibration. Typically, that would be bond-vibrations but these are accurately modelled by holonomic constraints.[61] This allows for using a time step of 2 fs, significantly increasing the accessible time scales compared to 0,5 fs or 1 fs time steps. The LINCS constraints are another modification to the coordinates and velocities of a system.

External forces can be added to added to atoms to mimic interactions from outside the model. A simple example is constant force pulling, where the force is applied to the center of mass of two groups of atoms.

Force field

The potential V determines a molecule's dynamics and is defined by a force field. Different functional forms have been applied in the past. [62, 63, 64, 65, 66] For simulations of biomolecules, class I force fields[67] with additive terms for bonded and non-bonded interactionsare most commonly used.[68] The most popular force field families for biomolecular simulations are AMBER[69] and CHARMM[70]. The force field details below mostly apply for both families but are centred around AMBER force fields.

$$\begin{aligned} V_{\text{pot}} &= V_{\text{non-bonded}} + V_{\textbf{bonded}} \\ V_{\text{non-bonded}} &= V_{\text{lennard-jones}} + V_{\text{coulomb}} \\ V_{\text{bonded}} &= V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}} \end{aligned} \tag{3.5}$$

The non-bonded interactions are separated in Van der Waals interactions, modelled with a Lennard-Jones potential and the coulomb potential for electrostatic interactions. Both are pairwise-additive , which leads to a faster simulation speed compared to higher-order terms.

The Lennard-Jones potential contains a $1/r^{12}$ repulsive and a $1/r^{6}$ attractive term:

$$V_{\text{lennard-jones}}(r_{ij}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \tag{3.6}$$

The $1/r^{12}$ term has been chosen for computational efficiency and overestimates the initial repulsion compared to a more physically motived Buckingham potential with a $1/e^{\mathrm{r}}$ repulsive term. Consequently, the Lennard-Jones potential of atoms connected by three bonds (1-4 interactions) is scaled down by a factor to avoid artifacts.[71] Interactions between atoms with fewer bonds are completely excluded. The parameters $C_{ij}^{(12)}$ and $C_{ij}^{(6)}$ are combined from parameters from either atom using the geometric or arithmetic mean. Parameters are typically fit on condensed phase properties. [64]

Pair-wise potentials in a non-optimised form would scale with $O(N^2)$ for N particles. Due to the short range of Van der Waals interactions, a cut-off is applied and pairs with distances larger than the cut-off are not explicitly calculated. Neighbour-lists are constructed with a buffer region slightly larger than the cut-off to reduce the number of neighbour searches. An energy and pressure correction for long-range Van der Waals interactions is applied dependent on the cut-off distance and system density.

Electrostatic interactions are modelled with the Coulomb potential:

$$V_{\mathrm{coulomb}}(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_r r_{ij}} \tag{3.7}$$

The variable q denotes charges, ϵ_0 is the vacuum dielectric constant and ϵ_r the relative dielectric constant. The same scaling problem applies here as well and a short-range interactions within cut-off are directly calculated. Because of the $1/r$ distance scaling, the contribution of long-range electrostatics is much larger and needs to be accurately modelled. Another complication factor is that simulations use periodic boundary conditions to avoid surface artifacts. The calculation of long-range electrostatics would thus scale even worse than $O(N^2)$ but the calculation of interactions in periodic images can be split into a sum of direct space and reciprocal space contributions. Using the particle-mesh Ewald method[72], the reciprocal space calculation is further optimised to a scaling of $O(N\log(N)$ by approximation the charge distribution on a grid and performing a fast Fourier transformation on this data. The energy can then be summed over the grid in reciprocal space and transformed back. Different charge models exist, for example Mulliken charges or RESP charges[73], which are fit on QM electrostatic potentials for specific conformations of small molecule units.

The bonded potential is split into two, three and four atom interactions:

$$\begin{aligned} V_{\mathrm{bond}}(r_{ij}) &= \frac{1}{2}k^b_{ij}(r_{ij} - r^0_{ij})^2 \\ V_{\mathrm{angle}}(\theta_{ijk}) &= \frac{1}{2}k^a_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2 \\ V_{\mathrm{dihedral}}(\phi_{ijkl}) &= k^d_{ijkl}(1 - \cos(n\phi_{ijkl} - \phi^0_{ijkl})) \end{aligned} \tag{3.8}$$

A simplified dihedral potential can be used where ϕ^0_{ijkl} is either 0 ° or 180 °, which can be parametrised with a single k^d_{ijkl} because $\cos(nx+\pi) = -\cos(nx)$. For the non-pairwise terms, distributing the force on individual atoms has an analytic solution. [74]. Equilibrium bond and angle values are obtained from experimental structures and their force constants fit to vibrational frequency data of reference compounds.[71, 64]. Dihedrals are often fit on QM dihedral screens by minimising the difference between relative QM energies and MM energies without a dihedral term.[75] As a result, bonded parameters depend on the set of nonbonded parameters used for the dihedral fit.

Application

MD simulations with class I force fields are able to capture structural dynamics of molecules in the μs to ms range due to their efficient functional form. Using enhanced sampling techniques to capture non-reactive transitions[76], rare events occuring on longer can be selectively sampled. From the trajectories, experimental observables can be calculated and molecular mechanisms proposed. Applications include drug discovery, protein folding, electrophysiology and structural biology.[68] However, as a fundamental limitation, the functional form does not allow for chemical reactions. This has sparked the development of variations that allow for the most common chemical reactions in biomolecular systems, including protonation.[77, 78]

3.2 Quantum chemistry methods

Motivation

The task of assigning an energy to a system of atoms with positions $\boldsymbol{x}$ is also fundamental to quantum mechanics. Solving the eigenvalue problem known as the time-independent Schrödinger equation,

$$\hat{H}|\Psi\rangle = E|\Psi\rangle, \tag{3.9}$$

the system energy can be obtained. $\hat{H}$ is the Hamiltonian operator and $|\Psi\rangle$ the wave function for the system. The Hamiltonian operator can further be decomposed into nuclear n and electronic e contributions and their interaction to the kinetic energy $\hat{T}$ and the potential energy $\hat{V}$:

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{nn} + \hat{V}_{ne} + \hat{V}_{ee} \tag{3.10}$$

Using the Born-Oppenheimer approximation, the nuclei positions can be treated as fixed. Hence, $\hat{T}_n$ is zero and $\hat{V}_{nn}$ is a constant, leading to following simplification:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} \tag{3.11}$$

For molecular systems, the Schrödinger equation can not be solved exactly. Various methods exist to find an approximate solution, including density functional theory (DFT).

DFT

The idea behind DFT is to model the electron density of a system instead of individual electrons. According to the Hohenberg-Kohn theorems[79], the wave function is uniquely determined by the electron density. In theory, ground-state properties could thus be calculated from the electron density. Density functional calculations became practically feasible when Kohn and Sham proposed to calculate the energy of a system of non-interacting electrons with the same electron density as the interacting system.[80] The Hamiltonian for a system of non-interacting electrons can be solved iteratively in a self-consistent field approach, providing the total energy for the interacting system. However, electron exchange and correlation interactions are only approximately modelled in density functional theory. A higher accuracy can be gained from including the exact exchange energy from Hartree-Fock theory. The combined methods are known as hybrid functionals and contain a variable fraction of exact exchange.

The set of functions that represent an atomic orbital can be chosen such that their influence is eliminated, which is called the complete basis set limit. A computation with such a basis set would be computationally expensive and a range of finite basis sets that introduce only a small error have been developed over the decades.[81] Taken together, DFT uses an empirical functional form based on quantum mechanics with a much higher accuracy than the previously described force fields that comes with a higher computational cost.

Application

Having calculated the energy and possibly some derivatives of it for a set of positions $\boldsymbol{x}$ using the iterative self consistent field method, multiple computational experiments can be performed. If the goal is to simply calculate the single-point energy and forces because the current structure is the target of investigation, the experiment is finished. A minimisation can be performed, typically using a quasi-Newton method to find a local minimum. [82] Several minimisations could be chained with an internal coordinate frozen to a particular value in a relaxed potential energy scan. In case the target structure is not a minimum but a saddle point, as in transition state search, the lowest value eigenvalue of the Hessian can be followed.

Alternatively, the DFT-calculated energies and forces can be used for molecular dynamics. Multiple approaches, such as Born-Oppenheimer MD or Car-Parrinello MD, exist. Due to the computational cost, the molecular system size tends to be limited and timescales are typically in the ps to ns range.[83] For cases where an improved accuracy is only necessary in specific regions of the system, *e.g.* enzymes, a combined QM/MM approach can be used.[37]

3.3 Kinetic Monte Carlo

Motivation

The motivation for using the kinetic Monte Carlo method follows the excellent reviews of Gillespie[84] and Andersen *et al.*[44] with the simplification of assuming unimolecular reactions with a single molecule as product.

Reactive MD simulations, *e.g.* Born-Oppenheimer MD, could be used to count how often reactions occur within a certain time interval and determine the reaction rate r:

$$\frac{\Delta s}{\Delta t} = r(s) \tag{3.12}$$

Here, s is the number of molecules of the simulated species. For unimolecular reactions, the reaction rate of individual molecules is treated as independent and the reaction rate is $r(s) = ks$ for the rate constant k. Unfortunately, most reactions occur rarely within the time scales of reactive MD simulations and the rate constant can not reliably be determined.

However, if there was a way to know the probability of a reaction event within a time interval, it would be unnecessary to perform molecular simulations to model the number of molecules of a certain species. This is called the fundamental premise of stochastic chemical kinetics:

$$a_j(s)dt \triangleq \text{the probability, given S(t) = s, that one } R_j \text{ reaction will occur somewhere inside a constant volume in the next infinitesimal time [t,t+dt]} \tag{3.13}$$

$S(t)$ denotes the state of the system at time t, specifically the number of reactive molecules. The constant value is only relevant for bimolecular reactions and for the unimolecular case, $a_j(s) = c_j s$. Hence, the deterministic reaction rates for systems with many molecules relates to the probabilistic propensity.

Using the propensity, it would be possible to find following probability:

$$P(s,t|s_o,t_o) \triangleq \text{Prob}\{S(t) = s, \text{ given } S(t_0) = x_0\} \tag{3.14}$$

A chemical master equation can be constructed to propagate the system over time by calculating the probability of entering and leaving a state:

$$\frac{\delta P(s,t|s_o,t_o)}{\delta t} = \sum_j^M [a_k(s-1)P(s-1,t|s_0,t_0) - a_k(s)P(s,t|s_0,t_0)] \tag{3.15}$$

For applications with a large state space or many reactions, the system of equation becomes too large to compute. Instead, a single trajectory of S(t) is simulated with the goal of finding a new probability:

$$p(\tau,j|s,t) \triangleq \text{the probability, given S(t)=s, that the next reaction in the system will occur in the infinitesimal time interval [t + } \tau\text{,t+}\tau\text{+dt), and will be an } R_j \text{ reaction.} \tag{3.16}$$

Assuming reactions can be modelled as Poisson point processes, the above probability is given as:

$$p(\tau,j|s,t) = a_j(s)\ \exp(-\tau \sum_i a_i(s) \tag{3.17}$$

Algorithm

Using two standard uniform random numbers u_1 and u_2, the time until the next reaction τ and the identity of the next reaction j can be calculated as:

$$\begin{aligned} \tau &= \frac{1}{a_0(s)} ln(\frac{1}{u_1}) \\ j &= \text{the smallest integer satisfying} \sum_{j'}^{j} a_{j'}(s) > u_2 a_o(s) \end{aligned} \tag{3.18}$$

This leads to the stochastic sampling algorithm[42], also called rejection-free KMC algorithm:

0. Initialise the system with the time $t = t_0$ and the state $s = s_0$.

1. For the given state, evaluate all $a_j(s)$.
2. Calculate τ and j as above.
3. Effect the reaction by updating time and state.
4. Repeat steps 1-3, else end the simulation.

Application

Many KMC simulations are performed on crystalline material because system states can be mapped on a lattice and matched to a limited list of reactions with associated propensities. [44] Consequently, a KMC step consists mainly of event table look-ups for gathering the full list of propensities and no reaction rates have to be computed during the simulation. A fundamental limitation to this approach is that all possible reactions have to be explicitly defined in the first place, which limits the potential insight from these simulations.

A more explorative approach is possible using adaptive KMC[47], where the event list is generated by performing transition state searches during the simulation. The transition state search can be conducted in different ways, including dimer search[48] and high-temperature MD[49]. Even though various optimisation approaches exist, this type of KMC simulation is much slower than lattice KMC.

3.4 Reaction rate calculation

So far, it has only been assumed that a propensity, as defined in the fundamental premise of stochastic chemical kinetics, exists. A relation has been defined to the experimentally observable reaction rate has been made and for unimolecular reactions in a system with a reactive single molecule, the propensity is equal to the rate constant.

Arrhenius described an empirical relation[85] between activation energy E_a and rate constant using a pre-exponential factor A:

$$k = A \exp(\frac{-E_a}{RT}) \tag{3.19}$$

Following a statistical mechanics perspective, Eyring found an expression for A[86] that forms the basis for transition state theory (TST):

$$\begin{aligned} A &= \kappa_0 \frac{k_B T}{h} \\ k_{TST} &= \kappa_0 \frac{k_B T}{h} \exp(\frac{-E_a}{RT}) \end{aligned} \tag{3.20}$$

Several extensions to TST exist that use different expression for the transmission coefficient κ_0[87], including harmonic TST[44], where vibrational modes at the reactant and transition state are calculated to estimate the likelihood of barrier crossing events. In practice, κ_0 is often estimated to be in the range of 1 - 10.[44]

This simplifies the task of finding a rate constant to finding a reaction barrier. Many QM-based methods exist to find transition states from energy minima, including the nudged elastic band method and gradient based search methods like the Dimer method.[88, 82, 44] While a vibrational analysis can be performed to confirm the putative transition state as a saddle point, meaning the hessian has one negative eigenvalue[82], there is no guarantee that this transition state is along a minimum energy path that carries most of the reaction flux.

3.5 HAT GNN

Motivation

The hydrogen atom transfer (HAT) barrier calculation in proteins is complicated by the fact that a range of different hydrogen donors and acceptors exist in the 20 natural amino acids. Additionally, the minimum energy path for HATs in isolated amino acids is likely to have an increased barrier in a protein from reorganising the protein environment in a way that facilitates the HAT along that path. This energy contribution expected to be different for different HAT paths, thus, the path with the most reaction flux can depend on the protein environment. DFT calculations of HATs in every protein environment would be prohibitively expensive and a HAT GNN has been developed recently to emulate reactions for individual conformations.[45] It predicts barriers that can be used to calculate reaction rates using the Eyring equation mentioned above.

When combined with a method that generates conformations, *e.g.* MD, the HAT GNN predicts barriers for paths that are accessible in the current conformational ensemble.

Architecture

As input, the HAT GNN takes the atom positions and elements of the reactant and product structure. Instead of being encoded as hydrogen, the HAT hydrogen is treated as pseudo-element at the start and end position. Using two message and update blocks from the PaiNN architecture[89] alternately, an atom embedding for the hydrogen at the start and end position is generated. This embedding is concatenated and passed to a multilayer perceptron (MLP) with two layers and 128 nodes each, followed by the output node.

Training

The original HAT GNN dataset consists of three different parts: a synthetic dataset of one or two amino closed-shell acids positioned for a HAT from the position of one hydrogen to another, tabulated force field collagen trajectory data with capped cut-outs of amino acids with hydrogens in close proximity to each other and an optimised dataset with geometry optimised structures from either of the first datasets.

For the synthetic and trajectory dataset, hydrogen start and end position are the positions of the respective hydrogens, *i.e.* the hydrogen at the end position is removed for the start structure and vice versa. The transition state is determined by linear interpolation between both hydrogen positions. As only the HAT hydrogen is mobile/unfrozen for the DFT calculations, data generated in this way is hereafter referred to as layer 0 unfrozen data.

HAT barriers correlate with the translation distance (Fig. 3.1,5.13a) and are sampled with a focus on the important low translation distance region. Structures with atoms within 0.8 Å of the interpolated HAT path are discarded because the transition state would have a geometric clash.

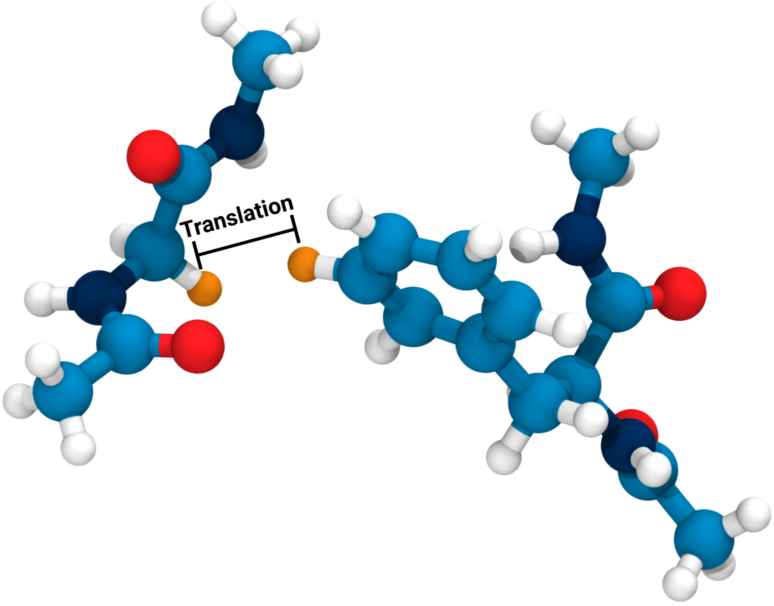


Figure 3.1: **HAT translation distance.** For a HAT from the Cα of a NME-Gly-Ace peptide to the Cϵ of a NME-Phe-Ace peptide. The translation distance, *i.e.* the distance from hydrogen start to end position, indicated in orange, is shown.

The optimisation is performed with all atoms frozen but the HAT acceptor and donor and their bonded hydrogen atoms (layer 1 unfrozen data). A local transition state search is performed to find the transition state structure and validate it with frequency calculations. For all DFT calculations, the hybrid BMK functional[90] with 6-31+G(2df,p) basis set is used.

Training on the layer 0 unfrozen barriers results in a test data MAE of 2.4 kcal/mol for translation distances below 2 Å. In a transfer learning step without freezing parts of the model, the model is trained on predicting layer 1 unfrozen barriers from the initial non-optimised structures. For this experiment, the test data MAE is 3.64 kcal/mol.

The initial dataset neglected the local structure of radical atoms because molecules were generated with closed-shell geometries. In a subsequent work, the Grappa[91] force field is used to simulate protein radicals for an improved dataset.[92] To this end, a collagen fibril was modelled with 200 randomly abstracted hydrogens. HATs from six replicates were gathered and their layer

1 unfrozen barriers calculated. Transfer learning was again used to retrain the original model with this extended dataset for a final test set MAE of 3.24 kcal/mol for translation distances below 2 Å. In contrast to the initial dataset with barriers as low as 12 kcal/mol, the lowest calculated barrier is at 20 kcal/mol.

3.6 Grappa

Motivation

Reactive simulations can lead to products that are not parametrised in tabulated force fields. Even general force fields[93, 94] may lack certain parameters for a molecule and would require a parametrisation process involving QM calculations. Espaloma[95, 96], a machine learning model for predicting classical MM parameters, relies on machine-learned atom embeddings instead of tabulated atom type interactions. With this approach and given a diverse training dataset, a range of molecule force field parameters can be reliably predicted.

Limitations of Espaloma led to the development of Grappa[91] under my supervision. Grappa can deal with open-shell data, quickly parametrises large molecules, directly writes parameter files readable by GROMACS[56] and works with different non-bonded parameter sets.

Architecture

As input for molecule parametrisation, Grappa uses solely the atom elements, partial charges, the number of neighbours and the ring membership based on an atom connectivity graph (Fig. 3.2). In practice, this information is parsed from the molecular system definition of the respective force field. For GROMACS, the topology file is used. Atom positions are not used for parameter prediction.

Atom embeddings $\boldsymbol{\nu}$ are constructed in a Graph Attention Network that resembles the transformer architecture[97]. The attention mechanism is constrained to the graph edges, hence updates are kept local and the atom embedding contains only informations on atoms within a certain number of edges, *i.e.* bonds.

Another transformer is used to obtain bonded parameters from the atom embeddings. Bonded parameters ξ should be invariant under certain permutations:

$$\begin{aligned} \xi_{ij}^{(\mathrm{bond})} &= \xi_{ji}^{(\mathrm{bond})} \\ \xi_{ijk}^{(\mathrm{angle})} &= \xi_{kji}^{(\mathrm{angle})} \\ \xi_{ijkl}^{(\mathrm{dihedral})} &= \xi_{lkji}^{(\mathrm{dihedral})} \end{aligned} \tag{3.21}$$

Permutation invariance is achieved in two steps: first, a positional encoding that is invariant under the respective permutation is appended to the atom embedding, *e.g.*

$$(\nu_i,\, \nu_j,\, \nu_k) \mapsto (\nu_i \oplus 0,\, \nu_j \oplus 1,\, \nu_k \oplus 0)\,, \tag{3.22}$$

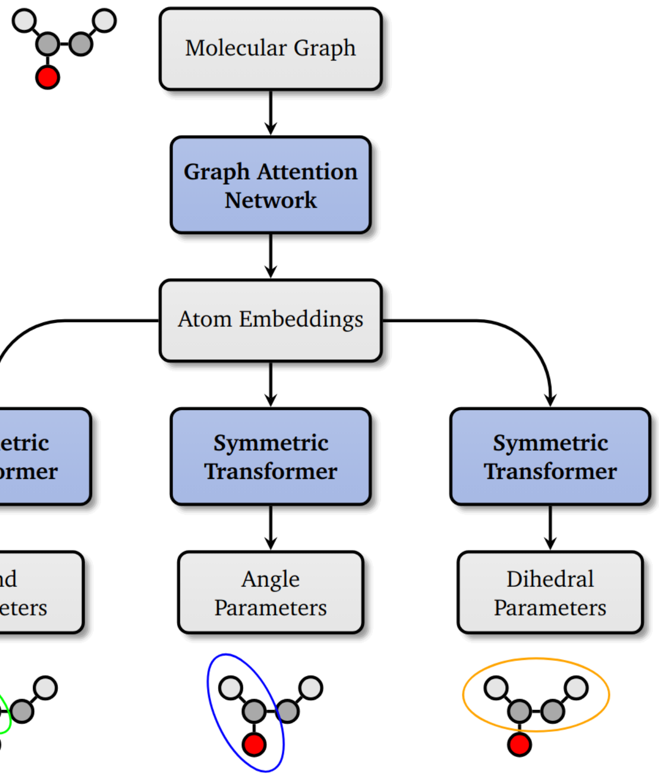


Figure 3.2: **Grappa architecture overview.** Bonded parameters are predicted from the molecular graph via atom embeddings. Molecule-specific quantities are represented in grey, functions in blue. Reused with permission from [91].

second, the modified atom embeddings are pooled over all invariant permutations, resulting in permutation invariant parameter scores $\boldsymbol{z}$. These scores are then mapped to the range of sensible parameters, *e.g.* $(0,\infty)$ for the bond force constant. Furthermore, parameters are initialised according to the parameter distribution in a supplied force field to get a good initial guess, if available, which leads to better convergence during training.

Training

For training, the Grappa architecture is embedded in a larger workflow (Fig. 3.3). A set of molecule conformations is used to calculate MM energies and forces for the MM parameters predicted by Grappa. The QM energies and forces associated with the molecule conformations are then used to calculate a loss:

$$\begin{aligned}\mathcal{L} =&\mathrm{MSE}(E_{\mathrm{MM}}, E_{\mathrm{QM}}) + \lambda_{\mathrm{F}}\mathrm{MSE}(\nabla_{\boldsymbol{x}}E_{\mathrm{MM}}, \nabla_{\boldsymbol{x}}E_{\mathrm{QM}}) + \\ &\lambda_{\mathrm{MM}}\mathrm{MSE}(\xi, \xi_{\mathrm{ref}}) + \lambda_{\mathrm{dih}}\|\xi^{(\mathrm{dih})}\|_2^2\end{aligned} \tag{3.23}$$

Apart from the hyperparameter λ_{F} that weights energy and force, additional regularisation of the difference to reference MM parameters and the magnitude of dihedral parameters is possible by tuning λ_{MM} and λ_{dih}. Finally, the loss is backpropagated through both the symmetric transformers and graph attentional neural network to update learnable parameters. It should be noted that while molecular conformations are used for training Grappa, and thus determine the weights and biases of the model, the parameter prediction is independent of the three-dimensional structure of the target molecule.

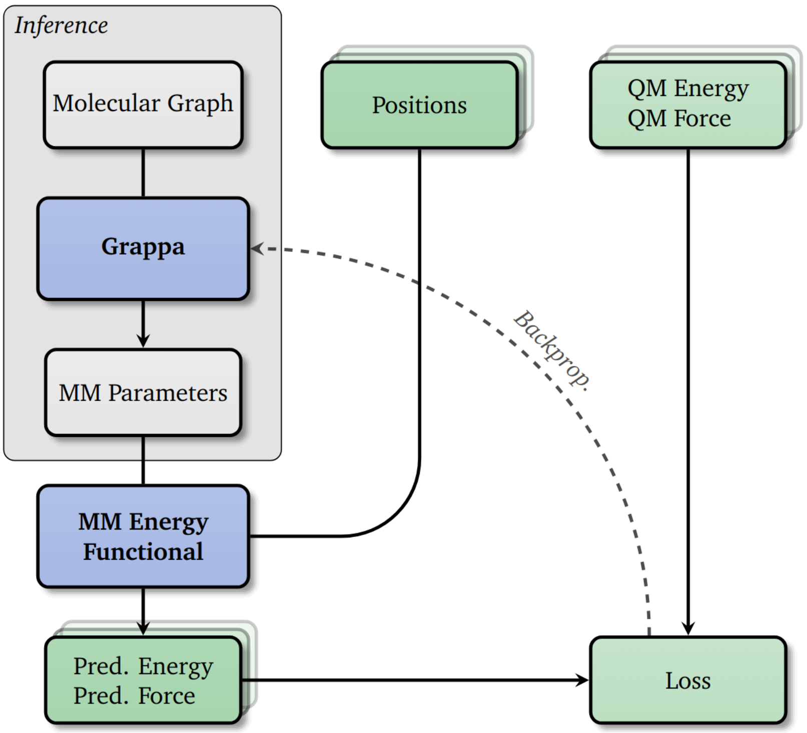


Figure 3.3: **Grappa training workflow.** The Grappa inference architecture (grey box) is integrated into a training workflow for learning MM parameters from molecule positions with associated QM energies and forces. Conformation-specific quantities are represented in green. Reused with permission from [91].

Since one set of MM parameters is only valid for a certain non-reactive state, formation energies do not have to be reproduced. Instead, energy differences of conformations are the optimisation target and thus the mean of target and predicted energies is subtracted before the loss evaluation.

In the original publication, Grappa was trained on a combination of SPICE[98] and QCArchive dataset[99]. The included molecules cover small drug-like molecules, peptides and nucleic acids.

3.7 Simulation details

Molecular dynamics

General

Unless noted otherwise, molecular dynamics simulations are performed with GROMACS[56] versions 2019 to 2024. The Grappa version 1.4.0 force field is used together with Amber non-bonded parameters originating from ff94[64]. As water model, TIP3P[100] is used. Typically, 2 fs time steps with LINCS[61] constraints on H-bonds are employed. Simulations are kept at 300 K using the v-rescale thermostat[60] and, except for NVT equilibrations, 1 bar using the Parrinello-Rahman barostat[101]. The Coulomb and Lennard-Jones cutoffs are at 1 nm.

NMR validation

The J-coupling benchmark simulations of $(\text{Xaa})_n$ and Gly-Xaa-Gly peptides, where Xaa denotes different amino acids, are performed with a box size with 1.4 nm padding around the peptide. The pH of J-coupling experiments was at 2, which is modelled by protonating all carboxy groups. Charges are neutralised with Na^+ and Cl^- ions with the parameters of Joung and Cheatham[102]. After short equilibration simulations, triplicates with 500 ns simulation length were simulated for each of the 13 peptides. Apart from the Grappa 1.4.0 simulations with Amber charges, the same experiment was repeated with Grappa 1.3.0, Grappa 1.4.0 with AM1-BCC charges[103], Espaloma 0.3[96] and Amber ff14SB[75].

Protein folding

Simulations of the CLN025 variant of chignolin were performed at the experimental melting temperature of 340 K in duplicates of 4 μs simulation time. The padding distance is 20 Å around the experimental structure (PDB accession code: 5awl[104]) and ions are added to firstly neutralise the protein charge and secondly reach a NaCl concentration of 0.1 M.

Radical training

A total of 146 radical peptides were generated through homolysis at the Cα-C bond or through hydrogen abstraction. Instead of in GROMACS, simulations to obtain training conformations were performed in OpenMM[105] with 1 fs time steps and a intermediate Grappa model fine-tuned on QM optimisation data of radical peptides. A padding of 10 Å around the radical was used and the salt concentration set to 0.1 M. For keeping the system temperature at 300 K and 1000 K, respectively, a Langevin integrator was used. 100 conformations were sampled, with 100 ps inbetween them. The first 50 ps are simulated at 1000 K and the 50 ps before saving a conformation are simulated at 300 K.

Collagen fibril HAT sampling

For sampling HAT in the collagen fibril, a model of type I collagen in *Rattus norvegicus* with HLKNL cross-links was generated from ColBuilder[106]. It is comprised of about 320 000 protein atoms with 41 individual triple helices and spans one central overlap as well as half a gap region on either side. The aromatic residues, phenylalanine and tyrosine, were randomly mutated to 20% DOPA and 26.6% tyrosine and either DOPA anion.

The protein model is aligned along the z axis, placed in a 16.3 nm x 17.7 nm x 95.0 nm triclinic box, solvated, neutralised and the NaCl concentration set to 150 mM. Afterwards, a 10 ns NVT equilibration with restrained protein heavy atoms and a 10 ns NPT equilibration with 1 nN pulling force per protein chain is conducted.

Homolysis sites are sampled from recently conducted extensive simulations, weighted by the number of sampled breaks.[14] Three sampled breaks are combined and proposed for further simulation if they are more than 40 Å apart and not connected. Out of the proposed break groups, four variations with diverse homolysis locations and including cross-link sites were taken.

The three homolysis reactions per break group are simultaneously effected and a 10 ps slow-growth simulation performed to receive the product state structure. Simulations with radicals use the Grappa 1.4.1-radical force field that is developed in this work. Subsequently, another 10 ns NPT equilibration is performed. The Simulation of one break group with a loose fragment crashed and was discarded.

For sampling HATs, three types production simulations were run: 500 ps simulations with 10 fs stride,*i.e.* write-out frequency, 5 ns simulations with 100 fs stride and a subsequent 50 ns with 1 ps stride. For every NPT equilibration, three replicates were used for the production runs, starting after 6 ns, 8 ns and 10 ns of the NPT equilibration. Two individual simulations crashed and were discarded, for a total of 7 simulated systems. The 50 ns simulations were only performed for the triplicates of one break group.

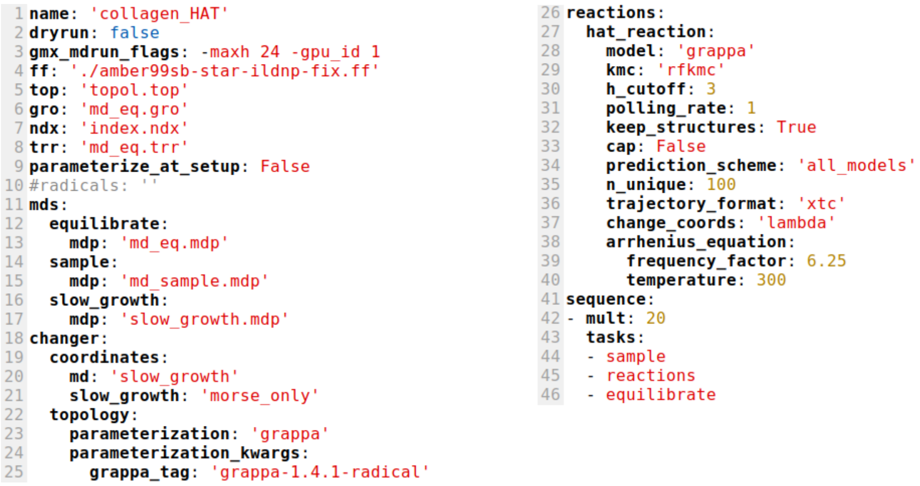
KIMMDY

PYD homolysis sites

Starting from the same model of type I collagen in *Rattus norvegicus*, another model was created with N- and C-terminal PYD crosslinks with a connectivity of 9.C-5.B-944.B and 1047.C-1047.A-98.B, respectively. All aromatic residues were mutated to a DOPA-anion and the two variants randomly distributed. The same equilibration scheme as above was used up to the homolysis reaction. This time, combinations of breaks at three of the eight PYD short arm Cα-Cβ bond (compare Fig. 7.1b) were sampled with the same connectivity and distance criteria as above. A total of 12 different homolysis patterns were generated, with every PYD cross-link being included in three simulation systems. Again, three replicates were used by starting from different snapshots of the post-break NPT equilibration.

