Julius-Maximilians-Universität Würzburg
Fakultät für Chemie und Pharmazie

Mechanistic Insights into SARS Coronavirus Main Protease by Computational Chemistry Methods

Dissertation

zur Erlangung des naturwissenschaftlichen Doktorgrads
der Julius-Maximilians-Universität Würzburg

vorgelegt von Alexander Paasche
aus Schweinfurt

Würzburg 2013

Eingereicht am: ....

bei der Fakultät für Chemie und Pharmazie

1. Gutachter: ....

2. Gutachter: ....

der Dissertation

1. Prüfer: ....

2. Prüfer: ....

3. Prüfer: ....

des öffentlichen Promotionskolloquiums

Tag des des öffentlichen Promotionskolloquiums: ....

Doktorurkunde ausgehändigt am: ....

Für Christina & Emma

Abstract

The SARS virus is the etiological agent of the severe acute respiratory syndrome, a deadly disease that caused more than 700 causalities in 2003. One of its viral proteins, the SARS coronavirus main protease, is considered as a potential drug target and represents an important model system for other coronaviruses. Despite extensive knowledge about this enzyme, it still lacks an effective anti-viral drug. Furthermore, it possesses some unusual features related to its active-site region.

This work gives atomistic insights into the SARS coronavirus main protease and tries to reveal mechanistic aspects that control catalysis and inhibition. Thereby, it applies state-of-the-art computational methods to develop models for this enzyme that are capable to reproduce and interpreting the experimental observations. The theoretical investigations are elaborated over four main fields that assess the accuracy of the used methods, and employ them to understand the function of the active-site region, the inhibition mechanism, and the ligand binding.

The testing of different quantum chemical methods reveals that their performance depends partly on the employed model. This can be a gas phase description, a continuum solvent model, or a hybrid QM/MM approach. The latter represents the preferred method for the atomistic modeling of biochemical reactions. A benchmarking uncovers some serious problems for semi-empirical methods when applied in proton transfer reactions.

To understand substrate cleavage and inhibition of SARS coronavirus main protease, proton transfer reactions between the Cys/His catalytic dyad are calculated. Results show that the switching between neutral and zwitterionic state plays a central role for both mechanisms. It is demonstrated that this electrostatic trigger is remarkably influenced by substrate binding. Whereas the occupation of the active-site by the substrate leads to a fostered zwitterion formation, the inhibitor binding does not mimic this effect for the employed example. The underlying reason is related to the coverage of the active-site by the ligand, which gives new implications for rational improvements of inhibitors.

More detailed insights into reversible and irreversible inhibition are derived from *in silico* screenings for the class of Michael acceptors that follow a conjugated addition reaction. From the comparison of several substitution patterns it becomes obvious that different inhibitor warheads follow different mechanisms. Nevertheless, the initial formation of a zwitterionic catalytic dyad is found as a common precondition for all inhibition reactions.

Finally, non-covalent inhibitor binding is investigated for the case of SARS coranavirus main protease in complex with the inhibitor TS174. A novel workflow is developed that includes an interplay between theory and experiment in terms of molecular dynamic simulation, tabu search, and X-ray structure refinement. The results show that inhibitor binding is possible for multiple poses and stereoisomers of TS174.

Zusammenfassung

Das Schwere Akute Respiratorische Syndrom (SARS) wird durch eine Infektion mit dem SARS Virus ausgelöst, dessen weltweite Verbreitung 2003 zu über 700 Todesfällen führte. Die SARS Coronavirus Hauptprotease stellt ein mögliches Wirkstoffziel zur Behandlung dar und hat Modellcharakter für andere Coronaviren. Trotz intensiver Forschung sind bis heute keine effektiven Wirkstoffe gegen SARS verfügbar.

Die vorliegende Arbeit gibt Einblicke in die mechanistischen Aspekte der Enzymkatalyse und Inhibierung der SARS Coronavirus Hauptprotease. Hierzu werden moderne computerchemische Methoden angewandt, die mittels atomistischer Modelle experimentelle Ergebnisse qualitativ reproduzieren und interpretieren können. Im Zuge der durchgeführten theoretischen Arbeiten wird zunächst eine Fehlereinschätzung der Methoden durchgeführt und diese nachfolgend auf Fragestellungen zur aktiven Tasche, dem Inhibierungsmechanismus und der Ligandenbindung angewandt.

Die Einschätzung der quantenchemischen Methoden zeigt, dass deren Genauigkeit teilweise von der Umgebungsbeschreibung abhängt, welche als Gasphasen, Kontinuum, oder QM/MM Modell dargestellt werden kann. Letzteres gilt als Methode der Wahl für die atomistische Modellierung biochemischer Reaktionen. Die Vergleiche zeigen für semi-empirische Methoden gravierende Probleme bei der Beschreibung von Proton-Transfer Reaktionen auf. Diese wurden für die katalytische Cys/His Dyade betrachtet, um Einblicke in Substratspaltung und Inhibierung zu erhalten. Dem Wechsel zwischen neutralem und zwitterionischem Zustand konnte hierbei eine zentrale Bedeutung für beide Prozesse zugeordnet werden. Es zeigt sich, dass dieser „electrostatic trigger“ von der Substratbindung, nicht aber von der Inhibitorbindung beeinflusst wird. Folglich beschleunigt ausschließlich die Substratbindung die Zwitterionbildung, was im Zusammenhang mit der Abschirmung der aktiven Tasche durch den Liganden steht. Dies gibt Ansatzpunkte für die Verbesserung von Inhibitoren.

Aus *in silico* screenings werden genauere Einblicke in die reversible und irreversible Inhibierung durch Michael-Akzeptor Verbindungen gewonnen. Es wird gezeigt, dass unterschiedlichen Substitutionsmustern unterschiedliche Reaktionsmechanismen in der konjugierten Additionsreaktion zugrunde liegen. Die vorangehende Bildung eines Cys^-/His^+ Zwitterions ist allerdings für alle Inhibierungsmechanismen eine notwendige Voraussetzung.

Letztendlich wurde die nicht-kovalente Bindung eines Inhibitors am Beispiel des TS174-SARS Coronavirus Hauptprotease Komplexes untersucht. Im Zusammenspiel von Theorie und Experiment wurde ein Prozess, bestehend aus Molekulardynamik Simulation, Tabu Search und Röntgenstruktur Verfeinerung ausgearbeitet, der eine Interpretation der Bindungssituation von TS174 ermöglicht. Im Ergebnis zeigt sich, dass der Inhibitor gleichzeitig in mehreren Orientierungen, als auch in beiden stereoisomeren Formen im Komplex vorliegt.

Preface

Die atomistische Modellierung von Enzymen ist eine der großen Herausforderungen für die Computerchemie. Durch die immer größer werdende Vielfalt des theoretisch-chemischen „Werkzeugkastens“ bietet sie, jenseits der Grenzen des Experiments, immer mehr individuelle Wege sich einer Fragestellung zu nähern. Im Rahmen einer Doktorarbeit muss dieser Spielraum aber auch gewährt werden. Gerade hierfür danke ich an erster Stelle meinem Doktorvater Prof. Bernd Engels. Er hat mir diesen akademischen Freiraum nicht nur gegeben, sondern auch durch seine fachliche Unterstützung, das Ermöglichen von Tagungsbesuchen und Fortbildungen und vor allem durch sein Vertrauen in mich, meine Fähigkeiten um ein Vielfaches erweitert. Auch für die Freiheit, neben der Promotion einem wirtschaftswissenschaftlichen Abendstudium nachzugehen, bedanke ich mich recht herzlich.

Prof. Tanja Schirmeister gebührt mein besonderer Dank für die fachliche Unterstützung aus dem pharmazeutischen Bereich sowie Prof. John Ziebuhr für Diskussionen und Beiträge aus der biologischen Perspektive. Prof. Reinhold Fink danke ich für seine Hilfe bei quantenchemischen und programmiertechnischen Fragestellungen und viele angenehme Gespräche jenseits der fachlichen Dinge. An Prof. Matthias Breuning geht mein Dank für die unkomplizierte Zusammenarbeit und die Möglichkeit, meine theoretischen Kenntnisse in seinem experimentellen Themenfeld praxisnah anwenden zu dürfen.

Für die gute Zusammenarbeit an gemeinsamen Projekten danke ich auch Uwe Dietzel, Dr. Markus Schiller sowie Lukas Pason und Sebastian Brickel. Ein großes Dankeschön gebührt Zarah Falk für ihre Geduld und Hilfe, die sprachlichen Schlaglöcher meiner Arbeit im Englischen auszubessern. Dass das Thema SARS offensichtlich eine große Anziehungskraft hat, zeigt das große Interesse bei Praktikanten und Bachelorstudenten. Meinen Praktikanten Max Rieger, Jan Mies, Thomas Kramer, Michael Krapf, Thomas Herdmann, sowie meinen Bachelorstudenten Simon Schäfer und Andreas Zipper danke ich für ihre engagierte Mitarbeit.

Weiterhin haben viele nette Menschen der Arbeitsgruppe dazu beigetragen, dass ich auf eine wunderschöne Promotionszeit in Würzburg zurückblicken kann. Besonders zu erwähnen sind hier Uschi Rüppel für Ihre Unterstützung in allen Lebenslagen sowie Thomas Schmidt, Christoph Grebner oder Johannes Becker für viel gemeinsamen Spaß bei Computerbasteleien und sonstigem groben Unfug, der den wissenschaftlichen Alltag immer wieder aufgelockert hat. Ferner danke ich allen anderen Kollegen der Arbeitsgruppe für die schöne Zeit, auch den „alten Hasen“ Milena Mladenovic und Sebastian Schlund, von denen ich gerade am Anfang meiner Promotion viel gelernt habe.

Da das Beste bekanntlich zum Schluss kommt, gebührt dieser Platz meiner Frau Christina. Ohne deine bedingungslose und liebevolle Unterstützung wäre ich heute nicht da wo ich bin!

Würzburg, Juni 2013

Table of contents

1 Introduction

When one hears the word “SARS” most people think of the fearsome epidemic that swept the world at the beginning of the 21th century. In fact, the spread of the previously unknown human SARS coronavirus in 2003 has received much attention worldwide and during that time had a great impact on public life. Nearly 10 years later, SARS has lost its visibility and does not fill the headlines of newspapers or covers of magazines any more, but its impact on the scientific community is still vital. In 2011, an impressive number of more than 500 publications have been published containing the keyword “SARS”, and moreover the statistics indicate that we can expect a sustainable interest in this topic for several years (Fig. 1-1).[1,2] The enduring interest in this topic shows that scientists are still learning their lessons from the SARS epidemic. This situation is resonated by the overall number of publications related to the severe acute respiratory syndrome, which has exceeded 10.000 in the year 2010.

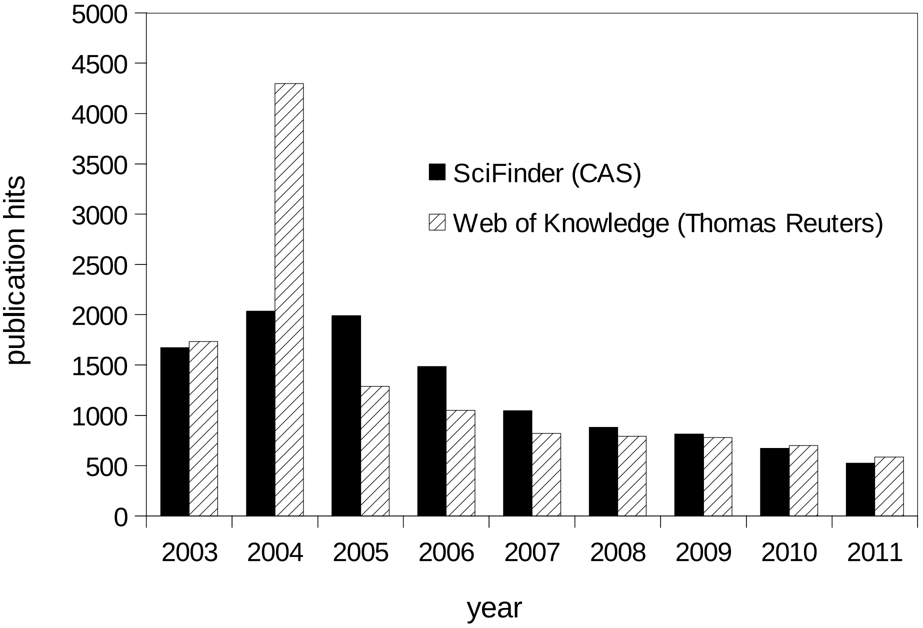


Fig. 1-1: Number of SARS-related publications with respect to publication year, estimated by two commonly used scientific database search engines.

But what makes SARS different from other diseases? One major reason might be that it has been considered as something being completely novel.[3,4] Pneumonia-associated human *coronaviradea* have been unknown before the human SARS coronavirus (SARS-CoV) was identified as the etiological agent of the SARS epidemic.[5–7] This fact especially contributed to the frightening reputation of SARS as a deadly disease. The lack of an effective therapy against this new illness led to a fatality rate of nearly 10% among infected people. During the outbreak and development of the disease, the world health organization WHO counted 8096 probable cases of SARS from which 774 people have died.[8]

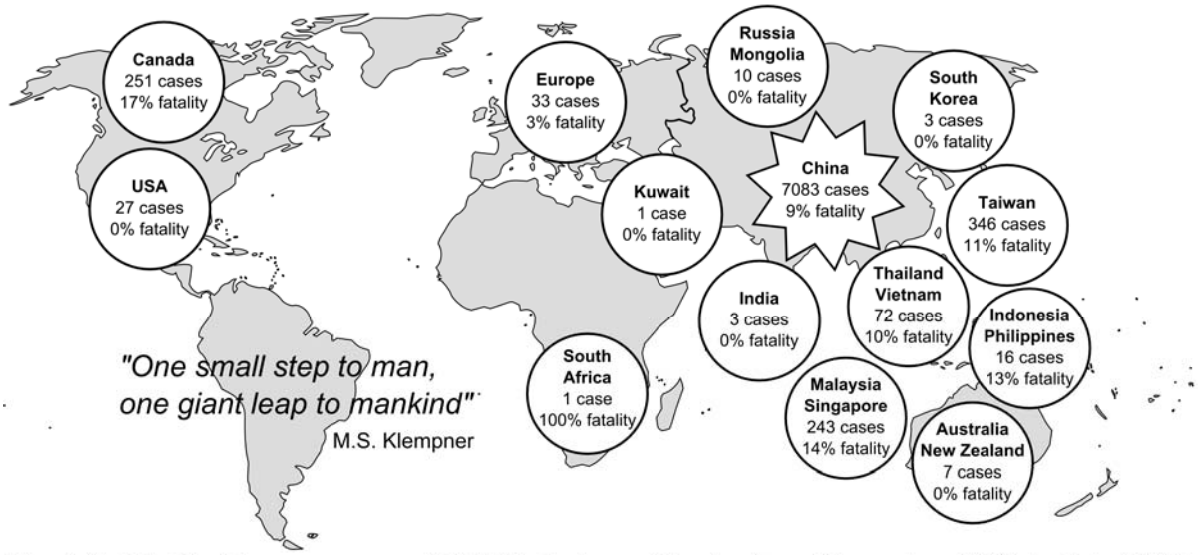


Fig. 1-2: Worldwide appearance of SARS during epidemic from November 2002 to July 2003. Number of probable cases and deaths are based on statistics of the world health organization WHO.[8]

The clinical symptoms of a SARS infection are fever and other indications similar to the common cold, like dizziness, cough or headache. About one quarter of affected patients have to be treated in intensive care units, partly with need for mechanical ventilation due to respiratory failure. SARS infections have a mean incubation time of 6 days and show in their severity a remarkable dependence on the age and constitution of the infected individual. This means that older people are more often affected by an adverse outcome than younger and healthy people.[9]

The origin of SARS-CoV seems to stem from wild life animals. Himalayan palm civets, which are retailed on life-animal markets in China, were found to be infected by SARS-CoV-like viruses. Therefore it seems likely that SARS-CoV first entered the human population at these markets, since people there live in close proximity with the animals mentioned above.[10] However, surveillance studies indicate that palm civets play only a role as intermediate hosts and bat species are the natural reservoir of SARS-CoV.[11,12] In the context of this aspect, SARS is probably the most studied example for a virus jumping from animals to humans to date.[13] After crossing the interspecies barrier, the virus better adopted to the human host by mutation and as a result, SARS-CoV transmissions from human to human were possible (Fig. 1-3).[14] This was the key issue that made SARS-CoV so successful from the viral perspective. The major route of transmission is caused by droplet infection.[15] However, despite its bad reputation depicted by the public media, the average number of SARS secondary infections caused by a single affected individual is lower than that of HIV, smallpox, or influenza. Therefore SARS-CoV transmissions are relatively inefficient compared to these viruses.[16,17]