For production runs, 500 ps equilibration and 500 ns sampling simulations were conducted with the latter having a stride of 10 fs. The slow-growth simulation is over 10 ps. All simulations use the C-rescale[107] barostat. Simulations are performed with the Grappa 1.4.1-radical force field.

The KIMMDY configuration options are shown in Figure 3.4. The 36 simulations with 20 reactions each would sample at most 720 reactions. Due to crashes, 600 (83.3%) reactions were actually sampled.



```yaml
name: 'collagen_HAT'
dryrun: false
gmx_mdrun_flags: -maxh 24 -gpu_id 1
ff: './amber99sb-star-ildnp-fix.ff'
top: 'topol.top'
gro: 'md_eq.gro'
ndx: 'index.ndx'
trr: 'md_eq.trr'
parameterize_at_setup: False
#radicals: ''
mds:
  equilibrate:
    mdp: 'md_eq.mdp'
  sample:
    mdp: 'md_sample.mdp'
  slow_growth:
    mdp: 'slow_growth.mdp'
changer:
  coordinates:
    md: 'slow_growth'
    slow_growth: 'morse_only'
  topology:
    parameterization: 'grappa'
    parameterization_kwargs:
      grappa_tag: 'grappa-1.4.1-radical'
reactions:
  hat_reaction:
    model: 'grappa'
    kmc: 'rfkmc'
    h_cutoff: 3
    polling_rate: 1
    keep_structures: True
    cap: False
    prediction_scheme: 'all_models'
    n_unique: 100
    trajectory_format: 'xtc'
    change_coords: 'lambda'
    arrhenius_equation:
      frequency_factor: 6.25
      temperature: 300
sequence:
- mult: 20
  tasks:
  - sample
  - reactions
  - equilibrate
```

Figure 3.4: **KIMMDY PYD simulation configuration file.** Configuration options in a YAML file. Lines starting with '#' are treated as comments. Note the KIMMDY task sequence definition in the last lines.

Diverse homolysis sites

KIMMDY simulations with diverse homolysis sites are performed from the equilibrated radical structures of the collagen fibril HAT sampling (see above). For production runs, the MD settings are the same as above. All simulations use the C-rescale barostat. Simulations are performed with the Grappa 1.4.1-radical force field.

The configuration options are set as in the PYD simulations above but only 10 reactions were sampled per system. The seven simulated systems were simulated with triplicates for a sum of 21 simulations. Out of 210 expected reactions, 146 (70%) were performed because some systems encountered errors during the simulation.

Chapter 4

Generating a protein radical force field

4.1 Radical MD parameters are necessary for modeling HAT

Protein force fields (see 3.1) have been under development for many decades[71, 64, 108, 75, 69] to achieve better agreement with QM and experimental data. Observables of interest in MD studies often depend on a fine balance between conformations[109, 70] because energy differences, according to the Boltzmann distribution, have an exponential effect on the population of states. Typically, conformations are discernible in the backbone dihedral space (Fig. 4.1a). This is why particular effort went into improving the agreement with QM or experimental observables by adjusting backbone dihedral parameters directly[108, 75, 69, 70]

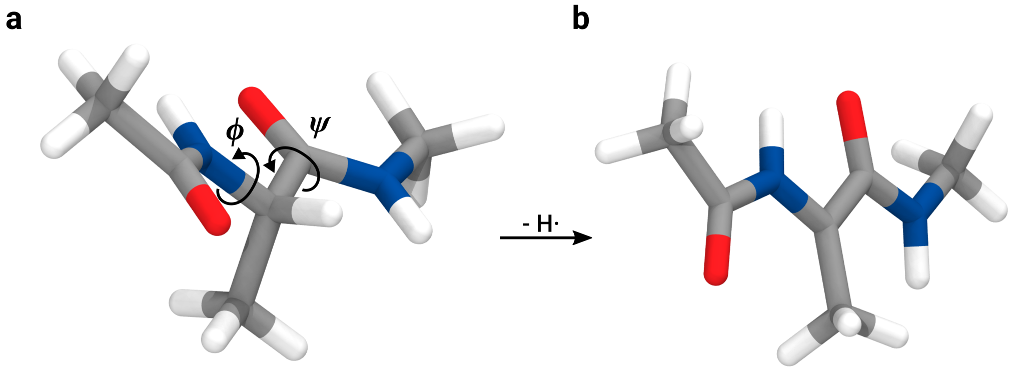


Figure 4.1: **QM optimized structures of alanine and its radical.** Both panels show QM optimised structures of an alanine peptide with acetyl and N-methyl capping groups. **a**, Natural, closed-shell peptide. ϕ and ψ backbone dihedrals are shown. **b**, Cα radical peptide. The radical peptide is formed by hydrogen abstraction at the natural amino acid Cα atom.

For modelling protein radical species, which are the reactants of protein HAT reactions, accurate radical force field parameters are crucial. Apart from the general need for accurate parameters to correctly model the occupancy of different conformations, the local environment of radical atoms may include different neighbouring atoms. These neighbouring atoms are potential H-donors to the radical atom. HAT reaction rates correlate exponentially with the distance of reactants[45], thus, accurate conformational sampling and local environments are crucial for modelling reactions. Even small force field inaccuracies or artifacts may drastically alter the relative rates of competing HAT reactions. With the exception of the glycine Cα radical [110, 111], no protein radicals have been parametrised so far, highlighting the necessity to create a high-quality force field for protein radicals to perform reactive KIMMDY simulations of HATs in collagen.

4.2 Protein radicals adopt untypical conformations

4.2.1 About different 200 radical amino acids need to be modelled

Radicals in collagen are the product of either homolytic cleavage or HAT with both reactions resulting mostly in different species (Fig. 4.2a,b). The number of possible radical species is larger than the number of amino acids since homolytic cleavage produces a Cα radical with a broken Cα-C bond, as well as a radical ketone group connected to the nitrogen of the following amino acid which likely warrants reparametrisation of that amino acid. Other backbone bonds are unlikely to be homolytically cleaved under physiological conditions [14]. Even more peptide radical species can be formed by hydrogen abstraction. For every heavy atom bound to a hydrogen a different radical species can be formed. Taken together, roughly 200 different radical amino acid species need to be parametrised.

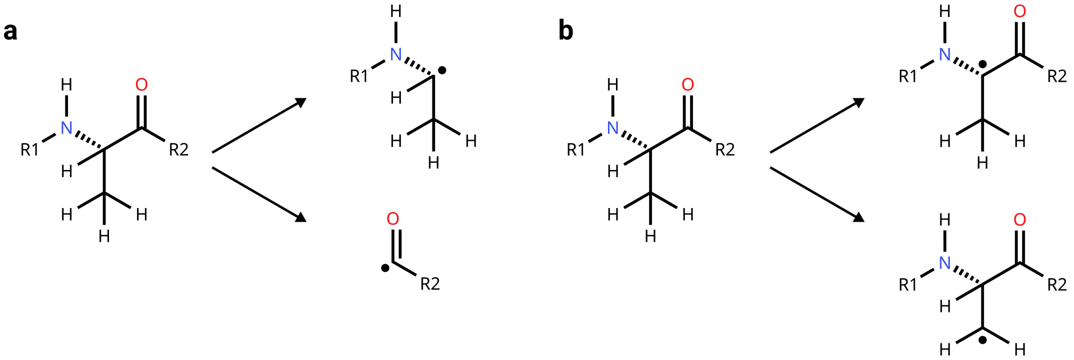


Figure 4.2: **Radical species of alanine.** **a**, Homolytic cleavage leads to a Cα-radical and a radical ketone group. **b**, HAT radicals are different if the H-abstraction occurs at different heavy atoms.

Albeit changes to a protein force field to model the different radicals would likely be similar for different species because they are produced by the same mechanism, this work would require huge efforts in the framework of traditional force field development, which moves at a pace of few changes to select parameters per publication.[108, 75] In a first step, changes to the molecular structure upon radical formation are investigated to get an understanding of the necessary parametrisation steps.

4.2.2 Tetrahedral carbons adopt a planar geometry as radicals

To illustrate differences in the molecule geometry between closed-shell amino acid and their radical counterparts, the alanine Cα radical, from H-abstraction, is taken as example (Fig. 4.1a,b). As a backbone radical it can have a large impact on the surrounding geometry. The carbon radical changes from a tetrahedral geometry to planar, making this an example with rather large differences compared to other positions. Interestingly, the radical can be approached from both sides of the plane created by its bonded neighbours. The planarity leads to a different relative position of the side chain compared to the backbone. This could lead to slight structural rearrangements or folding of longer side chains back on to the backbone for a side chain to backbone HAT.

4.2.3 H-abstraction has an impact on internal coordinates

Changes in the molecule geometry also lead to changes in the internal coordinates, *i.e.* bond, angle and dihdral values. Because of their relation to bonded MM parameters, analysis of the internal coordinate changes upon H-abstraction or homolytic cleavage gives first insights on the magnitude of parameter changes to model radicals.

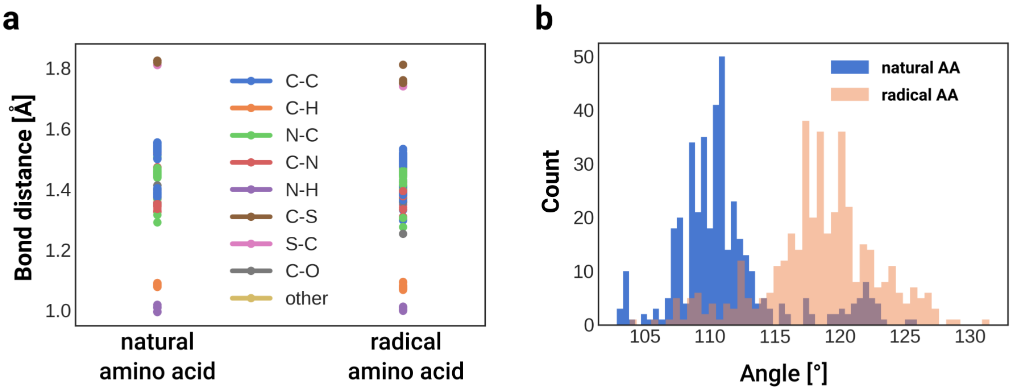


Figure 4.3: **Bonds and angles are affected by radical formation. a**, Bond distances at radical atoms by involved elements for QM optimized structures of radical amino acids and at the same positions for their corresponding natural amino acids. **b**, Angles around radical atoms for QM optimized structures of radical amino acids and at the same positions for their corresponding natural amino acids. Radical amino acids were formed by H-abstraction.

Bond distances slightly decrease at the radical atoms (Fig. 4.3a). As the bond distance itself, changes upon radical formation depend on the involved elements. Bonds involving hydrogen atoms are barely affected, while most heavy atom to heavy atom bonds decrease in distance. In contrast, angles (Fig. 4.3b) mainly increase from around 110° to 120°. This corresponds to the previously mentioned change from tetrahedral to planar geometry. Small populations can also be found at angles below 115° and above 125°. Taken together, bond distances and angles show clear trends when a nearby radical is introduced. Introducing radicals can, however, also change internal coordinates further away from the radical atom. Bond and angle force constants are comparably high, so small distance or angle errors would result in large energy errors. The derivation of such paremeters for radicals, however, is challenging because they are traditionally fit to experimental vibrational frequencies and structures[64], which are not available for radicals.

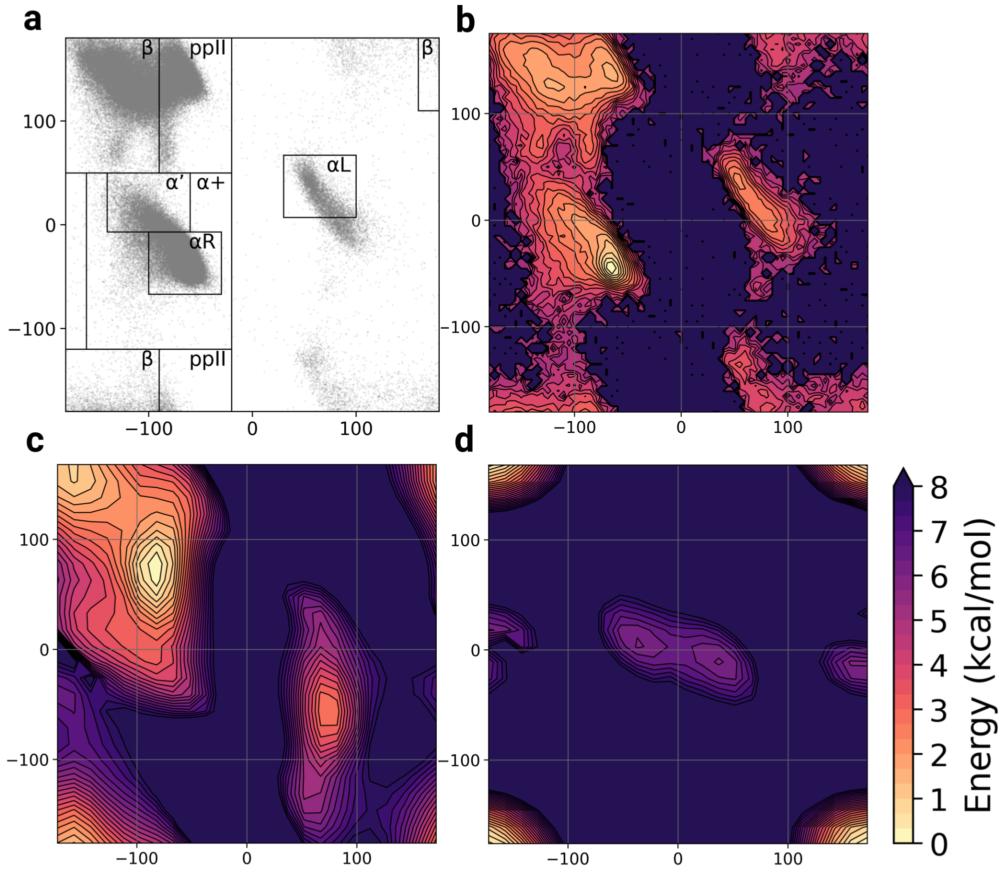


Figure 4.4: **Ramachandran plots.** **a**, Backbone dihedral values of 500 X-ray structures deposited in the PDB[112]. **b**, Free-energy surface of the data presented in panel a. **c**, QM potential energy surface of the alanine dipeptide backbone dihedrals *in vacuo*. QM potential energy surface of the alanine dipeptide Cα radical backbone dihedrals *in vacuo*.

H-abstraction heavily influences dihedrals. The backbone dihedral space has several minima, which can be distinguished in a Ramachandran plot[113], also called ϕ/ψ plot. Here, the ϕ is defined as $C_{(-1)}$-N-Cα-C and the ψ as N-Cα-C-$N_{(+1)}$ dihedral angle, where the subscripts denote membership to the previous or upcoming amino acid. These minima are defined from dihedral values of X-ray structures deposited in the PDB (Fig. 4.4a). Conversion of binned counts via Boltzmann inversion to a free-energy surface (FES) leads to an interpretable energy surface (Fig. 4.4b).

In comparison, *in vacuo* potential energy surfaces (PES) of alanine and the alanine Cα radical (Fig. 4.4c,d) are much simpler. The alanine PES has minima in the β region, between ppII and αR region and below the αL region. These differences to the protein FES arise from solvation and entropic effects.[69] The alanine Cα radical PES is even simpler, with a single main minimum at (180,180). As the Ramachandran plot is periodic in both x- and y-axis, this minimum is shown at the four corners of the plot. A further minimum with relatively high energy is located around (0,0). Ramachandran plots for other Cα radicals look similar, indicating a general trend. The backbone dihedral parametrisation, which required the most attention for natural amino acid[108, 75, 69] results indicate that Cα radical backbone dihedrals, which parametrisation. Again, radicals other than Cα radicals are expected to also influence backbone dihedrals and other dihedrals around them, highlighting the complexity of the parametrisation task.

4.2.4 Radical van der Waals parameters are difficult to obtain

Partial charges of atoms in a molecule undergoing homolytic cleavage or H-abstraction must change to keep integer charges of the resulting fragments. Traditional charge parametrisation for Amber force fields is done with RESP charges[114] using HF/6-31G* functional and basis set. This choice of functional and basis set is known to over-polarize but was intentional to account for the polarization in water compared to *in vacuo*. Fortunately, this also leads to computationally inexpensive QM optimizations for partial charge calculations.

In Amber force fields, van der Waals parameters are completely identical for all types of carbon and nitrogen atoms, indicating little adaptation to the chemical context. Historically, protein van der Waals parameters were derived from condensed-phase simulations and fit to reproduce experimental observables[64], which is not possible for radicals due to lacking experimental data.

4.3 Graph neural networks can facilitate radical parametrisation

Since there is no clear allocation of energy contributions to MM terms, it is possible to partially compensate deficiencies in one set of parameters with other contributions. Hence, fitting bonded parameters given a certain set of suboptimal non-bonded parameters may still yield an accurate force field. A recently introduced method, Espaloma[95, 96], uses a graph neural network (GNN) to fit bonded parameters for an extensive set of small molecules, peptides and nucleic acids. Espaloma uses the molecular graph and hand-crafted chemical features, such as formal charge, hybridization and aromaticity, to create atom embeddings. These embeddings are then used to predict bond, angle and dihedral parameters by minimizing the energy and force loss compared to QM data. Intriguingly, the method predicts classical MM parameters independent of the molecule conformation, as conformations are only used for training but not as an input to the model during inference. This allows for using Espaloma parameters in the same way as a classical force field without loosing simulation speed.

Given a suitable QM data set of peptide radicals, a method like Espaloma can also predict the parameters required for reactive simulations of HATs in collagen. However, implementation details limit the applicability of Espaloma. First, Espaloma is designed to predict charges[115] and was trained on AM1-BCC[103] charges which is itself a approximation of RESP charges. Contrastingly, a method predicting bonded parameters with non-bonded parameters as input would be more versatile and allow for testing different charge models for protein radicals. Second, handcrafted chemical features may be helpful in providing an inductive bias for data-efficient training but also add limitations to the model if they are not straightforwardly available. Third, formal charges can lead to non-physical parameters for mesomeric structures[96] and in the current implementation, open-shell molecules can not be processed in Espaloma. Finally, parameter prediction for large proteins is prohibitively slow.

To address these limitations, a new method, Grappa[91], was developed in a master's thesis by Leif Seute under my supervision (see 3.6). Grappa uses non-bonded parameters as input feature and does not depend on handcrafted chemical features. This design, together with an efficient implementation that is less prone to limitations inherited from specialised software dependencies solves many shortcomings of Espaloma. In the sections below, we show that Grappa is a suitable method for parametrising proteins. A major benefit of starting our force field development and validation with natural amino acids is that QM datasets are available for training and several benchmarks exist to evaluate the force field performance. Furthermore, a highly accurate protein force field is of general interest to the MD community. In a second step we use the now validated Grappa to parametrise amino acid radicals.

4.4 Validation of Grappa for proteins

The results presented in this subsection originate from collaborative work of Leif Seute and myself.

4.4.1 Grappa accuracy is state-of-the-art

For comparison with Espaloma, Grappa was trained on a combination of SPICE[98] and QCArchive dataset[99] used in the most recent Espaloma publication[96]. Grappa consistently outperforms Espaloma in energy and force RMSE by a slight margin (Table 4.1). Both the accuracy of Espaloma and Grappa far outperform traditional force fields, including the general force field Gaff-2.11[93] and the specialized force fields ff14SB[75](protein) and OL3[118](RNA). The mean predictor shows average absolute deviations from the sample mean and serves as a reference to assess the distribution of energies and forces in the datasets.

Even though training, validation and test data were strictly separated, an argument could be made that the training data for both GNN methods is more

Dataset	Test Mols	Confs	RMSE	Grappa	Espaloma	Gaff-2.11	ff14SB, RNA.OL3	Mean Predictor
SPICE-Pubchem	1411	60853	*Energy*	**2.3**	**2.3**	4.6		18.4
			Force	**6.1**	6.8	14.6		23.4
SPICE-DES-Monomers	39	2032	*Energy*	**1.3**	1.4	2.5		8.2
			Force	**5.2**	5.9	11.1		21.3
SPICE-Dipeptide	67	2592	*Energy*	**2.3**	3.1	4.5	4.6	18.7
			Force	**5.4**	7.8	12.9	12.1	21.6
RNA-Diverse	6	357	*Energy*	**3.3**	4.2	6.5	6.0	5.4
			Force	**3.7**	4.4	16.7	19.4	17.1
RNA-Trinucleotide	64	35811	*Energy*	**3.5**	3.8	5.9	6.1	5.3
			Force	**3.6**	4.3	17.1	19.7	17.7

Table 4.1: Accuracy of Grappa, Espaloma and tabulated general (Gaff-2.11) and specialized (ff14SB, OL3) force fields. All force fields use ff99SB VdW parameters and all use AM1-BCC partial charges except for ff14SB, which uses RESP charges. States in all datasets were MD sampled at temperatures of 300 K or 500 K. RMSEs of zero-centred energies are in kcal/mol, componentwise RMSEs of forces in kcal/mol/Å. SPICE-Pubchem is a small-molecule dataset, SPICE-DES-Monomers and SPICE-Dipeptide peptide datasets and RNA-Diverse and RNA-Trinucleotide are RNA datasets. The ωB97M-D3(BJ) functional[116] and def2-TZVPPD[117] basis set were used. Reused with permission from [91].

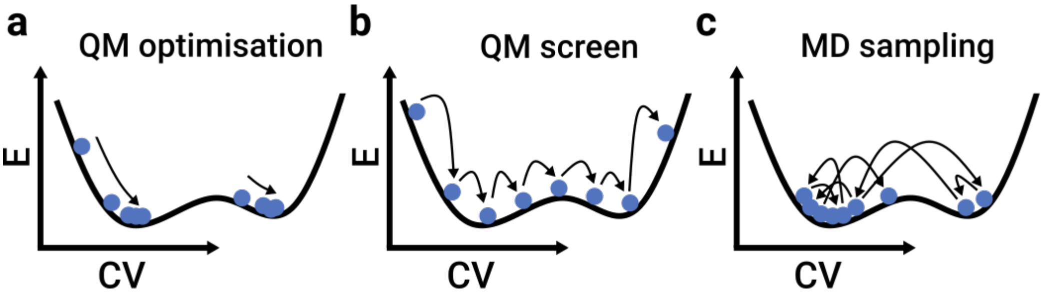


Figure 4.5: **Sampling types for QM datasets. a**, Using QM geometry optimisations leads to few points outside minima, corresponding to initial starting structures and many structures very close to minima. **b**, QM screens result in equally spaced data points along the screened coordinate, irrespective of the underlying PES. **c**, MD sampling yields Boltzmann distributed samples. E - Energy, CV - Collective variable.

similar to the test data than what the general or specialized tabulated force fields were fit on. Consequently, accuracy differences in the test set would be biased towards the GNN methods. A good dataset, however, should be representative of the real world applications of the methods under investigation. In this regard, use of the SPICE and QCArchive datasets are a major step forward because they contain single-point QM energies and forces of MD sampled states. Given that MD states are sampled from the Boltzmann distribution, accessible conformational states are included and weighted by the probability of their occurrence. Most applications of protein simulations are in aqueous solution, hence the system should be solvated (compare Fig. 4.4) to sample conformations that occur frequently during the application. This consideration is relatively novel in force field fitting.[69] Fortunately, the system needs to be solvated only during the sampling and not for QM calculations because the bonded parameters depend on solute atoms only. QM optimized structures are helpful additions to the MD sampled datasets to improve sampling at minima and relaxed QM screens can be beneficial to learn barriers or more generally the edges of probable region in state space (Fig. 4.5). To sum up, the datasets used for training and testing Grappa are far more extensive and closer to MD applications than previous datasets used for fitting MM parameters.

One remaining assumption is that learning MM parameters from the QM energies and forces of small peptides is enough to apply these parameters to proteins. This assumption has been long-standing and many successful traditional force fields, especially of the Amber family, have been developed under this assumption. Recent CHARMM force fields[119, 70] challenged it and developed parameters with a greater emphasis on reproducing experimental data, for example with ϕ/ψ torsional correction maps (CMAPS) based on experimentally measured occupancies of different regions of the Ramachandran plot. These 'empirical force fields' may not be highly accurate compared to QM energies and forces but are still very valuable for studies relating to experimental observables. Another indication for this behavior can be seen when comparing the energy and force RMSEs for Gaff-2.11 and ff14SB on the SPICE-Dipeptide dataset. Both are similarly accurate but ff14SB is arguably more suited to reproduce experimental observables and be used as a protein force field. With steadily increasing computational resources, the assumption may no longer be necessary: Whole protein[120] or protein fragment[121, 41] QM datasets are becoming available. By fine-tuning on these datasets, Grappa can be instrumental in exploring advantages from training on these kinds of datasets.

4.4.2 Grappa captures peptide backbone J-coupling best

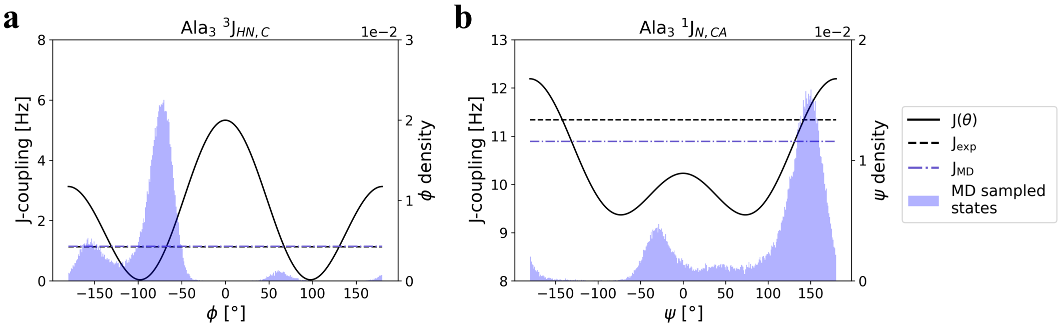


Figure 4.6: **Relation between dihedral and J-coupling value.** **a**, Visualisation of the Karplus equation for the $^3J_{HN,C}$ coupling of Ala$_3$. **b**, Visualisation of the Karplus equation for the $^1J_{\mathrm{N,CA}}$ coupling of Ala$_3$. Both panels show the density of MD sampled states, the J(θ) function, the reference J$_{\mathrm{exp}}$ value, and the calculated average J$_{\mathrm{MD}}$ value, which is obtained by averaging $J(\theta)$ over the MD sampled states.

A key metric for protein force field quality is its agreement with backbone J-coupling data. As mentioned previously, protein backbone parameters determine the occupancy of different conformational states and thus strongly influence key observables. J-coupling refers to interactions of spins connected through chemical bonds. Depending on bond distances and angles of the connecting bonds, the coupling strength varies. In the Karplus equation[122] this dependence is used to relate backbone dihedral values to J-coupling values of backbone atoms through empirical constants. Empirical constants A, B, C and Δ are calculated for specific molecules but are treated as transferable to a certain extent[96], introducing a source of error in the relation.

$$J(\theta) = A\cos^2(\theta + \Delta) + B\cos(\theta + \Delta) + C \tag{4.1}$$

To evaluate the peptide backbone parameters, an extensive MD simulation of the molecule of interest is performed. For each snapshot, the dihedral observable is calculated and a J-coupling value derived using the Karplus equation. Finally, the average computationally calculated J-coupling value J_{MD} is compared to the experimental $\mathrm{J}_{\mathrm{exp}}$ value using the RMSE or χ^2 value. Two representative J-coupling relations, one with strong agreement, the other with a rather large error, are shown in Figure 4.6.

$$\chi^2 = \frac{1}{N_{\mathrm{obs}}} \sum_{\mathrm{obs}} \frac{(J_{\mathrm{MD}} - J_{\mathrm{exp}})^2}{\sigma^2_{\mathrm{model}}} \tag{4.2}$$

Several dihedral distributions can lead to the same average J_{MD} value, making the validation underdetermined. However, if several J-couplings are calculated for the same dihedral, the space of possible dihedral distributions is limited. Further assuming a sensible ϕ/ψ dihedral surface reduces the validation to assessing the relative populations of mainly the β, ppII and αR region. This is the limited but very valuable information that can be gained from the comparison of backbone J-coupling data.

For the J-coupling analysis, we use the benchmark dataset and analysis pipeline established by the Open Force Field Initiative[123]. The same benchmark has also been used by the Espaloma authors[96], which facilitates comparison. It includes the peptides Ala_3, Ala_4, Ala_5, Gly_3, Val_3, GAG, GEG, GFG, GKG, GLG, GMG, GSG and GVG for a total of 121 observables. In agreement with NMR experiments, simulations are performed at pH 2. For every peptide, triplicates of 500 ns simulation time are simulated. Seven different J-couplings, namely ${}^1J_{\mathrm{N,CA}}$[124], ${}^2J_{\mathrm{N,CA}}$ [125], ${}^3J_{\mathrm{HA,C}}$ [126], ${}^3J_{HN,CB}$ [127], ${}^3J_{\mathrm{HN,C}}$ [127], ${}^3J_{\mathrm{HN,HA}}$ [127], ${}^3J_{\mathrm{HN,CA}}$[128] with a modified Karplus equation, are analyzed. The same J-couplings can be evaluated for different residues of the same molecule.

Grappa and Espaloma both reproduce J-coupling values more accurately than the tabulated ff14SB force field (Fig. 4.7a). Glycine, valine and alanine have off-diagonal points, indicating weaker agreement with $\mathrm{J}_{\mathrm{exp}}$ than other residues. In the case of alanine, this could be attributed to the fact that a high fraction of all observables is for alanine. Espaloma and Grappa results show a similar pattern of J_{MD} values. This points to both architectures learning similar parameters from the underlying data. The χ^2 value for Grappa improves more than the RMSE comapred to Espaloma. Because the χ^2 value contains a correction for experimental uncertainty $\sigma^2_{\mathrm{model}}$, this means that Grappa performance improved for J-couplings with low uncertainty, while the overall squared error to the reference $\mathrm{J}_{\mathrm{exp}}$ values remains similar. The Ala_3 PES (Fig. 4.7b) for Espaloma and Grappa is notably broader than for ff14SB. Areas of the Ramachandran plot found in X-ray structures (compare Fig. 4.4a,b) are generally more stable in contrast to the sharp ff14SB minima. Hand-crafted force fields like ff14SB are known to overstabilize folded conformations[129], an artifact that could be related to the observed narrow minima.

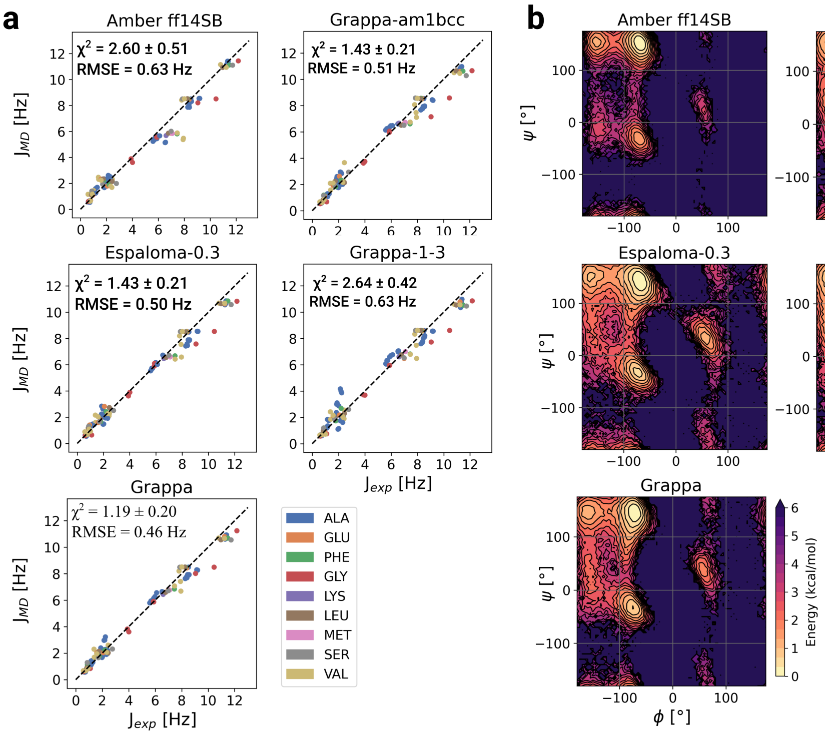


Figure 4.7: **Peptide backbone parameters are evaluated with J-couplings.** **a**, Experimental and computational J-coupling values for 13 peptides. Points are colored by the amino acid to which the dihedral belongs. As quantitative measure, RMSE and χ^2 value are included. The classical force field ff14SB, Espaloma, Grappa and two Grappa variants, Grappa-1.3 and Grappa-am1bcc are evaluated. ff14SB, Grappa and Grappa-1.3 use RESP charges, while Espaloma and Grappa-am1bcc use AM1-BBC charges. **b**, Free-energy surfaces of the Ala$_3$ peptide for the same force fields, calculated from the population density observed in MD.

A main difference between Espaloma and Grappa in the previous results is the charge model. Grappa is compatible with charge models it encountered during training, *i.e.* RESP, AM1BCC and CHARMM ESP charges. To further understand the differences between Espaloma and Grappa, the same AM1BCC charge model is used for a direct comparison (Fig. 4.7a). RMSE and χ^2 value are almost identical between Espaloma and Grappa-am1bcc. Hence, the previously observed differences between Espaloma and Grappa can be attributed to the charge model.

Furthermore, the training dataset was extended with single-point QM energies and forces of peptides sampled in the gas-phase at 300 K, 500 K and 1000K to train the Grappa-1.3 model. The performance of this model is considerably worse than Grappa models trained on the original dataset (Fig. 4.7a) and only on-par with ff14SB. Still, the PES of Ala_3 appears only slightly different (Fig. 4.7b).

Taken together, different GNN methods for predicting bonded parameters reach the same accuracy in the employed J-coupling benchmark if they are trained on the same dataset and use the same non-bonded parameters. Grappa is compatible with different non-bonded parameter sets by design and can thus be combined with high-quality RESP charges to reproduce backbone J-coupling far better than tabulated force fields, the current gold standard. The training dataset influences the performance of Grappa models and even the addition of peptide data sampled in gas-phase instead of in solution leads to a degraded force field quality.

4.4.3 Grappa reproduces important features on dihedral surfaces

Backbone dihedrals are the most important observable to judge the quality of a protein force field. A popular approach is thus to tune only the dihedral parameters to improve agreements with experiments. For example, Amber force fields up to ff14SB were mainly modified at their backbone dihedral parameters to improve the agreement with QM screens and experimental data[75], producing highly symmetric backbone dihedral PESs. Reproducing diagonal elements is not possible with this approach, because there is no cross-term for dihedrals. CHARMM 22 first introduced grid-based cross-terms, called CMAPs, to reproduce QM dihedral screens[94] and the same approach has been used in Amber ff19SB to create residue-specific CMAPs. However, force field parameters other than dihedral parameters also influence the dihedral PES and their optimisation can improve the agreement with experimental data. Without using CMAPs, Grappa reproduces key features of implicit solvent QM dihedral screens (Fig. 4.8). Most notably, the αR basin has a diagonal shape, an additional minimum is found between ppII and αR region at (-90°, 80°) and transition regions between minima are more shallow.

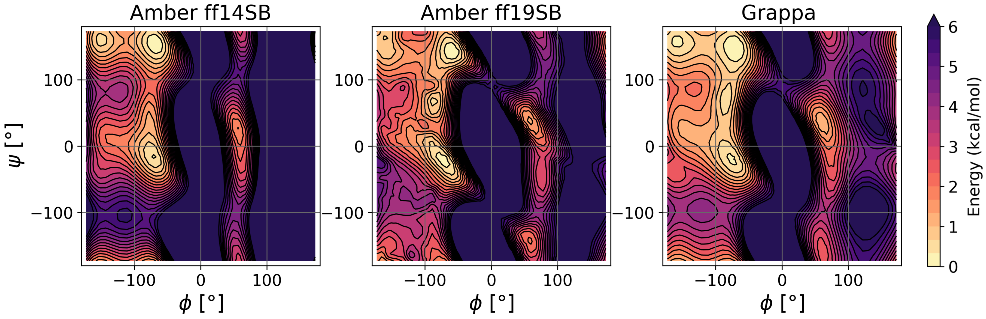


Figure 4.8: **Implicit solvent dihedral screens.** Implicit solvent QM backbone dihedral screens of Ace–Ala–Nme are shown for ff14SB, ff19SB and Grappa. As implicit solvent model, GB-Neck2[130] was used. Own figure reused with permission from [91].

MM dihedral screens in OpenMM[105] can be decomposed by the different energy contributions (Fig. 4.9). Large differences between ff14SB and Grappa

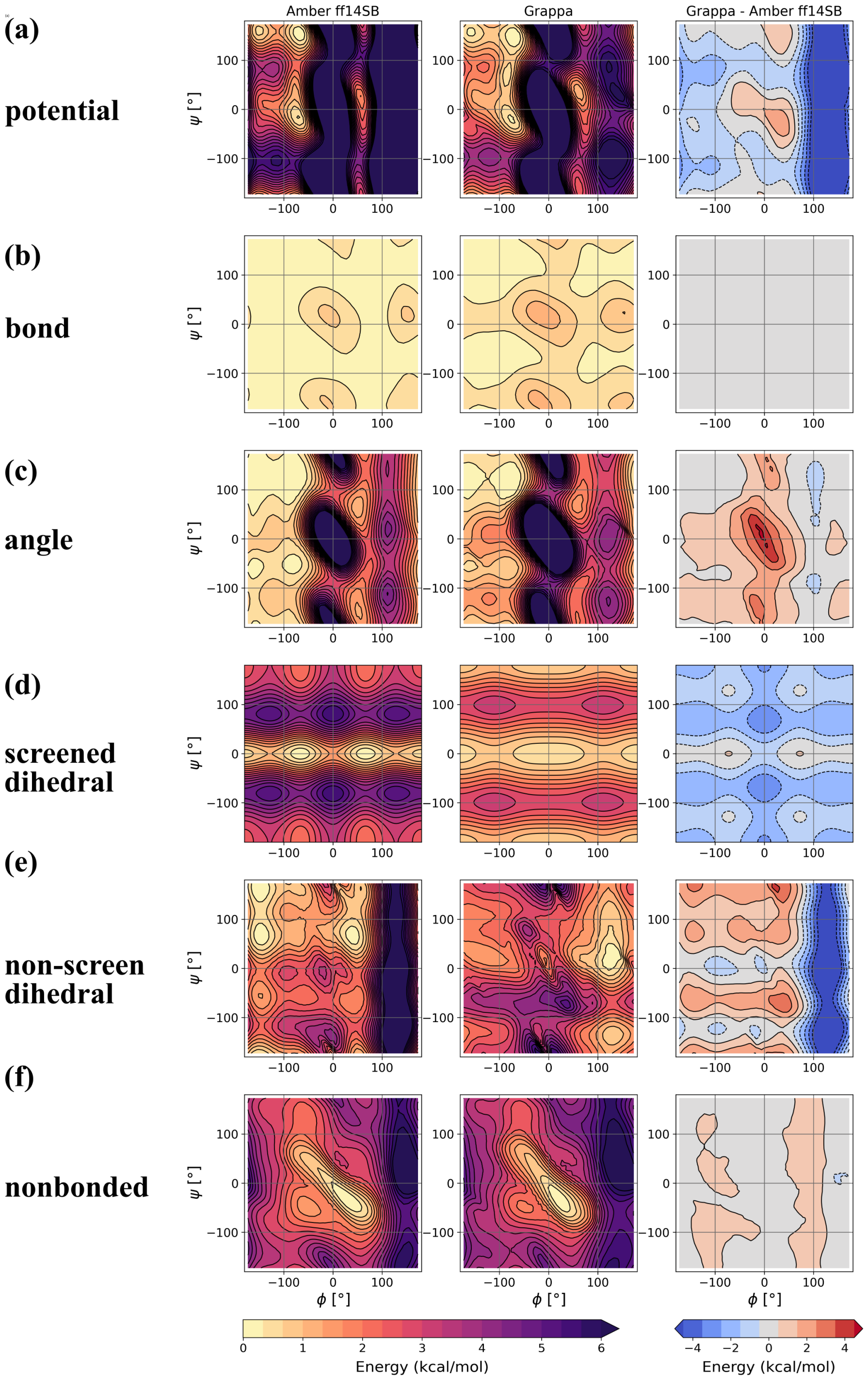


Figure 4.9: **Energy contributions to a QM dihedral screen.** Backbone dihedral screens are performed in implicit solvent for Ace-Ala-Nme for ff14SB and Grappa. The contribution type is mentioned in the panel. Own figure reused with permission from [91].

in the PES of a capped alanine peptide can not only be found for the screened ϕ and ψ dihedrals but also for the angle and non-screened dihedral contributions. It can be seen that the angle contributions differ especially around (0°,0°) and are responsible for diagonal elements in the ϕ/ψ space. The non-screened dihedral surface also shows several differences, for example lower energies in the $\phi > 90°$ region. Grappa training data does contain dihedral screens, but no particular emphasis has been put on reducing the loss for these screens compared to optimisation or MD sampled data. In comparison, tabulated force fields relied more on QM dihedral screens for parameter fitting. The sampling strategy for Grappa could have lead to an underestimation of the energy in the $\phi > 90°$ region because it is barely sampled. However, a ML based simultaneous prediction of bond, angle and dihedral parameters can make use of non-trivial correlations between different parameters to create a highly accurate force field without requiring explicit special treatment of dihedral screens.

Interestingly, ff19SB could not outperform ff14SB in a NMR J-coupling benchmark[69]. Although Grappa has not been compared directly to ff19SB due to the long MD simulations required to calculate J_{MD}, it seems likely that Grappa is more accurate. From this perspective, residue-specific CMAPS, which have not been implemented in the popular MD engines OpenMM and GROMACS[56] six years after the publication of ff19SB, may not be necessary for high-quality force fields. Further evidence could be provided by using Grappa to fit CMAP parameters on the Espaloma dataset and comparing the performance of both Grappa force fields. In general, Grappa offers the possibility to evaluate different bonded and non-bonded functional forms, including class II force fields[63], in a comparable setting and could be further used as a force field development tool.

4.4.4 Grappa is accurate for folding small proteins

A major challenge in force field parameterisation is to model protein folding. Accurate folding simulations would require the force field to model unfolded and folded states while disfavoring misfolded states. Tertiary contacts and hence non-bonded interactions are important for protein folding[109] but bonded parameters also have a strong influence on folding energies as studies with different force fields that share their non-bonded parameter set show[109, 70].

We apply Grappa to folding free energy simulations of CLN025[104] (PDB accession code: 5awl), a variant of the small protein chignolin. Following the simulation setup of Shabane *et al.*[131], we simulate two replicates for 4 μs at the experimental melting temperature of 340 K, starting from the folded structure. At this temperature, the folding free energy is per definition 0 kT. Multiple folding and unfolding events occur but CLN025 remains folded during most of the simulations (Fig. 4.10b). This behavior is also reflected in the free energy profile along the backbone RMSD (Fig. 4.10a) with a deep minimum around 1 Å from the folded structure. Still, the calculated folding free energy of -3.0 kT is closer to the experimental value than the -4.0 kT previously reported with ff99SB with the same non-bonded parameters and water model[131]. Folded and unfolded states are defined in a three-state model with an intermediate state and boundaries at 1.5 Å and 4.5 Å.

$$\Delta G = -\log\left(\frac{\mathrm{p(RMSD} < 1.5\,\text{Å})}{\mathrm{p(RMSD} > 4.5\,\text{Å}}\right) \tag{4.3}$$

Bonded parameters can hardly be evaluated solely by folding free energies.

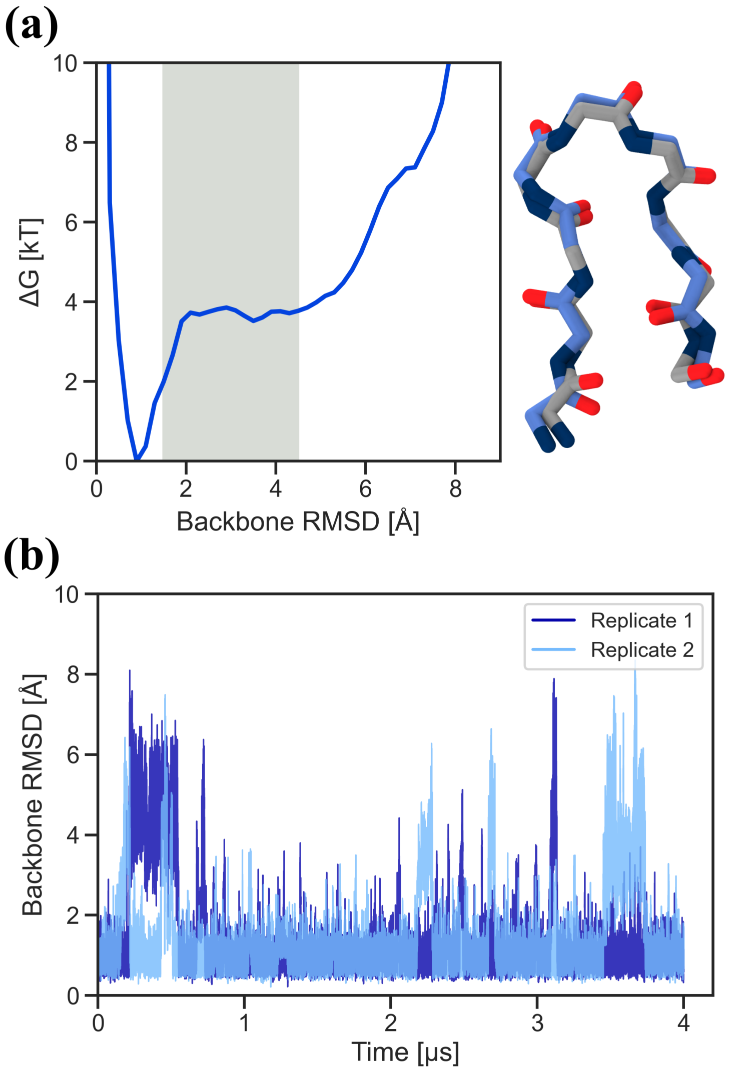


Figure 4.10: **Folding free energy simulations with Grappa. a**, Folding free energy profile for CLN025. The intermediate state between folded and unfolded state is shaded in gray. On the right, the CLN025 X-ray structure is shown in gray and the structure representing the center of the largest cluster from Grappa MD simulations is shown in blue. **b**, RMSD for two replicates of the CLN025 folding simulations. Own figure reused with permission from [91].

Rather, the whole parameter set with non-bonded and water parameters has to fit together. The water model OPC[132] is known to stabilize the unfolded state for CLN025 simulations[133] and force fields with CHARMM non-bonded parameters tend to overstabilize proteins less than with Amber non-bonded parameters[134]. Grappa bonded parameters overstabilize folded proteins less than traditional tabulated parameter sets. Hence, combining Grappa with CHARMM non-bonded parameters (Fig. 4.11) leads to folding free energies further away from the experimental value than with the reference tabulated force field (0.65 kT vs 0.1 kT). We also note that other approaches simulated CLN025 for an order of magnitude longer to calculate folding free energies.[133, 70] Certainly, our results are influenced by a sampling effect but longer simulations are beyond the scope for showing possible applications for a novel force field.

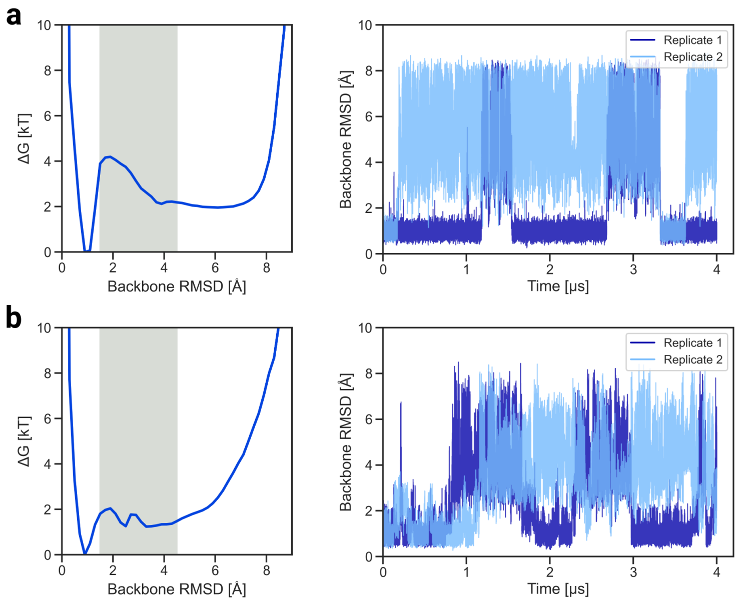


Figure 4.11: **CLN025 folding free energy profiles with CHARMM non-bonded parameters. a**, Folding free energy profile and RMSD for CHARMM 36m **b**, Folding free energy profile and RMSD for Grappa with CHARMM non-bonded parameters.

Pinpointing which parameter set should account for a certain effect would be difficult. Preferably, bonded parameters should be fit for explicit solvent single-point QM energies and forces based on a certain water model and non-bonded parameter set or ideally all parameters optimised together. Piecing together force field parameters can lead to cherry-picking and is mainly useful for finding areas of application where this combination reproduces experimental data accurately but not for validation of a parameter subset. To sum up, we show that Grappa produces reasonable results in protein folding simulations. Simulations with Grappa parameters show a reduced stability of CLN025 compared to tabulated force field parameters, which may be desirable in some simulation contexts but not in others.

4.4.5 Grappa accuracy is determined by the dataset and functional form

In the previous sections, Grappa has been shown to be superior to the gold standard for classical force fields, *i.e.* tabulated force fields. The success of Grappa can be attributed to training on a large QM energy and force dataset and predicting all bonded parameters simultaneously. However, a small addition of data points from a less relevant region of the dihedral PES to the dataset already negatively impacts the agreement with J-coupling data. This points to little robustness towards the conformational distribution in the training datasets.

The functional form of classical force fields is a tradeoff between speed and accuracy. A more expressive E(3) equivariant GNN[66] that predicts energies and forces based on the molecular coordinates achieves a 10 times higher accuracy on a small molecule benchmark[135] than Grappa. On the other hand, classical force fields are roughly three orders of magnitude faster than E3NNs[91], while most of the computational cost goes into long-range electrostatics calculations, which are not modeled in E3NNs[41], and typically less than 1% is spent on bonded energy. This accuracy and speed gap creates a large design space for reevaluation of established functional forms[63] and novel bonded functional forms.

4.5 Treatment of non-bonded interactions

A prerequisite for predicting bonded parameters for peptide radicals is to have a suitable non-bonded parameter set. As a first approach, the VdW parameters are left unchanged and only partial charges are reparametrised. Amber amino acid partial charges were derived from simultaneously fitting partial charges to Ace-X-Nme peptides in the αR and β basin.[73] Here, due to the simpler backbone dihedral landscape of radical amino acids, only optimised structures at the (180°,180°) minimum are used.

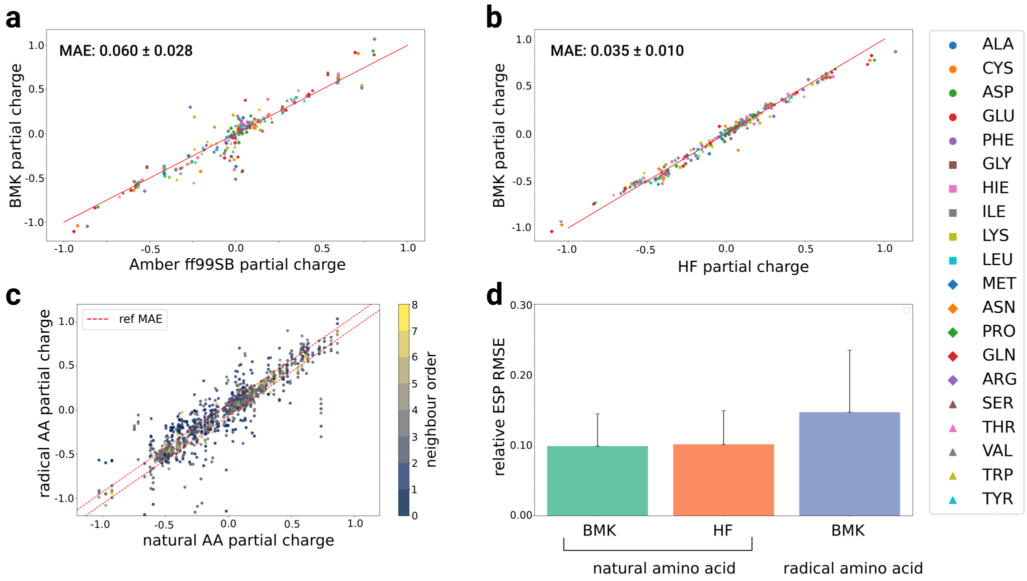


Figure 4.12: **Partial charge fitting for radicals.** **a**, Partial charge differences per atom for the 20 canonical amino acids between ff99SB RESP partial charges and BMK/6-311+G(2df,p). ff99SB partial charges were derived using the HF/6-31G∗ functional and basis set. **b**, Partial charge differences per atom for the 20 canonical amino acids between HF/6-31G∗ and BMK/6-311+G(2df,p). **c**, Partial charge differences per atom between natural amino acids and radical amino acids created through H-abstraction. Scatter points are coloured by the number of bonds to the radical atom. **d**, Relative ESP RMSE[136] for natural and radical amino acid charge fits.

For fitting partial charges of Ace-X-Nme radical peptides in Antechamber[137], both the standard functional and basis set for RESP partial charge derivation, HF/6-31G∗ and BMK/6-311+G(2df,p), which is suitable for modeling radicals, were tested (Fig. 4.12a,b). The resulting partial charges are strongly correlated and differences to the partial charges of the ff99SB force field can be attributed to conformational differences. Hence, BMK/6-311+G(2df,p) derived RESP charges should be compatible with HF/6-31G∗ derived RESP charges. A third partial charge set, based on heuristically moving the charge of an abstracted hydrogen onto the previously bonded heavy atom, or restoring integer charge of two fragments after homolytic cleavage by adjusting the partial charges of the atoms involved in homolysis, is also tested. In first tests on radical QM optimisation and QM screen data, no large differences in final energy and force RMSE of combined non-bonded Grappa bonded parameters were found between the charge sets. For ease of implementation in the adaptive KMC framework, the heuristic partial charge set is currently in use.

Partial charges of the heuristic model do not change for any non-radical atom while the BMK RESP charges can change for atoms with several bonds between them and the radical (Fig. 4.12c). Thus, it seems promising to further investigate the impact of an improved radical non-bonded parameter set. A collaboration with the authors of ffparaim[138], a software that calculates non-bonded parameters from QM calculations, is ongoing to calculate radical non-bonded parameters.

4.6 Grappa predicts radical parameters

Following the lessons from the previous peptide parameter validation, I created three peptide radical datasets with QM optimisations, QM dihedral scans and MD sampled single-point energies and forces. For all datasets, Ace-X-Nme peptides are used, where X denotes a peptide radical amino acid. No force field existed for the peptide radicals of interest before this work, so a first Grappa model was trained on the Espaloma dataset and the first two peptide radical datasets. This intermediate model was then used for simulating radicals for 10 ns each. 100 equally spaced frames per trajectory were used for single point QM calculations in Psi4[139] using BMK/6-311+G(2df,p)[90] functional and basis set. A final model, Grappa-radical, is trained on the Espaloma dataset and the peptide radical datasets.

Grappa-radical is the only force field applicable to peptide radical molecules, thus, accuracy can only be evaluated relative to other datasets with a similar mean predictor. Overall, the accuracy is on a high level but slightly worse on radical structures compared to the other test sets (Table 4.2). Grappa-radical consistently has a higher accuracy than any classical force field on any of the shown datasets. The test sets, and due to the 80:10:10 split also the training sets, are smaller for peptide radicals than for natural peptides.

As shown in section 4.2, peptide radicals have a simpler backbone dihedral PES than natural amino acids and may have been a simple target for parametrisation. The benchmark shows that Grappa is able to accurately predict parameters for peptide radicals, albeit slightly less accurate than for closed-shell molecules. A focus of the parametrisation effort was the backbone dihedral parameters but it is possible that radicals in the side chain have a

Dataset	Test Mols	Confs	RMSE	Grappa	Grappa-radical	Mean predictor
SPICE-Pubchem	1411	60853	*Energy* *Force*	**2.3** **6.1**	2.4 6.4	18.4 23.4
SPICE-DES-Monomers	39	2032	*Energy* *Force*	**1.3** **5.2**	1.4 5.5	8.2 21.3
SPICE-Dipeptide	67	2592	*Energy* *Force*	**2.3** **5.4**	2.5 5.7	18.7 21.6
RNA-Diverse	6	357	*Energy* *Force*	**3.3** **3.7**	3.5 4.4	5.4 17.1
RNA-Trinucleotide	64	35811	*Energy* *Force*	**3.5** **3.6**	3.7 4.3	5.3 17.7
Peptide-Radical-Opt	17	818	*Energy* *Force*		**3.7** **9.1**	6.9 7.7
Peptide-Radical-Scan	1	1235	*Energy* *Force*		**2.9** **4.9**	5.0 3.9
Peptide-Radical-MD	16	1598	*Energy* *Force*		**2.4** **8.7**	5.3 19.7

Table 4.2: Accuracy of Grappa and Grappa-radical on MD sampled datasets from the Espaloma datasets and on peptide radical datasets. RMSEs of zero-centered energies are in kcal/mol, component-wise RMSEs of forces in kcal/mol/Å.

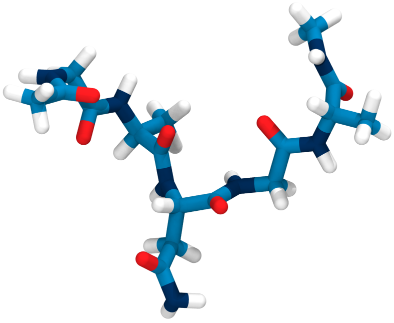


Figure 4.13: **A force field artifact in an amide group.** Rendering of a short peptide containing asparagine. The asparagine side chain amide group located towards the bottom of the figure has too close hydrogen atoms. A QM optimised structure would have a planar amide nitrogen with 120° angle between the hydrogens.

complex local environment that is currently not adequately sampled. Also, an improved non-bonded parameter set could increase the accuracy.

Facing the same issue of missing experimental data for the validation of the radical parameters as previous authors[110, 111], only the general plausibility of the simulated structures and agreement with QM data can be assessed. The preliminary Grappa model trained without MD sampled radical data predicted parameters with an artifact for the amino acids asparagine and glutamine, where the hydrogen positions at the side chain amide group are too close when a solvent anion is nearby (Fig. 4.13). This has not been observed for Grappa-radical for which the Cα radical structure minimum is captured accurately (Fig. 4.14). Artifacts of hydrogen placement would be problematic for reactive HAT simulations because suboptimal hydrogen positions would lower the barrier for H-abstraction and lead to high rates for non-reactive moieties. In summary, peptide radical parameters cannot be validated to the same extent as natural peptide parameters because experimental data is missing. Despite this limitation, Grappa-radical QM benchmark results look promising and no major artifacts have been observed, yet.

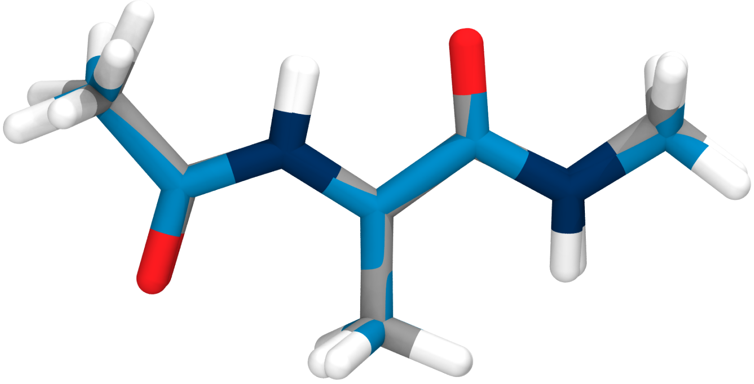


Figure 4.14: **Grappa captures the alanine dipeptide Cα radical minimum.** Rendering of a Ace-Ala-NME peptide with Cα radical. The QM reference structure is colored in silver and the Grappa-radical structure is colored in cyan. The heavy atom RMSD between both structures is 0.04 Å.

The development of a force field for hundreds of different protein radical species is a milestone and is the foundation not only for HATs after mechanoradical formation but also for other biological systems with open-shell amino acids, *e.g.* radical enzymes[140, 141] and those affected by molecular ageing[142]. Without the recent development of GNN-based force field fitting, this development would not have been possible. With the appropriate datasets, the same Grappa extension approach could be utilised to model diverse amino acid modifications in a range of settings.

Chapter 5

Optimising the reaction search

5.1 A new method for reaction search is proposed

Before moving on to the full reactive molecular simulation method with KMC, MD and ML components (compare Fig. 2.1), the combination of the MD and ML component is investigated. Here, HAT reactions and their associated rates are predicted with a HAT GNN[45] based on the conformational ensemble of non-reactive MD simulations. The combined MD/ML method can therefore be seen as a transition state search for multiple competing HAT reactions or as event list generation step from a KMC perspective.

This approach should prove to be more accurate than previously used heuristics[53, 54] because the HAT GNN emulates QM energy barrier calculations. Energy barriers are calculated hundreds of times, which would have necessitated computationally expensive calculations without machine learning models at hand.

The duration of the event list generation step determines how long an iteration of the KMC algorithm takes. To get a comparison for the method timing, a reference for typical KMC and adaptive KMC applications is established in the following section. Then, we investigate different approaches for efficient MD simulations of collagen conformational ensembles. Finally, the combined MD/ML method is optimised on a collagen HAT dataset and the timing compared to other methods.

5.2 Adaptive KMC is several orders of magnitude slower than regular KMC

Packages for standard KMC reach a speed of 1 million reactions per second on a standard consumer CPU.[46] In comparison, one of the first applications of adaptive KMC, using the Dimer method[88] for finding transitions in Al(100) surface models with hundreds of atoms, evaluated new states at 2 KMC steps per hour.[47] The novelty in adaptive KMC is that the event list is constructed during the KMC simulation for every state by performing a reaction search. Due to the symmetry of the system, many states are equal and their event list can be reused. This led to a speed-up to 600 KMC steps per hour.

In biological systems, very few radicals are equal (an example would be the Cγ or Cδ atoms of phenylalanine or tyrosine) and the overall number of radical positions is high, indicating that many reaction searches will have to be conducted for extensive collagen HAT simulations. Hence, the two components of the reaction search step, ensemble generation and reaction emulation, need to be efficient.

5.3 HAT reactions are unlikely to occur during MD timescales

Following the definition from Gillespie[84], the propensity (or rate constant for 0th order reactions) function for all modelled reactions a must be known for the interval [t, t + τ + $d\tau$] with t defined as current time and τ as the time to the next reaction.

$$a_j(s)dt \triangleq \text{the probability, given S(t) = s, that one } R_j \text{ reaction will occur somewhere inside a constant volume in the next infinitesimal time [t,t+dt]} \tag{5.1}$$

Using Monte Carlo theory, the total propensity

$$a_0(s) \triangleq \sum_{i=1}^{M} a_i(s) \tag{5.2}$$

is related to the time to the next reaction τ

$$\tau = \frac{1}{a_0(s)} ln(\frac{1}{u_1}), \tag{5.3}$$

where u_1 is a uniformly distributed random number [0,1]. This means the ensemble generation via MD must be as long as the order of τ or converge to constant propensities to know a_j for the required time interval.

No experimental HAT rates for collagen or proteins in general are available. One study measured cysteamine thiyl radical transfer to amino acid side chains[143] in a range of 10^3 - 10^5 M$-$1 s^{-1} at 310 K and measurements on intramolecular HAT in alkanes[144, 23] can be extrapolated to up to 10^2 s^{-1} at 300 K. In free amino acid experiments, HAT rates were estimated to be in the order of 10^6 s^{-1}.[33] Intramolecular HAT barriers of a glycine dipeptide radical were calculated to be as low as 7.5 kcal/mol in a QM study, but the reactions would require significant rearrangements that may not be possible inside a protein.[145] Using the Eyring equation for a barrier of 7.5 kcal/mol at 300 K with a transmission coefficient of 1 results in a rate of 2.15*10^7 s^{-1}, slightly out of range for MD simulations of large molecular systems. Accounting for tunnelling effects[146] would further raise the rates by two to three orders of magnitude. However, the training data from collagen fibril simulations of the GNN model that is used here scarcely contains that low barriers[45], indicating that times to the next reaction above the achievable MD sampling time will likely occur. Hence, MD sampling should be long enough to yield constant propensities/rates, which is difficult to achieve given that MD sampling is typically not ergodic[76]. The convergence of rates is therefore a main component for testing the MD/ML method.

5.4 MD simulation

5.4.1 Simulating the first equilibrated collagen fibril ensemble takes multiple days

As we want to calculate the HAT rates for a given reaction as an ensemble property, it is useful to start sampling with a starting structure drawn from the equilibrium distribution. For the modelled collagen fibril this is computationally expensive because it contains 33 triple helices in the gap region and is 67 nm in length for a total of about 200 000 protein atoms and 2.6 million atoms overall (Fig. 5.1a,b).