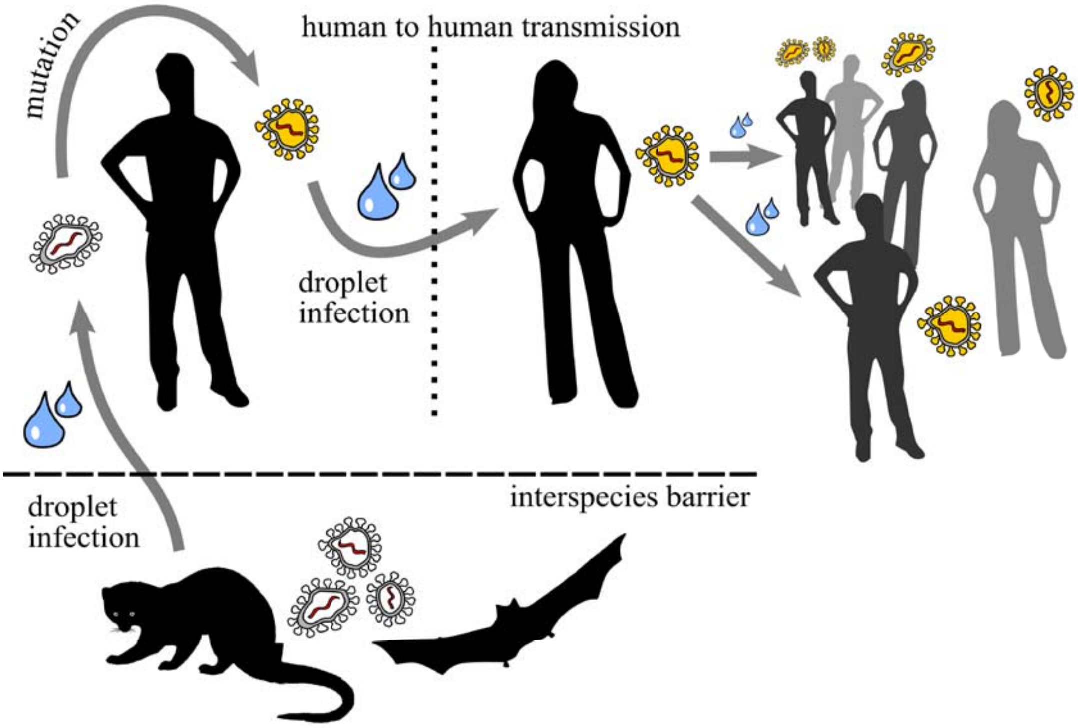


Fig. 1-3: The "giant leap to mankind":[14] Crossing of the interspecies barrier (dashed lines) and enabling of human to human transmission (dotted lines) through mutation of the SARS coronavirus. Virions of zoonotic origin are indicated in white, mutated virions in orange.

Nevertheless, SARS remains a deadly disease. Although it seems unlikely that SARS in its original form from 2003 will re-emerge again, the global epidemic of the severe acute respiratory syndrome played, and stills plays an important role in the history of infectious diseases. It is a warning example of how our modern society enables viruses to overcome continents within a very short time, igniting worldwide epidemics. Due to lessons learned from the SARS outbreak, successful intervention is today more understood as a global thread than a national issue. The need for fast response by international health communities has led to a better preparedness against future global outbreaks of SARS or other infectious diseases. [18] It has further led to bioethical discussions, for example, how to balance individual freedoms against the common good – an important question when health officials are faced with short term decisions to prevent the spread of a disease. [19] It is also worth to think about SARS in the context of bioterrorism prevention, which is a much attended topic since 2001. But because most emerging diseases have zoonotic origin, nature can clearly be seen as the greatest "bioterrorist".[20]

So far, the SARS coronavirus displays many features that attract scientists´ and academic researchers´ interest. Impressive progress has been made since 2003, but one of the most important questions at the end of the day still is: could we handle SARS today? The surprising answer is: not at all. Even today, no effective treatment of SARS is known.[21] The recent global alert from the world health organization about the emergence of a novel SARS-like coronavirus[22–24] is a further warning that, even after ten years, *coronaviridea* still state a real and ubiquitous threat that will not simply disappear in the near future. Especially this fact will ensure that scientists all over the world remain motivated to learn about them independently whether 10.000 scientific papers have been published or not.

2 The SARS Coronavirus Main Protease

2.1 A Promising Drug Target against SARS?

The term *coronavirus* is a description for a virus that belongs to the family of *coronaviridae,* which are named after the characteristic shape of their virions (Fig. 2-1). Their circular appearance of the viral envelope, with ray-like arranged spikes at the surface, reminds one of a crown and looks similar to a solar corona from which the name was adopted.[25]

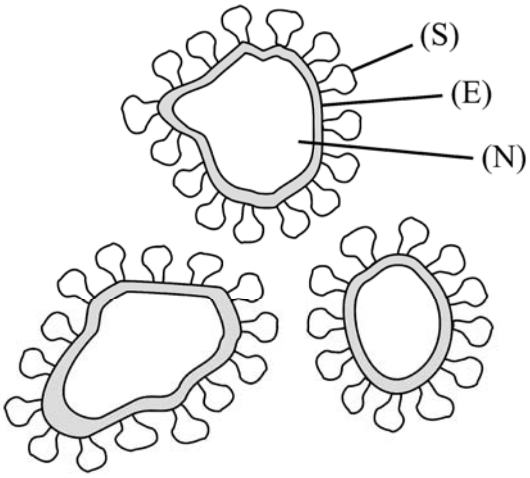


Fig. 2-1: Virions of the SARS coronavirus, composed of the spikes (S), viral envelope (E) and nucleocapsid (N).

The virus envelope is composed of different layers consisting of membrane (M-protein), envelope (E-protein), and nucleocapsid (N-proteins) proteins. Located at the surface of the virions are spikes, which resemble trimers of the spike proteins (S-protein) and are used by the virus to enter the host cell.[26] Within their role as transmembrane proteins, the S-proteins facilitate SARS-CoV to attach and fuse the viral cell membrane with the host cell membrane. This is one of the key steps during the infection of the target cell.[27] It has earlier been recognized that the S-proteins of human SARS-CoV possess a high affinity towards the cell-surface molecule angiotensin-converting enzyme 2 (ACE2), which are therefore used as receptors for the virus at the host cell.[28,29] ACE2 is found predominantly at the surface of endothelial cells. Since receptor binding to ACE2 is a unique feature of human SARS-CoV, compared to SARS coronaviruses from zoonotic origin, like bats or palm civets,[12] it seems plausible that this aspect is the crucial factor that enables human-to-human transmission. This is further underlined by the fact that sequences of the S-proteins differ for human SARS-CoV and animal SARS-CoV. The adaptation of the virus with respect to the receptor binding might therefore be the “giant leap to mankind”[14] that ignited the SARS epidemic in 2003.

A successful receptor binding involves the occupation of three receptors that are addressed by the spike protein trimers. The reason for this can easily be derived from intuition, since three is the minimum number to ensure a perpendicular orientation of the viral cell membrane towards the host cell membrane.[30,31] Nevertheless, this functional relationship is not proven at all, but seems likely, since similar motifs are also conserved in other virus families.[25]

Whenever a virion has been attached to the receptors, fusion of the cell membranes is initiated. The motif present in the SARS-CoV spikes is classified as Class I of membrane fusion proteins and therefore shares features with other viruses, like HIV.[32] It is notable that the mechanisms occurring at the interface between virion and the host cell can be used as drug targets to prevent the viral entry as an antiviral strategy.

After fusion of the viral and the host cell membrane has proceeded, the nucleocapsid is allowed to be inserted into the cellular cytoplasma. Within this step, the viral genome is released. Instead of DNA, the coronaviral genome is encoded in single positive RNA strands. Hence, SARS-CoV finds itself classified as a positive single stranded RNA virus within the international virus taxonomy.[33] Historically, the genome of the SARS coronavirus has been sequenced and published only a few months after the emergence of the SARS epidemic in 2003.[4] One outstanding feature of the *coronaviridae* family is their exceptionally large genome, which is so far the largest among all known RNA viruses.[34] They further tend to undergo rapid mutation and very efficient replication, which enables a fast evolution process for these viruses.[35]

Once the RNA genome has entered the infected cell, it is translated into two large polypeptides denoted as pp1a and pp1ab, whereas pp1ab is an extended version of ppa1.[36]

The newly synthesized polypeptide chains contain arrays of several nonstructural proteins (nsps) and need to be processed into 16 mature products, which serve for different functions within the viral lifecycle. These are partly essential factors for the replication of SARS-CoV, however, not all functions are known. Other parts of the translated genome are important for the viral structure build-up or have accessory functions.[36]

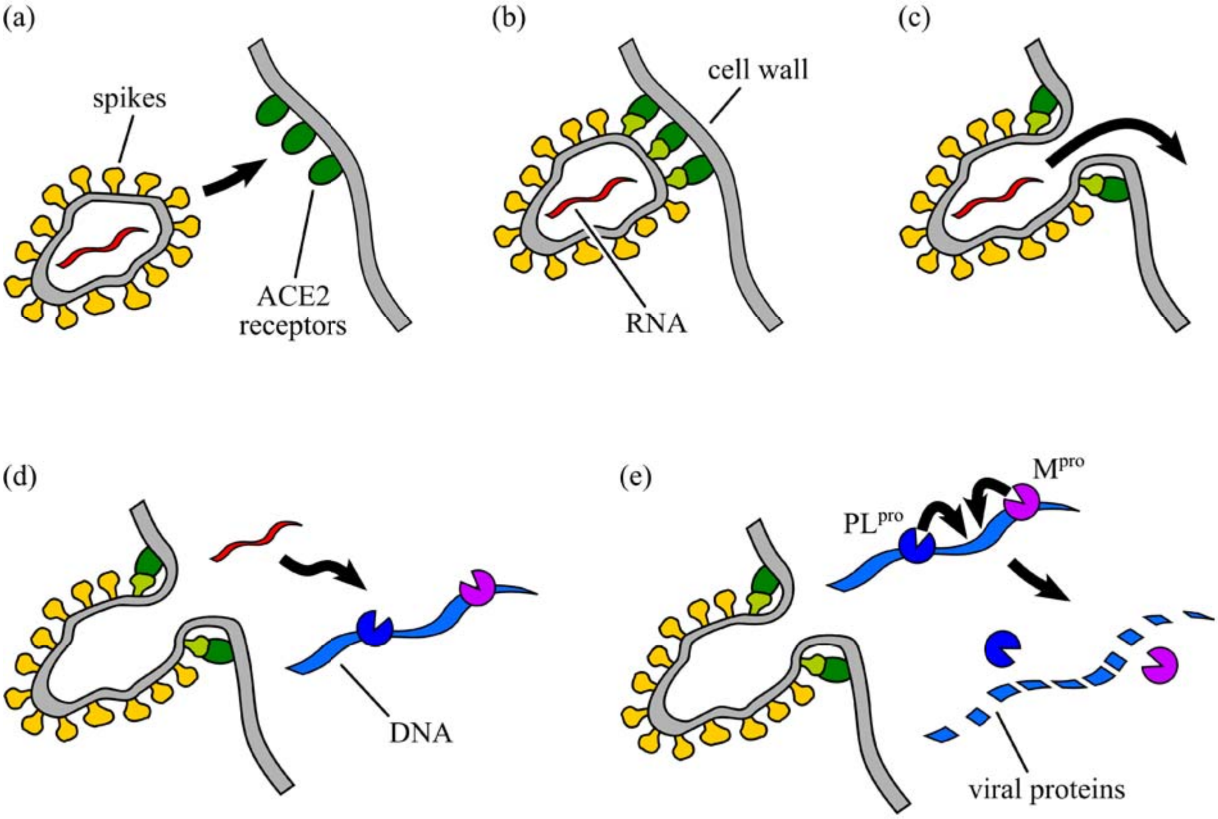


Fig. 2-2: Illustration of viral cell entry and infection by the SARS coronavirus. The crucial steps are denoted as follows: (a) A virion particle approaches the host cell membrane. (b) Attachment of the viral spike proteins (yellow) at the cell surface via binding to the angiotensin-converting enzyme 2 receptors (green) of the target cell. (c) Fusion of the cell membranes and release of the viral RNA (red) into the host cell. (d) Transcription of the viral RNA to DNA (blue) and transcription into protein by the infected host cell. (e) Autocatalytic cleavage of the polyprotein and release of the papain-like protease (dark blue) and main protease (magenta), followed by the proteolytic processing of the polyprotein into mature viral proteins (dashed blue) that are needed for the replication of new virions.

The cleavage of the polypeptide chains at the correct position into the 16 nonstructural proteins is an essential step for the viral lifecycle, since it converts the large amino acid sequences pp1a and pp1ab into functional viral proteins. This step is usually denoted as proteolytic processing and is managed by two important proteins in case of the SARS-CoV. Due to their function, these proteins are generally classified as proteases or proteinases.

The first important one is the papain-like protease (PL^{pro}) which belongs to the group of cysteine proteases and cuts off the nsp1, nsp2, nsp3, and nsp4 proteins from the large polypeptide chains.[37] The terminus cysteine protease categorizes the catalytic mechanism of this protease, since it involves a cysteine amino acid at the active-site which participates in the cleavage mechanism of the processed peptide bond. Beside this mechanism, there are currently eight other families of proteases known, namely aspartic, glutamic, metallo, asparagine, serine, threonine, and the family of mixed or unknown proteases.[38]

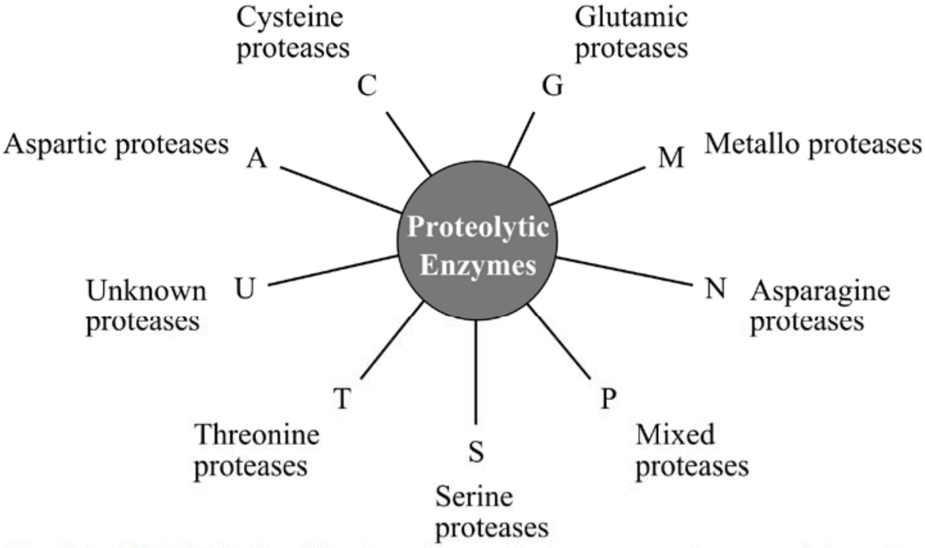


Fig. 2-3: MEROPS classification of proteolytic enzymes (proteases) into nine families, based on amino acid sequence similarities.[38]

In the case of SARS-CoV, the PL^{pro} domain resides within nsp3 and cleaves itself from the polypeptide in an autocatalytic way. This means that at the beginning, the enzyme catalysis proceeds slowly, since the catalytic active proteases have first to release themselves. As a consequence, the concentration for this enzyme is relatively low at the beginning of the process. Later on, the concentration of the PL protease increases very fast because it catalyzes its own release.

The remaining 11 cleavage sites (nsp5 to nsp16) of the polypeptide chains ppa1 and pp1ab are cleaved by the main protease (M^{pro}), which is the second important protein for proteolytic processing. It is also often denoted as 3CL-like protease (3CL^{pro}) since it shares cleavage-site specificity with the analogue 3C protease in picornavirus.[39] The M^{pro} domain resides in nsp5 and undergoes a similar autocatalytic cleavage process as the PL^{pro}. It has recently been discussed whether the M^{pro} comprises different quaternary structures to increase catalytic activity at low concentrations.[40]

The proteolytic processing of the polyproteins to mature replicase proteins, conducted by the PL^{pro} and M^{pro}, provides all active components of the viral replicase complex[41] and can therefore be considered as the “replication machinery” of the SARS coronavirus. Everything that perturbs, prevents, or blocks these two crucial proteases in their natural function can, therefore, be used as an ansatz for antiviral therapy. Beside the development of vaccines or antibodies, this strategy can be elaborated towards a highly specific and effective treatment of the viral infection by attacking either PL^{pro} or M^{pro}. This can typically be conducted by chemical substances. Substances that consciously act in this way are typically denoted as (protease) inhibitors, where the targeted enzymes are described as the so-called drug targets.

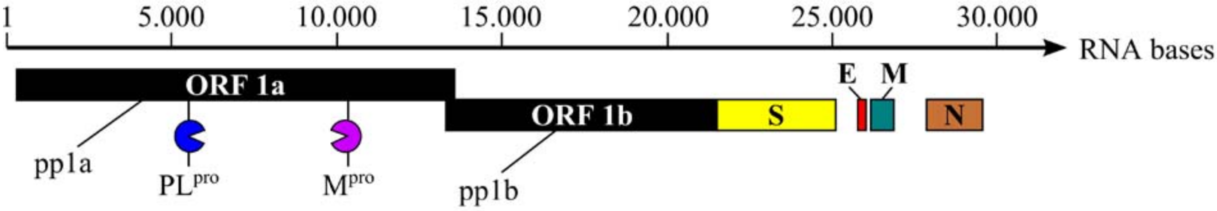


Fig. 2-4: Genome of SARS-CoV,[3] comprising open reading frames (ORFs) for the encoding of non-structural proteins (black bars) and structural proteins (colored bars). The papain-like protease (PL^{pro}, dark blue) and main protease (M^{pro}, magenta) reside within the non-structural protein region of polyprotein pp1a. Important structural proteins are in particular the spike protein (S, yellow), the envelope protein (E, red), the membrane protein (M, cyan), and the nucleocapsid protein (N, brown).

To date, relatively less effort has been spent on the PL^{pro} as a drug target and its significance as a potential aim for antiviral therapy has first been addressed throughout the last years.[42,43] For different reasons, much more attention has been paid to the M^{pro} as a drug target. One factor is the differently pronounced homology of the two proteases towards other known proteins of the coronavirus family. The PL^{pro} is not conserved in human coronavirus 229E (HCoV 229E), transmissible gastroenterities coronavirus (TGEV), mouse hepatitis virus (MHV), or avian infectious bronchities virus (IBV). On the other hand, the M^{pro} shares nearly 50% amino acid sequence identity with the proteases of the viruses mentioned above and moreover other *coronaviridae* of this kind.[44] This fact made it possible to construct a three-dimensional structure model of SARS-CoV M^{pro} from two homologue proteases[45] that was published just 4 months before the first "real" crystal structure was reported at the end of 2003.[46] It has also been found that M^{pro} provides similar substrate specificities than transmissible gastroenteritis virus (TGEV) and several other *coronaviridae.*[41]

Taking these things together, the SARS-CoV M^{pro} is generally considered as a very promising drug target. The fact that main proteases comprise the best characterized enzymes among the coronavirus family,[44] gives helpful assistance for the drug development process. Since the substrate specificity of SARS-CoV M^{pro} is moreover well conserved for *coronaviridae*,[47] it serves also as a valuable model system and states, therefore, an interesting and potential object for research.

2.2 The Anatomy of a Viral Cysteine Protease

Knowledge of the molecular structure of a drug target is the first step towards a rational drug design. Further it is an essential issue for the understanding of the underlying mechanisms facilitating enzyme catalysis on the molecular level. Since SARS-CoV M^{pro} has recently been recognized as a possible and promising drug target against the SARS virus, impressive research effort has been spent to get structural insights into this viral cysteine protease.

The most important source of information concerning the structure of an enzyme are X-ray structures that are obtained from X-ray crystallography experiments. They can deliver "pictures" of atomistic resolution. Basis for the generation of structure information via X-rays is the availability of a crystal. Although this technique has already been known from the beginning of the 20th century, its application in proteins was established much later. This proved to be a breakthrough for the field of structural biology, and was honored by several nobel prizes.[48] Owing to the fact that protein structure determination via X-ray crystallography can be routinely done today, an extensive set of structures is available for SARS-CoV M^{pro} and its mutants. **Tab. 2-1** gives an overview of important SARS-CoV M^{pro} structure models, without any ligand or substrate, that have been published between the years 2003 and 2012.

published	PDB code	quaternary structure	resolution	pH value	author
2003	1UJ1	dimer	1.9 Å	6.0	Yang et al.[46]
2003	1UK2	dimer	2.2 Å	8.0	Yang et al.[46]
2003	1UK3	dimer	2.4 Å	7.6	Yang et al.[46]
2005	1Z1I	monomer	2.8 Å	-	Hsu et al.[49]
2006	2GZ9	monomer	2.2 Å	6.0	Lu et al.[50]
2007	2GT8	monomer	2.0 Å	6.5	Lee et al.[51]
2007	2H2Z	monomer	1.6 Å	6.0	Xue et al.[52]
2007	2DUC	dimer	1.7 Å	5.8	Wang et al.[53]

Tab. 2-1: Overview about published X-ray structures of SARS-CoV M^{pro}. The PDB code refers to the unique four letter code that is used to identify the structure in the protein databank of the Research Collaboratory for Structural Bioinformatics (RCSB).[54] The quaternary structure column indicates whether the published structure comprises a monomer or dimer structure within the unit cell. Resolution values refer to the employed wavelength of the respective X-ray diffraction experiment. The pH value refers to the crystallization conditions.

Each structure can be identified by a unique four letter code (PDB code), which is assigned by the protein database of the Research Collaboratory for Structural Bioinformatics (RCSB).[54] Typically, published structures are deposited in this archive and are made available for researchers all over the world.

The available X-ray structures form the basis for understanding the secondary, tertiary, and quaternary structure of SARS-CoV M^{pro}. The protein structure models are generally interpreted from an electron density map that is obtained from the X-ray diffraction experiment. The quality of the electron density map depends mainly on the quality of the employed protein crystal but also on the wavelength of the X-rays that is used for the experiment. This is typically denoted as the resolution of the structure. Hence, the quality of an interpreted X-ray structure model depends partly on the resolution of the structure. In general, smaller wavelengths are better suited to obtain pictures on the atomistic scale than at higher values. A major limitation, however, is that shorter wavelength possess the risk of damaging the protein structure due to their higher energy. Therefore employed X-ray wavelength are typically chosen in the magnitude of chemical bond lengths, which range on the scale of one or two Ångstrøms. Another notable experimental factor is the pH value. Protein structures typically contain titratable amino acids (e.g. histidines) that can change their protonation state as a function of the pH value of the environment. For this reason, the structure of a protein might be sensitive to different pH conditions that are employed during the crystallization process, since changes in the protonation state of certain amino acid influences the conformation of the protein backbone. In the case of SARS-CoV M^{pro}, the set of available structures partly differ in their crystallization conditions (**Tab. 2-1**). This gave rise to studies involving the analysis of how the protein structure depends on the pH value.[46] Another important issue is the quaternary structure of the published structure model. Since enzymes often tend to form aggregates, like dimers, exact knowledge of the aggregate structure helps to recognize functional issues that might be related to quaternary structure and is an important starting point for molecular modeling studies.

The SARS-CoV M^{pro} has been shown to exist in monomeric, dimeric, and octameric form. Fig. 2-5 gives an overview of the different quaternary structures that have been observed and proven by X-ray structures.

The monomeric form plays a minor role and is mainly studied for the sake of understanding the dimerization process and mechanism of SARS-CoV M^{pro}.[55] Since the active-site in the monomeric form has collapsed, it is catalytically inactive.

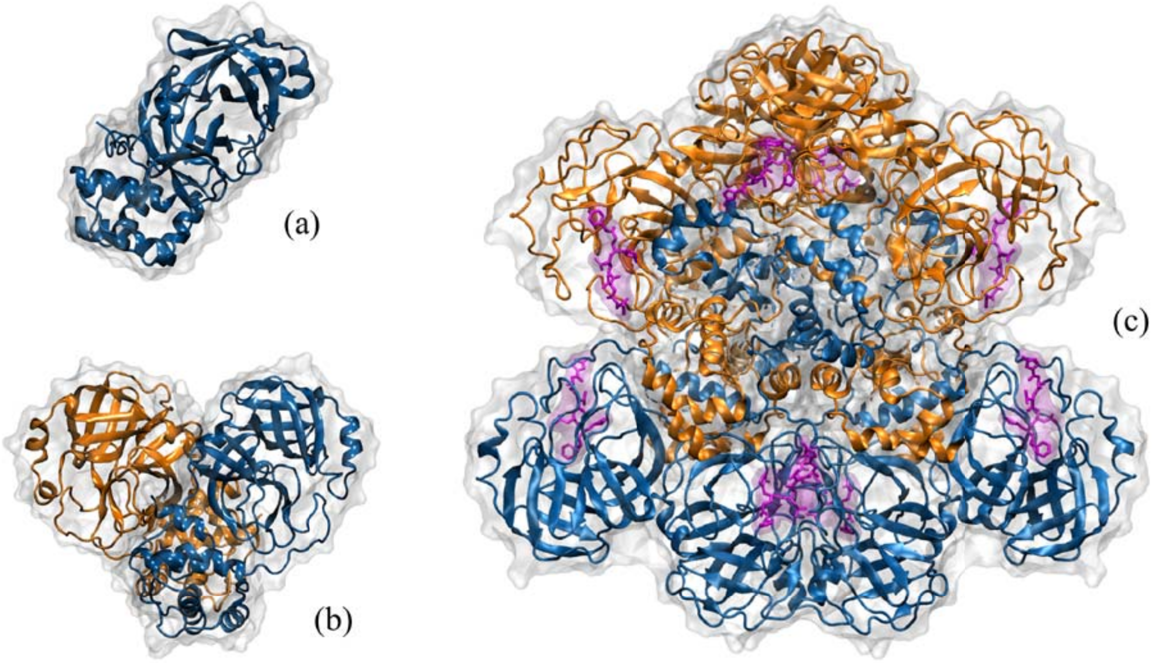


Fig. 2-5: Overview of the different quaternary structures of SARS-CoV M^{pro} that are available as X-ray structures. The monomeric form (a) is only available as a mutant (PDB code 2QCY)[55] due to the high dimerization tendency of SARS-CoV M^{pro}. The dimeric form (b) represents the most important quaternary structure and comprises an asymmetric homodimer (PDB code 1UK2)[46] that is assembled from two protomers A (blue) and B (orange). The octameric form (c) contains 8 protomers with 4 of them in an active-state (alternated colored in blue and orange) and has been co-crystallized (PDB code 3IWM)[40] with an inhibitor (magenta).

The dimeric form is built up of two identical monomers (protomers) that are arranged perpendicular to each other[46] and represent the most important quaternary structure form of SARS-CoV M^{pro}. It has been found that only in the dimer state catalytic activity is present.[56] The structure of the homodimer depends partly on the pH value and is found to be asymmetric, which means that there are small conformational differences between the two protomers. Due to this antisymmetry, only one of the two protomers is in a catalytic active state.[46,57] There are discussions in the literature whether SARS-CoV M^{pro} can employ a ping pong mechanism for substrate cleavage (also denoted as double-displacement mechanism),[58] where substrate binding and product release happen in an alternating manner, like a ping pong ball that bounces from one side to another. In the case of SARS-CoV M^{pro}, these sides would refer to the two catalytic sites of protomer A and B. It is an interesting question whether there is communication between the two active-sites, as discussed for other homodimeric enzymes,

[59] but this has not been investigated so far.

The tendency of SARS-CoV M^{pro} to form dimers has been extensively studied throughout literature. Although several experimentally measured dimerization constants are available, they are not always in line, and reported values are distributed unevenly across the studies. Dimerization constants K_D for the monomer-dimer equilibrium, as defined in equation 1,[60] span from 100 µM [56] and 227 µM [61] down to 190 nM.[62] However the range of 5 to 13 µM seem to be most likely as it has been yielded from a complementary use of three different experimental methods.[60]

$$K_D = \frac{[monomer] \cdot [monomer]}{[dimer]} \tag{1}$$

One consequence of the dimerization equilibrium is a dependence between catalytic activity and protein concentration in solution. Enzymatic activity of SARS-CoV M^{pro} is predominantly given at concentrations above 2mg/mL, whereas lower concentrations lead to increased dissociation into the inactive monomers, thus resulting in a loss of enzymatic activity.[60] This behavior might be useful to control enzymatic activity during the early stages of an infection and represents an ideal example for activity regulation through quaternary structure.[63]

The octameric form of SARS-CoV M^{pro} was discovered several years after the dimeric form and its role for catalysis is not as clear yet. It consists of eight monomeric units that form a domain-swapped dimer after addition of an N-terminal extension peptide and self-assembles into an octameric structure after removal of the extension peptide.[40] Since the octamer seems to comprise of four active-sites in catalytic active conformation, it could provide superior catalytic activity to the dimeric form, especially at low concentrations where monomer-dimer equilibrium would be unfavorably shifted towards the catalytically inactive monomeric form. Therefore, the octameric form could play an important role for catalysis. Nevertheless, since the authors themselves state that the functional mechanism seems to be far more complicated than initially thought, these hypotheses still need to be proven. Hence, the existence of an octameric structure *in vivo* presents an open question.[64]

By taking a closer look at the tertiary structure of SARS-CoV M^{pro}, it becomes apparent that it shares some features with chymotrypsin, which is a well-known digestive enzyme.[65] In detail, the structure of SARS-CoV M^{pro} is assembled from 306 amino acids and has an atomic mass of 33.8 kDa.[46] The sequence of amino acids, also denoted as residues, forms three structural domains, as shown in Fig. 2-6.

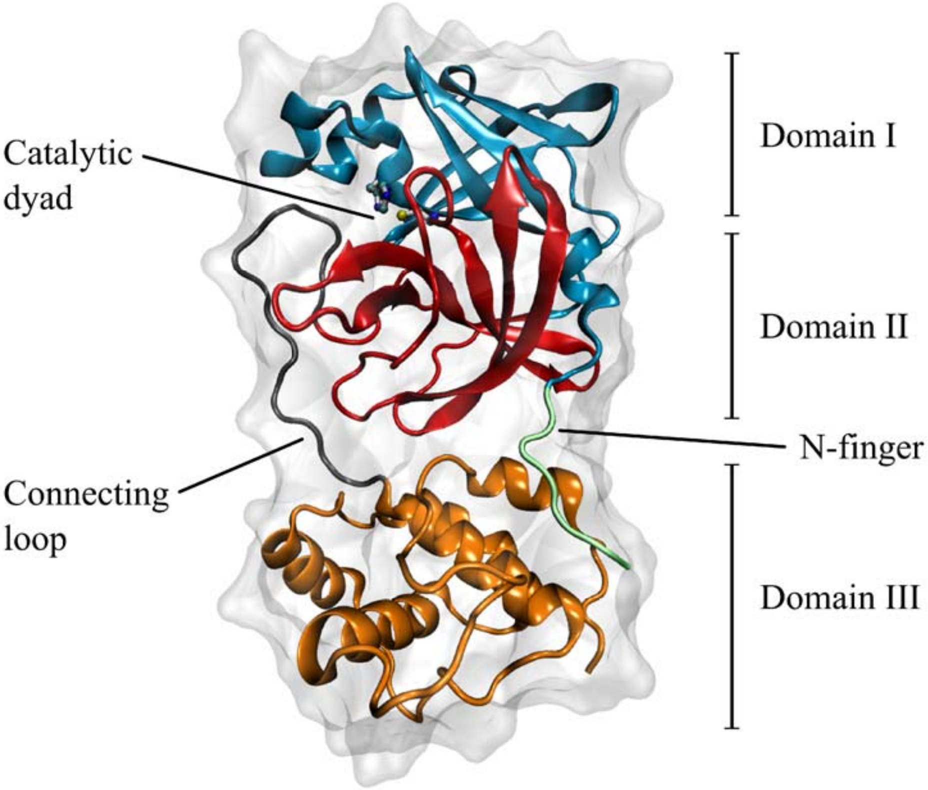


Fig. 2-6: Tertiary structure of the SARS-CoV M^{pro} monomer. The structural motifs are indicated with different colors and contain the N-finger (green), domain I (blue), domain II (red), connecting loop (grey), and domain III (gold). Schematic drawing is based on X-ray structure 2H2Z.[52]

The domain I (residues 8-101) and the domain II (residues 102-184) contain six-stranded β-barrels that are arranged in an anti-parallel manner. The domain III (201-306) is connected to the domain II by a long loop of fifteen amino acids (residues 185-200) and contains the secondary structures of five α-helices that form together a globular cluster-like arrangement. [45] The similarity of SARS-CoV M^{pro} with chymotrypsin holds only for the first and second domain, since domain III is a unique feature of the M^{pro}. Therefore, the fold of SARS-CoV M^{pro} is often described as an augmented serine protease.[45] Despite the similar folding, it is notable that SARS-CoV M^{pro} belongs to a different protease class than chymotrypsin, since it is a cysteine protease and not a serine protease (Fig. 2-3). The third domain seems to be of functional importance for the dimerization process. Theoretical investigations showed that domain III provides specific electrostatic and hydrophobic interactions at the interfaces between the protomer A and the protomer B in the dimeric state.[66,67] The predominant role of domain III as a key factor for dimerization is also confirmed by experimental results[68] and gives a perspective for inhibitors that can modulate or even prevent the dimerization process. [69] The fact that SARS-CoV M^{pro} is exclusively active in the dimer state is further controlled by a second factor, the so-called N-finger (residues 1-7), which denotes the first seven N-terminal amino acids next to domain I. The N-finger is squeezed in between the domain III of the parent protomer A and the domain II of the neighboring protomer B, thus results in close interactions with the active-site of the second protomer.[46] These interactions are important to keep the catalytic pocket of protomer B in an active conformation and "switch on" its catalytic activity. Deletion experiments with a truncated form of SARS-CoV M^{pro} confirm that in absence of the N-finger, dimerization is still observed, but catalytic activity is completely lost.[61,70] Therefore, the N-finger is not an essential factor for the formation of the dimer, but indispensable for the catalytic activity.