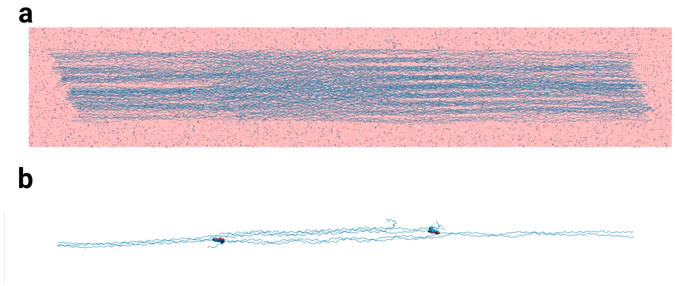


Figure 5.1: **Collagen fibril structure.** **a**, Rendering of the collagen fibril in water. The collagen triple helices are displayed in cyan and ions as cyan and blue spheres. **b**, Two triple helices connected through trivalent PYD cross-links depicted as spheres. The collagen gap region is at the edges of the fibril and the overlap region in the centre between the cross-links.

Starting from a deposited non-solvated collagen structure[106, 147], general ensemble properties, such as system energy, temperature and pressure, are quick to converge[148] after previous solvation and energy minimization, taking less than 1 ns to reach the target value (Fig. 5.2a,b,c,d). The collagen fibril is simulated under mechanical stress to model the experimental setting[13], thus, the fibril end-to-end distance is a system specific observable of interest. It also converges within 1 ns (Fig. 5.2e) but can show a drift over tens of nanoseconds[14].

A further relaxation effect is the solvation of the protein because water is placed based on geometric considerations and is unlikely to be optimally distributed. To avoid structural disruptions of the protein, protein atoms are initially constrained to allow water to populate cavities. Following a safe equilibration protocol of 10 ns simulations in NVT and then NPT ensemble takes two days on a A100 or four days on a RTX 2080 consumer GPU. Hence, the first simulation of an equilibrated collagen fibril can only be started after multiple days of simulation. The simulation that would be used for calculating HAT rates could only then be started, potentially taking several more hours.

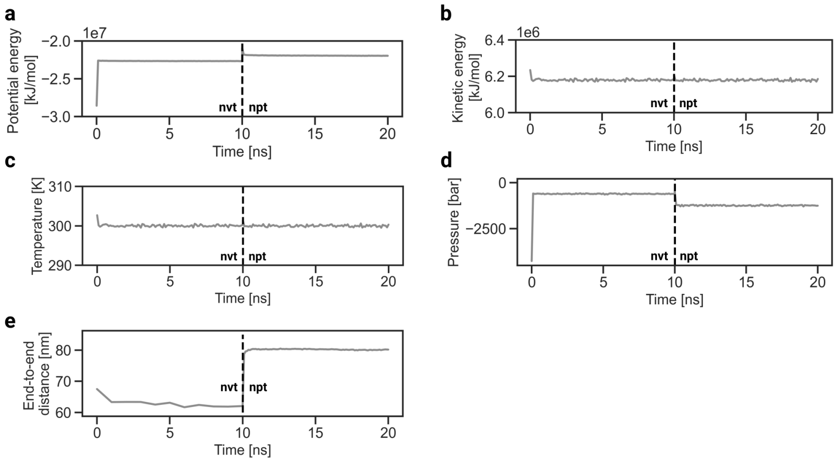


Figure 5.2: **Collagen fibril equilibration simulations.** General and system specific equilibrium properties **a**, potential energy, **b**, kinetic energy, **c**, temperature, **d**, pressure and **e**, end-to-end distance for 10 ns NVT and NPT simulations. The pressure is displayed as more negative than the reference value of 1 bar because of interactions with the pulling code.

5.4.2 Subsequent simulations start from equilibrated reactant structures

Luckily, it is not necessary to equilibrate a structure from a non-solvated state every time because the reactant state should be similar to the product state. When starting from an equilibrated reactant structure with MD parameters for the product structure, it should be possible to quickly generate an equilibrated product structure. A previous solution[55], in a different method framework than adaptive KMC, used a MD simulation with smaller time steps after homolytic cleavage for relaxation but this leads to severe artifacts (Fig. 5.3) with potentially multiple Morse bonds breaking at one homolysis site.

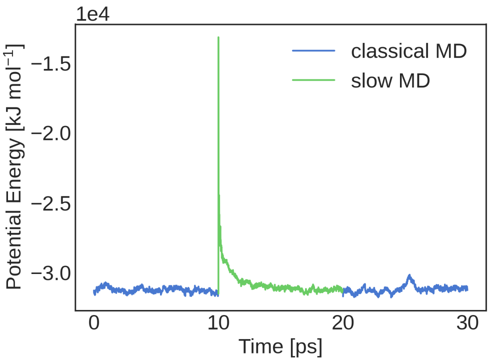


Figure 5.3: **Simulating after a reaction can lead to artifacts.** In a previous KIMMDY implementation a bond was chosen to be homolytically cleaved according to the KMC algorithm. Then, that bond was removed from the simulation parameters. Afterwards, a simulation with small time steps (shown in green) was started to relax the structure from the reactant state to the product state. Data generated from a single triple helix simulation.

Energy minimisation after a reaction is an option because properties like end-to-end distance and solvation are unaffected by the local minimisation. HAT mainly involves the movement of a hydrogen atom through unoccupied space and would be a good target for minimization. Other reactions that involve larger movements, *e.g.* homolytic cleavage from large pulling forces, may not reach a minimum without steric clashes from an energy minimisation. Still, for HAT reactions and assuming the general ensemble properties to be equilibrated after 500 ps of simulation in the NPT ensemble after an energy minimisation, a starting structure for reaction sampling can be obtained within two to three hours.

5.4.3 Slow-growth simulations generate equilibrated product from reactant structures

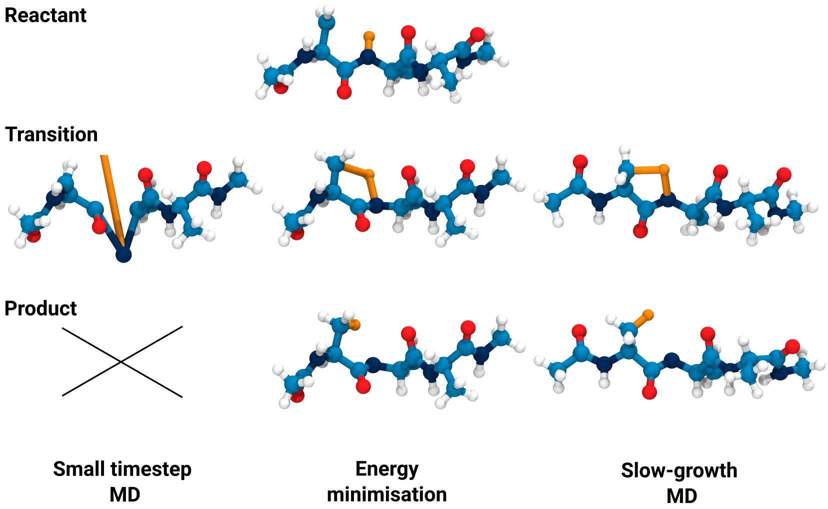


Figure 5.4: **Transition schemes from reactant to product state.** A HAT reaction in Ace-Ala$_3$-NME is modelled alchemically with small time step MD, energy minimisation and slow-growth MD. The HAT hydrogen is depicted in orange, as well as the bonds to H-donor and H-acceptor atom. Using small time step MD for HAT leads to simulation failure because the hydrogen is no longer bound to the H-donor nitrogen and strong VdW repulsion drives the hydrogen away. Energy minimisation and slow-growth MD lead to similar transition and end structures. All atoms are propagated through time during the second option.

Another MD simulation method to change a molecular system from one state to another is called "slow-growth". Here, force field parameters for both states are linearly interpolated by using a coupling parameter λ, which amounts to slowly changing from the state A Hamiltonian to state B Hamiltonian,[149]

$$H_\lambda = (1 - \lambda)H_A + \lambda H_B. \tag{5.4}$$

The slow-growth method is typically used to calculate free energy differences using thermodynamic integration.[149] Applications of thermodynamic integration include drug discovery or interacting molecules in general[150]. It should be noted that the pathways between both states are alchemical and do not accurately model the reaction pathway.

Starting from a reactant state A conformation, it is thus possible to obtain a product state B conformation in a continuous MD trajectory that only differs from a classical simulation at the reaction centre (Fig. 5.4). In theory, general equilibrium properties are not affected at all and the system only needs to relax due to PES differences in both states. This would reduce the task of optimising the equilibration and sampling times for a reaction to only one relevant property, *i.e.* the convergence of reaction rates. In small example systems, this method works reliably to obtain product state structures.

In practice, there is no straightforward way to interpolate perfectly between reactant and product state of chemical reactions in GROMACS. Bonded terms can be individually accessed and start and end parameters defined. However, non-bonded parameters can only be manipulated per atom, meaning that non-bonded interactions of a particle are either entirely turned off or on. Options exist to distinguish between intramolecular and intermolecular non-bonded interactions but are tailored toward the usual drug discovery application. non-bonded parameter changes are crucial because interactions of atoms connected by a certain number of bonds are partially or completely excluded and changing bonds also impact these exclusions. An alternative is to define all changing non-bonded interactions as pair interactions, which can be interpolated.

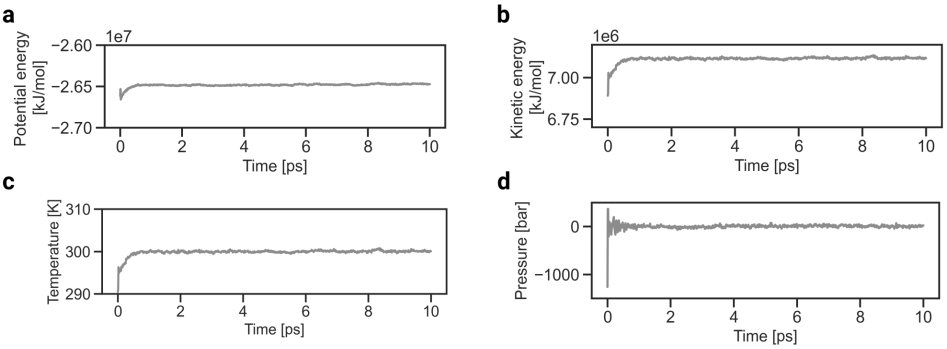


Figure 5.5: **Collagen fibril slow-growth simulations.** General equilibrium properties **a**, potential energy, **b** kinetic energy, **c** temperature and **d** pressure for a 10 ps slow-growth simulation. Note the different time scale and y-axis limits compared to Figure 5.2.

Even for a reasonable start and end parameter set difficulties related to non-bonded interactions can arise. Their highly non-linear nature can lead to sudden jumps in energy especially at the start and end of a slow-growth simulation, which lead to the development of soft-core potentials.[149] Unfortunately, we found no way to use these soft-core potentials for alchemical transitions for reactions. Another detail likely related to the slow-growth implementation in GROMACS is the reinitialisation of thermostat and barostat in the beginning of the simulation (Fig. 5.5), leading to a short equilibration period. All in all, despite the currently not realised potential of the slow-growth implementation in GROMACS for alchemical transitions for reactions, it is still a convenient approach and implemented in KIMMDY.

5.5 HAT GNN prediction

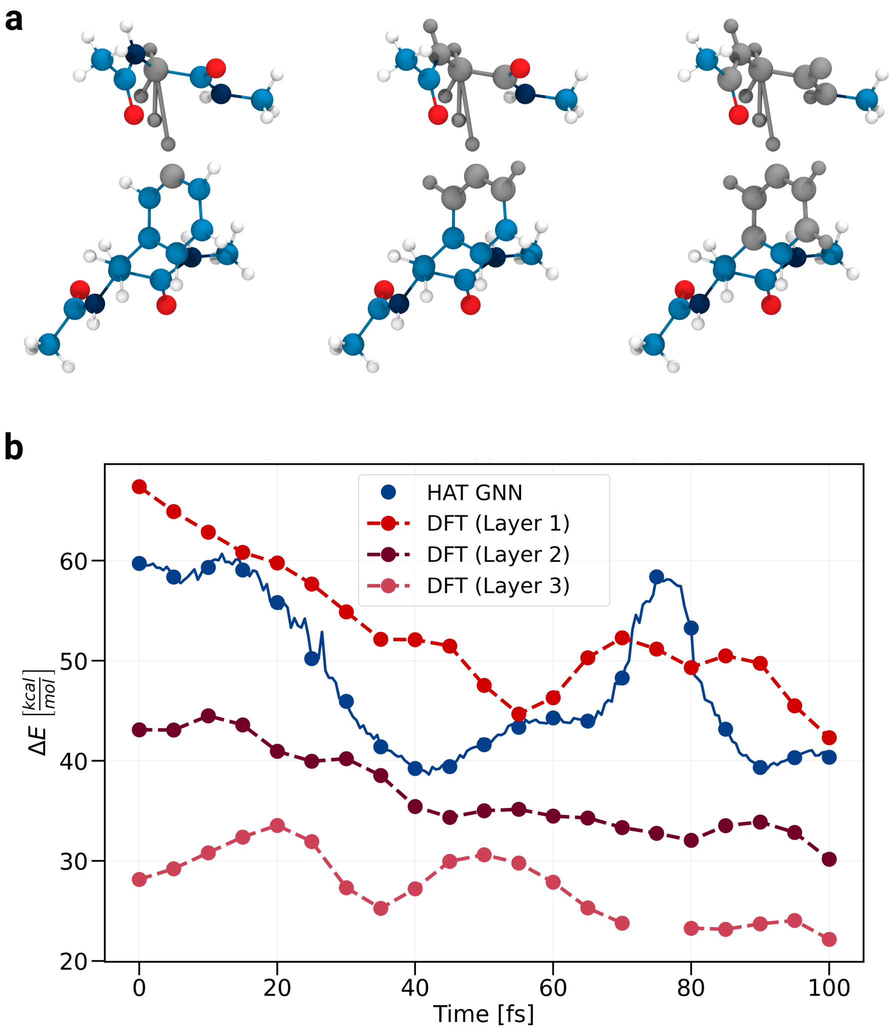


Figure 5.6: **HAT barrier prediction. a**, Start, transition and end structure for a HAT from a glycine Cα to the Cϵ radical atom of phenylalanine. Atoms that are free to move during a QM saddle point optimisation are shown in gray for freezing atoms after bond layer 1 (left), layer 2 (centre), layer 3 (right). The shown structures correspond to overlays of input structures for the HAT GNN and interpolated transition state, *i.e.* the reacting hydrogen is depicted in the three states while all other atoms remain in the same position. **b**, Predicted barriers for the reaction depicted in panel a over 100 fs. Prediction are conducted every 0.5 fs for the the HAT GNN and every 5 fs for QM optimisations of start, transition and end structure. Data and visualisation for panel B were generated by Evgeni Ulanov.

5.5.1 The HAT GNN delegates conformational sampling to MD

A hydrogen atom transfer (HAT) GNN has recently been developed for predicting reaction rates based on MD structures (see 3.5).[45] This is the method that will receive the generated ensemble and predict HAT rates for the reactive molecular simulations. The application is conceptually different from homolytic cleavage[12] because the rates are not derived from an ensemble property like bond distances but are predicted per frame using a machine learning model. Hence, the model is called multiple times per reaction for a potentially high computational cost and the rates need to converge, not the property that is used to derive the rates.

In contrast to transition state search with the aim of finding the minimum energy path from reactant state global minimum to global state minimum and the energy barrier with it,[151] the HAT GNN learns the barrier for an instantaneous transition to the product state given a reactant state structure. The reaction rate is calculated from the barrier using the Arrhenius or Eyring equation (see 3.4). This setup simplifies the transition state search to a local saddle point optimisation with a reduced mobility of atoms not directly involved in the reaction. The overall system flexibility is accounted for through the MD sampled input structures for the HAT GNN. Prediction of barriers based on reactant structures (or only physicochemical properties) has been shown for small organic molecules. [152] In the context of protein HAT such a prediction would be extremely difficult because the GNN would have to learn the geometric restraints on transition states from the conformational flexibility of HAT sites within the protein environment. A local saddle point optimisation is implemented by freezing all atoms except for donor- and acceptor heavy atoms and the hydrogens connected to them during the optimisation (Fig. 5.6a, left).

To get an understanding on the implications of freezing atoms during the QM optimisations, a model system is investigated with a different set of frozen atoms (Fig. 5.6b). The molecular system was generated by abstracting random hydrogens in a collagen fibril and simulating it for a short period of time. The particular reaction was chosen for having a low barrier during the simulation. The prediction target for the HAT GNN is the layer 1 QM calculation. In this example, the MAE is 6.25 kcal/mol, double the error than reported for the HAT GNN model test set of 3.15 kcal/mol. Interestingly, the error is not systematic but the HAT GNN fluctuates around the layer 1 DFT calculation. Both layer 2 and 3 have drastically lower barriers with MAEs of 11.13 kcal/mol and 19.85 kcal/mol, leading to rate differences in the order of 10^8 and $10^{14}\,\mathrm{s}^{-1}$, respectively. Also, the barriers of layer 2 and 3 QM calculations fluctuate less than layer 1 and the HAT GNN prediction, which reduces the required frequency of predictions to accurately sample the rate over time. The choice for training on layer 1 QM calculations was made to prevent artifacts from atom movements that would not be possible in the whole fibril structure[45] but this comes at the cost of severely limiting concerted heavy-atom movement during the hydrogen atom transfer from H-donor to H-acceptor. As a consequence, the chance of finding the minimum energy path is slim and conformational sampling is delegated to the MD simulation (Fig. 5.7) although the functional form may not allow for structures that stabilize the transition state.

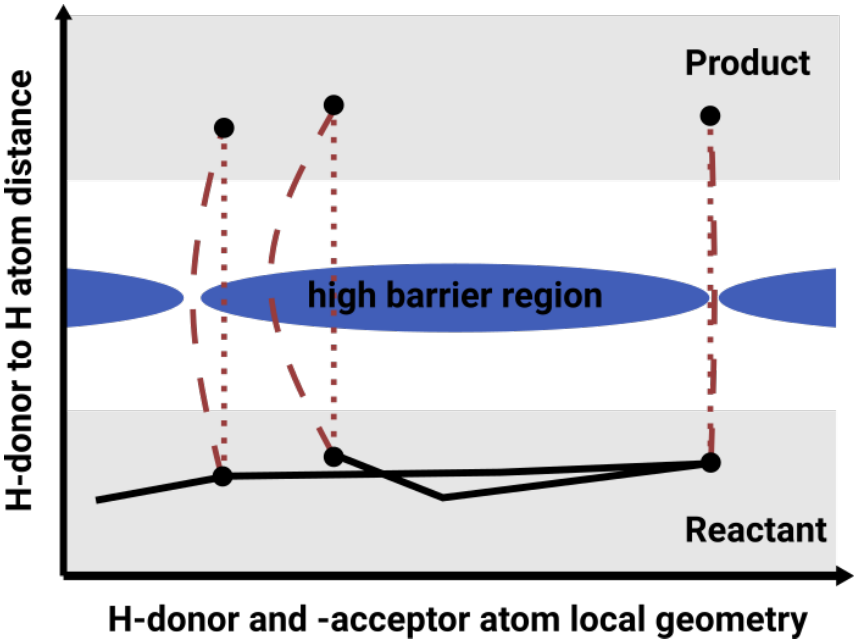

Product
high barrier region
Reactant
H-donor to H atom distance
H-donor and -acceptor atom local geometry

Figure 5.7: **Combined MD and ML transition path search with the HAT GNN.** The scheme shows a conformational landscape in the space that can be sampled by classical MD *i.e.* the local geometry of H-donor and acceptor on the x-axis and the reaction coordinate that separates between reactant and product state on the y-axis. Reactant and product state are shaded in grey. A MD trajectory (black line) with multiple structures (black dots) is the starting point for HAT rate predictions (red lines). For predictions with a HAT GNN trained on barriers calculated with many frozen atoms, the local geometry does not change and only the HAT hydrogen moves from H-donor to H-acceptor (vertical dotted red line), resulting in a transition path through high barrier regions (blue). Optimisations with a more flexible local geometry would find low-energy transition paths (dashed red line) for different reactant conformations.

A combined MD and ML transition path search also means that the starting structure of the ML prediction is not necessarily in the global minimum of the reactant state (Fig. 5.8a,b). Hence, the HAT GNN barrier $E_{\mathrm{ML,HAT}}$ is not the total barrier E‡, because the $E_{\mathrm{MD,conf}}$ contribution is missing. As the HAT GNN is trained to predict barriers from QM optimised minima to transition states, with some atoms frozen, the constant rate r_i for the reaction i can be seen as sum of rates for different states, weighted by the probability of being populated,

$$r_i = \sum_{j}^{states} p_j r_{i,j}. \tag{5.5}$$

An adjacent argument can be made using a potential of mean force[153] approach. The probability of a state is related to its energy by

$$p_j = \frac{1}{Z} e^{-\beta Ej}; \qquad Z = \sum_{j}^{states} e^{-\beta Ej}. \tag{5.6}$$

The energy difference between two states can thus be expressed as difference in their probabilities

$$\Delta E = E_j - E_k = -kT \; ln \frac{p_j}{p_k}. \tag{5.7}$$

For the practical use in simulations, states i and j can be defined as bins along a coordinate.[153, 154] Defining $\Delta E = E_{MD,conf} = E_j - E_{min}$, the missing part of the energy barrier in Figure 5.8 could be calculated. The above description of states in equilibrium and the potential of mean force approach lead to a similar prefactor of p_j and $\frac{p_j}{p_{min}}$ for the calculation of the constant rate r_i, respectively.

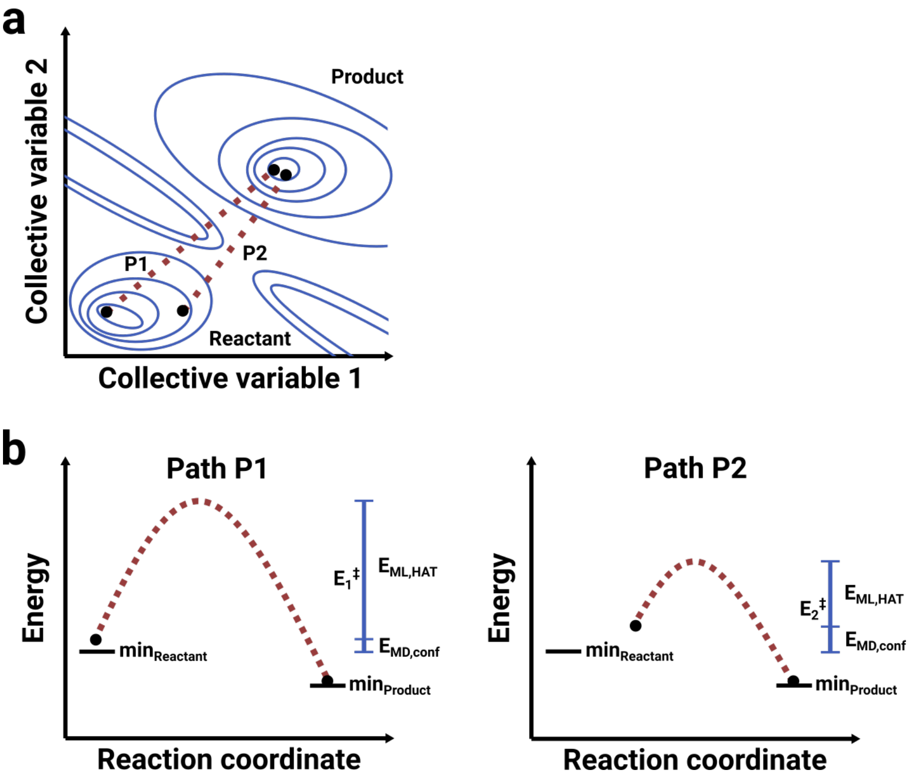


Figure 5.8: **The HAT GNN barrier does not account for the whole activation energy.** **a**, 2D energy landscape showing reactant and product state. MD sampled states (black dots) are the starting points of HAT rate predictions (red dotted lines) along different paths. **b**, Two reaction paths along the reaction coordinate. The total barrier E‡ is different and its constituents $\mathrm{E}_{\mathrm{MD,conf}}$ and $\mathrm{E}_{\mathrm{ML,HAT}}$ contribute a varying fraction.

Interestingly, both approaches do not only consider rates from crossing the barriers on the minimum energy path but for the ensemble of reactive trajectories in transition tubes[155]. Since the rates are related to barriers exponentially, the lowest energy path likely contributes most to the constant rate.

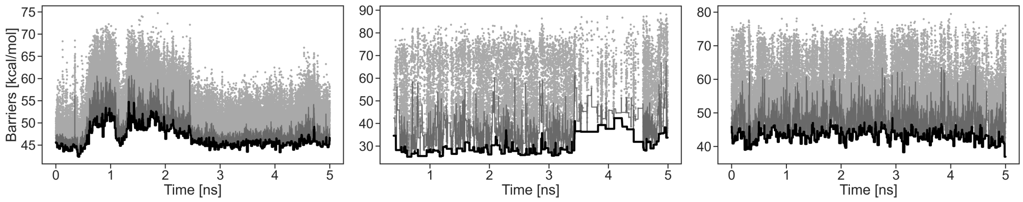


Figure 5.9: **HAT rate fluctuations over time.** Three reactions were chosen for their high number of HAT GNN predictions (translation distance < 3 Å) in a 5 ns MD trajectory with a sampling frequency of 100 fs. The predictions are shown as light grey dots. A 20 frame running minimum, corresponding to predicting the lowest barrier within 1 ps of atom movement, is shown in dark grey and a 200 frame running minimum in black. Data generated from a collagen fibril simulations with randomly abstracted H-atoms.

To sum up, the previous development of a HAT GNN allows for the prediction of instantaneous barriers from MD trajectories. The current model is trained on optimised structures with only layer 1 unfrozen atoms, delegating all conformational sampling to MD. Further investigations are necessary to determine the ideal stiffness, *i.e.* constraints used in the training dataset of the GNN, for an efficient prediction of low-energy transition paths. Learning atom movements within 1-10 ps would already smoothen the predicted barriers down to very few required predictions (Fig. 5.9). Further theoretical work is also necessary for connecting the new possibility of predicting instantaneous barriers over time to the existing theory on transition paths.

5.5.2 Many samples are necessary to find low barriers

Next, we want to survey the dependence of mean reaction rates on the number of sampled conformations. Using a collagen fibril system with six radicals, 100 000 frames spanning a simulation time of 55 ns are analysed (Fig. 5.10).

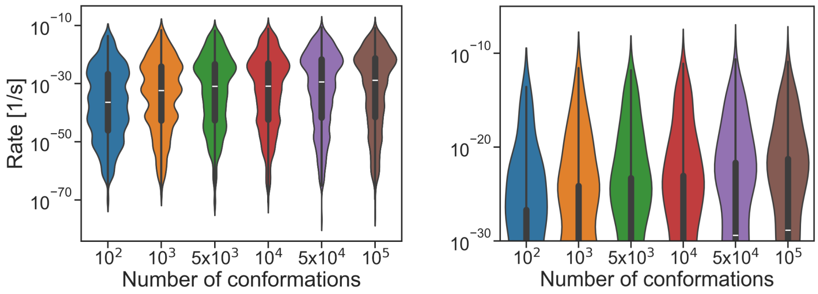


Figure 5.10: **Impact of conformations on mean rate.** For three simulations with 500 reactions the mean rate is calculated per reaction. From the initial 100 000 frames, every nth frame is taken for predictions with a reduced number of conformations. Inside the violin plot that represents the rate distributions, a boxplot is shown with the white line denoting the median, the box showing the interquartile range and whiskers indicating the most extreme value. The right panel is a cutout of the left panel.

From 1000 conformations upwards the median increases by 2 orders of magnitude for a tenfold increase of sampled number of conformations without reaching a plateau. In contrast, the highest rate, 90%ile and 95%ile are equal for the last two observations, indicating the start of a plateau for the most relevant part of the rate distribution.

Viewing the data on the linear instead of logarithmic scale makes the diminishing effect of additional sampling clear but certainly the obtained reaction rates can not be seen as independent of the sampling size. However, it would be exceedingly computationally expensive to obtain orders of magnitude more conformations to further quantify the sampling effect. Given these constraints, the prediction of rates from 50 000 conformations is a reasonable trade-off between accuracy and speed.

With the number of conformations set, the impact of the overall sampling time can be examined. Three collagen fibril datasets with the same radicals but different simulation time and sampling frequency were examined. The 500 ps simulation has a sampling frequency of 10 fs, and the 5 ns and 50 ns simulations have a one order of magnitude higher writing frequency each for a total of 50 000 frames per dataset. All three datasets show a similar distribution of barrier of mean rates per reaction (Fig. 5.11). The sampled minimum barriers are all around 30 kcal/mol but the number of reactions doubles for a tenfold increased simulation time.

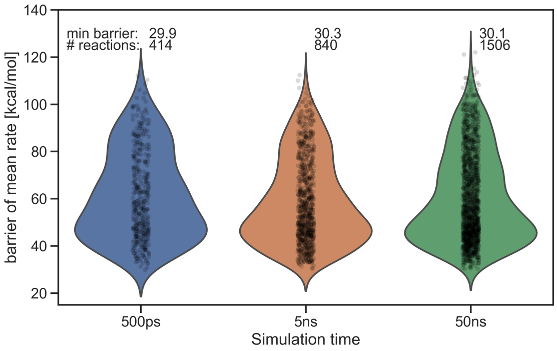


Figure 5.11: **Distributions of HAT prediction datasets.** For three datasets of 500 ps, 5 ns and 50 ns simulation time and 50 000 frames each, the barrier of mean rate distributions are shown. The barrier of mean rate is defined as $\Delta E_i = -RT \ln\left(\frac{\frac{1}{N}\sum_j r_{i,j}}{k_B T/h}\right)$. Individual barriers are shown as black dots.

A reaction-wise comparison of the datasets reveals that mostly the same reactions are sampled irrespective of the simulation time (Fig. 5.12a). Especially for the comparatively shorter simulations, almost all reactions are also found in the other simulation, *i.e.* the intersection of both sets of reactions. For the longer simulation, a few reactions with low-barriers are exclusively sampled but most are above 40 kcal/mol. When all conformations with a translation distance above 2 Å are filtered out, the number of sampled reactions halves (Fig. 5.12b). For these most relevant barriers, the MAE between datasets is 4 - 5 kcal/mol, only slightly above the HAT GNN MAE of 3.15 kcal/mol. Taking the computational cost of simulating for 10 or 100 times longer into consideration, the 500 ps simulations seem long enough for capturing possible reactions of a state. However, this choice may lead to underrepresentation of reactions that require large-scale molecular movements.

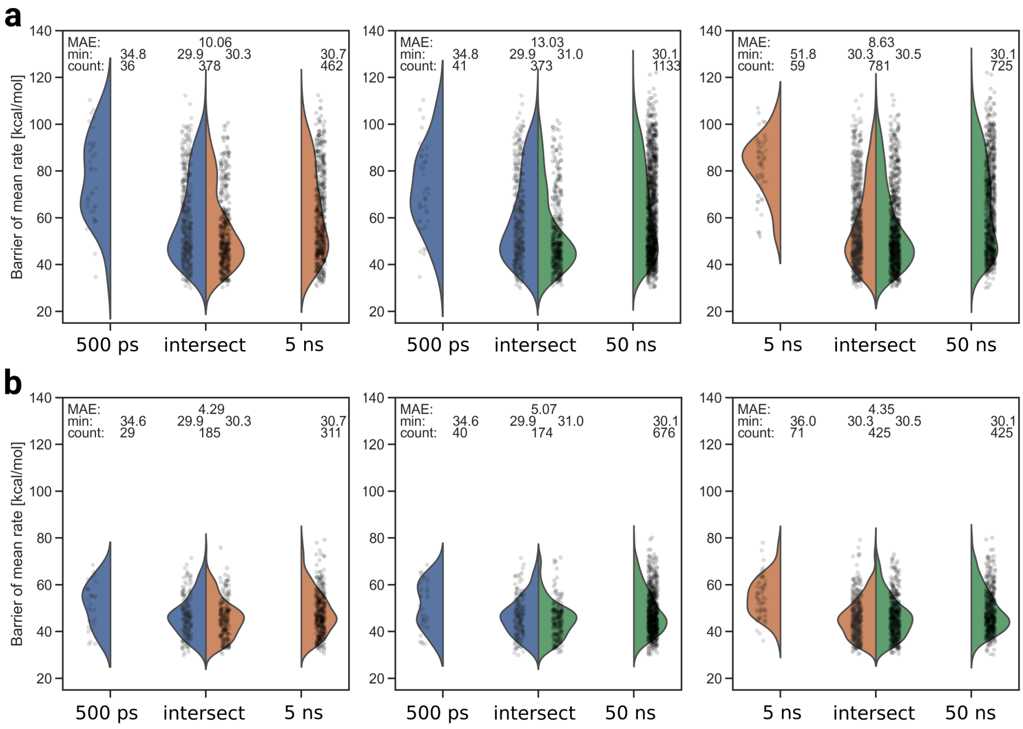


Figure 5.12: **Reaction-wise rate comparison.** Barrier of mean rates for comparing two simulation datasets, respectively. For each dataset, the barriers are split into those not observed in the other simulation (left,right) and the intersection with observations in both datasets (center). **a**, Barrier of mean rates for all predicted reactions. **b**, Barrier of mean rates for reactions with at least one conformation with a translation distance below 2 Å.