The residues that are responsible for the catalytic activity are cysteine 145 (Cys145) and histidine 41 (His41), which form the catalytic dyad of SARS-CoV M^{pro}. They are located in the cleft between domain I and domain II and cleave the peptide bond, whenever a substrate is bound to the substrate-binding site. Since the thiol-containing cysteine residue is the main factor for the peptide bond cleavage mechanism, SARS-CoV M^{pro} is classified as cysteine protease.[38] The essential role of Cys145 for catalytic activity has been proven by mutation experiments that revealed a 40-fold decreased catalytic activity when cysteine is replaced by a serine residue[71] and further by complete loss of catalytic activity, when His41 is knocked out by an alanine residue.[72]

The catalytic activity of native SARS-CoV M^{pro} is considered to be on a high level[52] and can be expressed by the Michaelis constant K_M, that defines the substrate concentration $[S]$ at which the half of the maximum reaction rate $\upsilon_{\max}$ is reached.[73] The Michaelis constant is a crucial characteristic of the Michaelis-Menten equation, which is an often applied model in enzyme kinetics. It describes the simplified relationship between substrate concentration $[S]$ and reaction rate of product formation υ. It holds only for the assumption that the enzyme concentrations is relatively low compared to substrate concentration.[74] Equation 2 gives a definition of the Michaelis-Menten model that describes the equilibrium between enzyme E, substrate S, reversible bound enzyme-substrate complex E···S, and product P. The rate constants k_S, k_{-S}, and k_{cat} describe how fast each unidirectional step proceeds.

$$\mathrm{E} + \mathrm{S} \underset{k_{-S}}{\overset{k_S}{\rightleftarrows}} \mathrm{E}\cdots\mathrm{S} \xrightarrow{k_{cat}} \mathrm{E} + \mathrm{P} \qquad (2)$$

The Michaelis-Menten equation, according to the simplified kinetic model above, is given in equation 3, where *d*[*P*] is the infinitesimal change in product concentration within time *dt*.

$$\frac{d[P]}{dt}=\upsilon=\frac{\upsilon_{max}[S]}{K_M+[S]} \tag{3}$$

As apparent from the equation 3, the Michaelis constant K_M behaves inverse to the reaction rate of product formation υ, thus in case of a smaller K_M value, the maximum rate υ_{max} is reached earlier with respect to increased substrate concentration [*S*] than in case of larger K_M values. For the SARS-CoV M$^{\mathrm{pro}}$, the Michaelis constant K_M has been found to be 306 µM[75] for substrate with consensus cleavage sequence, but have also been reported to be in the range between 1 µM and 404 µM for related substrate sequences.[56,76,77] It is noteworthy that the catalytic activity is influenced by the pH value of the environment and that only for a pH value of 7.6 a high proteolytic activity can be observed for the SARS-CoV M$^{\mathrm{pro}}$, which was derived from experimental pH-activity profiles.[71] The reason for this behavior has also been investigated on the atomistic level, where significant structural changes can be observed. Therefore, the activity of SARS-CoV M$^{\mathrm{pro}}$ can be considered as pH-triggered.[46] The origin of the structural rearrangements are mainly due to two histidine residues that can switch their protonation state as a function of the environment pH.[78] Especially the protonation state of histidine 163 (His163) plays an important role, since it has an aromatic interaction (π-stacking) with phenylalanine 140 (Phe140) that keeps the substrate binding site in an active conformation. The protonation of His163 at low pH values is supposed to disrupt this aromatic interaction that leads to a collapse of the binding pocket.[51] Hence, the catalytic activity is reduced or abrogated at pH values below 6.5. The essential role of Phe140 has further been shown by mutation experiments, where Phe140 is replaced by an alanine residue which leads to a collapsed oxyanion hole and thus inactivation of the SARS-CoV M$^{\mathrm{pro}}$.[79]

In general, SARS-CoV M$^{\mathrm{pro}}$ shows a high substrate specificity, therefore only substrates with a specific amino acid sequence are recognized and cleaved. Highest catalytic activity is observed for the sequence Ser-Ala-Val-Leu-Gln-↓-Ser-Gly-Phe, starting from N-terminus to C-terminus,[80] whereas the arrow ↓ indicates the cleavage site. The binding pockets of the enzyme (and the corresponding amino acids of the substrate) can be numbered according to the description scheme of Schechter and Berger,[81] which means that the substrate binding pockets are numbered in an ascending order, starting from the cleavage site with S1 (P1) towards the N-terminus of the substrate. The numbering towards the C-terminus of the substrate is indicated by similar numbers and include bars, like S1' (P1'). The nomenclature scheme is shown in Fig. 2-7.

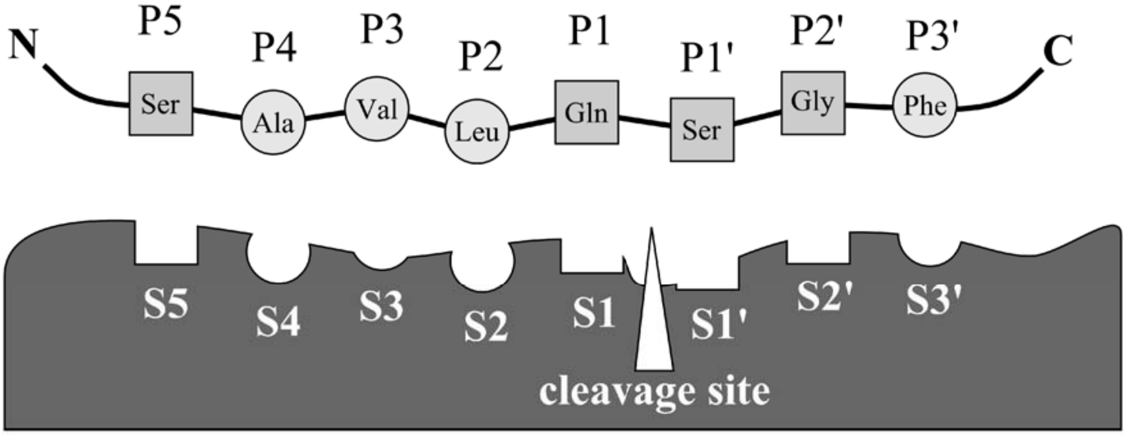


Fig. 2-7: Nomenclature of the substrate binding site and the corresponding amino-acid sequence after Schechter and Berger,[81] showing a cleavable substrate sequence of SARS-CoV M^{pro}.[80] Hydrophilic residues are indicated with squared shapes, hydrophobic residues as circular shapes.

The substrate binding site S1 shows an absolute specificity for a glutamine residue in P1 position, a feature that is also conserved for other main proteases of the coronavirus family.[47] Recent studies showed that histidine and methionine are also tolerated in this position, but with significant decreased catalytic efficiency.[82] The binding pocket is built up from the residues His163, Phe140 and backbone atoms of Met165, Glu166, and His172.[46]

The S2 site prefers hydrophobic residues with larger side-chains, like leucine or phenylalanine.[47] The binding pocket is formed by His41, Asp48, Pro52, Tyr54, and Met165 and is more tolerant with respect to variations at the P2 position than the S1 pocket.[45]

The S3 binding pocket is more an assumed interaction than a real pocket. In earlier studies, the assumption was made that the cleavage site is unspecific at the S3 position, since the corresponding amino acid side chain points toward the solvent, as derived from the crystal structure of a substrate-analogue bound inhibitor.[46] More recent reports indicate that the S3 pocket plays also a role for the substrate binding[80] and could comprise an important electrostatic interaction between enzyme and substrate.[83] The S3 site contains the residues Glu47 and Glu166 from which one of them is supposed to interact with a valine or arginine residue in P3 position, whereas the latter one leads to even increased activity.[82]

The S4 binding pocket is a cavity formed by the amino acids Met165, Leu167, Thr190, and

Gln192[45] that preferably binds to a small hydrophobic residue like alanine or valine, but also tolerates a cysteine residue in P4 position.[82]

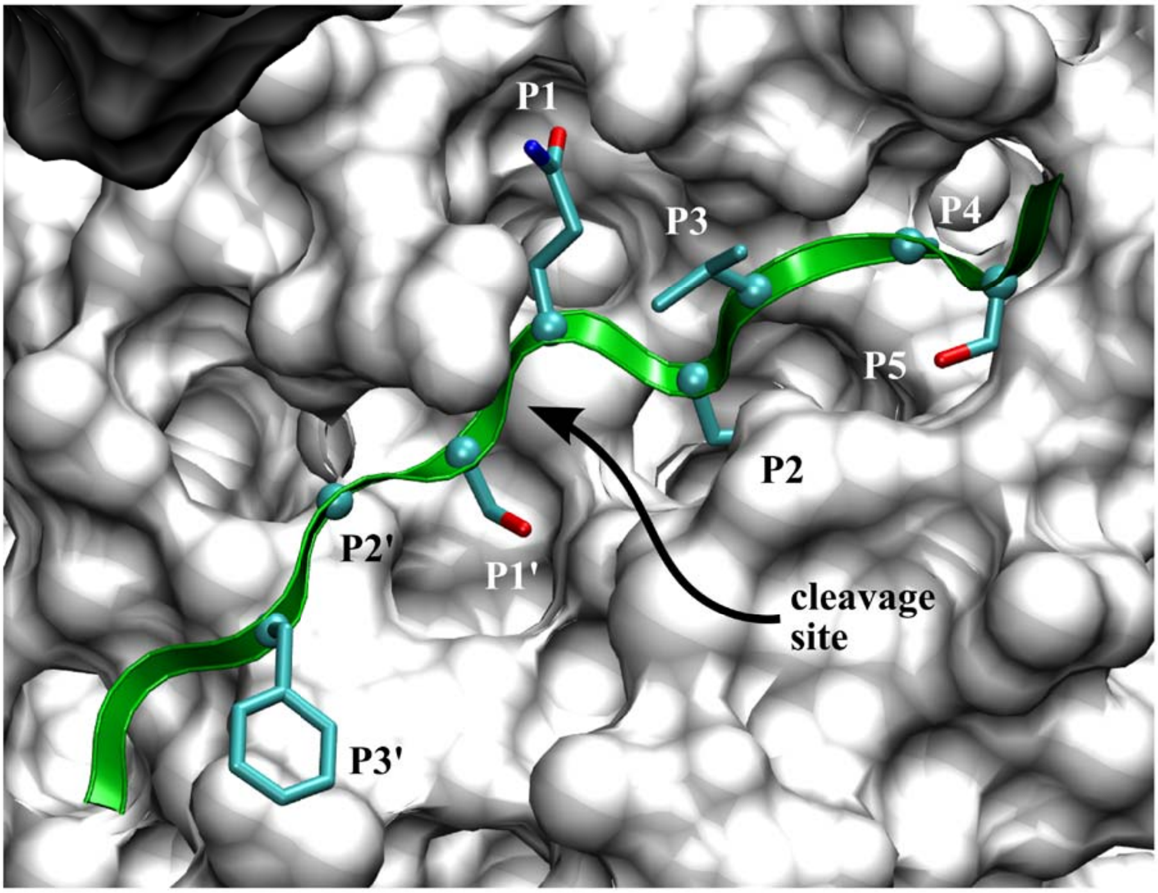


Fig. 2-8: Topology of the substrate binding site of SARS-CoV M^{pro} (white) in complex with a cleavable substrate, based on X-ray structure 2Q6G.[72] The peptide backbone of the substrate is depicted as ribbon structure (green) for the sake of clarity. The side chains of the substrate sequence are denoted by the Schechter and Berger nomenclature.[81] They are embedded in the respective binding pockets.

The S5 site typically hosts a serine residue and has also a broad tolerance towards other amino acids, although residues that have a propensity to form β-sheets tend to give higher catalytic activity.[82] The S5 pocket is described by the residues Pro168, Thr190, and Glu192.[46]

The C-terminal side of the binding pockets starts with the S1' binding site that is formed by Thr25, Leu27 and Cys145.[72] It is considered as a shallow pocket and accommodates only small residues in the P1' position, like serine, glycine, and alanine.

Compared to the latter one, the S2' binding pocket is more narrow and deep and consists of the residues Thr26, Asn2, Tyr118, Asn119, and Gly143.[72] This site is typically occupied by a glycine residue in P2' position but shows no strong preference in its specificity.[82]

The S3' pocket does not exist in SARS-CoV M^{pro}, since the P3' residue of the substrate points towards the solvent and therefore no specificity can be observed for this position.[72]

Beside the binding pockets, there is a further important site in SARS-CoV M^{pro}, the so-called oxyanion hole. The oxyanion hole is crucial for the substrate cleavage, since it stabilizes the tetrahedral intermediate that occurs during the peptide bond cleavage reaction by specific hydrogen bond interactions.[84] The oxyanion hole of SARS-CoV M^{pro} is formed by the backbone amides of Gly143, Ser144, and Cys145,[57] accommodating the negatively charged carbonyl oxygen during catalysis. Since the oxyanion hole collapses in the catalytically inactive monomer state,[55] it is one of the reasons why SARS-CoV M^{pro} is exclusively active in the dimer state. The classical model of substrate recognition and binding is the key-lock theory.[85,86] It describes enzyme (lock) and substrate (key) as complementary and static parts that must fit properly together in order to fulfill their biological function. However, in case of SARS-CoV M^{pro}, the classical key-lock principle does not hold, since substrate binding follows an induced fit mechanism,[87] which has been deduced from observed differences in the crystal structures of an inhibitor-bound and unbound form of SARS-CoV M^{pro}.[51] In induced fit theory, the key-lock principle is extended towards the model of a flexible enzyme, where the substrate fits in like a hand in a glove.[88] Therefore, the active-site of SARS-CoV M^{pro} undergoes several changes whenever a substrate or substrate-like molecule is bound. This mainly accounts for the S2 and the S4 binding pockets, whereas the S1 pocket and the oxyanion hole remain nearly unchanged during substrate binding.[89] Interestingly, substrate binding itself is not influenced by the dimerization of SARS-CoV M^{pro}[63] and can also proceed for a catalytically incompetent conformation of the active-site.[51] Nevertheless, catalytic activity is only given in the dimer state as explained above.