It should be noted that this extensive sampling of more than 1000 different reactions for a total of 150 000 frames produced minimum barriers of mean rate of around 30 kcal/mol. Compared to the 21 frames analysed for a single reaction in Figure 5.6b, which contained a 30 kcal/mol barrier for the layer 2 calculations and a 23 kcal/mol barrier of the layer 3 calculations, this is unexpectedly high. In the extensive dataset, radicals were created by homolytic cleavage of bonds that are likely to break and thus good at stabilising radicals. The exploration of layer freezing effects was on a random H-abstraction radical at a phenyl group, which is an unstable radical. Still, it seems unlikely that the sampled HAT barriers include the minimum energy path for the respective simulations. Absolute rates and barriers are therefore not meaningful and the interpretation should be focussed on relative rates and barriers.

5.5.3 The number of HAT predictions can be reduced

From the previous section it is clear that a large number of conformations needs to be generated per reaction to find low barriers. However, not every conformation is equally likely to harbour a low-barrier reaction. From the HAT GNN training data[45] and rate predictions in the collagen fibril (Fig. 5.13a) a correlation between translation distance (see 3.5) and HAT barrier is apparent. This effect can also be seen for individual reactions (Fig. 5.13b).

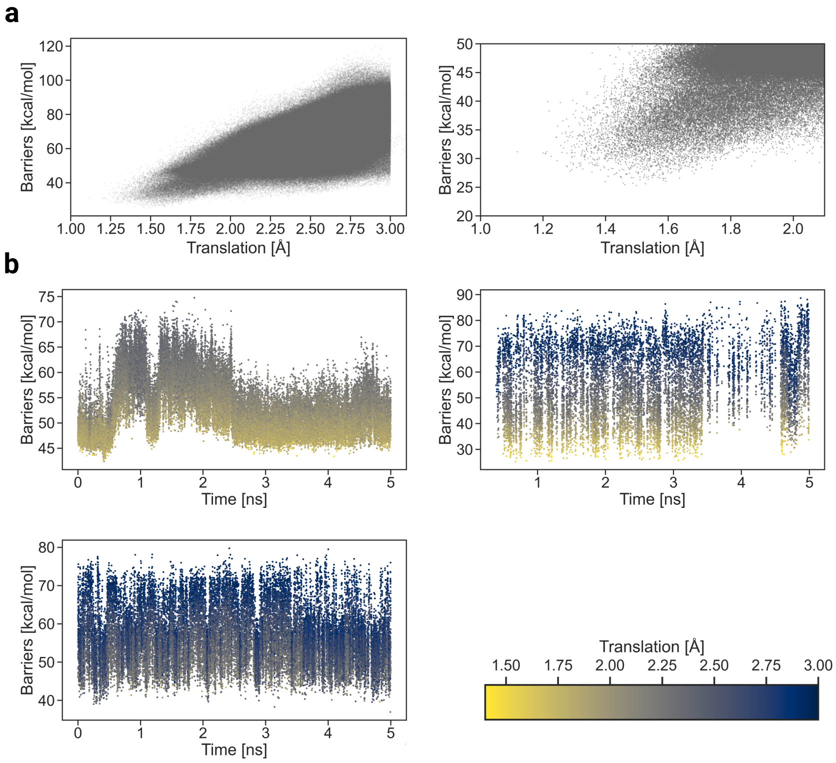


Figure 5.13: **HAT barriers depend on the translation distance. a**, Barrier over translation distance for 960 000 individual predictions. The right panel is a cutout region of the left panel. **b** Barrier over time for the same examples shown above. Individual predictions are shown as points and are coloured by the translation distance.

The barrier-translation correlation can be used to perform predictions only for a limited number of small translation distance conformations per reaction. In this approach, the conformations that were not used for predictions are assumed to have a rate of zero. For a collagen fibril test dataset with 7 5 ns simulations with 50 000 conformations each, most reaction rates are accurately predicted using only a fraction of the conformations (Fig. 5.14a,b). For example, predicting the rates for the 100 conformations with the smallest translation distance, *i.e.* 0.2 % of the conformations, 90 % of the rates are within 1.1 % of the rate predicted from using all conformations.

In the test dataset, and likely most applications due to the exponential relation between barrier and reaction rate, a few reactions contribute most of the total rate $r_{tot} = \sum r_i$. The choice of which reaction happens in the KMC algorithm depends on the fraction $\frac{r_i}{r_{tot}}$ and thus it is most important that the relatively high rates are accurately predicted. The rate fraction of reactions with the 10 highest rates per simulation are accurately reproduced using only the 100 conformations with smallest translation distance (Fig. 5.15a). Because the rate fractions are a probability, the predictions can be compared with the

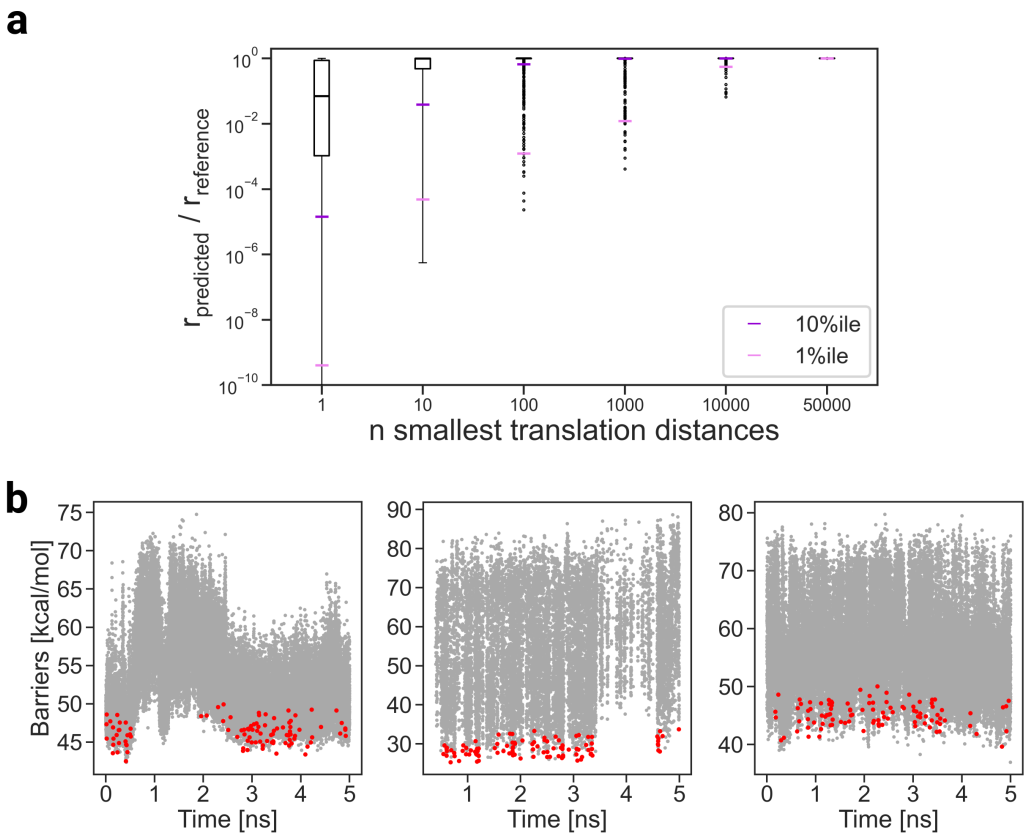


Figure 5.14: **Rate predictions of conformations with small translation distances.** **a**, Predicted mean rate for n smallest translation distances divided by the mean rate for all frames. Boxplots show the median and IQR, whiskers show the most extreme values within 1.5 times the IQR and outliers are shown as points. Data from 7 collagen fibril MD simulations with 50 000 frames and a total of 1084 reactions. **b**, Individual barrier predictions for three example reactions mentioned above. The barriers with the 100 lowest translations are shown in red.

divergence function of the Brier score[156],

$$d(p,q) = \sum_{i}^{reactions} (p_i - q_i)^2. \tag{5.8}$$

Compared to the baseline of predicting an equal probability for every reaction, which leads to a median divergence of 0.45, the prediction based on 100 conformations has a small median divergence compared to predicting on all conformations of 0.002 (Fig. 5.15b). To sum up, this mode of prediction is very accurate and reduces the number of predictions by several orders of magnitude. Unless mentioned otherwise, predictions for the 100 smallest translation distances per reaction are used after this section. The exact number of predictions per reaction is a hyperparameter and should be determined anew for every system and sampling time.

A further speed-up can be gained from implementing an idea from Xu *et al.*[48], who proposed to the limit rate calculations to rates within a fraction of

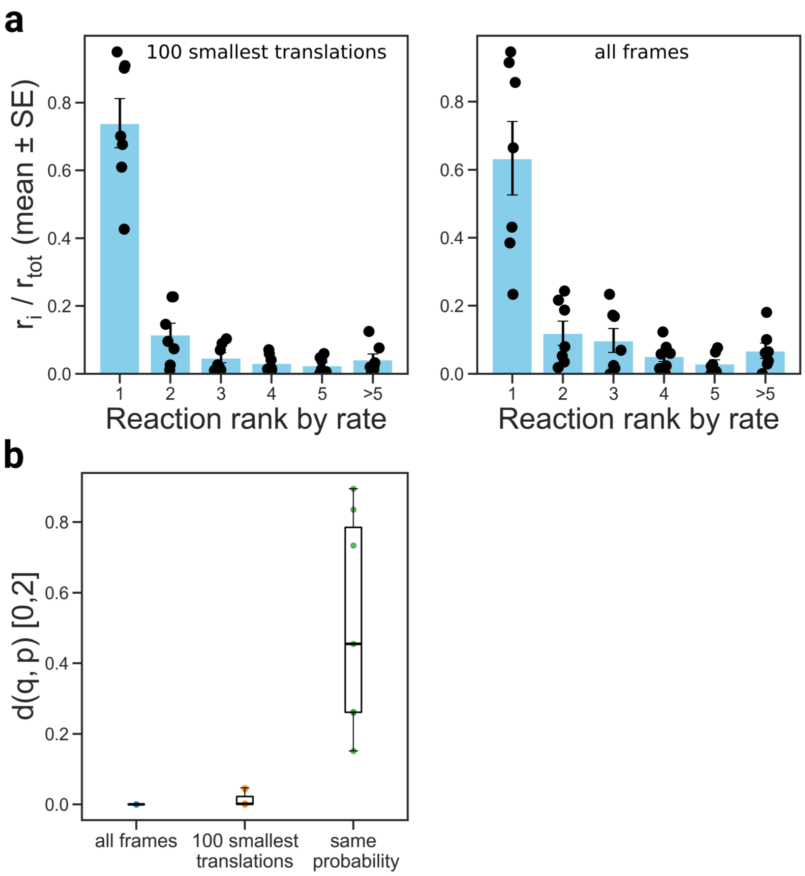


Figure 5.15: **Reaction probabilities are reproduced by efficient predictions.** **a**, The mean rate for a reaction i divided by total rate, *i.e.* the reaction probability in KMC, is shown. For the dataset of 7 simulations with 50 000 frames, reactions are ranked by probability for each simulation and aggregated over the simulations. On the left panel, the probabilities are shown for predicting on the 100 smallest translations, on the right for predicting on all frames. **b**, Divergence function of the Brier score for predicting on all frames, on the 100 smallest translations and for assigning all reactions the same probability. The divergence function can take values between 0 and 2.

the highest rate. For example, reactions with a rate of lower than a billionth of the highest rate have a likelihood to be chosen of less than one in a billion. For a transmission coefficient of 1 and at 300 K, this translates to a barrier difference of 12.35 kcal/mol using the Eyring equation. The barrier can not be known before the prediction but the HAT GNN is using an ensemble prediction with a MAE of 3.15 kcal/mol compared to a single model MAE of 3.55 kcal/mol. Given the single model barrier is predicted to be higher than 12.35 + 2*SD kcal/mol, it is not necessary to use the other nine models for the ensemble prediction.

5.6 The reaction search efficiency is improved

Having reduced the number of predictions the HAT GNN needs to do to for the same results, we want to see how this translates into the elapsed real time of a HAT GNN call. The timing is tested on simulations with 50 000 frames and six radicals. In the case of predicting rates for all HATs with hydrogens within 3 Å of a radical, 1 million rates are predicted. Using the ensemble prediction, this makes a total of 10 million predictions. About 4 reaction rates are predicted every frame for an average of 2 predictions every three frames per radical. The prediction of all sampled HATs takes 117 h on 20 CPU cores with a RTX 2080 GPU (Fig. 5.16). Taking only the conformations with the 100 smallest translation distances per reaction reduces the number of predicted rates to 8 000 - 10 000, and the prediction time to 13 h. Out of these 13 hours, all but a half hour are spent on the pre- or post-processing of the trajectory data.

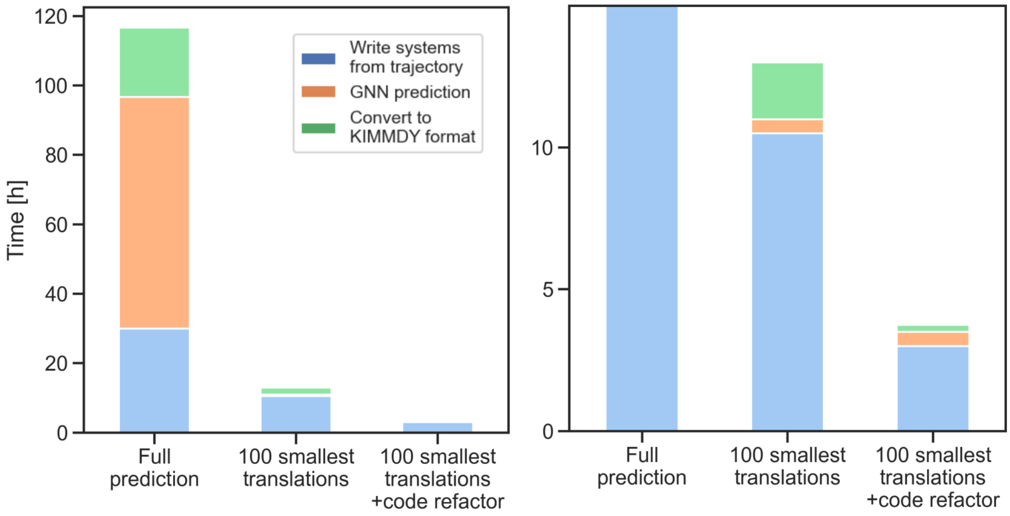


Figure 5.16: **HAT GNN walltime.** The walltime is split into pre-processing, *i.e.* parsing the MD trajectory and writing out subsystems centred around a radical and nearby hydrogen, the HAT GNN prediction and converting the results for the KMC simulation. Three different settings were tested, including the full prediction on all conformations with a hydrogen to radical translation distance below 3 Å, taking only the 100 smallest translation distances and the same with a refactored HAT GNN code. Data generated from a collagen fibril MD simulation with 50 000 frames and six radicals.

Most of the time is spent on accessing the trajectory file with a size of more than 50 GB through MDAnalysis[157]. The number of files the trajectory file is accessed can be reduced, which leads to a final walltime of less than 4 h. Still, most of the time is spent on preparing structures for predicting from the MD trajectory. A faster library could be used to further improve on the HAT GNN speed.

To sum up, HAT reaction rates for a MD sampled state of the collagen fibril can be predicted within 4 h. Choosing a sampling time of 500 ps leads to some reactions not being sampled compared to 50 ns simulations, likely those depending on larger motions for the hydrogen to come close to the radical. However, simulating for 500 ps for both the equilibration and sampling would lead to an overall time of the KMC step of 12 h. Compared to previous adaptive KMC applications[48], this is three to four orders of magnitude slower, significantly limiting the number of reactions that can be sampled. This increased computational cost is the price for accounting for the conformational flexibility of biomolecules and without this sampling, the transition path search would be incomplete.

Nevertheless, the current standard for reactive structure-based modelling in proteins, QM/MM[37], is typically used to study single reactions, albeit with higher accuracy and chemical transition mechanism modelling. The data generated from QM/MM simulations could be used to train a neural network in an approach similar to the HAT GNN but would be too slow by itself in a KMC setting. With this HAT sampling setup, we have all necessary parts for adaptive KMC simulations of HATs in collagen. Hundreds of competing HAT reactions can be sampled within hours to choose one reaction according to its probability.

5.7 Further optimisation

Multiple approaches to reduce the amount of necessary sampling for adaptive KMC have been published.[48] As mentioned previously, rates from previous KMC steps can be reused because they sampled the same state. In the context of HATs after homolytic cleavage, at least two radicals exist in a system. For every additional homolytic cleavage, two more radicals would be sampled. Hence, at least half of the reactions can be reused for every adaptive KMC step. Additionally, if the same forward and backward reactions are sampled multiple times because their rates are much higher than for other reactions, these states could be aggregated as ”superbasins”. Reactions out of either state would then be considered simultaneously and reactions within the same superbasin ignored.

Also, improved MD sampling, for example by enhanced sampling techniques[76, 158] could be used to sample more diverse conformations. In that case, the bias energy needs to be considered in the barrier calculation because states are no longer Boltzmann distributed.

Chapter 6

Implementing a transferable adaptive KMC software

Jannik Buhr, Kai Riedmiller and myself contributed equally to KIMMDY, the software described in this section. Parts of the text, that was solely written by myself, are also used for a manuscript that is currently in preparation with the authors Eric Hartmann, Jannik Buhr, Kai Riedmiller, Evgeni Ulanov, Boris Schüpp, Denis Kiesewetter, Daniel Sucerquia, Camilo Aponte-Santamaría and Frauke Gräter.[92]

6.1 Motivation

Now that the combined MD/ML method for HAT reaction search has been established, it is integrated into a KMC scheme for reactive molecular simulations. Mechanoradicals in collagen were first studied in an early version of KIMMDY that used a hybrid MD/KMC algorithm and could only perform the first homolysis step in a cascade of reactions following radical generation.[55, 12] In this work, the reactive repertoire of KIMMDY is extended to HAT reactions but more reactions would have to be modelled for an accurate description of the molecular system, *e.g.* the hydrolysis of peptide bonds that is competing with homolysis. Reactive pathways in other condensed phase systems with intramolecular reactions or that do not satisfy the assumption of well-mixedness can also be investigated with KIMMDY. Hence, we decided to create a transferable adaptive KMC method that facilitates the extension to new reactions and combines MD and KMC simulations with a firm theoretical background.

6.2 KIMMDY is an adaptive KMC method for reactive biomolecules

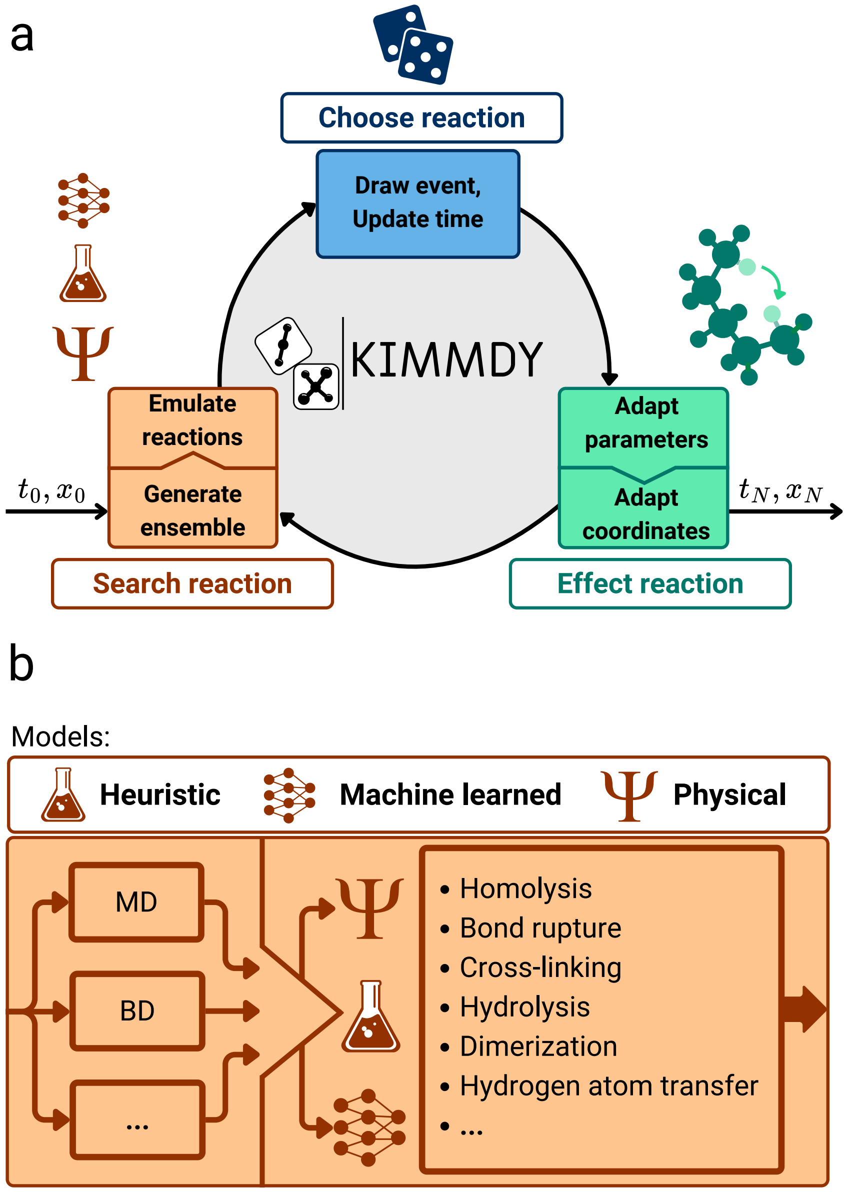


Figure 6.1: **A KMC step in KIMMDY. a**, Propagating a system to the next state and time using the KMC algorithm is divided into three modules in KIMMDY. The first module generates a list of possible reactions with the associated rates from a MD ensemble using a prediction model. A reaction and time step is chosen according to the KMC algorithm and finally, parameters and coordinates of the molecular system are adapted to the new product state. **b**, The search reaction module has multiple options for both ensemble generation and reaction emulation. The prediction model can be heuristic, machine-learned or physical. Visualised by Denis Kiesewetter in the context of Hartmann *et al.*[92].

6.2.1 Adaptive KMC

KIMMDY simulates a stochastic walk through the reactive state space[84] of biomolecules. It uses the rejection-free KMC (rfKMC) algorithm[84, 43], which consists of the following steps:

1. From an initial state, create an event list of all N possible transition rates,
2. Draw a uniform random number u_1 and select the event for which $F(p_{i-1}) < u_1 \leq F(p_i)$, where F is the cumulative function and p_i the probability of event i.
3. Draw a uniform random number u_2 to update the time according to $\Delta t = -\frac{\ln u_2}{R}$, where R is the total rate.
4. Carry out the event i.

Adaptive KMC[47, 48, 49] is a variation of rfKMC where the event list for a state is calculated only if it is populated during the KMC simulation. This approach is beneficial if the state space is too vast or the number of possible events per state too large to precompute the transition probabilities. In biopolymers and in general soft matter systems, most atoms exhibit unique reactivities because they are embedded in a certain structural context with electrostatics, solvent accessibility and steric effects. This dependence on the environment necessitates to calculate reaction rates individually for every set of reactant atoms, rendering their reactive simulation an ideal application for adaptive KMC. For a detailed description of KMC methods and especially adaptive KMC, see 3.3. In KIMMDY, a KMC step is divided into three modules (Fig. 6.1a). First, possible reactions are sampled by generating a conformational ensemble of the current state, which then is used to predict reaction rates to neighbouring states either by a heuristic, physical or machine learned model (Fig. 6.1b). The second module comprises selecting the event and a corresponding update time according to the KMC algorithm. Finally, MD simulation parameters and coordinates are adapted according to obtain the product state.

6.2.2 Sample reaction

For reaction sampling, a molecular system, described by coordinates and MD parameters, is simulated using MD to predict reaction rates for transitions to chemical states not sampled in the simulation. One approach to calculate the event list from the conformational ensemble generated with MD or by other means is by using ensemble averages of properties and relate them to reaction rates by physics-based or empirical models. This is the case for homolytic cleavage, where average bond distances are used to calculate the force on a bond and from this the dissociation energy barrier using a Morse potential.[55] For HAT, we use a machine-learned model to predict transition rates from individual snapshots to calculate a constant average rate per reaction over the whole ensemble. This has the additional benefit of accounting for entropic effects by sampling how frequent highly reactive conformations are visited. For the HAT application, using the conformation ensemble instead of a single structure representing a state, the prediction task is significantly simplified from emulating a complete transition state search to local optimization of the reacting atoms. Still, the MD sampling problem[158] may lead to an underestimation of reaction rates. The simulation setup is automated within KIMMDY and relies on user supplied simulation parameters. KIMMDY is designed as a framework to be extended to diverse reactions. To this end, a plugin architecture providing a stable interface is available.

6.2.3 Choose reaction

This module takes an event list and chooses a reaction. It then associates with this event a time update from all predicted reactions. We implemented different KMC algorithms in a modular fashion in KIMMDY. Here, only the rfKMC algorithm with adaptive event list generation is used.

6.2.4 Effect reaction

To effect the chosen reaction, the corresponding reaction recipe is applied to change the molecular system topology, parameters and coordinates to the product state. Recipes define reactions through elementary 'recipe steps'. 'Bind' and 'break' reference the two involved atoms and modify the MD bond definitions. Angles, dihedrals and pairs are modified accordingly. Changes of force field parameters are either handled by supplying a force field that has parameters for all reaction products or by re-parameterising the bonded parameters with the general machine-learned Grappa force field[91] combined with heuristics for the non-bonded parameters (Sec. 3.6). To generate the product coordinates, 'Place' moves an atom in a certain snapshot to a new position and, as an alternative, 'Relax' starts a MD simulation with the slow-growth feature of GROMACS[56] to interpolate smoothly between reactant and product parameters. For each predicted rate, the associated snapshot in the analysed MD trajectory is defined in the recipe to identify snapshots with high rate conformations. Recipes with identical recipe steps are aggregated because they denote a transition to the same product state. Finally, an equilibration MD simulation is performed to generate a start structure for the next reaction sampling step. This has the benefit of sampling from the product state Boltzmann distribution. Thus, KIMMDY models state transitions as a Markov process.

6.3 A framework architecture imposes the control flow and facilitates extensibility

The goal of KIMMDY as a software is to automate or give useful defaults for most parts of the adaptive KMC application. Only parts that need to be changed to enable a certain application shall be modified by the user. In the context of adaptive KMC, this means the sequence of sample, choose and effect reactions is set. Defaults for the choose reaction and effect reaction module exist but can be modified for special applications. The search reaction module specific to the reaction and can be defined without modifying the main KIMMDY code.

A framework architecture[159] is ideal to comply with these requirements. Key features of a framework architecture are inversion of control and extensibility that set a framework apart from a standard library. Inversion of control means that the control flow, here search, choose and effect reaction is defined by the framework and can not be changed by the user. *I.e.* if a reaction has been chosen, parameters and coordinates will always be adapted before the next reaction search starts. We allow the user to extend KIMMDY by providing abstract base classes for reaction rate prediction and force field parametersiation and using user-supplied MD configuration files. The exact input and output that is required for each module is programatically defined. Also, a library functionality is implemented to facilitate extensions by the user. Further utilities and resources help the user in using and understanding KIMMDY and retain a certain code quality.

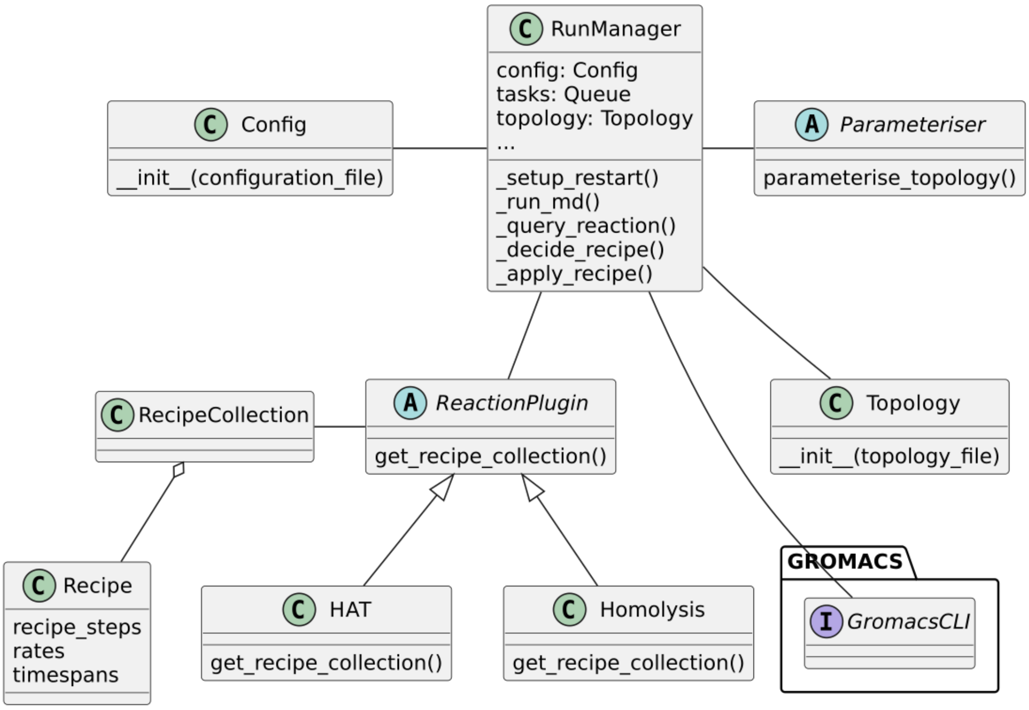


Figure 6.2: **A class diagram of KIMMDY.** Classes are represented as grey boxes with a C icon. Abstract base classes have a A icon and are a template for classes and define an interface through the definition of an abstract method. Interfaces are indicated with an I icon and packages outside of KIMMDY denoted as a transparent directory. Classes have attributes that are variables stored in a class object and methods that are functions defined within the context of a class. Both are shown as non-exhaustive lists below the class name. Lines indicate a relation between classes. Generalisations are special relations that indicate the general abstract class with an arrow. Aggregration is shown with a diamond and indicates that one or multiple class objects are aggregated in another class.

In KIMMDY, the RunManager is the central class (Fig. 6.2). It receives the run configuration and builds a queue of tasks. Tasks are called sequentially and receive the necessary information from the RunManager, *e.g.* a representation of the system connectivity and force field parameters, called topology, or a parameteriser that updates force field parameters. The ReactionPlugin is an abstract base class that provides the interface for reaction rate prediction extensions. It is a generalisation of the implemented reaction classes HAT and Homolysis. A Reaction is defined as Recipe with instructions on how to effect the reaction, called recipe_steps, associated rates and timespans for which the rates are valid. Recipes are aggregated in a RecipeCollection that is returned from a ReactionPlugin to the RunManager. The choose and effect recipe steps are functions of the RunManager.

6.4 Flow of control

Configuration file

While the control flow is fixed, the code that is run for a specific module is specified in the configuration file in YAML format. In YAML, files are a list of key-value pairs that are potentially nested or have scalars as value. Reaction rate prediction modules are defined under the 'reactions' key and MD simulation configurations can be defined under the 'mds' key. The defined modules can then be used to construct a sequence of MD simulations and reaction modules. Furthermore, KIMMDY run settings and simulation files can be defined in the configuration file.

Initialisation

The configuration file is parsed and validated during initialisation of the Config class. All files mentioned in the configuration have to exist on the filesystem. The GROMACS[56] version has to fit to the specified options, *e.g.* compatibility with PLUMED[160]. The specified Parameterisers and ReactionPlugins have to be installed.

For initialisation of the RunManager, a queue of tasks is generated from the sequence in the configuration file. Files containing information that tasks may need, like the topology file, are parsed and stored as an attribute of the RunManager.

RunManager

The RunManager manages the control flow. It starts the next task from the queue, provides it with input data and parses the output. The output data is processed to update RunManager attributes, add new tasks to the queue or update a list of current files. Tasks are methods of the RunManager.

6.5 Extensibility

Sampling interface

Currently, the ensemble generation is tailored around the MD engine GROMACS[56]. GROMACS allows for unbiased or biased simulations, especially using a version patched with PLUMED[160]. Biased simulations have an external force to drive a molecular system in certain conformations. Simulations with MLIPs[161] are planned for an upcoming release. We interact with GROMACS using the command-line interface, which has the benefit of being fairly stable between releases. However, different environments may require different GROMACS binary names and prefixes to the command that can be specified in the KIMMDY configuration file. MD runs can be completely customized within the boundaries of GROMACS because KIMMDY takes user-specified GROMACS configuration files (mdp files) as input.

In theory the ensemble generation could be performed with any method that generates a collection of conformations, including ML methods based on Boltzmann generators[162, 163] or flow matching[164]. Most of these methods require a template of the molecular system in the state for which an ensemble should be generated. For proteins, the predicted ensembles are either trained on classical MD simulations[164] or by finding multiple solutions to predicting experimentally resolved structures based on the amino acid sequence[165]. While this helps to generate ensembles more quickly, the fundamental limitation of long-timescale structural modelling to account for chemical reactions is not solved. Thus, KIMMDY synergises with novel ensemble generation methods for reactive modelling and integrating them into KIMMDY may lead to faster adaptive KMC steps.

Reaction plugins

The reaction module is implemented as a plugin and has to contain a class that inherits from the ReactionPlugin abstract base class. In KIMMDY, an entry point is defined for reaction plugins. Referring to this entry point in a python package makes KIMMDY recognize the reaction plugin.

The reaction class has to return a list of reaction recipes from an ensemble of structures (Fig. 6.3). Hence, it performs parts of the ”search reaction” step, specifically the ”emulate reactions” component. How a reaction plugin derives the rates is outside of the scope of KIMMDY but it contains functions and classes to help with writing the extension (see 6.6). So far, rate prediction models based on experimental heuristics, physical models and neural networks have been applied. The operations on the system connectivity and coordinates to effect the reaction are expressed as recipe steps (see 6.6).

Parametrisation plugins

After each reaction, the product state needs to be parametrised. In KIMMDY, two options are available: basic parametrisation and Grappa[91] parametrisation (see 3.6). All topology changes are based on the chosen reaction recipe.

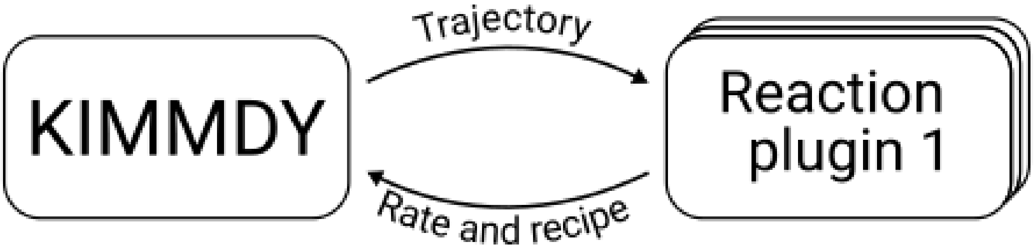


Figure 6.3: **The reaction plugin interface.** A ensemble of structures, called trajectory, is passed to the reaction plugin that has to derive rates and reaction recipes from it. "KIMMDY" refers to the software framework that does not include specific realisations of the reaction search. The reaction search code for a specific reaction is interfacing with KIMMDY through a plugin architecture.

Usage of the basic parametriser assumes that the product molecule parameters are contained in the force field included in the topology file, not modifying the topology beyond the instructions in the recipe.

Grappa parametrises around the atoms involved in the reaction. The whole structure is parametrised with Grappa either before a KIMMDY run or at the start, producing consistent parameters for any reaction. A basic distinction defined in the KIMMDY configuration file is between reactive and non-reactive molecules. By default, solute atoms are defined as reactive and are consequently parametrised with Grappa. In this case, solvent atoms are defined as non-reactive and have classical force field parameters.

KMC algorithms

Different KMC algorithms are available in KIMMDY and can be specified in the configuration file. In this work, only the adaptive variant of the rfKMC algorithm is used (see above and 3.3). All implemented KMC variants have in common that they have an event list of reactions with propensities/rate constants a_i (Fig. 6.4a). The probability of choosing an event is equal to $\frac{a_i}{\sum_j a_j}$ (Fig. 6.4b).

6.6 Library functionality

Recipes

A recipe provides the language for a reaction plugin to define which reaction occurs with which rate. It contains the rates and associated timespans of validity, as well as recipe steps as instructions for the effect reaction module. Recipe steps define the breakage and formation of a bonds and modification of all bonded and non-bonded interactions with it. Also, atoms can be placed at some coordinates or a specific MD simulation for relaxation requested that is executed as the next task.

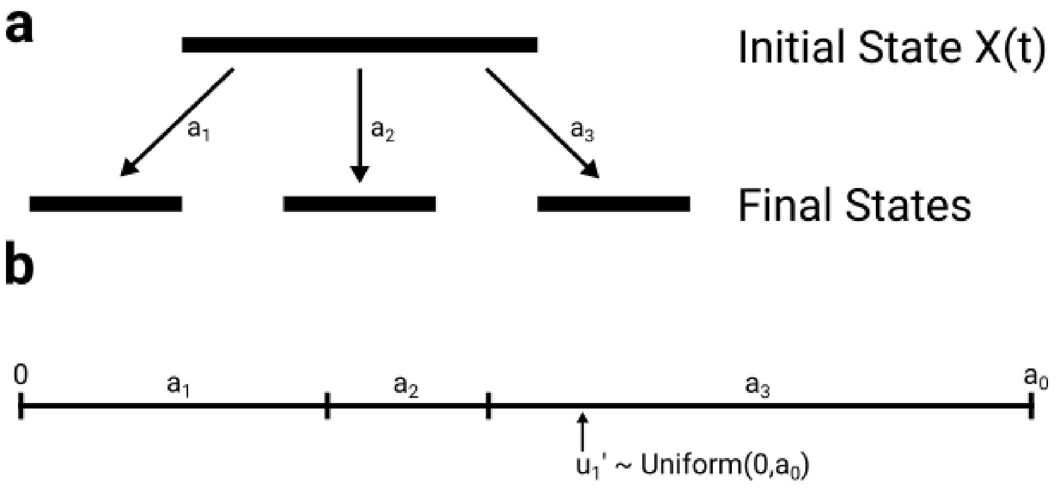


Figure 6.4: **Choosing a reaction.** **a**, An event list denotes all possible transitions to different final states. **b**, The selection probability depends on the fraction a reaction constant contributes to the total rate.

Parameter interpolation

For relaxing the system coordinates after a reaction, we use the slow-growth method implemented in the GROMACS free-energy module. It interpolates between product and reactant state for continuous parameter and coordinate changes. Subsequently, a short equilibration is necessary to ensure the states are Markovian.

Currently, exact continuations of MD simulations in GROMACS are only possible from checkpoint files, which offer very limited options for changing the molecular system. We use only reactant coordinates and velocities to continue MD simulations, which leads to a short period where thermostat and barostat equilibrate again. Another source of discontinuities is the parameter interpolation of non-bonded parameters, especially for Van der Waals interactions, that can quickly lead to high energies and forces due to the exponential nature of the potential. In practice, the impact of these discontinuities appears limited for applications tested so far but it should always be assessed for new reactions and systems.

Topology class

The topology class allows access to all individual bonded interactions and non-bonded parameters as well as the force field defined in the user supplied GROMACS topology file. This is helpful for parametrisation and the effect reaction module in general. Reaction plugins may also retrieve topology information for the rate prediction.

6.7 Utilities and resources

Checkpoints and restarting

A checkpoint system faciliates long KIMMDY simulations. Conveniently, GROMACS simulations can be restarted from checkpoints and the end of a simulation is a state where the current system state is written explicitly in files. Thus, we implemented KIMMDY restarting during and after GROMACS simulations. Files previously written by KIMMDY are parsed as well, to reproduce the current state of the KIMMDY simulation faithfully. All tasks started after the last GROMACS task can not be used and are discarded.

HPC infrastructure

KIMMDY simulations are easily parallelisable and require extensive resources, typically available on HPC systems. We added utilities to run KIMMDY on HPC systems using SLURM[166] and expose prefixes to the GROMACS commands necessary for running it in an MPI setting.

Analysis

We provide some analysis functions to visualise KIMMDY simulations. Energy terms that are recorded in the GROMACS energy file can be displayed over all KIMMDY steps. Radical populations can be shown by number of occurrences or KMC time in a plot or as a structure in VMD[167]. A visualisation of the migration pathway is implemented with PyMOL[168] and is shown in Figure 7.2.

Testing

The promise of a framework is for the user to only having to care about the specific extension they need for their application. While an understanding of the method is beneficial, the implementation should be reliable. Due to the flexibility of the framework, the code contains more abstractions and is more complex than necessary for a bespoke implementation and implementing an adaptive KMC algorithm may be easier than understanding the whole code base of KIMMDY. Thus, it is essential to test units and whole use cases for KIMMDY. We have implemented 120 tests to identify regressions but further work is needed to increase the coverage of KIMMDY tests.

Documentation and tutorials

Documentation and tutorials are available on `https://graeter-group.github.io/kimmdy/`. Sections of the documentation are automatically generated from docstrings in the code and explain how to interact with certain elements of the code. Others are curated and revolve around setting up or applying KIMMDY. Tutorials give a first usage example of KIMMDY.

File parsing

Functions to read (and write) a GROMACS topology and other GROMACS files, PLUMED files, JSONs and CSVs are implemented.

Chapter 7

Adaptive KMC in the collagen fibril

7.1 Motivation

KIMMDY is finally applied to HAT reactions in the collagen fibril. The main aim of this chapter is to find out whether DOPA can directly scavenge radicals from common homolysis sites. This would complete the radical detoxification mechanism from homolysis to hydrogen peroxide generation via DOPA or other phenoxy groups.

Before moving on to the analysis of HATs involving DOPA, specifics of the simulated model are detailed and an overview of the sampled reactions provided. As the simulations freely sample any possible HAT reaction, other radical scavengers in collagen are also investigated. Furthermore, enantiomerisation at Cα atoms resulting in the formation of D-amino acids is observed.

7.2 Collagen fibrils break at cross-links

Collagen is a structural protein that forms large hierarchical structures from triple helices. As a main component of several tissues, it is comparatively easy to perform mechanical experiments on. These can be modelled computationally with constant-force MD simulations (Fig. 7.1a,b) using an experimentally determined structure (PDB accession code: 3hr2[147]) and structural modelling[106]. Main features, such as gap and overlap regions are only accounted for in relatively large structures, making it computationally expensive to simulate for extended time scales.

Interestingly, previous studies have shown that pulling forces, in a collagen fibril model spanning one gap and overlap region, concentrate in few specific bonds, especially in or nearby enzymatically formed cross-links[169] between neighbouring triple helices.[12, 14] This force concentration has been associated with homolytic cleavage, producing mechanoradicals. The short arm (LY2) Cα-Cβ bond of PYD is one of these bonds acting as a sacrificial bond that breaks with high rates but keeps the two triple helices connected through the other two arms.[14]

In this work, the divalent cross-link HLKNL and the trivalent cross-link

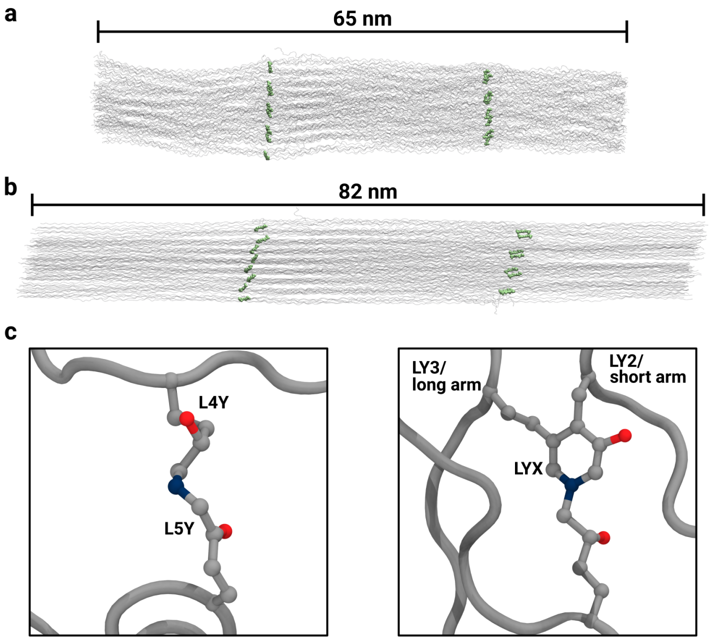


Figure 7.1: **Structure of the collagen fibril.** **a**, Relaxed atomistic model of type I collagen with a cartoon backbone representation. The HLKNL cross-links between gap (outer) and overlap (inner) region are coloured in green. **b**, Collagen model under 1 nN force per strand. The whole fibril extends from 65 nm to 82ṅm and the cross-links orient towards the pulling direction. **c**, HLKNL cross-link (left panel) and PYD cross-link (right panel). Force field amino acid names are shown with L5Y including the nitrogen in HLKNL. LYX in PYD includes the aromatic ring, LY2 and LY3 only the atoms up to the ring system.

PYD (Fig. 7.1c) are modelled (see 3.7). While general structural properties of collagen fibrils are accurately captured[106], the available collagen fibril structures represent a limited set of possible conformations and cross-link positions. For example, mechanical stretching has been shown to alter the structure of collagen fibrils.[170] Slight changes like these are not accounted for in the collagen fibril model but change the local environment of sacrificial bonds like the short arm Cα-Cβ bond of PYD. This can have a strong impact on the observed HAT reactions. Still, a stochastically exact time trajectory of reactive states for a plausible collagen fibril can be obtained using KIMMDY. This means not all HATs that could be observed in experiments will have high rates in the simulations shown here because the necessary conformations are not sampled but it should be possible to identify high rate HATs for the given set of accessible conformations.

7.3 Numerous different HAT reactions occur in collagen

To understand how the simulated HATs behave, they are first shown from a structural perspective. The HAT pathways analysed here start at four short arm PYD Cα-Cβ homolysis sites (Fig. 7.2). At one site, a DOPA anion residue is nearby and scavenges the radical (Fig. 7.2 left zoom-in). Apart from the initial LY2 Cβ radical, the pyridine-bound hydroxy group of PYD and a aspartate Cβ group also harbour a radical at one point of the simulation. A visualisation of HAT directions at a different homolysis site (Fig. 7.2 right zoom-in) shows back-and forth reactions with various groups, again including DOPA and PYD, as well as arginine, glycine and proline.

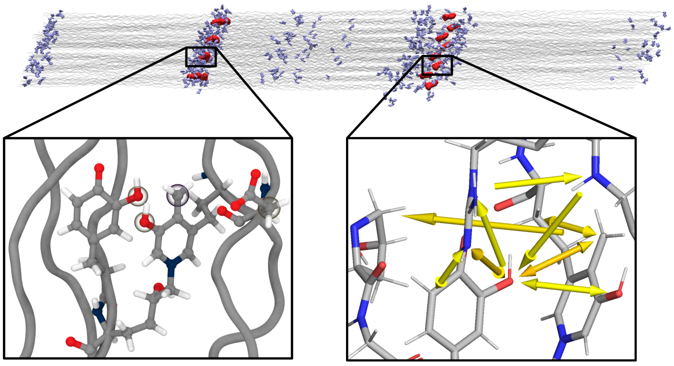


Figure 7.2: **HAT reactions in collagen.** In a collagen model with PYD cross-links (red) and DOPA anions (blue) shown, radical positions and migration pathways are visualised. The zoom-ins show different homolysis sites and were chosen because they include DOPA reactions. The structure was chosen for visual clarity and does not represent a high-rate structure for any particular reaction. Any actual distances between H-donor and H-acceptor in the ensemble for which a reaction was chosen may differ from this representation.

Overall, during a total of 600 HATs, many different reactions occur. Consistent with the HAT GNN validation on alkanes[92] and QM calculations[145], HATs occur seldom between bonded atoms (1-2 transfer) or for reactants that would have to form small ring systems during the transition state (Fig. 7.3)a. 1-6 HATs are unexpectedly rare and most HATs are between atoms that are many bonds apart or intermolecular. Most reactions were found only once during the 36 different simulations with 12 different homolysis patterns and 3 repeats (Fig. 7.3b), indicating a large diversity of possible reactions, also for the same chemical state. Almost every amino acid is involved in a HAT reaction as H-donor (Fig. 7.3c). Missing reaction involvement can be due to the amino acid being absent in the collagen fibril (tryptophan, cysteine), far away from homolysis sites or because other amino acids have higher rates as H-donors.

In addition to their diversity, HAT reactions take place in a large volume surrounding the original homolysis site. Their minimal distance from the Cα or Cβ atom of LY2 can reach more than 15 Å (Fig. 7.4). A sphere with a 15 Å radius in collagen would include roughly 1600 atoms, reaching the limits of traditional QM/MM reactive modelling[37]). From the same data it is apparent that most radical atoms are still Cα or Cβ atoms of LY2 after 20 reactions. Both sites are stable radicals and many HAT reaction products appear to be more reactive than the remaining Cα and Cβ atoms, leading to few sites with many consecutive reactions and other, stable sites.

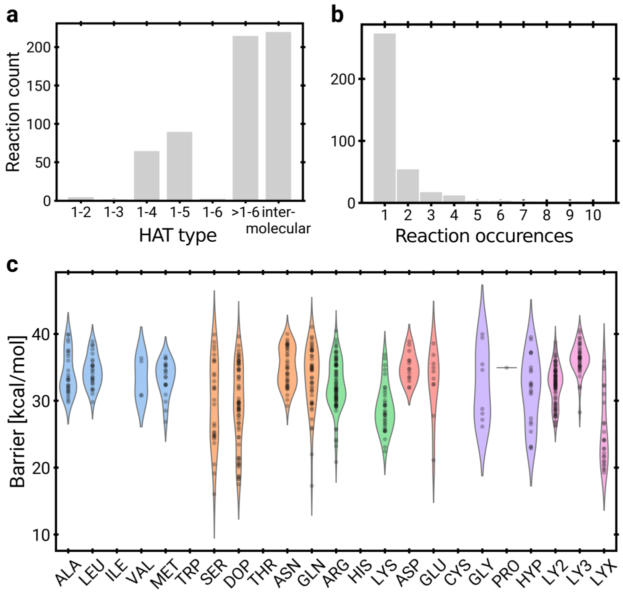


Figure 7.3: **Characterisation of HAT reactions.** **a**, Reaction count defined by the number of bonds between H-donor and H-acceptor. 1-2 HATs are between bonded atoms, 1-3 HATs have a single atoms between them *et cetera*. **b**, Number of times the same reactions occurred over all 36 simulations in the presented dataset. **c** HAT barrier distribution of reactions chosen with the KMC algorithm by H-donor amino acid. PYD is split into the three amino acids LY2, LY3 and LYX. Amino acids are sorted by physicochemical properties.

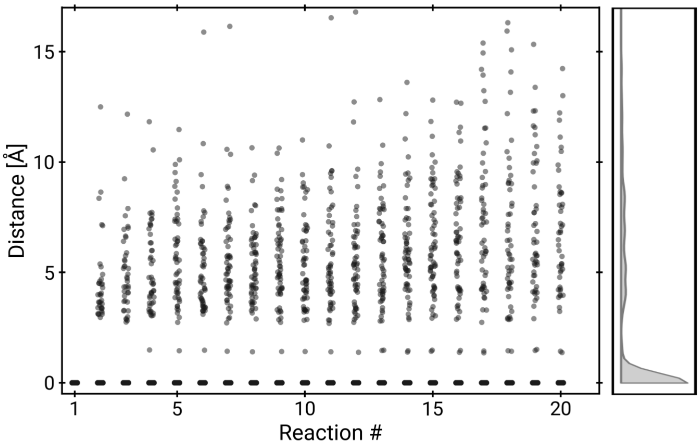


Figure 7.4: **Transfer distance of radicals.** The distance of radical atoms at a certain reaction number to the nearest Cα or Cβ atoms of LY2 that were involved in homolytic cleavage. The distribution is shown separately for every reaction step, starting with n=288 radicals and is decreased towards higher reaction numbers due to a few simulations ending before sampling all 20 reactions. The right panel shows the density for the 20th reaction.

7.4 DOPA radicals are kinetically accessible

The initial goal of this work was to establish DOPA as a kinetically accessible radical scavenger in the collagen fibril. DOPA is generated through post-translational oxidation of phenylalanine and tyrosine residues in collagen.[13] In the previous section, DOPA has been identified as a H-donor in HAT reactions. A radical scavenger should furthermore donate its hydrogen in a low barrier reaction but not easily accept hydrogen atoms again. This way, nearby radicals are quickly scavenged and then withheld from the environment until further detoxification steps occur.

To test this property of DOPA, reactions involving the hydroxy group as a H-donor are tested for lower barrier than non-DOPA reactions and hydroxy group H-acceptor reactions are tested for higher barriers (Fig. 7.5a). As expected, reactions with a DOPA hydroxy group H-donor have lower barriers and thus higher rates than non-DOPA reactions. The distribution appears to be long-tailed with many barriers around and below 20 kcal/mol. At 300 K using the Eyring equation, rates in the seconds to minutes range can be obtained. The median barrier difference is 4.5 kcal/mol, which amounts to 1800 times faster reactions. In contrast, H-acceptor DOPA hydroxy group reactions have no higher barrier than the reference. The median is even slightly lower than for non-DOPA reactions. A low HAT reaction barrier for reactions to a radical scavenger is not necessarily inconsistent, because HATs could frequently occur between neighbouring scavengers.

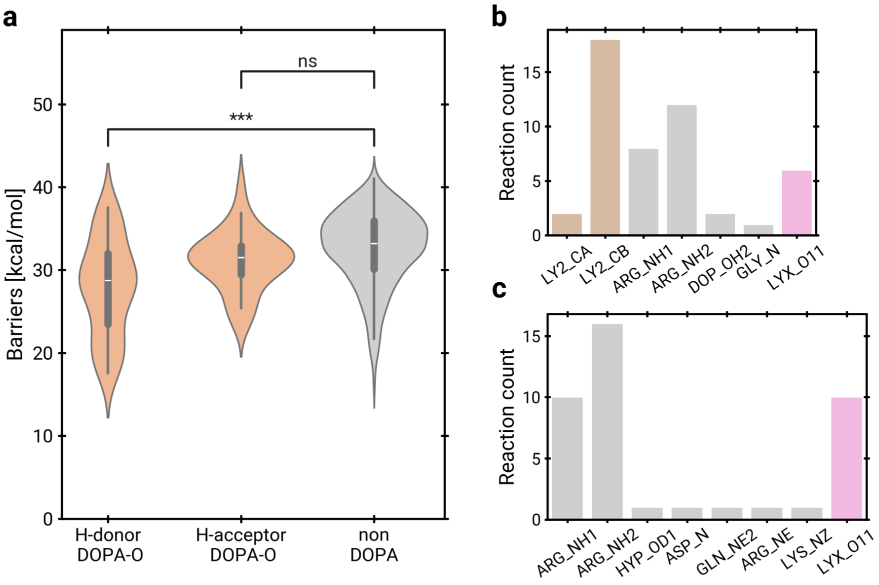


Figure 7.5: **DOPA is a radical scavenger.** **a**, Distributions of DOPA hydroxy group H-donor, H-acceptor and reactions without DOPA involvement are shown. The Welch's t-test is performed to test differences in the distributions and their significance is shown. The number of samples is n=49, n=41 and n=467, respectively. **b**, H-acceptors for reactions with DOPA hydroxy group H-donors. Homolysis radical H-acceptors are depicted in brown, PYD H-acceptors in pink and others in grey. **c**, H-donors for reactions with DOPA hydroxy group H-acceptors.

For reactions with a DOPA hydroxy group H-donor, the H-acceptor was often one of the two homolysis radicals, LY2 Cα and Cβ (Fig. 7.5b), confirming a direct transfer from the homolysis site to DOPA. H-donors for a DOPA hydroxy group H-acceptor include surprisingly many nitrogen atoms and, notably, the PYD hydroxy group at the aromatic ring (LYX O11) (Fig. 7.5c).

The high number of reactions involving nitrogen atoms is unexpected, as many have BDEs above the average protein BDE.[171] The thermodynamic property BDE correlates with reaction rates[172, 173], making the scenario of high BDE and low barrier HATs unlikely. Rather, the HAT GNN could overestimate HAT rates involving nitrogen compared to HATs with oxygen or carbon. Another possibility is the simplified protonation state modelling. pKa differences after HATs are not accounted for because the protonation state is only determined at the beginning of the KIMMDY simulation. Further investigation, ideally with QM data and accounting for pKa shifts is necessary to illuminate the role of nitrogen atoms for protein HAT.

Nevertheless, DOPA hydroxy group radicals are kinetically accessible and scavenge nearby radicals with high rates. The main aim of this work has thus been achieved. Caveats still exist, especially the generally low rate predictions of the HAT GNN (see 5.5.1) that obscure whether likely reaction paths have actually been found or are just exceedingly rare on timescales sampled by MD. The dataset generated here contains 6 000 000 evaluated structures and can increase the HAT GNN accuracy on application scenarios by supplementing the original dataset of 20 000 structure-barrier pairs if the QM energy barriers are calculated.

7.5 PYD and lysine also scavenge radicals

Previously shown data indicates that not only HATs to DOPA have low barriers and thus high rates (compare Fig. 7.3c). In a multiple testing scenario, lysine and the LYX part of the cross-link PYD also have a significantly lower barrier (Fig. 7.6). Lysine has been found to have the most stable Cα radical[171] but in the KIMMDY simulations most reactions involved the Cγ and Nζ atoms. In LYX, the hydroxy group in the aromatic ring acts as a frequent H-donor. This moiety has not been investigated as potential radical scavenger before but is ideally positioned, since it is always exceptionally close to the PYD short arm Cα and Cβ atoms (compare Fig. 7.2). If the hydroxy group is deprotonated upon homolytic cleavage of the Cα-Cβ bond, the Cβ atom, aromatic ring and deprotonated hydroxy group would form a resonant system, further stabilising the previously mentioned stable Cβ radical. Thus, PYD with a deprotonated aromatic ring hydroxy group has been identified as a plausible further radical scavenger in the collagen fibril.

Resonant systems stabilise radicals and are hence major objects of this investigation. In these systems, the unpaired electron density is distributed over multiple atoms and can engage in HAT reactions from there. For example, the DOPA radical anion (compare Fig. 1.4) can react with a H-donor at both oxygens. This effect is currently not accounted for in the HAT GNN. The same applies for the previously mentioned different protonation states that have a sizeable influence on BDEs and reaction rates. An improved HAT GNN model should include these effects.

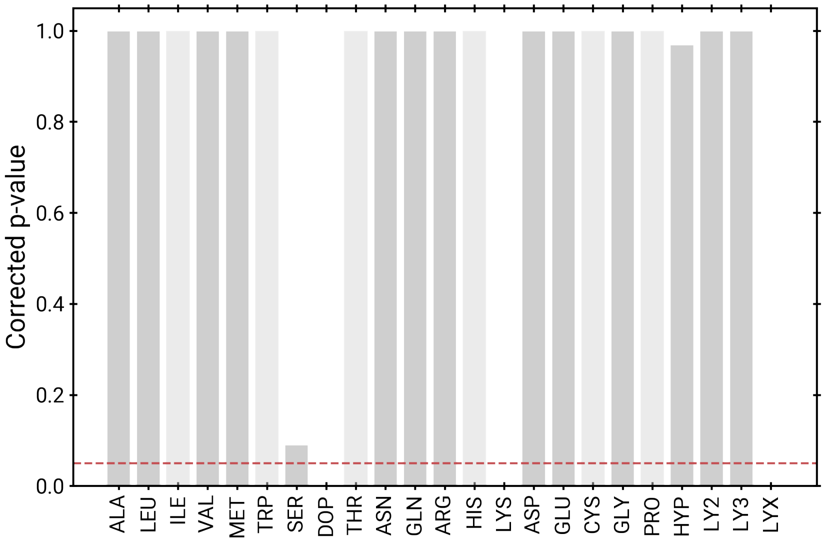


Figure 7.6: **Testing for further radical scavengers.** In a multiple testing scenario, the Welch's t-test is applied to test differences in the distribution of HAT barriers by H-donor amino acids. The p-value is corrected using the Benjamini and Hochberg approach to correcting the false discovery rate.[174] Amino acids with less than two HAT observations are assigned a p-value of 1 (light grey). All observed corrected p-values are shown in grey.

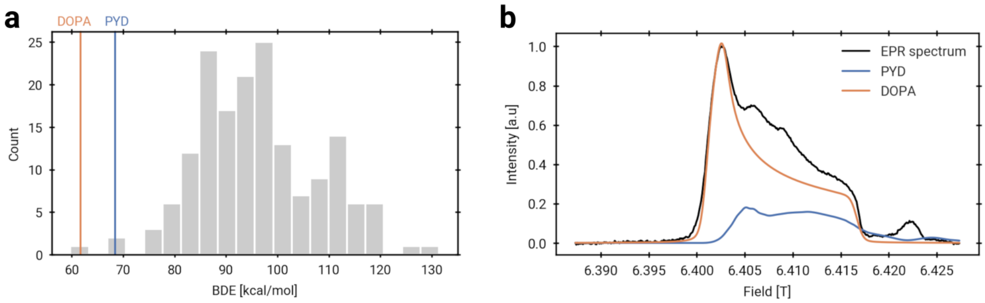


Figure 7.7: **Further investigation of the PYD radical.** **a**, BDEs of the DOPA anion and PYD with deprotonated aromatic ring hydroxy group compared to other protein BDEs from Treyde *et al.*[171]. **b**, EPR signal of a rat achilles tendon stretched at 14.7 N for 1000s with a simulated DOPA anion and deprotonated PYD spectrum. Experimental measurement from Kurth *et al*[13]. Data gathered, calculated and visualised by Daniel Sucerquia.

7.6 QM data and EPR experiments corroborate the scavenger role of PYD

Compared to other protein BDEs[171], PYD with a deprotonated hydroxy group has an exceptionally low BDE on par with the lysine Cα atom and less than 10 kcal/mol above the DOPA anion hydroxy group (Fig. 7.7a). This shows the strong radical stabilisation capabilities of PYD. Encouragingly, the EPR absorption spectrum of said PYD radical is within a region that has an experimental signal (Fig. 7.7b). The region above 6.405 T has an intensity that is currently not explained by the DOPA anion and can only partially be explained by other DOPA species. The peak of the PYD signal does not perfectly match the second-highest experimentally measured peak at 6.406 T but this could be attributed to environment effects and slight differences in the spin delocalisation.

As the PYD radical is closer to sites of homolytic cleavage but does not stabilise radicals as well as the DOPA anion, one explanation would be that PYD initially stabiles mechanoradicals until HAT transfer leads to a DOPA radical. The available experimental data does not prove the existence of stable PYD radicals in collagen but there is also no reason to deny this hypothesis. An experiment to shed light on the existence of stable PYD radicals would be time-resolved EPR[175] to identify a shift of a PYD radical population to the DOPA radical population.

7.7 Only homolytic cleavage products of PYD are near DOPA

Another observation on PYD is the different structural response after homolytic cleavage occurs. For PYD, a broken short arm Cα-Cβ bond still leaves the long arm as a connection to the same triple helix. The structure changes only locally. For divalent cross-links and backbone breaks, the broken ends move apart because the strain from one side is lifted, similar to a broken rubber band. The concentration of aromatic residues at the border between fibril gap and overlap region (compare Fig. 7.2) results in the mechanoradicals from backbone and divalent cross-link breaks being further away from potential radical scavengers (Fig. 7.8). Hence, non-PYD mechanoradicals are potentially less likely to be captured and can cause more side-reactions than PYD mechanoradicals. This kind of co-localisation is a pattern for ROS sensing systems.[6]

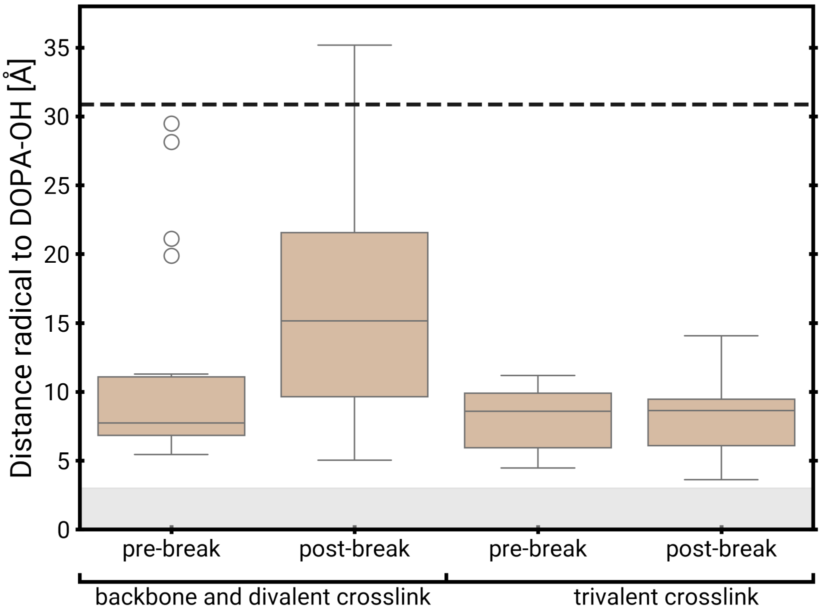


Figure 7.8: **PYD mechanoradicals are close to DOPA.** The mechanoradical distance to the hydroxy group of DOPA is shown after backbone and divalent cross-link, as well as trivalent cross-link homolytic cleavage. Distances are shown for a single snapshot directly after the homolysis reaction for n=72 radicals for non-trivalent cross-link radicals and n=32 for trivalent cross-link radicals. Distances in the area up to 3 Å can lead to HAT reactions and are depicted in grey. The average Cα distance to the hydroxy group of DOPA in the collagen fibril model is shown as dashed line.