The chemical step of peptide bond cleavage is conducted by the catalytic dyad of SARS-CoV M^{pro}. It is formed by the residues cysteine 145 (Cys145) and histidine 41 (His41). They are located in the cleft between domain I and domain II and sterically arranged next to each other.[46] This is a remarkable feature of SARS-CoV M^{pro} since it stands in contrast to typical serine proteases, like chymotrypsin, that comprise a catalytic triad consisting of a serine, histidine, and aspartic acid residue.[90,91] In place of the third catalytic residue, SARS-CoV M^{pro} possesses a stable water molecule (H_2O@His41) that is well-conserved in X-ray structures[46,49–53] and has further shown to be a dynamically stable component of the active-site.[78] There are several hypotheses in the literature about the functional role of this water molecule. Since the water molecule is trapped in between His41 and an aspartic acid residue (Asp187), it has been proposed that substrate binding leads to a "squeeze out" of this water molecule and thus induces the formation of a more efficient catalytic triad in which aspartic acid can contribute.[67,92] Another proposal is that H_2O@His41 assists the catalytic function of His41 simply by stabilizing its conformational orientation during catalysis.[93] Fig. 2-9 gives an overview about the catalytic dyad, the water molecule, and its embedding into neighboring residues.

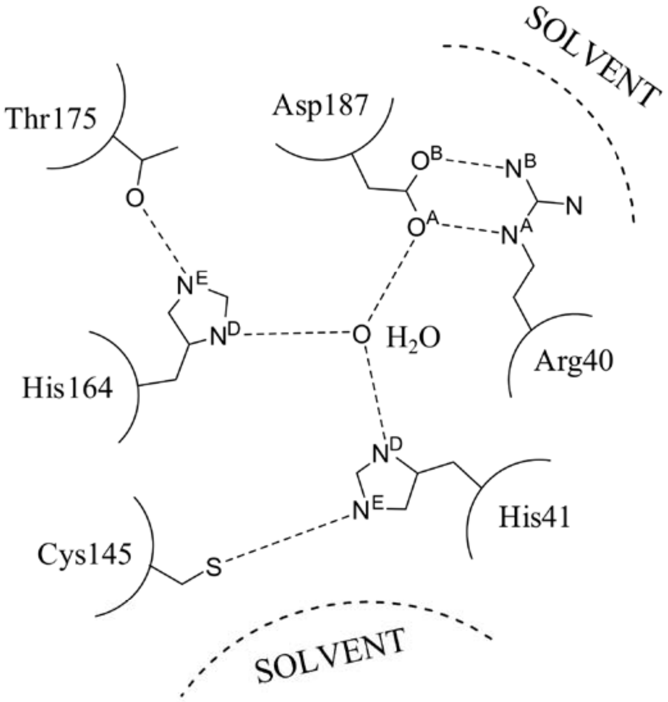


Fig. 2-9: Schematic drawing of the catalytic dyad Cys145/His41 of SARS-CoV M^{pro} and some selected neighboring residues that might play a role for enzyme catalysis. Hydrogen atoms have been omitted for the sake of clarity. The water molecule in the middle (H2O@His41) possesses a stable position and is trapped between His41, His164, and Asp187.

A further remarkable feature of the catalytic site is the protonation state of the Cys/His dyad. Experimental studies on the active-site residues of SARS-CoV M^{pro} estimate the pK_a values of Cys145 in the range of 7.7-8.3 and His41 in the range of 6.2-6.4.[58,71] This means that they rest in an uncharged state with a protonated thiol moiety. Since SARS-CoV M^{pro} is a cysteine protease, this finding distinguishes it from typical enzymes of this protease class, which normally possess a thiolate/imidazolium ion pair for their catalytic Cys/His dyad.[94–97] However, the existence of exceptions from this general characteristic has been shown few years before SARS-CoV M^{pro} was discovered at the example of a picornavirus protease that is also a cysteine protease but does not form an ion pair at its catalytic site.[98,99] Interestingly, the fold of picornavirus 3C protease has a close similarity to the serine protease chymotrypsin, which is in line with observations at SARS-CoV M^{pro}, for which a similar structural relationship is found.[100] There are also more recent examples of cysteine/histidine containing active-sites that reside in an uncharged state.[101] Hence, the paradigma that cysteine proteases usually possess a thiolate/imidazolium ion pair, seems to vanish step by step.[102] The current understanding of the catalytic mechanism indicates that SARS-CoV M^{pro} exhibits mechanistic features that are quite different to the archetypical proteases papain and chymotrypsin, but also shares common features with them.[58,71] The common enzymatic mechanisms for these two "prototype proteins" are the ion-pair mechanism for papain,[103] and the general-base catalysis for chymotrypsin,[104] respectively. Concerning the peptide cleavage reaction mediated by SARS-CoV M^{pro}, experimental results indicate that an ion-pair seems clearly to be involved into the mechanism. This is similar to papain, but on the other side the catalytic activity is found to be independent of the chemical reactivity of the substrate.[58] This is counter intuitive, since substrates with increased reactivity should react faster with the thiolate moiety of cysteine and thus proceed faster within the enzyme catalysis cycle. Derived from pK_a measurements that are indicative for a general-base mechanism,[71] the existence of an "electrostatic trigger" is postulated[58] that denotes the formation of an ion pair of SARS-CoV M^{pro}'s catalytic dyad as a first reaction step and seems furthermore the rate-determining step for enzyme catalysis. This hypotheses could conclusively explain why the catalytic activity is independent from the reactivity of the substrate.

2.3 The Search for Inhibitors

Proteases are generally involved in many useful and important physiological processes like digestion, wound healing or cell growth, but also occur in many undesirable contexts like viral infections, cancer, or Alzheimer disease. The inhibiton of proteases has therefore potential therapeutic use for the treatment of diseases.[105] In case of SARS-CoV M^{pro} it was recognized that inhibitors against this cysteine protease are ideal drugs for the treatment of the SARS disease and could furthermore possess antiviral activity against a broad spectrum of *coronaviridae*,[106] due to the high conservation of the main protease motif among the coronavirus family.[47]

The inhibition of a protease can be seen as a competing process to its original function, like peptide bond cleavage, whereas the inhibitor should have sufficient potential to prevent the enzyme from fulfilling its biological role. Inhibitors that target the active-site and interfere with the catalytic mechanism can hereby achieve a high specificity in their intervening role and typically act as so-called competitive inhibitors.[107]

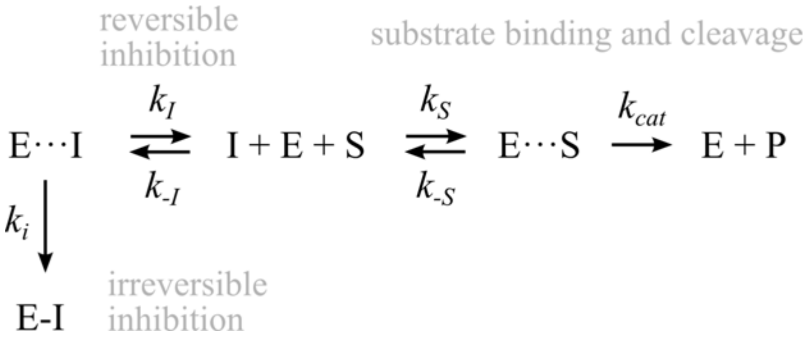
The inhibitors are classified by their mode of action, which means that they can act in a reversible way or an irreversible way.[107] In case of reversible inhibition, the binding of the inhibitor I forms a more or less loosely bound inhibitor-enzyme complex E···I that stands in a simple competition with the substrate binding. This can be described by rate constants k_I and k_{-I} in an analogue way to the reversible substrate binding complex E···S within the Michaelis-Menten model (equation 2). The ratio between backward and forward reaction k-$_I$ and k_I is usually expressed as dissociation constant K_i.

$$K_i = \frac{k_{-I}}{k_I} \tag{4}$$

In case of irreversible binding, the inhibition process contains a further one way step, described by the rate constant k_i, that leads to a completely inactivated enzyme-inhibitor complex E-I that is also denoted as covalent complex.[105] The overall inhibitory power is often expressed by the rate constant of second order k_2 that includes the rate of covalent inhibition step k_i as well as the dissociation constant of the non-covalent complex K_i.[108]

$$k_2 = \frac{k_i}{K_i} \tag{5}$$

Equation 6 gives an overview about the different competing processes concerning the enzyme inhibition.[109] Since a strict discrimination between reversible and irreversible is not always possible, there are two further types of inhibitors. The first one is denoted as tight binding inhibitor that possesses in principle a non-covalent inhibition, but has such a strong binding affinity that it appears like an irreversible one. The second one is the type of slow binding inhibitor that induce conformational changes and need, hence, a longer time to induce an inhibition effect at the target enzyme.[110]



(6)

For drug development, the pathway of designing reversible inhibitors is generally more preferably than irreversible ones. The problem concerning irreversible inhibition is related to risks in long-term treatments that are hardly to predict,[110] and possible side effects.[111] Therefore, the development of non-covalent reversible inhibitors is much more desirable for the pharmaceutical industry whenever a chronic treatment is necessary. Nevertheless, this does not hold for acute treatments like antiviral or cancer therapy.[110] In case of these short-term uses, the application of irreversible protease inhibitors is much more acceptable. The orphan drug Emricasan (Pfizer AG) is a successful example that has entered clinical use for transplantation issues,[112] which represents the first irreversible caspase protease inhibitor that has entered the clinical development.[113] Irreversible inhibitors can even have advantages with respect to reversible ones, since they are less sensitive to mutations of the target enzyme. Mutation normally leads to a lower binding affinity and thus to a reduced efficiency of the reversible inhibitor, whereas the irreversible inhibitor still has potency to inactivate the enzyme, even with a lowered occupancy at the active-site.[114]

For the development of irreversible cysteine protease inhibitors, or protease inhibitors in general, one can follow a typically applied scheme. It consist of finding a suitable amino acid sequence, which provides a high affinity to the targeted enzyme, and insertion of an electrophilic group, the so-called “warhead”. The warhead can be attacked by the nucleophilic active-site residue, e.g. by the thiol group of the cysteine.[115] The insertion into the chemical scaffold of the inhibitor is typically conducted at the position where the substrate cleavage proceeds.

In case of inhibitor development against the SARS-CoV M^{pro}, a broad spectrum of reports is available that probe different types of warheads. They include peptidic, non-peptidic, or peptidomimetic inhibitor structures. The aim of these inhibitors is the irreversible inactivation of the Cys145 residue at the active-site of SARS-CoV M^{pro}. It is normally achieved by formation of a covalent bond between the nucleophilic sulphur atom of the thiol group and the electrophilic building block of the employed inhibitor warhead. The chemical groups that are used for this purpose span from chloromethyl ketones,[46] aldehydes,[116] α,β-unsaturated esters,[106] triazole esters,[117] and phthalhydrazines[89] to epoxides,[51] aziridines[118] and nitriles.[119] Fig. 2-10 gives an overview about the diversity of chemical groups that have been reported as inhibitor warheads.

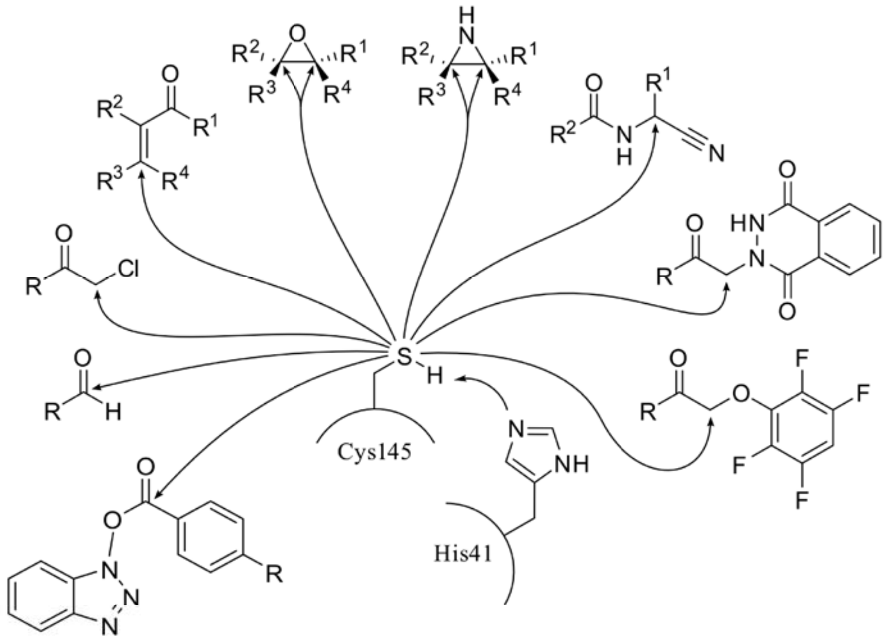


Fig. 2-10: Overview about different warheads that can be used for an active-site directed irreversible inhibition of SARS-CoV M^{pro}.

Beside the common group of active-site directed inhibitors, there are also examples of other inhibition strategies against SARS-CoV M^{pro}, like the use of boronic acids that target serine residues,[76] an octapeptide that prevents dimerization, [120] or the use of zinc- or mercury-complexation.[121] Interestingly, some natural products like extracts from black or green teas, have also been shown to have inhibitory activity against SARS-CoV M^{pro}.[122]

The effectiveness of an inhibitor can in general be expressed in terms of the inhibition dissociation constant K_i or the IC_{50} value. The inhibition constant K_i is defined as an equilibrium constant that expresses the binding affinity of an inhibitor to the enzyme. It includes the ratio of the individual rate constant for the dissociation and binding reactions. The often used IC_{50} value is a measure for the inhibitor concentration at which 50% inhibition of the enzyme is achieved. In case of competitive inhibition, the relationship to K_i is given by equation 7 that depends further on the Michaelis constant K_M and the concentration of the substrate [S].[123] For irreversible inhibition, IC_{50} further depends on the incubation time between inhibitor and enzyme.

$$IC_{50}=\left(1+\frac{[S]}{K_M}\right)K_i \tag{7}$$

The development of SARS-CoV M^{pro} inhibitors has yielded an impressive number of reports about possible inhibitor scaffolds that possess reversible or irreversible binding modes. The progress in inhibitor design has been reviewed several times,[124,125] also with emphasis on patents.[126] Tab. 2-2 tries to give an overview about important SARS-CoV M^{pro} inhibitors, their inhibition characteristics in terms of K_i or IC_{50} values, and whether structural data about their respective enzyme-inhibitor complexes are published or not. Furthermore pure *in silico* screening methods have been applied to identify novel inhibitors,[127,128] as well as the new and promising approach of dynamic ligation screening (DLS) from Schmidt and coworkers.[129]

In case of irreversible binding, a covalent linkage between inhibitor and enzyme exists, what is a good precondition for X-ray structure determination of the enzyme-inhibitor complex. Therefore several structures are available in the PDB database of the Research Collaboratory for Structural Bioinformatics (RCSB),[54] which give valuable insights into active-site directed inhibitor binding at the SARS-CoV M^{pro}.[46,51,52,82,89,106,116,117,130–133] Historically, the first inhibitor-complex structure contained a substrate-analogue, peptide-based chloromethyl ketone inhibitor (CMK) that was published together with the first X-ray crystal structure of SARS-CoV M^{pro}.[46] It is noteworthy that CMKs are chemically highly active and undergo reactions with the target enzyme as well as with gastric or enteric proteases. Therefore they are mainly used for gaining structural insights and not as therapeutic drugs.[106] Nevertheless, there is also an example of a reversible or non-covalent inhibitor, for which structural data exists.[50]

year	warhead (wh) or substance class	K_i	k_i	IC_{50}	PDB code	author
2003	chloromethyl ketone (wh)	-	-	2 mM	1UK4	Yang et al.[46]
2004	small molecule inhibitor	-	-	0.5 µM	-	Blanchard et al.[134]
2004	keto-glutamine analogue	-	-	0.6 µM	-	Jain et al.[135]
2005	α,β-unsaturated ester (wh)	6.7 µM	0.0022 s^{-1}	-	2AMD	Yang et al.[106]
2005	tannic acid	-	-	3 µM	-	Chen et al.[122]
2005	cinanserin	-	-	323 µM	-	Chen et al.[136,137]
2005	peptide anilide	0.03 µM	-	0.06 µM	-	Shie et al.[138]
2005	isatin derivative	-	-	0.95 µM	-	Chen et al.[139]
2005	α,β-unsaturated ester (wh)	0.52 µM	-	-	-	Shie et al.[140]
2005	α,β-unsaturated ester (wh)	-	0.01 min^{-1}	45 µM	2ALV	Ghosh et al.[130]
2005	etacrynic acid derivative (wh)	35.3 µM	-	-	-	Kaeppler et al.[141]
2005	epoxide (wh)	18 µM	0.035 s^{-1}	-	2A5I	Lee et al.[131]
2005	aziridine (wh)	k_{obs}/[I] = 311 M^{-1}/min^{-1}			-	Martina et al.[118]
2006	small molecule inhibitor	-	-	0.3 µM	2GZ8	Lu et al. [50]
2006	benzotriazole ester (wh)	7.5 nM*	0.0011 s^{-1}	-	-	Wu et al.[142]
2006	aldehyde (wh)	0.05 µM	-	-	2GX4	Yang et al.[116]
2007	epoxide (wh)	-	-	-	2GTB	Lee et al.[51]
2007	α,β-unsaturated ester (wh)	1.9 µM	0.035 s^{-1}	-	2HOB	Xue et al.[52]
2007	α,β-epoxyketone (wh)	2.2 µM	0.004 s^{-1}	-	2OP9	Goetz et al.[132]
2007	phthalhydrazide (wh)	0.25 µM	-	-	2Z3C	Yin et al.[89]
2007	1,2-diol	0.07 µM	-	-	-	Shao et al.[143]
2008	small molecule inhibitor	2.9 µM	-	-	-	Schmidt et al.[129]
2008	benzotriazole ester (wh)	1.38 µM	0.013 s^{-1}	-	2V6N	Verschueren et al.[117]
2008	aldehyde (wh)	-	-	37 µM	-	Akaji et al.[144]
2009	small molecule inhibitor	-	-	2.5 µM	-	Kuo et al.[145]
2011	α,β-unsaturated ester (wh)	-	-	330 µM	3AVZ	Akaji et al.[133]
2011	small molecule inhibitor	9.1 µM	-	38.6 µM	-	Hanh et al.[146]
2011	aldehyde (wh)	2.2 µM	-	-	2SN8	Zhu et al.[147]
2013	nitrile (wh)	-	-	4.6 µM	3VB3	Chuck et al.[82]

Tab. 2-2: Overview about selected published inhibitors of the SARS-CoV M^{pro}. The warhead describes the chemical moiety that is used as electrophilic building block in case of an irreversible inhibition. For reversible inhibitors, the substance class is given and their inhibition potency listed by the respective K_i or IC_{50} values. For irreversible inhibitors, rate constant k_i of covalent bond formation is given, if available. Whenever an X-ray structure of the enzyme-inhibitor complex is available, the PDB code four-letter identifier of the Research Collaboratory for Structural Bioinformatics (RCSB)[54] is listed. *Reported K_i values of this work could not be reproduced by Verschueren et al.[117] who measured significant reduced inhibition potency in the micromolar range.

The similarity between SARS-CoV M^{pro} and other chymotrypsin-like proteases offered a good starting point for the inhibitor development and led to the proposal of AG7088 as a possible inhibitor.[45] AG7088 is a well-known drug (Rupintrivir, Pfizer AG) that was originally developed as an irreversible inhibitor against the picornavirus 3C protease. It showed furthermore antiviral activity against a broad spectrum of human rhinoviruses,[148] the cause of the common cold.[149] Nevertheless, AG7088 turned out to be inactive as inhibitor of SARS-CoV M^{pro},[140] but was used as basis for the development of several improved inhibitor scaffolds that contain an α,β-unsaturated ester function (Michael system ester) as electrophilic warhead.[106,140]

The continuous development of α,β-unsaturated esters has been last over the years, but did not lead to significant improvement in terms of their inhibition characteristics as the comparison between the work of Akaji,[133] Xue,[52] Ghosh,[130] Shie,[140] and Yang[106] in Tab. 2-2 reveals. This fact seems also to hold for all other irreversible inhibitors. The "high score" of inhibition potency was initially achieved by Wu et al. in 2006,[142] who reported a K_i value of 7.5 nM for a benzotriazole ester. It represented a SARS-CoV M^{pro} inhibitor that successfully entered the nanomolar range. Later on, these results were found to be not reproducible in the work of Verschueren et al.,[117] who also reported about a series of benzotriazole ester inhibitors. Therefore, the reliability of these values is questionable. A similar story also happened to the reversible inhibitor cinanserin, which was initially reported to possess an IC_{50} value of 5 µM,[136] but was later corrected to have a significant less inhibition potency of 323 µM.[137]

Taking these data together, no inhibitors seriously crossed the border beyond the micro molar range towards the nano molar region. In order to achieve effective inhibitory activity, the inhibitor dissociation constant must reach the sub-nanomolar range (K_i < 0.1 nM) if it aims to inhibit effectively for a longer period.[115] As apparent from Tab. 2-2, clearly none of the so far reported SARS-CoV M^{pro} inhibitors does fulfill this condition. Since there have been passed ten years of active research and development efforts in this field now, the question arises whether there is a specific reason for the missing progress.

3 Aim of this Work

As stated in the last chapter, the progress in inhibitor development for the SARS-CoV M^{pro} can be considered as moderate with respect to the elapsed time since 2003 and the enormous research efforts that have been spent by many contributing researchers. The goal of this work, therefore is to gain more insight into this field from a theoretical point of view and to reveal whether there exist some underlying reasons for this.

There are already some hints in literature that the SARS coronavirus main protease is rather distinctive from the typical known cysteine proteases. Some very interesting observations are the unusual resting state of the catalytic dyad, which does not form an ion pair,[71] or the fact that ester substrates are processed equally as fast as amide substrates by the SARS-CoV M^{pro}, even though these two chemical substances classes differ considerably in their non-enzymatic reactivity by a factor of 2000.[58] These findings have been explained with the postulate of an “electrostatic trigger”, which could be a rate-determining factor for the overall substrate cleavage process. Although not stated by the authors, the existence of such a rate-determining trigger could not only limit the rate of the substrate cleavage, but might also limit the rate of the irreversible inhibition.

For the full understanding of processes like enzyme inhibition, insights on an atomistic level are necessary to obtain a “picture” of what is happening inside the molecular machinery. Whereas such “pictures” are often difficult to derive from experimental work, like X-ray crystallography, spectroscopy, or other experimental methods, the discipline of computational chemistry offers a broad spectrum of tools that can be used to build up sophisticated models. These models can be used to probe hypotheses about a specific chemical problem and do often overcome the limitations of experiments. Nevertheless, all theoretical models possess frontiers of accuracy and are challenged with limited computational power, even in the era of multi-processor systems and supercomputers. Therefore, this work will not only focus on the respective problems, but also on a critical assessment of the accuracy for the applied approaches.

Since substrate and inhibition processes always involve an exchange of hydrogen atoms, the consideration of proton transfers will be a central part of this work. The question of, how well these can be described for a protein model on the atomistic scale is analyzed and is the first step towards reliable modeling. For this reason, a state-of-the art computational model of the whole SARS-CoV M^{pro} is set up and subsequently used for the treatment of the different problems on a theoretical level. In particular, this work tries to answer the following questions:

1) How accurate are the commonly used quantum chemical methods for the description of proton transfers inside of the protein model?
2) What are the characteristic features of the SARS-CoV M^{pro} active-site and what can be concluded about the substrate cleavage and inhibition mechanism?
3) How do Michael acceptor based inhibitor warheads react with the thiol moiety of the active-site?
4) To which extend can theoretical methods support X-ray structure refinement of enzyme-inhibitor complexes in difficult cases?

It is furthermore aimed to extend the current knowledge in a complementary way, since there are only very few examples in the literature[150] that give insights into the underlying chemistry of substrate cleavage and inhibition reactions of the SARS-CoV M^{pro} on the atomistic level.

The results and discussions of this work are organized in four sub chapters, according to the questions above. To give the reader some principle background about the employed theoretical approaches, the next chapter summarizes some basic facts about current computational chemistry methodology in the context of proteins. The last chapter gives an overall conclusion about the achieved insights and tries to answer the questions stated above.

4 Describing Proteins by Computational Chemistry

With the availability of modern computers, computational chemistry methods have evolved into a routinely applied tool for chemists in many areas and also have become applicable to molecules of biological interest with the beginning of the 21st century.[151]

Walther Thiel, one of the important pioneers in theoretical and computational chemistry,[152] once offered a simplified categorization of the three stakeholders in this field, consisting of "pure" theoreticians that develop and evaluate methods and their limitations, computational chemists that use the state-of-the-art methods to gain insight into a chemical problem, and experimentalists who want to characterize their specific problems with standard computations.[153] Employing this categorization, this work lies between the first and the second category, since it applies a variety of different theoretical methods to grasp insights into a particular (bio)chemical problem, but also evaluates the performance of the used methods.

One common challenge for nearly all computational chemistry methods is related to the computational effort, or costs, that are needed to calculate the desired property. In the practical use of computational chemistry it is known that this bottleneck can be decisive in choosing whether an appropriate method is available for a given problem or not. In all cases the numerical "workload" correlates somehow with the size of the atomistic system. Unfortunately, this is only in very rare cases a linear relationship, moreover scaling behaviors of N^2 to N^8 with respect to the number of atoms N, or often basis functions, are typical orders of magnitude. On the other hand, the correlation between accuracy, in the sense of physically well-founded, and computational effort, follows in the opposite direction. This means that only very exact results based on first principles are obtained exclusively by the higher scaling methods. Hence, only small systems can be treated accurately using first principles, but in the case of very few atoms, they can be even more exactly than the experiments. Since this is clearly not achievable for larger molecular systems, like enzymes, one has to find a way to reduce the numerical workload while keeping a satisfactory accuracy.

The most common route for this, is to introduce parameters that are empirically fitted in order to deliver the desired result. Thus, the extensive use of parameters has an ambivalent advantage. On the advantageous side, parameters not only deliver accuracy for larger system, but they can even push the limit of treatable system size beyond one million atoms, which makes atomistic simulations of whole virions possible.[154] On the other side, highly parametrized methods are naturally biased towards the reference values, they have been fitted and may not work for distinctive problems.

The “art of computational chemistry” can therefore be considered as using the right method for the right problem, while achieving the best compromise between accuracy and computational feasibility. This is clearly not possible by using the methods in a black box manner.

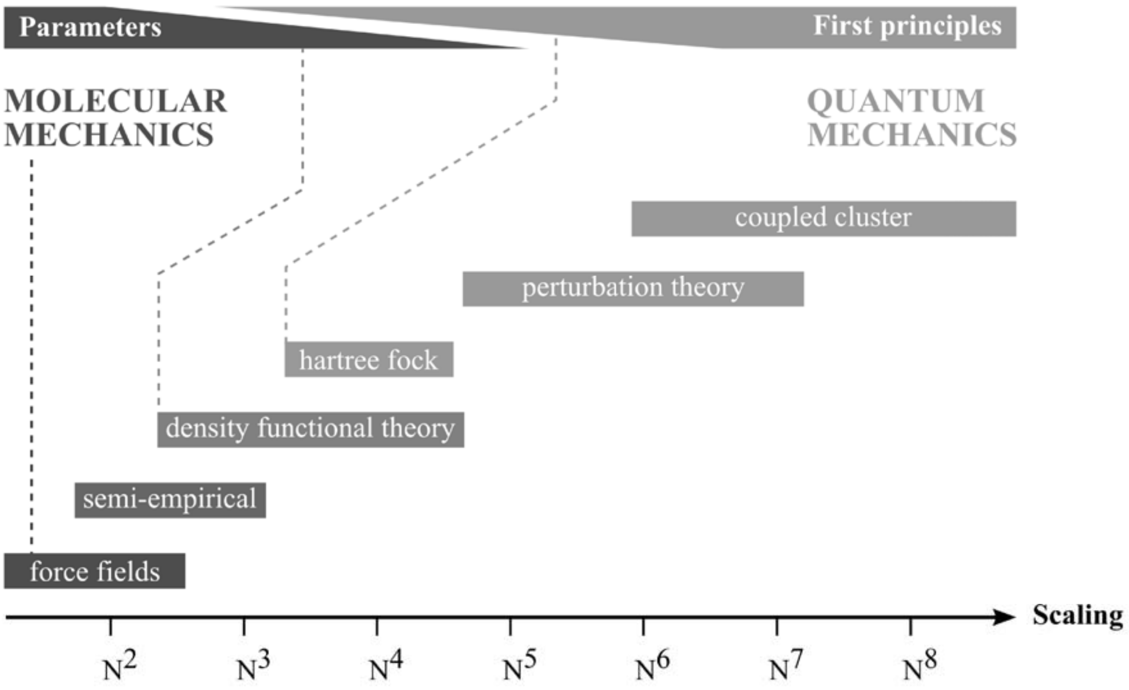


Fig. 4-1: Scaling behavior of computational chemistry methods with respect to the numerical workload. The latter restricts the size of the atomistic model that can be treated, as obvious from the low scaling (left hand side) and the high scaling (right hand side). N refers to the number of atoms (molecular mechanics) or basis functions (quantum mechanics). The parameters bar indicates how much a method's accuracy depends on the choice of empirical parameters, which does not mean that they have to be obligatory included in a method.

The following two sub chapters will, therefore, introduce the two main fields of computational chemistry, namely the molecular mechanical and the quantum mechanical approach, with a special emphasis on their practical use for treating problems on the atomistic scale. Hence, there is a need to explain the different methods on a level that enables one to understand the main differences, strength, and weaknesses. The aim of this chapter is to provide the reader some basic and essential ideas behind the herein applied methods, without giving too much exhaustive information on a single topic. Since the two striking factors, feasible number of atoms, and achievable accuracy for the given problem, usually exclude each other, a reasonable modeling of biomolecular reactions will naturally end up in the endeavor to combine the best of the two worlds. Methods that share this feature are called QM/MM approaches and are discussed in the third sub chapter.

4.1 The Molecular Mechanical Approach

The basic idea behind molecular mechanics is, to consider molecules as an ensemble of atoms or groups that are connected by elastic bands or springs. In the setting of this framework, the molecule is held together by harmonic forces that can either be caused by directly bonded interactions, like changes in bond length, angles, and dihedrals, or by non-bonded interactions, like van der Waals forces and coulomb interactions.

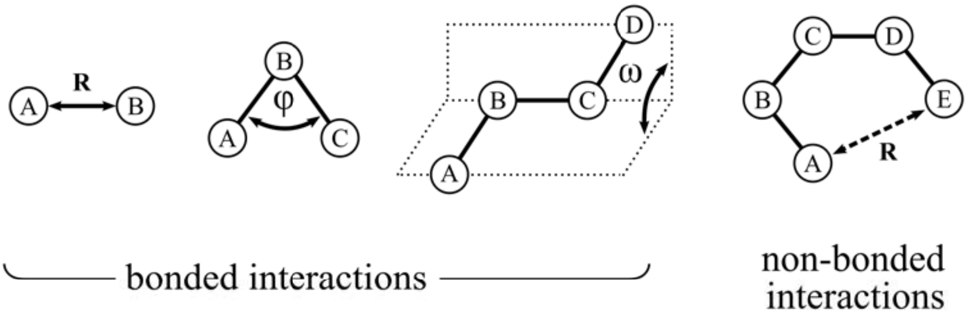


Fig. 4-2: Interactions that are typically considered in force field methods. Bonded interactions account for stretching, bending, and torsion of atoms, non-bonded interactions are calculated only for atoms that have a minimum of four bonds between them.