7.8 D-amino acids are HAT products

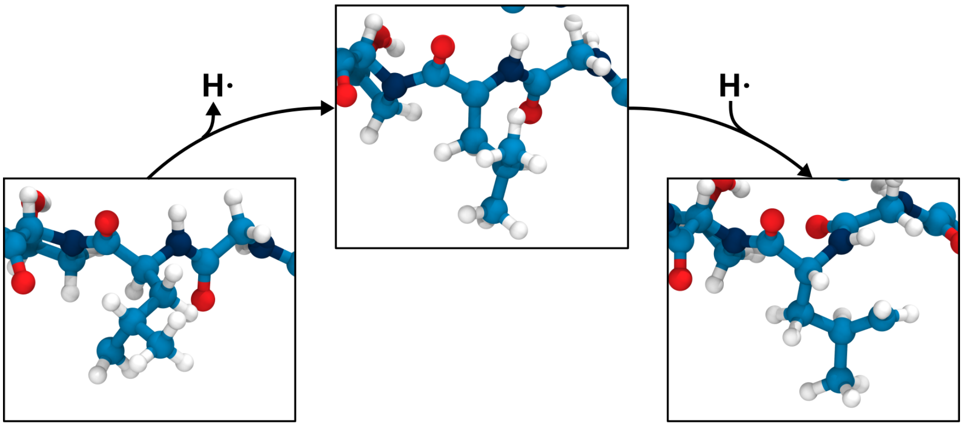


Figure 7.9: **Mechanism of D-amino acid formation.** A L-amino acid (left panel) donates its Cα hydrogen to a nearby radical, creating a resonance stabilised Cα radical (centre panel). In a consecutive HAT reaction, the Cα radical receives a hydrogen atom. The hydrogen atom can be donated from either side of the N-Cα-C plane, which results in either a L- or D-amino acid.

Recently, D-amino acids have been brought to attention as interesting target for basic research and therapeutic applications with huge risks attributed to the creation of mirror bacteria. [176]. D-amino acids are enantiomers of the natural L-amino acids, determined by the chiral Cα atom. Small quantities of D-amino acids are generated by molecular ageing processes. Aspargine and aspartate racemize via a cyclic succinimide intermediate[177] and other mechanism can lead to racemisation in a serine and leucine[178]. In bacteria, alanine is enzymatically racemised to be incorporated in the cell wall.[179] A degradation pathway exists and includes the D-amino acid oxidase.[180]

In chemical synthesis, reversible HAT has been used to selectively form enantiomers at chiral sites.[181, 182] In the KIMMDY simulations of HAT in collagen fibrils, formation of D-amino acids was also observed (Fig. 7.9). The D-amino acid is formed by two consecutive HAT reactions at a Cα atom with an initial H-donor and then H-acceptor role.

HATs at backbone atoms were uncommon in the 600 sampled reactions, contributing only 49 reactions with backbone H-donor and 35 reactions with backbone H-acceptor atoms (Fig. 7.10a,b). Out of these, seven reactions were at a Cα stereocenter at a total of 4 different sites. At three sites, including those with multiple reactions, the D-amino acid was formed. In the case of multiple reactions at the same site, always the D-amino acid was formed.

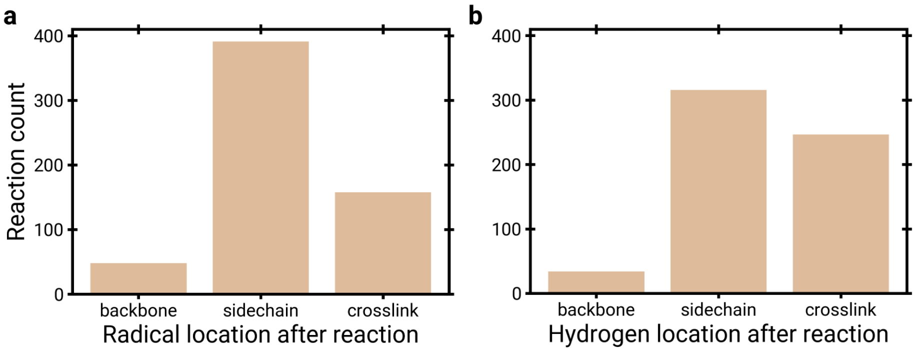


Figure 7.10: **Location of heavy atoms involved in HAT. a**, H-donor location, divided into backbone, side chain and cross-link atoms. **b**, H-acceptor location as described in panel a.

Further sampling is necessary to achieve a better understanding on the factors contributing to D-amino acid formation and whether certain amino acid positions preferably accept hydrogens from a direction that results in D-amino acids. D-amino acid formation appears to be the result of a minor fraction of HAT reactions but unless the absolute rate of HAT reactions in proteins is clarified, the abundance of D-amino acids due to this effect can not be clearly estimated.

The proposed mechanism is independent of the amino acid side chain and would give rise to an increased fraction of D-amino acids in proteins that are subject to mechanical stress, especially close to the weakest bonds in the molecule. D-amino acids can be measured experimentally using chiral column chromatography. [183] However, depending on the kind of sample, the limit of detection may be too high to measure these enantiomers.

7.9 Uncovering novel functions of trivalent cross-links

The discovery of a sacrificial bond in the trivalent cross-links of collagen[14] creates a new view on cross-links as a mechanochemically active instead of an inert compound that solely increases the mechanical strength of collagen[184]. Using a computational approach, PYD was identified as a radical scavenger with a low BDE at the pyridinium ring. Homolytic cleavage of PYD keeps radicals close to the interface of gap and overlap region that is rich in amino acids with aromatic moieties. This allows DOPA, a post-translational modification of tyrosine or phenylalanine, to scavenge mechanoradicals more efficiently. Side-reactions, such as conversion of amino acids to the D-enantiomer, are thereby minimised.

Taken together, KIMMDY simulations reveal a picture according to which collagen funnels radicals through specific and partially direct radical migration paths onto DOPA to avoid molecular damage.

Chapter 8

Conclusion and Outlook

This work is part of a larger effort to establish a method for long timescale reactive simulations in condensed phase systems. Hydrogen atom transfer is the second type of reaction brought to an application in simulations of the collagen fibril in a hybrid KMC/ MD setting. The combination of KMC and MD in adaptive KMC seems like a promising theoretical framework for the novel approach to sample a reactive state with classical MD and predict transition rates to neighbouring states. With the insights gained in this work, an improved HAT GNN could be trained that predicts absolute HAT rates and serve as a reference for future ML-based models that predict a diverse range of reaction as well as their rates.

Grappa, a machine learned classical force field, has been validated for proteins and trained on peptide radicals. Its accuracy compared to QM energies and forces is higher than the current gold standard, tabulated force fields. Grappa has the potential to predict even more accurate force field parameters if high fidelity bonded terms are implemented or the non-bonded parameters are improved upon.

Apart from the bespoke implementation of reactive HAT simulations, significant effort went into creating a general reactive simulation program, KIMMDY. KIMMDY has a framework architecture that allows users to implement custom reactions within a python plug-in and manages the control flow for adaptive KMC steps. Due to the modular design, KIMMDY can be combined with machine learning force fields or ML methods that directly generate conformational ensembles of proteins.

Reactive simulations in the collagen fibril show that DOPA is kinetically accessible after homolytic cleavage and direct transfers from the homolysis site to DOPA were observed. This provides a mechanistic explanation for the missing link between homolysis and detoxification of DOPA radicals in collagen. In addition, PYD was found to act as another radical scavenger. A small population of PYD radicals in stressed rat tendons is plausible from analysis of the available data. Further unique properties of PYD as a trivalent cross-link in a mechanochemistry setting are identified. To confirm these observations, time-resolved EPR measurements could be used to identify PYD radical populations directly after homolysis occurs.

A mechanism for D-amino acid enantiomerisation in proteins under mechanical stress is proposed and could be tested experimentally. Especially chiral column chromatography or a method with a lower limit of detection could provide a means to measure D-amino acid concentrations in proteins under mechanical stress. Alternatively, small peptides could be designed with radical initiators at a position that favours HAT reactions to a Cα atoms for a higher yield of D-amino acids than in collagen samples isolated from tissue.

Bibliography

[1] George A. Khoury, Richard C. Baliban, and Christodoulos A. Floudas. "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database". en. In: *Scientific Reports* 1.1 (Sept. 2011), p. 90. ISSN: 2045-2322. DOI: 10.1038/srep00090. URL: https://www.nature.com/articles/srep00090 (visited on 05/18/2025).

[2] Christopher T. Walsh, Sylvie Garneau-Tsodikova, and Gregory J. Gatto. "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications". en. In: *Angewandte Chemie International Edition* 44.45 (Nov. 2005), pp. 7342–7372. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.200501023. URL: https://onlinelibrary.wiley.com/doi/10.1002/anie.200501023 (visited on 05/18/2025).

[3] Matthias Mann and Ole N. Jensen. "Proteomic analysis of post-translational modifications". en. In: *Nature Biotechnology* 21.3 (Mar. 2003), pp. 255–261. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt0303-255. URL: https://www.nature.com/articles/nbt0303-255 (visited on 05/18/2025).

[4] Alfred L. Goldberg. "Protein degradation and protection against misfolded or damaged proteins". en. In: *Nature* 426.6968 (Dec. 2003), pp. 895–899. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature02263. URL: https://www.nature.com/articles/nature02263 (visited on 05/18/2025).

[5] Robert Harmel and Dorothea Fiedler. "Features and regulation of non-enzymatic post-translational modifications". en. In: *Nature Chemical Biology* 14.3 (Mar. 2018), pp. 244–252. ISSN: 1552-4450, 1552-4469. DOI: 10.1038/nchembio.2575. URL: https://www.nature.com/articles/nchembio.2575 (visited on 05/18/2025).

[6] Bryan C Dickinson and Christopher J Chang. "Chemistry and biology of reactive oxygen species in signaling or stress responses". en. In: *Nature Chemical Biology* 7.8 (Aug. 2011), pp. 504–511. ISSN: 1552-4450, 1552-4469. DOI: 10.1038/nchembio.607. URL: https://www.nature.com/articles/nchembio.607 (visited on 05/18/2025).

[7] Frans S. M. Van Kleef, Wilfried W. De Jong, and Herman J. Hoenders. "Stepwise degradations and deamidation of the eye lens protein -crystallin in ageing". en. In: *Nature* 258.5532 (Nov. 1975), pp. 264–266. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/258264a0. URL: https://www.nature.com/articles/258264a0 (visited on 05/18/2025).

[8] Michael G. Friedrich et al. "Spontaneous Cleavage at Glu and Gln Residues in Long-Lived Proteins". en. In: *ACS Chemical Biology* 16.11 (Nov. 2021), pp. 2244–2254. ISSN: 1554-8929, 1554-8937. DOI: 10.1021/acschembio.1c00379. URL: https://pubs.acs.org/doi/10.1021/acschembio.1c00379 (visited on 05/18/2025).

[9] Allen J Bailey, Robert Gordon Paul, and Lynda Knott. "Mechanisms of maturation and ageing of collagen". en. In: *Mechanisms of Ageing and Development* 106.1-2 (Dec. 1998), pp. 1–56. ISSN: 00476374. DOI: 10.1016/S0047-6374(98)00119-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S0047637498001195 (visited on 05/18/2025).

[10] Roger J.W. Truscott, Kevin L. Schey, and Michael G. Friedrich. "Old Proteins in Man: A Field in its Infancy". en. In: *Trends in Biochemical Sciences* 41.8 (Aug. 2016), pp. 654–664. ISSN: 09680004. DOI: 10.1016/j.tibs.2016.06.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968000416300603 (visited on 05/18/2025).

[11] Mary M. Caruso et al. "Mechanically-Induced Chemical Changes in Polymeric Materials". en. In: *Chemical Reviews* 109.11 (Nov. 2009), pp. 5755–5798. ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr9001353. URL: https://pubs.acs.org/doi/10.1021/cr9001353 (visited on 05/18/2025).

[12] Christopher Zapp et al. "Mechanoradicals in tensed tendon collagen as a source of oxidative stress". In: *Nature Communications* 11.1 (May 2020), p. 2315. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15567-4. URL: https://doi.org/10.1038/s41467-020-15567-4.

[13] Markus Kurth et al. "DOPA Residues Endow Collagen with Radical Scavenging Capacity**". en. In: *Angewandte Chemie International Edition* 62.24 (June 2023), e202216610. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.202216610. URL: https://onlinelibrary.wiley.com/doi/10.1002/anie.202216610 (visited on 06/13/2024).

[14] Benedikt Rennekamp et al. "Collagen breaks at weak sacrificial bonds taming its mechanoradicals". en. In: *Nature Communications* 14.1 (Apr. 2023), p. 2075. ISSN: 2041-1723. DOI: 10.1038/s41467-023-37726-z. URL: https://www.nature.com/articles/s41467-023-37726-z (visited on 06/13/2024).

[15] Ron O. Dror et al. "Biomolecular Simulation: A Computational Microscope for Molecular Biology". en. In: *Annual Review of Biophysics* 41.1 (June 2012), pp. 429–452. ISSN: 1936-122X, 1936-1238. DOI: 10.1146/annurev-biophys-042910-155245. URL: https://www.annualreviews.org/doi/10.1146/annurev-biophys-042910-155245 (visited on 05/18/2025).

[16] David Van Der Spoel. "Systematic design of biomolecular force fields". en. In: *Current Opinion in Structural Biology* 67 (Apr. 2021), pp. 18–24. ISSN: 0959440X. DOI: 10.1016/j.sbi.2020.08.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X2030141X (visited on 05/18/2025).

[17] Alfonso Gautieri et al. "Hierarchical Structure and Nanomechanics of Collagen Microfibrils from the Atomistic Scale Up". en. In: *Nano Letters* 11.2 (Feb. 2011), pp. 757–766. ISSN: 1530-6984, 1530-6992. DOI: 10.1021/nl103943u. URL: https://pubs.acs.org/doi/10.1021/nl103943u (visited on 05/19/2025).

[18] J E Scott. "Proteoglycan-fibrillar collagen interactions". en. In: *Biochemical Journal* 252.2 (June 1988), pp. 313–323. ISSN: 0264-6021, 1470-8728. DOI: 10.1042/bj2520313. URL: https://portlandpress.com/biochemj/article/252/2/313/25261/Proteoglycan-fibrillar-collagen-interactions (visited on 05/19/2025).

[19] Matthew D. Shoulders and Ronald T. Raines. "Collagen Structure and Stability". en. In: *Annual Review of Biochemistry* 78.1 (June 2009), pp. 929–958. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.biochem.77.032207.120833. URL: https://www.annualreviews.org/doi/10.1146/annurev.biochem.77.032207.120833 (visited on 05/19/2025).

[20] David R. Eyre and Jiann-Jiu Wu. "Collagen Cross-Links". In: *Collagen*. Ed. by Jürgen Brinckmann, Holger Notbohm, and P. K. Müller. Vol. 247. Series Title: Topics in Current Chemistry. Berlin, Heidelberg: Springer Berlin Heidelberg, Apr. 2005, pp. 207–229. ISBN: 978-3-540-23272-8 978-3-540-31472-1. DOI: 10.1007/b103828. URL: http://link.springer.com/10.1007/b103828 (visited on 05/10/2025).

[21] R. Singh et al. "Advanced glycation end-products: a review". In: *Diabetologia* 44.2 (Feb. 2001), pp. 129–146. ISSN: 0012-186X, 1432-0428. DOI: 10.1007/s001250051591. URL: http://link.springer.com/10.1007/s001250051591 (visited on 05/19/2025).

[22] Aleksandra Twarda-Clapa et al. "Advanced Glycation End-Products (AGEs): Formation, Chemistry, Classification, Receptors, and Diseases Related to AGEs". en. In: *Cells* 11.8 (Apr. 2022), p. 1312. ISSN: 2073-4409. DOI: `10.3390/cells11081312`. URL: `https://www.mdpi.com/2073-4409/11/8/1312` (visited on 05/19/2025).

[23] Noboru Yamauchi et al. "Thermal Decomposition and Isomerization Processes of Alkyl Radicals". en. In: *The Journal of Physical Chemistry A* 103.15 (Apr. 1999), pp. 2723–2733. ISSN: 1089-5639, 1520-5215. DOI: `10.1021/jp9844563`. URL: `https://pubs.acs.org/doi/10.1021/jp9844563` (visited on 04/29/2025).

[24] Luigi Anastasia et al. "Chemical structure, biosynthesis and synthesis of free and glycosylated pyridinolines formed by cross-link of bone and synovium collagen". en. In: *Organic & Biomolecular Chemistry* 11.35 (2013), p. 5747. ISSN: 1477-0520, 1477-0539. DOI: `10.1039/c3ob40945g`. URL: `https://xlink.rsc.org/?DOI=c3ob40945g` (visited on 05/19/2025).

[25] Yoji Kato, Takashi Nishikawa, and Shunro Kawakishi. "FORMATION OF PROTEIN-BOUND 3,4-DIHYDROXYPHENYLALANINE IN COLLAGEN TYPES I AND IV EXPOSED TO ULTRAVIOLET LIGHT". en. In: *Photochemistry and Photobiology* 61.4 (Apr. 1995), pp. 367–372. ISSN: 0031-8655, 1751-1097. DOI: `10.1111/j.1751-1097.1995.tb08624.x`. URL: `https://onlinelibrary.wiley.com/doi/10.1111/j.1751-1097.1995.tb08624.x` (visited on 05/20/2025).

[26] Kenneth J Rodgers and Roger T Dean. "Metabolism of protein-bound DOPA in mammals". en. In: *The International Journal of Biochemistry & Cell Biology* 32.9 (Sept. 2000), pp. 945–955. ISSN: 13572725. DOI: `10.1016/S1357-2725(00)00034-0`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S1357272500000340` (visited on 05/20/2025).

[27] Elizabeth J. Blaesi et al. "Metal-free class Ie ribonucleotide reductase from pathogens initiates catalysis with a tyrosine-derived dihydroxyphenylalanine radical". en. In: *Proceedings of the National Academy of Sciences* 115.40 (Oct. 2018), pp. 10022–10027. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.1811993115`. URL: `https://pnas.org/doi/full/10.1073/pnas.1811993115` (visited on 05/19/2025).

[28] Helmut Sies, Ryan J. Mailloux, and Ursula Jakob. "Fundamentals of redox regulation in biology". en. In: *Nature Reviews Molecular Cell Biology* 25.9 (Sept. 2024), pp. 701–719. ISSN: 1471-0072, 1471-0080. DOI: `10.1038/s41580-024-00730-2`. URL: `https://www.nature.com/articles/s41580-024-00730-2` (visited on 05/20/2025).

[29] J. E. Eastoe. "The amino acid composition of mammalian collagen and gelatin". en. In: *Biochemical Journal* 61.4 (Dec. 1955), pp. 589–600. ISSN: 0306-3283. DOI: `10.1042/bj0610589`. URL: `https://portlandpress.com/biochemj/article/61/4/589/52121/The-amino-acid-composition-of-mammalian-collagen` (visited on 05/20/2025).

[30] L Ala-Kokko et al. "Single base mutation in the type II procollagen gene (COL2A1) as a cause of primary osteoarthritis associated with a mild chondrodysplasia." en. In: *Proceedings of the National Academy of Sciences* 87.17 (Sept. 1990), pp. 6565–6568. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.87.17.6565`. URL: `https://pnas.org/doi/full/10.1073/pnas.87.17.6565` (visited on 05/20/2025).

[31] Johanna Myllyharju and Kari I Kivirikko. "Collagens and collagen-related diseases". en. In: *Annals of Medicine* 33.1 (Jan. 2001), pp. 7–21. ISSN: 0785-3890, 1365-2060. DOI: `10.3109/07853890109002055`. URL: `http://www.tandfonline.com/doi/full/10.3109/07853890109002055` (visited on 05/20/2025).

[32] James M. Mayer. "Understanding Hydrogen Atom Transfer: From Bond Strengths to Marcus Theory". en. In: *Accounts of Chemical Research* 44.1 (Jan. 2011), pp. 36–46. ISSN: 0001-4842, 1520-4898. DOI: `10.1021/ar100093z`. URL: `https://pubs.acs.org/doi/10.1021/ar100093z` (visited on 05/21/2025).

[33] Clare L. Hawkins and Michael J. Davies. "Reaction of HOCl with amino acids and peptides: EPR evidence for rapid rearrangement and fragmentation reactions of nitrogen-centred radicals". In: *Journal of the Chemical Society, Perkin Transactions 2* 9 (1998), pp. 1937–1946. ISSN: 03009580, 13645471. DOI: 10.1039/a802949k. URL: https://xlink.rsc.org/?DOI=a802949k (visited on 05/21/2025).

[34] Clare L. Hawkins and Michael J. Davies. "Generation and propagation of radical reactions on proteins". en. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1504.2-3 (Apr. 2001), pp. 196–219. ISSN: 00052728. DOI: 10.1016/S0005-2728(00)00252-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0005272800002528 (visited on 05/21/2025).

[35] Christine C Winterbourn. "Reconciling the chemistry and biology of reactive oxygen species". en. In: *Nature Chemical Biology* 4.5 (May 2008), pp. 278–286. ISSN: 1552-4450, 1552-4469. DOI: 10.1038/nchembio.85. URL: https://www.nature.com/articles/nchembio.85 (visited on 05/20/2025).

[36] A. Graham Pockley and Brian Henderson. "Extracellular cell stress (heat shock) proteins—immune responses and disease: an overview". en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1738 (Jan. 2018), p. 20160522. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2016.0522. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0522 (visited on 05/20/2025).

[37] Vyshnavi Vennelakanti et al. "Harder, better, faster, stronger: Large-scale QM and QM/MM for predictive modeling in enzymes and proteins". en. In: *Current Opinion in Structural Biology* 72 (Feb. 2022), pp. 9–17. ISSN: 0959440X. DOI: 10.1016/j.sbi.2021.07.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X21001068 (visited on 05/05/2025).

[38] Thomas P. Senftle et al. *The ReaxFF reactive force-field: Development, applications and future directions.* ISSN: 20573960 Publication Title: npj Computational Materials Volume: 2. Mar. 2016. DOI: 10.1038/npjcompumats.2015.11.

[39] Arieh Warshel et al. "Electrostatic Basis for Enzyme Catalysis". en. In: *Chemical Reviews* 106.8 (Aug. 2006), pp. 3210–3235. ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr0503106. URL: https://pubs.acs.org/doi/10.1021/cr0503106 (visited on 05/21/2025).

[40] Albert Musaelian et al. "Learning local equivariant representations for large-scale atomistic dynamics". In: *Nature Communications* 14.1 (Feb. 2023), p. 579. ISSN: 2041-1723. DOI: 10.1038/s41467-023-36329-y. URL: https://doi.org/10.1038/s41467-023-36329-y.

[41] Tong Wang et al. "Ab initio characterization of protein molecular dynamics with AI2BMD". en. In: *Nature* 635.8040 (Nov. 2024), pp. 1019–1027. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-08127-z. URL: https://www.nature.com/articles/s41586-024-08127-z (visited on 04/25/2025).

[42] Daniel T Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". en. In: *Journal of Computational Physics* 22.4 (Dec. 1976), pp. 403–434. ISSN: 00219991. DOI: 10.1016/0021-9991(76)90041-3. URL: https://linkinghub.elsevier.com/retrieve/pii/0021999176900413 (visited on 05/14/2025).

[43] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz. "A new algorithm for Monte Carlo simulation of Ising spin systems". en. In: *Journal of Computational Physics* 17.1 (Jan. 1975), pp. 10–18. ISSN: 00219991. DOI: 10.1016/0021-9991(75)90060-1. URL: https://linkinghub.elsevier.com/retrieve/pii/0021999175900601 (visited on 05/06/2025).

[44] Mie Andersen, Chiara Panosetti, and Karsten Reuter. "A Practical Guide to Surface Kinetic Monte Carlo Simulations". In: *Frontiers in Chemistry* 7 (2019). ISSN: 2296-2646. URL: https://www.frontiersin.org/articles/10.3389/fchem.2019.00202 (visited on 04/06/2023).

[45] Kai Riedmiller et al. "Substituting density functional theory in reaction barrier calculations for hydrogen atom transfer in proteins". en. In: *Chemical Science* 15.7 (2024), pp. 2518–2527. ISSN: 2041-6520, 2041-6539. DOI: 10.1039/D3SC03922F. URL: https://xlink.rsc.org/?DOI=D3SC03922F (visited on 06/13/2024).

[46] Max J. Hoffmann, Sebastian Matera, and Karsten Reuter. "kmos: A lattice kinetic Monte Carlo framework". en. In: *Computer Physics Communications* 185.7 (July 2014), pp. 2138–2150. ISSN: 00104655. DOI: 10.1016/j.cpc.2014.04.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S001046551400126X (visited on 05/05/2025).

[47] Graeme Henkelman and Hannes Jónsson. "Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table". en. In: *The Journal of Chemical Physics* 115.21 (Dec. 2001), pp. 9657–9666. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1415500. URL: https://pubs.aip.org/jcp/article/115/21/9657/453229/Long-time-scale-kinetic-Monte-Carlo-simulations (visited on 05/05/2025).

[48] Lijun Xu and Graeme Henkelman. "Adaptive kinetic Monte Carlo for first-principles accelerated dynamics". en. In: *The Journal of Chemical Physics* 129.11 (Sept. 2008), p. 114104. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2976010. URL: https://pubs.aip.org/jcp/article/129/11/114104/296009/Adaptive-kinetic-Monte-Carlo-for-first-principles (visited on 06/13/2024).

[49] Samuel T. Chill and Graeme Henkelman. "Molecular dynamics saddle search adaptive kinetic Monte Carlo". en. In: *The Journal of Chemical Physics* 140.21 (June 2014), p. 214110. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4880721. URL: https://pubs.aip.org/jcp/article/140/21/214110/566728/Molecular-dynamics-saddle-search-adaptive-kinetic (visited on 05/14/2025).

[50] Emanuel Karl Peter and Joan-Emma Shea. "A hybrid MD-kMC algorithm for folding proteins in explicit solvent". en. In: *Physical Chemistry Chemical Physics* 16.14 (2014), p. 6430. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/c3cp55251a. URL: https://xlink.rsc.org/?DOI=c3cp55251a (visited on 05/21/2025).

[51] A. Violi et al. "A fully integrated kinetic monte carlo/molecular dynamics approach for the simulation of soot precursor growth". en. In: *Proceedings of the Combustion Institute* 29.2 (Jan. 2002), pp. 2343–2349. ISSN: 15407489. DOI: 10.1016/S1540-7489(02)80285-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S1540748902802851 (visited on 05/21/2025).

[52] Gabriel Kabbe, Christoph Wehmeyer, and Daniel Sebastiani. "A Coupled Molecular Dynamics/Kinetic Monte Carlo Approach for Protonation Dynamics in Extended Systems". en. In: *Journal of Chemical Theory and Computation* 10.10 (Oct. 2014), pp. 4221–4228. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct500482k. URL: https://pubs.acs.org/doi/10.1021/ct500482k (visited on 05/21/2025).

[53] Feranmi V. Olowookere and C. Heath Turner. "An integrated off-lattice kinetic Monte Carlo (KMC)-molecular dynamics (MD) framework for modeling polyvinyl chloride dehydrochlorination". en. In: *Chemical Engineering Science* 302 (Feb. 2025), p. 120928. ISSN: 00092509. DOI: 10.1016/j.ces.2024.120928. URL: https://linkinghub.elsevier.com/retrieve/pii/S0009250924012284 (visited on 05/21/2025).

[54] Joseph W. Abbott and Felix Hanke. "Kinetically Corrected Monte Carlo–Molecular Dynamics Simulations of Solid Electrolyte Interphase Growth". en. In: *Journal of Chemical Theory and Computation* 18.2 (Feb. 2022), pp. 925–934. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.1c00921. URL: https://pubs.acs.org/doi/10.1021/acs.jctc.1c00921 (visited on 05/21/2025).

[55] Benedikt Rennekamp et al. "Hybrid Kinetic Monte Carlo/Molecular Dynamics Simulations of Bond Scissions in Proteins". en. In: *Journal of Chemical Theory and Computation* 16.1 (Jan. 2020), pp. 553–563. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.9b00786. URL: https://pubs.acs.org/doi/10.1021/acs.jctc.9b00786 (visited on 06/13/2024).

[56] Mark James Abraham et al. “Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1-2 (2015). Publisher: Elsevier B.V., pp. 19–25. ISSN: 23527110. DOI: 10.1016/j.softx.2015.06.001.

[57] Mark Abraham et al. “GROMACS 2025.1 Manual”. In: (Mar. 2025). Publisher: Zenodo Version Number: 2025.1. DOI: 10.5281/ZENODO.15006631. URL: https://zenodo.org/doi/10.5281/zenodo.15006631 (visited on 05/12/2025).

[58] R.W Hockney, S.P Goel, and J.W Eastwood. “Quiet high-resolution computer models of a plasma”. en. In: *Journal of Computational Physics* 14.2 (Feb. 1974), pp. 148–158. ISSN: 00219991. DOI: 10.1016/0021-9991(74)90010-2. URL: https://linkinghub.elsevier.com/retrieve/pii/0021999174900102 (visited on 05/12/2025).

[59] H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath”. en. In: *The Journal of Chemical Physics* 81.8 (Oct. 1984), pp. 3684–3690. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.448118. URL: https://pubs.aip.org/jcp/article/81/8/3684/565473/Molecular-dynamics-with-coupling-to-an-external (visited on 05/12/2025).

[60] Giovanni Bussi, Davide Donadio, and Michele Parrinello. “Canonical sampling through velocity rescaling”. en. In: *The Journal of Chemical Physics* 126.1 (Jan. 2007), p. 014101. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2408420. URL: https://pubs.aip.org/jcp/article/126/1/014101/186581/Canonical-sampling-through-velocity-rescaling (visited on 03/22/2024).

[61] Berk Hess. “P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation”. en. In: *Journal of Chemical Theory and Computation* 4.1 (Jan. 2008), pp. 116–122. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct700200b. URL: https://pubs.acs.org/doi/10.1021/ct700200b (visited on 03/22/2024).

[62] Kurt Kremer and Gary S. Grest. “Dynamics of entangled linear polymer melts: A molecular-dynamics simulation”. en. In: *The Journal of Chemical Physics* 92.8 (Apr. 1990), pp. 5057–5086. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.458541. URL: https://pubs.aip.org/jcp/article/92/8/5057/220968/Dynamics-of-entangled-linear-polymer-melts-A (visited on 05/13/2025).

[63] M. J. Hwang, T. P. Stockfisch, and A. T. Hagler. “Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules”. en. In: *Journal of the American Chemical Society* 116.6 (Mar. 1994), pp. 2515–2525. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja00085a036. URL: https://pubs.acs.org/doi/abs/10.1021/ja00085a036 (visited on 04/25/2025).

[64] Wendy D Cornell et al. *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.* Tech. rep. Publication Title: J. Am. Chem. Soc Volume: 117. 1995, pp. 5179–5197. URL: https://pubs.acs.org/sharingguidelines.

[65] Jörg Behler. “First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems”. eng. In: *Angewandte Chemie (International Ed. in English)* 56.42 (Oct. 2017), pp. 12828–12840. ISSN: 1521-3773. DOI: 10.1002/anie.201703114.

[66] Ilyes Batatia et al. “MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields”. In: *Advances in Neural Information Processing Systems.* Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 11423–11436. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/4a36c3c51af11ed9f34615b81edb5bbc-Paper-Conference.pdf.

[67] Kenno Vanommeslaeghe, Olgun Guvench, and Alexander D. MacKerell. “Molecular Mechanics”. en. In: *Current Pharmaceutical Design* 20.20 (May 2014), pp. 3281–3292. ISSN: 13816128. DOI: 10.2174/13816128113199990600. URL: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1381-6128&volume=20&issue=20&spage=3281 (visited on 05/13/2025).

[68] Scott A. Hollingsworth and Ron O. Dror. "Molecular Dynamics Simulation for All". en. In: *Neuron* 99.6 (Sept. 2018), pp. 1129–1143. ISSN: 08966273. DOI: 10.1016/j.neuron.2018.08.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627318306846 (visited on 05/13/2025).

[69] Chuan Tian et al. "Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution". In: *Journal of Chemical Theory and Computation* 16.1 (Jan. 2020). Publisher: American Chemical Society, pp. 528–552. ISSN: 15499626. DOI: 10.1021/acs.jctc.9b00591.

[70] Jing Huang et al. "CHARMM36m: an improved force field for folded and intrinsically disordered proteins". In: *Nature Methods* 14.1 (2017), pp. 71–73. ISSN: 1548-7105. DOI: 10.1038/nmeth.4067.

[71] Scott J. Weiner et al. "A new force field for molecular mechanical simulation of nucleic acids and proteins". en. In: *Journal of the American Chemical Society* 106.3 (Feb. 1984), pp. 765–784. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja00315a051. URL: https://pubs.acs.org/doi/abs/10.1021/ja00315a051 (visited on 04/25/2025).

[72] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems". en. In: *The Journal of Chemical Physics* 98.12 (June 1993), pp. 10089–10092. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.464397. URL: https://pubs.aip.org/jcp/article/98/12/10089/461765/Particle-mesh-Ewald-An-N-log-N-method-for-Ewald (visited on 05/13/2025).

[73] Piotr Cieplak et al. "Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins". en. In: *Journal of Computational Chemistry* 16.11 (Nov. 1995), pp. 1357–1377. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.540161106. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.540161106 (visited on 04/28/2025).