By assigning individual potential energy functions to each of these interactions, a force field can be set up that is simply obtained through adding up all the terms. The resulting energy expression *E* is generally a function of the molecular structure that assigns a certain energy value to its structural configuration. In most cases, this structure will not be in line with the ideal configuration, as defined by the harmonic parameters of the individual terms. Thus, the energy *E* can also be considered as some kind of molecular strain or steric energy, but has no physical meaning at all.[155] Although there are generally no clear rules, how to set up the energy expression in detail, most of the force fields rely on the following expression.[156]

$$E = E_{stretch} + E_{bend} + E_{torsion} + E_{vdW} + E_{coulomb} + E_{cross} \tag{8}$$

The first three terms belong to the bonded interactions and resemble the stretching energy $E_{stretch}$ that is caused by elongating two atoms out of their equilibrium distance, the bending energy E_{bend} that arises by bending the natural angle between three neighbouring atoms, and $E_{torsion}$ that accounts for changing the dihedral angle between four atoms (Fig. 4-2). Considering the stretching energy $E_{stretch}$, the model of a spring or elastic band directly leads to Hooke's law that describes the energy of a spring spanned between two atoms A and B by a harmonic expression that depends on the elongation $\mathbf{R^{AB}}$-$\mathbf{R_0^{AB}}$ from the equilibrium distance $\mathbf{R_0^{AB}}$, and a parameter k^{AB} as the spring constant. For the total number of bonds N in the system, the stretching energy is given by the following sum.[157]

$$E_{stretch} = \sum_{i=1}^{N} \frac{k^{AB}}{2} (\mathbf{R^{AB}} - \mathbf{R_0^{AB}})^2 \tag{9}$$

The same approach can be used to describe the bending energy E_{bend}, but with the difference that now three atoms A, B, C are involved that enclose an angle $\boldsymbol{\varphi}^{\mathbf{ABC}}$ and can be shifted with respect to the equilibrium angle $\boldsymbol{\varphi_0}^{\mathbf{ABC}}$. The resulting bending strains that depend on the given spring constants k^{ABC} can be summed up over all directly bonded angles M to the total bending energy E_{bend}.[157]

$$E_{bend} = \sum_{i=1}^{M} \frac{k^{ABC}}{2} (\varphi^{\mathbf{ABC}} - \varphi_{\mathbf{0}}^{\mathbf{ABC}})^2 \tag{10}$$

A special case of bending angles are out of plane angles that have advantages at pyramidalized arrangements of atoms,[156] like found in ammonia NH_3. Here, a distortion of the central nitrogen atom with respect to the plane, spanned by the three hydrogen atoms, leads only to marginal changes in the bond angles $\boldsymbol{\varphi}^{\mathbf{NHN}}$ and is therefore not ideally described. The use of a different angle definition that employs the angle between the spanned plane and the respective bonds, instead of the usually defined bond angles, accounts much better for this situation. Howevr, it can be similarly expressed by a harmonic energy term. In the molecular mechanics terminology, these angles are often described as improper angles or “impropers”.

Both expressions 9 and 10 can also be considered as truncated Taylor series that are terminated at the second order. From this point of view it becomes clear that they state only approximations of the "real" potentials, like the Morse potential that describes bond stretching much more reasonably than the harmonic potential. The latter one fails for the asymptotic behavior at larger distances $\mathbf{R}^{AB}$-$\mathbf{R}_0{}^{AB}$, which means that bonds cannot dissociate, and are therefore overestimated in their energy. Nevertheless, the harmonic potentials are mostly used in force field calculations due to their simplicity and sufficient accuracy near the equilibrium states $\mathbf{R}_0{}^{AB}$ and $\boldsymbol{\varphi}_0{}^{ABC}$.

The last type of the bonded interactions accounts for the internal rotation around bonds and involves four atoms A, B, C, D that define a torsion or twist angle $\boldsymbol{\omega}^{ABCD}$. Since the respective potential energy function has periodical character, this can be mathematically described by a Fourier series that includes the barrier heights V_n between the local minima. An illustrative example for this are the eclipsed conformations of an ethane molecule, which prefers a staggered conformation in the equilibrium state.[155,156] In most cases, three terms are sufficient to describe periodicities of 360° (n=1), 180° (n=2), and 120° (n=3), whereas the sum over all unique and directly bonded dihedral angles results in the total torsion energy $E_{torsion}$.

$$\begin{aligned} E_{torsion} = \sum [& \frac{1}{2} V_1^{\mathrm{ABCD}} (1 + \cos \omega^{\mathbf{ABCD}}) \\ & + \frac{1}{2} V_2^{\mathrm{ABCD}} (1 - \cos 2\omega^{\mathbf{ABCD}}) \\ & + \frac{1}{2} V_3^{\mathrm{ABCD}} (1 + \cos 3\omega^{\mathbf{ABCD}})] \end{aligned} \tag{11}$$

The non-bonded energy terms in equation 8 consist of the electrostatic interaction energy $E_{coulomb}$, and the remaining non-electrostatic interactions that are usually summarized in the van der Waals energy term E_{vdW}. The latter includes the weak intermolecular forces, like dispersion and dipol-dipol interactions,[73] and has the characteristic to be slightly attractive in the medium distance region but highly repulsive at very short distances due to the Pauli repulsion. One way to describe this behavior is pronounced in the Lennard-Jones potential that was proposed 1925,[158] and is still the most popular expression for E_{vdW} in modern force field implementations. It consists of a repulsive term that scales to the power of -12, and an attractive term that scales to the power of -6. At least, two parameters c_1 and c_2 are necessary to define this energy expression as a function of the interatomic distance $\mathbf{R}^{AB}$.[156]

$$E_{vdw}^{\mathrm{AB}} = \frac{c_1}{\left(\mathbf{R}^{\mathbf{AB}}\right)^{12}} - \frac{c_2}{\left(\mathbf{R}^{\mathbf{AB}}\right)^{6}} \tag{12}$$

Again, this expression does not meet the exact form of the real potential, but is a quite reasonable approximation for it. It has furthermore significant computational advances compared to the odd-numbered exponents of more accurate potential expressions.

The major part of interatomic interactions is due to electrostatic interactions between the various partial charges of the system. They can be straight forwardly expressed by a Coulomb potential. By assuming pairs of atom centered point charges q_1 and q_2, the electrostatic energy term $E_{coulomb}$ can be formulated as a function of the interatomic distance $\mathbf{R}^{\mathbf{AB}}$ and includes a dielectric constant ε.

$$E_{coulomb}^{\mathrm{AB}} = \frac{q_1 q_2}{\varepsilon \left|\mathbf{R}^{\mathbf{AB}}\right|} \tag{13}$$

The remaining energy term E_{cross} from equation 8 is a third type of interaction that neither belongs to the bonded interactions nor to the non-bonded interactions. It describes, moreover, possible couplings between the other terms and can look quite different, depending on the force field that is used. A recent example for biochemical force fields is described in the CMAP correction that includes cross terms E_{cross} to account for a proper treatment of the amino acid specific dihedral angles ϕ and ψ. They include a force constant $k_{n,m}$, the multiplicity n of ϕ, and m of ψ, as well as the phase of the cross term $\delta_{n,m}$.[159]

$$E_{cross} = k_{m,n}\left[1 + \cos\left(n\phi + m\psi - \delta_{n,m}\right)\right] \tag{14}$$

The CMAP correction is a development for the established CHARMM force field[160] and significantly improves the structural description of proteins.[161]

The so far discussed energy functions 9-13 can be found in most of the force field methods, but not necessarily. Individual force field methods may differ in their functional form, in their parameter sets, in the cross terms they use, and are often dedicated to a specific group of molecules. For biomolecular applications, AMBER,[162,163] CHARMM,[159,160] and GROMOS[164,165] are today's most widely used force fields. The molecular mechanical treatment of proteins has a long tradition in the field of computational chemistry, since the first simulation was already conducted in 1977 by McCammon et al.[166] who gave first insights into the dynamics of a folded protein.

Generally, the simulation of molecular dynamics requires not only an energy expression *E* of the considered system, but also equations that describe the motion of the particles. These introduce time *t*, mass m_i, and force *f* acting on a particle *i* at the position $\mathbf{R}^\mathbf{i}$ by Newton's second equation.

$$m_i \frac{d^2 \mathbf{R}^{\boldsymbol{i}}}{dt^2} = f_i(\mathbf{R}^{\boldsymbol{i}}) = \frac{dV}{d\,\mathbf{R}^{\boldsymbol{i}}} \tag{15}$$

Unfortunately, the differential expressions resulting from the equations of motions of a molecular mechanical system cannot be solved in a closed form.[167] They have to be integrated numerically, which can be achieved by time stepping algorithms. A commonly used procedure to solve this problem was developed by Verlet, and is thus known as the Verlet algorithm.[168]

$$\mathbf{R}^{\boldsymbol{i}}(t+\Delta t) = -\mathbf{R}^{\boldsymbol{i}}(t-\Delta t) + 2\,\mathbf{R}^{\boldsymbol{i}}(t) + f_i(\mathbf{R}^{\boldsymbol{i}})(\Delta t)^2 \tag{16}$$

The time step *Δt* influences the accuracy of the numerical integration and is usually set to one or two femtoseconds, which is reasonable for most of the molecular movements.

By applying all of these essential principles, it is possible to simulate a molecular system and keep track on the trajectories of the individual atoms. In biomolecular MD simulations, this is for example, routinely employed to calculate the root mean square deviations (RMSD) of protein backbones. It is a measurement for the average distance of the simulated backbone from the reference positions $\mathbf{R}^\mathbf{i}_\mathbf{ref}$. A possible reference point can be an experimentally determined protein backbone or the starting structure of the simulation.

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mathbf{R}^{\boldsymbol{i}} - \mathbf{R}^{\boldsymbol{i}}_{\boldsymbol{ref}})^2} \tag{17}$$

In the first case, the RMSD, which is calculated from the whole set *n* of individual atom positions $\mathbf{R}^\mathbf{i}$ along the trajectory, can give conclusions about the agreement between experimental and simulated structure, whereas the latter case indicates how inherently stable the simulation is. A critical assessment of the quality of molecular mechanical simulations is very important since errors can be introduced by various sources; like inaccurate or biased parameters, inadequate potential functions, or deficiencies in the applied algorithms as partly discussed above.[169]

Today's developments of molecular mechanical methods can either focus on pushing the technical limits with respect to system size and simulated time scale, or aim on conceptual improvements to increase accuracy and applicability. For the first topic, the recent trend of graphic processor unit (GPU) implementations[170–172] that can accelerate simulations by a factor of 10-100 with respect to classical MD codes is notable.[173,174] Apparent improvements from the conceptual point of view present polarizable force fields[175,176] that go beyond the picture of fixed point charges towards a polarizable environment. Since actual MD programs and methods have become quite mature today, future developments might also account for new applications and statistical analysis methods to extract more information from the enormous amount of data that can be generated by molecular mechanical dynamic simulations.[177,178]

4.2 The Quantum Mechanical Approach

Quantum mechanics is the fundamental theory that enables a first principle treatment of the electronic structure of a molecular system. Since chemical reactions can be modeled with the help of electron structure theory, it overcomes a major drawback of molecular mechanical approaches that lack a description of bond breakages and formations, except for exotic developments.[179] However, methods relying on quantum mechanics suffer from their unfavorable scaling behavior with respect to the number of atoms treated, which causes prohibitive high numerical workloads for systems that enter the scale of proteins. Whereas first MD simulations of protein were already conducted in the seventies,[166] the atomistic modeling of whole proteins with quantum mechanical methods is still in its infant stage, where methods employ parameters to reduce the scaling behavior[180] or technical advancements like acceleration by graphic processor units.[181] Therefore, the practical use of quantum mechanical methods is usually restricted to truncated protein systems, which is inherent problematic because important effects from the protein environment may be overlooked.[182] Another approach is to combine quantum mechanical with molecular mechanical methods, which have lead to the so called QM/MM hybrid approaches. The latter have evolved to the state-of-the-art method for the modeling of reactions in biomolecular systems[183–185] and are discussed in a separate chapter due to their importance. Nevertheless, the accuracy of QM/MM methods depends mainly on the quantum mechanical method that is used. In the following, a brief description of the commonly used quantum mechanical methods is given.

The three basic assumptions that are made in modern quantum chemistry are; the existence of a wave function $\Psi(\mathbf{r})$ from which all observable properties can be derived, the time independent Schrödinger equation, and the Born Oppenheimer approximation that states electron movements are separable from atom nuclei movements due to their negligible coupling in most cases. The last assumption leads to the famous non-relativistic electronic Schrödinger equation 18 that includes the Hamilton operator $\hat{H}_{el}$ and its energy eigenvalue E_{el}.

$$\hat{H}_{el}\Psi(\mathbf{r})=E_{el}\Psi(\mathbf{r}) \tag{18}$$

With these basic ingredients and several more sophisticated derivations, it is possible to describe molecules as many electron systems on first principles, without requiring parameters except for the physical constants. Since this framework derives the finally calculated properties right “from the beginning”, the quantum mechanical algorithms are also called *ab initio* methods in the computational chemistry terminology.

A central role in all quantum chemical calculations is played by the hamiltonian that is the operator $\hat{H}_{el}$ that yields the electronic energy E_{el} and is set up individually for each molecular system. The terms that arise in many electron systems depend on the number of electrons N, and the number of nuclei M as point charge representations for different atoms A with their different charges Z_A. They can be categorized into kinetic energy terms of the electrons $\hat{T}_{el}$ that are obtained through application of the Laplace operator ∇^2, the potential energy of the electrons in the field of the atom nuclei V_{eM}, and the electron-electron interactions V_{ee}. The last two terms include the distance of the electrons to the nucleus $\mathbf{r}_{\mathbf{iA}}$ and the interelectron distances $\mathbf{r}_{\mathbf{ij}}$.[186]

$$\hat{H}_{el}=-\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2-\sum_{i=1}^{N}\sum_{A}^{M}\frac{Z_A}{\mathbf{r}_{\mathbf{iA}}}+\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{\mathbf{r}_{\mathbf{ij}}}=\hat{T}_e+\hat{V}_{eM}+\hat{V}_{ee} \tag{19}$$

A further important part of the electronic Schrödinger equation 18 is the wave function $\Psi(\mathbf{r})$ that is generally a function of the electron coordinates $\mathbf{r}$. The wave function must fulfill several conditions to be valid, like being finite, continuous, differentiable, and unique at all points in space.[187] Whereas the wave function itself is usually a complex function, its square integral must result in a real number and is interpreted as the probability density of finding a particle within the given integration interval. Thus, for a normalized one electron wave function $\Psi_i(\mathbf{r}_i)$, the integration over the whole space must equal one. The following equation illustrates this relationship and is also known as the probability interpretation.

$$\int_{-\infty}^{+\infty} |\Psi_i(\mathbf{r}_i)|^2 = 1 \tag{20}$$

In electronic structure calculations, the wave function, which is at first instance a complete entity, is typically expressed in terms of a finite basis set expansion, which represents a more or less reasonable approximation. The quality of this approximation depends mainly on the number of employed basis functions but also on their type. The most widely used basis functions are of the gaussian type that are typically centered on the positions of the individual atoms. These may have drawbacks in the description of the wave function cusp at the nucleus, but this disadvantage is overcompensated by the fact that they can be integrated in computational procedures much more efficiently than other basis function types. A cartesian gaussian orbital contains a normalization constant *N*, and an exponent ζ that determines the radial shape or width of the function.[188]

$$\chi(\mathbf{x},\mathbf{y},\mathbf{z}) = N\,\mathbf{x}^{\boldsymbol{k}}\mathbf{y}^{\boldsymbol{m}}\mathbf{z}^{\boldsymbol{n}}\,e^{-\zeta\,\mathbf{r}^2} \tag{21}$$

The sum of *k*+*m*+*n* refers to the angular momentum *L* and determines the type of the orbital, for example representing s-, p-, d-orbitals for L=0, 1, 2. Due to the problem that gaussian functions look like quite different to “real” atomic functions, it is common practice to compose the basis function from multiple gaussian functions with different coefficients to better adopt them to the desired shape. This is called contraction. A contraction of an s-type basis function includes *N* number of contraction coefficients d_i.[151]

$$\Psi(\mathbf{r}) = \sum_i^N d_i e^{-\zeta_i \mathbf{r}^2} \tag{22}$$

Adding more (contracted) functions for one atom, e.g. through splitting of the valence orbitals into more functions with different exponents ζ, leads to split valence basis sets that are classified by the extent of splitting to double zeta, triple zeta, quadruple zeta, or higher orders.

Introducing polarization or diffuse functions, generally improves the description of the electron distribution.

Hartree Fock theory. Based on the hitherto introduced fundamental parts of quantum chemistry, a variety of methods exist that can reach different levels of accuracy in the description of the electronic structure. It should be stated here, that none of the methods deliver exact solutions of the Schrödinger equation 18 (except for the most simplest cases of a one electron system) and this is an inherent property of the many body problem. Hence, this has resulted in a broad spectrum of approximations that follow different philosophies within the various methods and theories.