[74] Michael P. Allen. "Computational soft matter: from synthetic polymers to proteins. 2: Lecture notes". en. In: NIC series volume 23. Num Pages: 1. Jülich: NIC, 2004. ISBN: 978-3-00-012641-3.

[75] James A. Maier et al. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* 11.8 (2015). Publisher: American Chemical Society, pp. 3696–3713. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00255. URL: https://doi.org/10.1021/acs.jctc.5b00255.

[76] Jérôme Hénin et al. "Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]". In: *Living Journal of Computational Molecular Science* 4.1 (2022). ISSN: 25756524. DOI: 10.33011/livecoms.4.1.1583. URL: https://www.livecomsjournal.org/index.php/livecoms/article/view/v4i1e1583 (visited on 04/29/2025).

[77] Noora Aho et al. "Scalable Constant pH Molecular Dynamics in GROMACS". en. In: *Journal of Chemical Theory and Computation* 18.10 (Oct. 2022), pp. 6148–6160. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.2c00516. URL: https://pubs.acs.org/doi/10.1021/acs.jctc.2c00516 (visited on 05/13/2025).

[78] Wei Chen et al. "Recent development and application of constant pH molecular dynamics". en. In: *Molecular Simulation* 40.10-11 (Aug. 2014), pp. 830–838. ISSN: 0892-7022, 1029-0435. DOI: 10.1080/08927022.2014.907492. URL: http://www.tandfonline.com/doi/abs/10.1080/08927022.2014.907492 (visited on 05/13/2025).

[79] P Hohenberg and W Kohn. "Inhomogeneous Electron Gas". en. In: (Sept. 1964).

[80] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". en. In: *Physical Review* 140.4A (Nov. 1965), A1133–A1138. ISSN: 0031-899X. DOI: 10.1103/PhysRev.140.A1133. URL: https://link.aps.org/doi/10.1103/PhysRev.140.A1133 (visited on 05/22/2025).

[81] Markus Bursch et al. “Best-Practice DFT Protocols for Basic Molecular Computational Chemistry**”. en. In: *Angewandte Chemie International Edition* 61.42 (Oct. 2022), e202205735. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.202205735. URL: https://onlinelibrary.wiley.com/doi/10.1002/anie.202205735 (visited on 05/14/2025).

[82] H. Bernhard Schlegel. “Geometry optimization”. en. In: *WIREs Computational Molecular Science* 1.5 (Sept. 2011), pp. 790–809. ISSN: 1759-0876, 1759-0884. DOI: 10.1002/wcms.34. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.34 (visited on 05/14/2025).

[83] Radu Iftimie, Peter Minary, and Mark E. Tuckerman. “*Ab initio* molecular dynamics: Concepts, recent developments, and future trends”. en. In: *Proceedings of the National Academy of Sciences* 102.19 (May 2005), pp. 6654–6659. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0500193102. URL: https://pnas.org/doi/full/10.1073/pnas.0500193102 (visited on 05/14/2025).

[84] Daniel T. Gillespie. “Stochastic Simulation of Chemical Kinetics”. en. In: *Annual Review of Physical Chemistry* 58.1 (May 2007), pp. 35–55. ISSN: 0066-426X, 1545-1593. DOI: 10.1146/annurev.physchem.58.032806.104637. URL: https://www.annualreviews.org/doi/10.1146/annurev.physchem.58.032806.104637 (visited on 04/06/2023).

[85] Svante Arrhenius. “Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren”. en. In: *Zeitschrift für Physikalische Chemie* 4U.1 (July 1889), pp. 226–248. ISSN: 2196-7156, 0942-9352. DOI: 10.1515/zpch-1889-0416. URL: https://www.degruyter.com/document/doi/10.1515/zpch-1889-0416/html (visited on 05/15/2025).

[86] Henry Eyring. “The Activated Complex in Chemical Reactions”. en. In: *The Journal of Chemical Physics* 3.2 (Feb. 1935), pp. 107–115. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1749604. URL: https://pubs.aip.org/jcp/article/3/2/107/203352/The-Activated-Complex-in-Chemical-Reactions (visited on 05/15/2025).

[87] Donald G Truhlar, Bruce C Garrett, and Stephen J Klippenstein. “Current Status of Transition-State Theory”. en. In: (Feb. 1996).

[88] Graeme Henkelman and Hannes Jónsson. “A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives”. en. In: *The Journal of Chemical Physics* 111.15 (Oct. 1999), pp. 7010–7022. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.480097. URL: https://pubs.aip.org/jcp/article/111/15/7010/475160/A-dimer-method-for-finding-saddle-points-on-high (visited on 05/05/2025).

[89] Kristof Schütt, Oliver Unke, and Michael Gastegger. “Equivariant message passing for the prediction of tensorial properties and molecular spectra”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 9377–9388. URL: https://proceedings.mlr.press/v139/schutt21a.html.

[90] A. Daniel Boese and Jan M. L. Martin. “Development of density functionals for thermochemical kinetics”. In: *The Journal of Chemical Physics* 121.8 (Aug. 2004). Publisher: AIP Publishing, pp. 3405–3416. DOI: 10.1063/1.1774975. URL: https://doi.org/10.1063%2F1.1774975.

[91] Leif Seute et al. “Grappa – a machine learned molecular mechanics force field”. en. In: *Chemical Science* 16.6 (2025), pp. 2907–2930. ISSN: 2041-6520, 2041-6539. DOI: 10.1039/D4SC05465B. URL: https://xlink.rsc.org/?DOI=D4SC05465B (visited on 04/25/2025).

[92] Eric Hartmann et al. “Reactive Biomolecular Simulations using Adaptive Kinetic Monte Carlo”. In: (2025).

[93] Junmei Wang et al. "Development and testing of a general amber force field". en. In: *Journal of Computational Chemistry* 25.9 (July 2004), pp. 1157–1174. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.20035. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.20035 (visited on 04/25/2025).

[94] K. Vanommeslaeghe et al. "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". en. In: *Journal of Computational Chemistry* 31.4 (Mar. 2010), pp. 671–690. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.21367. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.21367 (visited on 04/25/2025).

[95] Yuanqing Wang et al. "End-to-end differentiable construction of molecular mechanics force fields". In: *Chemical Science* 13.41 (2022). Publisher: Royal Society of Chemistry (RSC), pp. 12016–12033. DOI: 10.1039/d2sc02739a. URL: https://doi.org/10.1039%2Fd2sc02739a.

[96] Kenichiro Takaba et al. "Machine-learned molecular mechanics force fields from large-scale quantum chemical data". In: *Chem. Sci.* 15.32 (2024). Publisher: The Royal Society of Chemistry, pp. 12861–12878. DOI: 10.1039/D4SC00690A. URL: http://dx.doi.org/10.1039/D4SC00690A.

[97] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[98] Peter Eastman et al. "SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials". In: *Scientific Data* 10.1 (2023), p. 11. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01882-6.

[99] Open Force Field Initiative. *qca-dataset-submission.* 2020. URL: https://github.com/openforcefield/qca-dataset-submission.

[100] William L. Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *The Journal of Chemical Physics* 79.2 (July 1983), pp. 926–935.

[101] M. Parrinello and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied Physics* 52.12 (Dec. 1981), pp. 7182–7190. ISSN: 0021-8979. DOI: 10.1063/1.328693. URL: https://doi.org/10.1063/1.328693 (visited on 03/22/2024).

[102] In Suk Joung and Thomas E. III Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations". In: *The Journal of Physical Chemistry B* 112.30 (2008), pp. 9020–9041.

[103] Araz Jakalian, David B. Jack, and Christopher I. Bayly. "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". In: *Journal of Computational Chemistry* 23.16 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf pp. 1623–1641. DOI: https://doi.org/10.1002/jcc.10128. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10128.

[104] Shinya Honda et al. "Crystal Structure of a Ten-Amino Acid Protein". In: *Journal of the American Chemical Society* 130.46 (2008). _eprint: https://doi.org/10.1021/ja8030533, pp. 15327–15331. DOI: 10.1021/ja8030533. URL: https://doi.org/10.1021/ja8030533.

[105] Peter Eastman et al. "OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials". en. In: *The Journal of Physical Chemistry B* 128.1 (Jan. 2024), pp. 109–116. ISSN: 1520-6106, 1520-5207. DOI: 10.1021/acs.jpcb.3c06662. URL: https://pubs.acs.org/doi/10.1021/acs.jpcb.3c06662 (visited on 04/25/2025).

[106] Agnieszka Obarska-Kosinska et al. "ColBuilder: A server to build collagen fibril models". In: *Biophysical Journal* 120.17 (Sept. 2021). Publisher: Elsevier, pp. 3544–3549. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2021.07.009. URL: https://doi.org/10.1016/j.bpj.2021.07.009 (visited on 04/29/2025).

[107] Mattia Bernetti and Giovanni Bussi. “Pressure control using stochastic cell rescaling”. en. In: *The Journal of Chemical Physics* 153.11 (Sept. 2020), p. 114107. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/5.0020514. URL: https://pubs.aip.org/jcp/article/153/11/114107/199610/Pressure-control-using-stochastic-cell-rescaling (visited on 05/16/2025).

[108] K. Lindorff-Larsen et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins* 78.8 (June 2010), pp. 1950–1958. DOI: 10.1002/prot.22711. URL: https://doi.org/10.1002/prot.22711.

[109] Kresten Lindorff-Larsen et al. “How Fast-Folding Proteins Fold”. In: *Science* 334.6055 (2011). _eprint: https://www.science.org/doi/pdf/10.1126/science.1208351, pp. 517–520. DOI: 10.1126/science.1208351. URL: https://www.science.org/doi/abs/10.1126/science.1208351.

[110] Vincenzo Barone et al. *Development and Validation of Force-Field Parameters for Molecular Simulations of Peptides and Proteins Containing Open-Shell Residues.* Tech. rep. Publication Title: J Comput Chem Volume: 18. John Wiley & Sons, Inc, 1997, p. 17201728.

[111] István Komáromi et al. “Development of glycyl radical parameters for the OPLS-AA/L force field”. en. In: *Journal of Computational Chemistry* 29.12 (Sept. 2008), pp. 1999–2009. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.20962. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.20962 (visited on 04/25/2025).

[112] Simon C. Lovell et al. “Structure validation by C geometry: , and C deviation”. en. In: *Proteins: Structure, Function, and Bioinformatics* 50.3 (Feb. 2003), pp. 437–450. ISSN: 0887-3585, 1097-0134. DOI: 10.1002/prot.10286. URL: https://onlinelibrary.wiley.com/doi/10.1002/prot.10286 (visited on 04/25/2025).

[113] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. “Stereochemistry of polypeptide chain configurations”. en. In: *Journal of Molecular Biology* 7.1 (July 1963), pp. 95–99. ISSN: 00222836. DOI: 10.1016/S0022-2836(63)80023-6. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022283663800236 (visited on 04/25/2025).

[114] Junmei Wang, Piotr Cieplak, and Peter A Kollman. *How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? Keywords: additive force field; nonadditive force field; restrained electrostatic potential (RESP); torsional angle parameterization.* Tech. rep. 12. Publication Title: Journal of Computational Chemistry Volume: 21. 2000, pp. 1049–1074. URL: www.amber.ucsf.edu/amber/.

[115] Yuanqing Wang et al. “EspalomaCharge: Machine Learning-Enabled Ultrafast Partial Charge Assignment”. In: *The Journal of Physical Chemistry A* 128.20 (2024), pp. 4160–4167. DOI: 10.1021/acs.jpca.4c01287.

[116] Narbe Mardirossian and Martin Head-Gordon. “ B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation”. en. In: *The Journal of Chemical Physics* 144.21 (June 2016), p. 214110. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4952647. URL: https://pubs.aip.org/jcp/article/144/21/214110/313155/B97M-V-A-combinatorially-optimized-range-separated (visited on 04/25/2025).

[117] Florian Weigend and Reinhart Ahlrichs. “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy”. en. In: *Physical Chemistry Chemical Physics* 7.18 (2005), p. 3297. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/b508541a. URL: https://xlink.rsc.org/?DOI=b508541a (visited on 04/25/2025).

[118] Marie Zgarbová et al. “Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles”. In: *Journal of Chemical Theory and Computation* 7.9 (Sept. 2011). Publisher: American Chemical Society, pp. 2886–2902. ISSN: 1549-9618. DOI: 10.1021/ct200162x. URL: https://doi.org/10.1021/ct200162x.

[119] Jing Huang and Alexander D. MacKerell Jr. “CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data”. In: *Journal of Computational Chemistry* 34.25 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23354, pp. 2135–2145. DOI: https://doi.org/10.1002/jcc.23354. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23354.

[120] Tong Wang et al. “AIMD-Chig: Exploring the conformational space of a 166-atom protein Chignolin with ab initio molecular dynamics”. en. In: *Scientific Data* 10.1 (Aug. 2023), p. 549. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02465-9. URL: https://www.nature.com/articles/s41597-023-02465-9 (visited on 04/25/2025).

[121] Oliver T. Unke et al. “Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments”. In: *Science Advances* 10.14 (2024). _eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adn4397, eadn4397. DOI: 10.1126/sciadv.adn4397. URL: https://www.science.org/doi/abs/10.1126/sciadv.adn4397.

[122] Martin. Karplus. “**Vicinal Proton Coupling in Nuclear Magnetic Resonance**”. en. In: *Journal of the American Chemical Society* 85.18 (Sept. 1963), pp. 2870–2871. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja00901a059. URL: https://pubs.acs.org/doi/abs/10.1021/ja00901a059 (visited on 04/25/2025).

[123] Lily Wang et al. “The Open Force Field Initiative: Open Software and Open Science for Molecular Modeling”. In: *The Journal of Physical Chemistry B* 128.29 (July 2024). Publisher: American Chemical Society, pp. 7043–7067. ISSN: 1520-6106. DOI: 10.1021/acs.jpcb.4c01558.

[124] Julia Wirmer and Harald Schwalbe. “Angular dependence of \(^1J(N_i,C_\alpha_i)\) and \(^2J(N_i,C_\alpha_i-1)\) coupling constants measured in J-modulated HSQCs”. In: *Journal of Biomolecular NMR* 23.1 (May 2002), pp. 47–55.

[125] Keyang Ding and Angela M. Gronenborn. “Protein Backbone 1HN13C and 15N13C Residual Dipolar and J Couplings: New Constraints for NMR Structure Determination”. In: *Journal of the American Chemical Society* 126.20 (2004), pp. 6232–6233.

[126] Jin-Shan Hu and Ad Bax. “Determination of and 1 Angles in Proteins from 13C13C Three-Bond J Couplings Measured by Three-Dimensional Heteronuclear NMR. How Planar Is the Peptide Bond?” In: *Journal of the American Chemical Society* 119.27 (1997), pp. 6360–6368.

[127] Beat Vögeli et al. “Limits on Variations in Protein Backbone Dynamics from Precise Measurements of Scalar Couplings”. In: *Journal of the American Chemical Society* 129.30 (2007), pp. 9377–9385.

[128] Mirko Hennig et al. “Determination of Torsion Angle Restraints from 3J(C,C) and 3J(C,HN) Coupling Constants in Proteins”. In: *Journal of the American Chemical Society* 122.26 (2000), pp. 6268–6277.

[129] Kresten Lindorff-Larsen et al. “Systematic Validation of Protein Force Fields against Experimental Data”. In: *PLoS ONE* 7.2 (2012), e32131. DOI: 10.1371/journal.pone.0032131. URL: https://doi.org/10.1371/journal.pone.0032131.

[130] Hai Nguyen, Daniel R. Roe, and Carlos Simmerling. “Improved Generalized Born Solvent Model Parameters for Protein Simulations”. en. In: *Journal of Chemical Theory and Computation* 9.4 (Apr. 2013), pp. 2020–2034. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct3010485. URL: https://pubs.acs.org/doi/10.1021/ct3010485 (visited on 05/16/2025).

[131] Parviz Seifpanahi Shabane, Saeed Izadi, and Alexey V. Onufriev. “General Purpose Water Model Can Improve Atomistic Simulations of Intrinsically Disordered Proteins”. In: *Journal of Chemical Theory and Computation* 15.4 (2019), pp. 2620–2634.

[132] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. “Building Water Models: A Different Approach”. en. In: *The Journal of Physical Chemistry Letters* 5.21 (Nov. 2014), pp. 3863–3871. ISSN: 1948-7185, 1948-7185. DOI: 10.1021/jz501780a. URL: https://pubs.acs.org/doi/10.1021/jz501780a (visited on 04/25/2025).

[133] Anna-Lena M. Fischer et al. “The Role of Force Fields and Water Models in Protein Folding and Unfolding Dynamics”. en. In: *Journal of Chemical Theory and Computation* 20.5 (Mar. 2024), pp. 2321–2333. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/acs.jctc.3c01106. URL: https://pubs.acs.org/doi/10.1021/acs.jctc.3c01106 (visited on 04/25/2025).

[134] Paul Robustelli, Stefano Piana, and David E. Shaw. “Developing a molecular dynamics force field for both folded and disordered protein states”. en. In: *Proceedings of the National Academy of Sciences* 115.21 (May 2018). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1800690115. URL: https://pnas.org/doi/full/10.1073/pnas.1800690115 (visited on 04/25/2025).

[135] Dávid Péter Kovács et al. “Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE”. In: *Journal of Chemical Theory and Computation* 17.12 (2021), pp. 7696–7711.

[136] Brent H. Besler, Kenneth M. Merz, and Peter A. Kollman. “Atomic charges derived from semiempirical methods”. en. In: *Journal of Computational Chemistry* 11.4 (May 1990), pp. 431–439. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.540110404. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.540110404 (visited on 04/28/2025).

[137] Junmei Wang et al. “Automatic atom type and bond type perception in molecular mechanical calculations”. In: *Journal of Molecular Graphics and Modelling* 25.2 (Oct. 2006), pp. 247–260. ISSN: 10933263. DOI: 10.1016/j.jmgm.2005.12.005.

[138] QCMM group. *ffparaim*. 2022. URL: https://github.com/QCMM/ffparaim.

[139] D. G. A. Smith, L. A. Burns, and et al. Simmonett. “Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry”. In: *J. Chem. Phys.* (2020). DOI: 10.1063/5.0006002.

[140] Hugo Lebrette et al. “Structure of a ribonucleotide reductase R2 protein radical”. en. In: *Science* 382.6666 (Oct. 2023), pp. 109–113. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.adh8160. URL: https://www.science.org/doi/10.1126/science.adh8160 (visited on 04/28/2025).

[141] Martin Högbom, Britt-Marie Sjöberg, and Gustav Berggren. “Radical Enzymes”. en. In: *Encyclopedia of Life Sciences*. 1st ed. Wiley, Sept. 2020, pp. 375–393. ISBN: 978-0-470-01617-6 978-0-470-01590-2. DOI: 10.1002/9780470015902.a0029205. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0029205 (visited on 04/28/2025).

[142] E. R. Stadtman. “OXIDATION OF FREE AMINO ACIDS AND AMINO ACID RESIDUES IN PROTEINS BY RADIOLYSIS AND BY METAL-CATALYZED REACTIONS”. en. In: *Annual Review of Biochemistry* 62.1 (June 1993), pp. 797–821. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.bi.62.070193.004053. URL: https://www.annualreviews.org/doi/10.1146/annurev.bi.62.070193.004053 (visited on 04/28/2025).

[143] Thomas Nauser, Jill Pelling, and Christian Schöneich. “Thiyl Radical Reaction with Amino Acid Side Chains: Rate Constants for Hydrogen Transfer and Relevance for Posttranslational Protein Modification”. en. In: *Chemical Research in Toxicology* 17.10 (Oct. 2004), pp. 1323–1328. ISSN: 0893-228X, 1520-5010. DOI: 10.1021/tx049856y. URL: https://pubs.acs.org/doi/10.1021/tx049856y (visited on 04/29/2025).

[144] Wing Tsang, W. Sean McGivern, and Jeffrey A. Manion. “Multichannel decomposition and isomerization of octyl radicals”. en. In: *Proceedings of the Combustion Institute* 32.1 (2009), pp. 131–138. ISSN: 15407489. DOI: 10.1016/j.proci.2008.05.048. URL: https://linkinghub.elsevier.com/retrieve/pii/S1540748908002460 (visited on 04/29/2025).

[145] Damian Moran et al. "Rearrangements in Model Peptide-Type Radicals *via* Intramolecular Hydrogen-Atom Transfer". en. In: *Helvetica Chimica Acta* 89.10 (Oct. 2006), pp. 2254–2272. ISSN: 0018-019X, 1522-2675. DOI: 10.1002/hlca.200690210. URL: https://onlinelibrary.wiley.com/doi/10.1002/hlca.200690210 (visited on 06/13/2024).

[146] Baptiste Sirjean et al. "Tunneling in Hydrogen-Transfer Isomerization of *n* -Alkyl Radicals". en. In: *The Journal of Physical Chemistry A* 116.1 (Jan. 2012), pp. 319–332. ISSN: 1089-5639, 1520-5215. DOI: 10.1021/jp209360u. URL: https://pubs.acs.org/doi/10.1021/jp209360u (visited on 04/29/2025).

[147] Joseph P. R. O. Orgel et al. "Microfibrillar structure of type I collagen *in situ*". en. In: *Proceedings of the National Academy of Sciences* 103.24 (June 2006), pp. 9001–9005. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0502718103. URL: https://pnas.org/doi/full/10.1073/pnas.0502718103 (visited on 06/13/2024).

[148] Daniel R. Roe and Bernard R. Brooks. "A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations". en. In: *The Journal of Chemical Physics* 153.5 (Aug. 2020), p. 054123. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/5.0013849. URL: https://pubs.aip.org/jcp/article/153/5/054123/1065686/A-protocol-for-preparing-explicitly-solvated (visited on 04/29/2025).

[149] Vytautas Gapsys, Daniel Seeliger, and Bert L. De Groot. "New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations". en. In: *Journal of Chemical Theory and Computation* 8.7 (July 2012), pp. 2373–2382. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct300220p. URL: https://pubs.acs.org/doi/10.1021/ct300220p (visited on 05/05/2025).

[150] Antonia S. J. S. Mey et al. "Best Practices for Alchemical Free Energy Calculations [Article v1.0]". eng. In: *Living Journal of Computational Molecular Science* 2.1 (2020), p. 18378. ISSN: 2575-6524. DOI: 10.33011/livecoms.2.1.18378.

[151] Daniel Sheppard, Rye Terrell, and Graeme Henkelman. "Optimization methods for finding minimum energy paths". en. In: *The Journal of Chemical Physics* 128.13 (Apr. 2008), p. 134106. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2841941. URL: https://pubs.aip.org/jcp/article/128/13/134106/977389/Optimization-methods-for-finding-minimum-energy (visited on 05/02/2025).

[152] Toby Lewis-Atwell, Piers A. Townsend, and Matthew N. Grayson. "Machine learning activation energies of chemical reactions". en. In: *WIREs Computational Molecular Science* 12.4 (July 2022), e1593. ISSN: 1759-0876, 1759-0884. DOI: 10.1002/wcms.1593. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1593 (visited on 05/02/2025).

[153] Benoît Roux. "The calculation of the potential of mean force using computer simulations". en. In: *Computer Physics Communications* 91.1-3 (Sept. 1995), pp. 275–282. ISSN: 00104655. DOI: 10.1016/0010-4655(95)00053-I. URL: https://linkinghub.elsevier.com/retrieve/pii/001046559500053I (visited on 05/05/2025).

[154] D L Beveridge and F M DiCapua. "Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems". en. In: ().

[155] Weinan E and Eric Vanden-Eijnden. "Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events". en. In: *Annual Review of Physical Chemistry* 61.1 (Mar. 2010), pp. 391–420. ISSN: 0066-426X, 1545-1593. DOI: 10.1146/annurev.physchem.040808.090412. URL: https://www.annualreviews.org/doi/10.1146/annurev.physchem.040808.090412 (visited on 05/05/2025).

[156] Tilmann Gneiting and Adrian E Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". en. In: *Journal of the American Statistical Association* 102.477 (Mar. 2007), pp. 359–378. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214506000001437. URL: http://www.tandfonline.com/doi/abs/10.1198/016214506000001437 (visited on 05/05/2025).

[157] Naveen Michaud-Agrawal et al. “MDAnalysis: A toolkit for the analysis of molecular dynamics simulations”. en. In: *Journal of Computational Chemistry* 32.10 (July 2011), pp. 2319–2327. ISSN: 0192-8651, 1096-987X. DOI: 10.1002/jcc.21787. URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.21787 (visited on 05/05/2025).

[158] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. “Enhanced sampling techniques in molecular dynamics simulations of biological systems”. en. In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850.5 (May 2015), pp. 872–877. ISSN: 03044165. DOI: 10.1016/j.bbagen.2014.10.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304416514003559 (visited on 05/06/2025).

[159] Erich Gamma, ed. *Design patterns: elements of reusable object-oriented software.* eng. 39. printing. Addison-Wesley professional computing series. Boston, Mass. Munich: Addison-Wesley, 2011. ISBN: 978-0-201-63361-0.

[160] Gareth A. Tribello et al. “PLUMED 2: New feathers for an old bird”. en. In: *Computer Physics Communications* 185.2 (Feb. 2014), pp. 604–613. ISSN: 00104655. DOI: 10.1016/j.cpc.2013.09.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010465513003196 (visited on 05/07/2025).

[161] Ryan Jacobs et al. “A practical guide to machine learning interatomic potentials – Status and future”. en. In: *Current Opinion in Solid State and Materials Science* 35 (Mar. 2025), p. 101214. ISSN: 13590286. DOI: 10.1016/j.cossms.2025.101214. URL: https://linkinghub.elsevier.com/retrieve/pii/S1359028625000014 (visited on 05/07/2025).

[162] Frank Noé et al. “Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning”. en. In: *Science* 365.6457 (Sept. 2019), eaaw1147. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaw1147. URL: https://www.science.org/doi/10.1126/science.aaw1147 (visited on 05/07/2025).

[163] Leon Klein, Andreas Krämer, and Frank Noe. “Equivariant flow matching”. In: *Advances in Neural Information Processing Systems.* Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 59886–59910. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/bc827452450356f9f558f4e4568d553b-Paper-Conference.pdf.

[164] Nicolas Wolf et al. *Learning conformational ensembles of proteins based on backbone geometry.* Version Number: 1. 2025. DOI: 10.48550/ARXIV.2503.05738. URL: https://arxiv.org/abs/2503.05738 (visited on 05/07/2025).

[165] Hannah K. Wayment-Steele et al. “Predicting multiple conformations via sequence clustering and AlphaFold2”. en. In: *Nature* 625.7996 (Jan. 2024), pp. 832–839. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06832-9. URL: https://www.nature.com/articles/s41586-023-06832-9 (visited on 05/07/2025).

[166] Andy B. Yoo, Morris A. Jette, and Mark Grondona. “SLURM: Simple Linux Utility for Resource Management”. In: *Job Scheduling Strategies for Parallel Processing.* Ed. by Gerhard Goos et al. Vol. 2862. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 44–60. ISBN: 978-3-540-20405-3 978-3-540-39727-4. DOI: 10.1007/10968987_3. URL: http://link.springer.com/10.1007/10968987_3 (visited on 05/07/2025).

[167] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD: Visual molecular dynamics”. In: *Journal of Molecular Graphics* 14.1 (Feb. 1996), pp. 33–38. ISSN: 02637855. DOI: 10.1016/0263-7855(96)00018-5. URL: https://linkinghub.elsevier.com/retrieve/pii/0263785596000185.

[168] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015.

[169] David R. Eyre, Mary Ann Weis, and Jiann-Jiu Wu. “Advances in collagen crosslink analysis”. en. In: *Methods* 45.1 (May 2008), pp. 65–74. ISSN: 10462023. DOI: 10.1016/j.ymeth.2008.01.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S104620230800025X (visited on 05/10/2025).

[170] Thomas Kress and Melinda J. Duer. "Solid-State NMR Spectroscopy Investigation of Structural Changes of Mechanically Strained Mouse Tail Tendons". en. In: *Journal of the American Chemical Society* 147.11 (Mar. 2025), pp. 9220–9228. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.4c13930. URL: https://pubs.acs.org/doi/10.1021/jacs.4c13930 (visited on 05/25/2025).

[171] Wojtek Treyde, Kai Riedmiller, and Frauke Gräter. "Bond dissociation energies of X–H bonds in proteins". en. In: *RSC Advances* 12.53 (2022), pp. 34557–34564. ISSN: 2046-2069. DOI: 10.1039/D2RA04002F. URL: https://xlink.rsc.org/?DOI=D2RA04002F (visited on 05/11/2025).

[172] Jacob J. A. Garwood, Andrew D. Chen, and David A. Nagib. "Radical Polarity". en. In: *Journal of the American Chemical Society* (Oct. 2024), jacs.4c06774. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.4c06774. URL: https://pubs.acs.org/doi/10.1021/jacs.4c06774 (visited on 05/11/2025).

[173] Michela Salamone et al. "Bimodal Evans–Polanyi Relationships in Hydrogen Atom Transfer from C(sp^3)–H Bonds to the Cumyloxyl Radical. A Combined Time-Resolved Kinetic and Computational Study". en. In: *Journal of the American Chemical Society* 143.30 (Aug. 2021), pp. 11759–11776. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.1c05566. URL: https://pubs.acs.org/doi/10.1021/jacs.1c05566 (visited on 05/11/2025).

[174] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (Jan. 1995), pp. 289–300. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://academic.oup.com/jrsssb/article/57/1/289/7035855 (visited on 05/11/2025).

[175] Jörg Fischer et al. "Current Developments in Operando Electron Paramagnetic Resonance Spectroscopy". In: *CHIMIA* 78.5 (May 2024), pp. 326–332. ISSN: 2673-2424, 0009-4293. DOI: 10.2533/chimia.2024.326. URL: https://www.chimia.ch/chimia/article/view/2024_326 (visited on 05/11/2025).

[176] Katarzyna P. Adamala et al. "Confronting risks of mirror life". en. In: *Science* 386.6728 (Dec. 2024), pp. 1351–1353. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ads9158. URL: https://www.science.org/doi/10.1126/science.ads9158 (visited on 05/12/2025).

[177] Jennifer L. Radkiewicz et al. "Accelerated Racemization of Aspartic Acid and Asparagine Residues via Succinimide Intermediates: An ab Initio Theoretical Exploration of Mechanism". en. In: *Journal of the American Chemical Society* 118.38 (Jan. 1996), pp. 9148–9155. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja953505b. URL: https://pubs.acs.org/doi/10.1021/ja953505b (visited on 05/12/2025).

[178] Thomas V Stabler et al. "Amino acid racemization reveals differential protein turnover in osteoarthritic articular and meniscal cartilages". en. In: *Arthritis Research & Therapy* 11.2 (Mar. 2009), R34. ISSN: 1478-6354. DOI: 10.1186/ar2639. URL: https://arthritis-research.biomedcentral.com/articles/10.1186/ar2639 (visited on 05/12/2025).

[179] Andrew Ballard et al. "Racemisation in Chemistry and Biology". en. In: *Chemistry – A European Journal* 26.17 (Mar. 2020), pp. 3661–3687. ISSN: 0947-6539, 1521-3765. DOI: 10.1002/chem.201903917. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.201903917 (visited on 05/12/2025).

[180] A D'Aniello et al. "Biological role of D-amino acid oxidase and D-aspartate oxidase. Effects of D-amino acids." en. In: *Journal of Biological Chemistry* 268.36 (Dec. 1993), pp. 26941–26949. ISSN: 00219258. DOI: 10.1016/S0021-9258(19)74201-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S002192581974201X (visited on 05/12/2025).

[181] Yangyang Shen et al. “Site-Selective -C–H Functionalization of Trialkylamines via Reversible Hydrogen Atom Transfer Catalysis”. en. In: *Journal of the American Chemical Society* 143.45 (Nov. 2021), pp. 18952–18959. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.1c07144. URL: https://pubs.acs.org/doi/10.1021/jacs.1c07144 (visited on 05/12/2025).

[182] Roger Jan Kutta et al. “Multifaceted View on the Mechanism of a Photochemical Deracemization Reaction”. en. In: *Journal of the American Chemical Society* 145.4 (Feb. 2023), pp. 2354–2363. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.2c11265. URL: https://pubs.acs.org/doi/10.1021/jacs.2c11265 (visited on 05/12/2025).

[183] Hartmut Frank, Greme J. Nicholson, and E. Bayer. “Enantiomer labelling, a method for the quantitative analysis of amino acids”. en. In: *Journal of Chromatography A* 167 (Dec. 1978), pp. 187–196. ISSN: 00219673. DOI: 10.1016/S0021-9673(00)91157-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021967300911579 (visited on 05/12/2025).

[184] Baptiste Depalle et al. “Influence of cross-link structure, density and mechanical properties in the mesoscale deformation mechanisms of collagen fibrils”. en. In: *Journal of the Mechanical Behavior of Biomedical Materials* 52 (Dec. 2015), pp. 1–13. ISSN: 17516161. DOI: 10.1016/j.jmbbm.2014.07.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S175161611400201X (visited on 05/12/2025).