The historically most important method is the Hartree Fock method. It forms the basis for many highly accurate approaches and is a cornerstone in quantum chemistry. A central concept within the Hartree Fock theory is the treatment of the many body problem by an iterative mean field approach. The wave function is expressed in the mathematical form of a determinant (Slater determinant) that is built up from one electron functions and has the inherent property to be antisymmetric, thus to fulfilling the Pauli principle. The goal is to find the best wave function with the lowest energy for a given basis set. This can be achieved by applying the variational principle as a criteria to systematically improve trial wave functions. In terms of this, the best trial wave function of a set is simply the one that leads to the lowest energy. With the mathematical method of Lagrange multipliers, one can derive the Hartree Fock equations that include the effective one-electron energy Fock operator $\hat{F}_i$ and form a set of pseudo eigenvalue equations that include the orbital energies ε_i .

$$\hat{F}_i \phi_i = \varepsilon_i \phi_i \tag{23}$$

From this set of functions it is possible to determine the solution for one orbital ϕ_i , if all other orbitals are known, thus one electron is considered in the mean field of all other electrons. This is one crucial issue for the inaccuracy of the method, since dynamical correlation of the electrons is not included here and cannot be accounted for by simply treating the electron-electron repulsion in an average fashion. The inclusion of a basis set expansion finally leads to the Roothaan Hall equations that represent the Hartree Fock equations in an atomic orbital basis. They can be written as a matrix equation that includes the Fock matrix **F**, the expansion coefficient matrix **C** and the overlap matrix **S** of the basis functions.[186]

$$\mathbf{F}\mathbf{C}=\mathbf{S}\mathbf{C}\varepsilon \tag{24}$$

The Hartree Fock wave function is now obtained through an iterative process that is started by an initial orbital coefficient guess, followed by several cycles in which the orbitals are subsequently optimized with respect to lower energies of the wave function, until final convergence is reached. At this state, the electric field does not change any more and is called self-consistent. In this iterative self consistent field (SCF) process, the Fock matrix **F** has to be diagonalized at each step and two electron integrals have to be calculated. This is very time consuming and is usually the computational bottleneck of the method. Regarding the number of employed basis functions, the computational effort of the Hartree Fock method scales to the fourth power.[156]

As stated above, one of the main deficiencies of the Hartree Fock method is due to the lack of dynamic electron correlation. In other words, the probability of finding an electron at a particular position does not explicitly depend on the position of the remaining electrons, i.e. part of the electron interactions are not included. The missing correlation energy can be partly recovered by several so called *post-HF* methods that introduce the explicit correlation of electrons through different approaches.

Perturbation theory. One common way to introduce dynamic electron correlation is the application of perturbation theory as elaborated in the Møller Plesset perturbation theory of second order (MP2).[189] The essential improvement over Hartree Fock is the extension of the reference hamiltonian $\hat{H}_0$, for which the solution is known, by a pertubation operator $\hat{H}'$ that accounts for the correlation energy and can be scaled by a factor λ.[156]

$$\hat{H}=\hat{H}_0+\lambda\hat{H}' \tag{25}$$

When λ is zero, the hamiltonian H equals the reference Hamiltonian $\hat{H}_0$. Perturbation methods have the premise that the perturbation is small compared to the unperturbed solution. By applying this formalism to the Schrödinger equation and the wave function, one ends up with a series of expressions that include energy corrections of zeroth, first, second, and higher order that can be truncated at any level and correspond to Møller Plesset perturbation theories of zeroth (MP0), first (MP1), second (MP2), ect. order. It is notable that in terms of this formalism, the Hartree Fock theory already includes an energy correction of the first order, whereas the energy of zeroth order is just the sum of the orbital energies. Therefore, Hartree Fock can be considered as perturbation theory of the first order.

$$E(MP0)+E(MP1)+E(MP2)=E(HF)+E(MP2) \tag{26}$$

Dynamic electron correlation starts therefore at corrections of the second order due to Brillouin's theorem.[190] MP2 can be considered as the most economic variant with regard to perturbation theories of higher orders, since it scales to the fifth power and recovers already 80-90% of the correlation energy. MP3 already scales to the sixth power and accounts for 90-95% of the correlation energy that is only a small improvement with respect to the increase in computational demand compared to MP2.[156]

Coupled cluster methods. Besides perturbation theory, electron correlation can also be included in electronic structure calculations by coupled cluster methods. The most prominent example is the CCSD(T) methods which is sometimes denoted as the gold standard of quantum chemistry. This reputation comes from the fact that it reaches “chemical accuracy” with typical errors in the margin of 1 kcal/mol.[191] In other words, coupled cluster methods are the method of choice when highest accuracy is desired in electron structure methods.

The idea of coupled cluster methods is to include all possible electronic excitations of a given order *i*. These are generated by an excitation operator $\hat{T}$ from the reference wave function. They are hence multi-determinant approaches, since calculations do not only include a single ground state slater determinant, as in Hartree Fock, but also include all excited slater determinants of the *i*-th order. The highest possible order of excitation is of course equal to the number of electrons present in the system.

$$\hat{T}=\hat{T}_1+\hat{T}_2+\hat{T}_3+\ldots\hat{T}_i \tag{27}$$

Whenever an exciting operator, e.g. the first excitation operator $\hat{T}_1$, acts on the reference wave function, typically the Hartree Fock wave function Φ_{HF} , it generates a set of excitations that have expansion coeffients t_{jk} and span over all occupied and virtual orbitals. The expansion coeffients t_{jk} are called amplitudes in this context.

$$\hat{T}_1\Phi_{HF}=\sum_{j}^{occupied}\sum_{k}^{virtual}t_{jk}\Phi_{jk} \tag{28}$$

Due to Brillouin's theorem that essentially states that single excited determinants do not directly interact with the reference Hartree Fock determinant (respective integral expressions become zero),[190] coupled cluster methods can only make sense of double or higher excitations. Unfortunately, the inclusion of single and double excitation (CCSD) already scale to the sixth power with regard to the number of basis functions. Inclusion of triple excitations (CCSDT) is far away from being applicable to molecules, except for the smallest ones, due to its scaling with the eighth power. To make these methods available to bigger molecules, methodological developments must aim at reducing the unfavorable scaling behavior.

This, for example, can be done in the above mentioned CCSD(T) method, where the brackets indicate that the triple excitations are not treated explicitly but with perturbation theory. The latter one yields a correction energy and is simply added to the CCSD energy, which is less expensive than an explicit treatment and still a very good approximation.[192] Nevertheless, CCSD(T) is still one of the most costly approaches that scales between the sixth and seventh power. Moreover, it needs large basis sets to unfold its high accuracy. Therefore, recent developments like density fitting (also often referred to as resolution of identity approximation),[193,194] local correlation treatments,[195] or F12 theory,[196,197] further try to reduce this problem.

Another development in this approach is the CC2 method that is loosely referred to as an intermediate between coupled cluster and perturbation theory. It is especially well suited to treat excited states problems and scales only to the fifth power. Its reliability is of comparable quality to the MP2 method.[198,199]

Spin component scaling. Another important development for *post-HF* methods are spin component scaling technique (SCS), which is an empirical correction. It is based on the separate scaling of the correlation energy by its parallel $E_{\uparrow\uparrow+\downarrow\downarrow}$ and anti-parallel spin components $E_{\uparrow\downarrow}$ and includes two scaling parameters p_t and p_s that are fitted to benchmark molecule data sets.[200]

$$E_{corr}(\mathrm{SCS}) = p_s E_{\uparrow\downarrow} + p_t E_{\uparrow\uparrow+\downarrow\downarrow} \tag{29}$$

Since the method is only an empirical scaling of quantities that are already calculated in most of the quantum chemical calculations, it does not cause any considerable additional computational costs. It is therefore *ad-hoc* applicable to existing methods like MP2 (SCS-

MP2)[200] or CC2 (SCS-CC2).[201] The SCS-MP2 method has been reported to show good performance for the treatment of non-covalent biological relevant molecules.[202]

Semi-empirical methods. As shown for the spin component scaling methods, introducing parameters into *ab initio* quantum chemical methods can be rather efficient to increase the accuracy of a given method or, thinking the other way around, to achieve a given accuracy with reduced computational effort. The group of semi-empiric methods goes one step further than *ad-hoc* corrections and introduces parameters directly into the quantum mechanical methodology, which is usually the Hartree Fock method. The most time consuming operations within *ab initio* methods are typically the evaluation of two electron integrals and matrix diagonalization procedures. The general aim of semi-empirical developments are therefore a "debottlenecking" of the critical steps to speed up the calculations. Some clues from the theoretician's toolbox to achieve this include; the neglect of three or four electron integrals, approximation of integrals, considering only the valence shell, using minimal basis sets, introducing fitting parameters, and several more. Indeed, the scaling behavior can thereby be reduced to a quadratic or to a third power scaling with respect to the number of basis functions.[155]

Not only historically important, but still partly in use today, are the group of methods that employ a neglect of diatomic differential overlap (NDDO) approximation. The NDDO methods are one variant of the zero differential overlap (ZDO) approximation that is a crucial feature of semi-empirical methods and requires that the product of basis functions χ located at different atoms A and B be set to zero. This reduces the scaling behavior from the fourth to the second power for the two electron integrals.

$$\chi_A \chi_B dV = \begin{Bmatrix} =0 \, for \, A \neq B \\ \neq 0 \, for \, A = B \end{Bmatrix} \tag{30}$$

The errors made by these approximations clearly need to be recovered, which is done by expressions that include atomic parameters. The latter ones are assigned in fitting procedures to experimental data. One common way to do this is the Dewar-Sabelli-Klopman (DSK) approximation[203,204] that is found in all NDDO models. Here, the two center two electron integral γ_{AB} is expressed by the interatomic distance $\mathbf{R}_{AB}$ and two atomic parameters p_1^A and p_2^B.

$$\gamma_{AB}=\frac{1}{\sqrt{\mathbf{R}_{AB}^{2}+\left(p_1^A+p_2^B\right)^2}} \tag{31}$$

Well known representatives of this group are MNDO,[203,205,206] AM1,[207] and PM3.[208] Whereas the first one was already published in 1977 and is rarely used today, the last two are still continuously developed. Newly parametrized and improved variants are RM1,[209] a reparametrization of AM1, and PM6,[210] a successor of PM3. Although these methods differ only slightly in their terms and parameters, the resulting differences in molecular properties can be considerable.[211]

A "new kid on the block" of semi-empirical methods is the self-consistent-charge density functional tight binding method (SCC-DFTB).[212] It shares common features to the classical semi-empirical methods but has a fundamental different background, since it relies on density functional theory. From the practical point of view, this method seems to outperform the classical methods with respect to geometries, especially for biological systems, but has drawbacks in energy predictions of ions and radicals.[211,213,214]

Density functional theory. All methods that have been discussed so far are based on a wave-function based approach and Hartree Fock theory. They provide a systematic way to improve the recovery of correlation energy, but always at a cost of higher orders of magnitude for the scaling behavior with respect to the number of employed basis functions. Density functional theory (DFT) is an alternative approach in treating the electron structure of molecules that intrinsically includes electron correlation and scales more favorable than wave-function based methods. Due to this good cost/benefit ratio, it has become very popular in computational chemistry.

Density functional theory relies on the fact that a relationship between the electron density and the ground state energy exists, which can be derived. The density ρ(**r**) itself corresponds to the number of electrons *N* that must be obtained by integrating over all space **r**.

$$N=\int_{-\infty}^{+\infty}\rho(\mathbf{r})\,d^3\mathbf{r} \tag{32}$$

The breakthrough for DFT was achieved by Hohenberg and Kohn in 1964 who first proved that it was possible to express the kinetic energy of non-interacting electrons and their interactions in the form of functionals, as a formally exact treatment of the many electron problem.[215] Functionals $F[\rho]$ can be considered as functions that itself depend on other functions, which can be the density that depends on the wave function Ψ with minimum energy.[216]

$$F[\rho]=\min_{\Psi\rightarrow n}\langle\Psi|[\hat{T}+\hat{V}_{ee}]|\Psi\rangle \tag{33}$$

In other words, the electronic energy of a molecule can be calculated by finding a suitable solution of non-interacting electrons that defines a unique and observable electron density ρ(**r**) of the real system, and calculate the missing electron-electron interactions by a functional of this density. The energy in DFT is therefore defined as a set of functionals that contain kinetic energy of the single electrons $T_s[\rho(\mathbf{r})]$, the Hartree energy $J[\rho(\mathbf{r})]$ that accounts for the inter-electron repulsion, the external potential $V_{eM}[\rho(\mathbf{r})]$, and a remaining term that includes the rest of the interactions, defined as exchange correlation functional $V_{XC}[\rho(\mathbf{r})]$. [156]

$$E[\rho(\mathbf{r})]=T_s[\rho(\mathbf{r})]+J[\rho(\mathbf{r})]+V_{eM}[\rho(\mathbf{r})]+V_{XC}[\rho(\mathbf{r})] \tag{34}$$

Note that the herein found solution corresponds to a set of orbitals for a fictitious non-interacting system (Kohn Sham orbitals) that has the purpose of representing the density of the real system.

Whereas the first three terms of equation 34 can be derived straight forwardly, the "magic" of DFT is hidden in the exchange correlation functional $V_{XC}[\rho(\mathbf{r})]$. Although DFT is formally exact, the exact form for this expression is unknown, which prompted the need for approximated solutions. This offered room for ideas and a variety of developments that involve the use of different functional forms for $V_{XC}[\rho(\mathbf{r})]$ and often include the use of empirical parameters. Due to this, DFT approaches have no clear hierarchy and lack a systematically way to improve the method. One attempt to introduce this, is the metaphor of the Jacob's ladder of density functional theory.[217] It categorizes DFT methods in four rungs by their level of sophistication that usually correlates well with the achieved accuracy.[218] The different rungs start at the most easiest way for deriving the exchange correlation energy; climbing up the first being rung one of these called the local electron density (local density approximation, LDA). Climbing up we get to the second rung of generalized gradient approximation (GGA), then to the third rung of hybrid functionals that partly include Hartree Fock exchange, and end up at the fourth rung of double-hybrids that include further improvements.

Today's DFT methods are usually named after their exchange correlation functionals. Each of them is unique and, therefore, a large "zoo" of abbreviations exist. The most widely used is the B3LYP functional, named after Becke, Lee, Yang, and Parr.[219–222] It contains energy contributions from local spin approximation $E_{XC}^{LDA}, E_X^{LDA}, E_C^{LDA}$, generalized gradient approximation E_X^{GGA}, E_C^{GGA}, Hartree Fock E_X^{HF}, and further includes three empirical parameters a_0, a_x, a_c that are fitted to atomic properties.

$$E_{XC}^{B3LYP} = E_{XC}^{LDA} + a_0\left(E_X^{HF} - E_X^{LDA}\right) + a_X\left(E_X^{GGA} - E_X^{LDA}\right) + a_C\left(E_C^{GGA} - E_C^{LDA}\right) \tag{35}$$

The B3LYP functional is therefore a hybrid-GGA functional. More sophisticated parametrization schemes can be found in recent functionals like the family of Minnesota functionals that are named by the letter M and the year of their development, e.g. M05, M06, M08, or M11.[223–227]

Although there is a long line of developments in the field of DFT, the methods possess some typical problems and it seems that still not all of them are completely untapped.[228] Major issues of inaccuracy are non-covalent interactions, underestimation of reaction barriers, problems with dissociated ions, or charge transfer in complexes. All of them are based on an inaccurate description of electron delocalization.[229] Some of these deficiencies can be overcome by the use of empirical dispersion correction terms,[230,231] or application of scaling factors that correct systematic errors, e.g. in the prediction of vibrational frequencies.[232,233]

Nevertheless, DFT methods remain a very good choice due to their favorable scaling behavior between the third and the fourth power with respect to the number of basis functions. They have therefore largely contributed to the more widely use of computational chemistry methods by giving access to much larger molecules. Recent technical developments with regard to GPU accelerated implementations push the limit further away to even bigger system sizes that enable simulations of whole proteins with density functional theory methods.[181,234]

Continuum Solvent models. Biological systems like proteins usually reside in water, which often imposes the description of this solvent as an obligation for realistic modeling. Since inclusion of explicit solvent molecules multiplies the size of the system, and therefore often exceeds the feasible computational limits, an alternative and cheap description of the solvent is needed. Continuum solvent models meet this need by treating the solvent particles in a continuous way by means of a distribution function that simulates the environment in an averaged manner. This general concept is also denoted as an implicit solvent modeling. In terms of the quantum mechanical formalism, this can be expressed by an effective hamiltonian $\hat{H}_{eff}$ that is built up from the hamiltonian of the solvated system $\hat{H}_S(\mathbf{r})$ and an interaction potential V_{int} that includes a solvent response function Q. The solvent response function depends on a couple of position vectors $\vec{s}, \vec{s}\,'$ that replace the whole set of solvent coordinates **s**.[235]

$$\hat{H}_{eff}(\mathbf{r},\mathbf{s})=\hat{H}_S(\mathbf{r})+V_{\text{int}}[\mathbf{r},Q(\vec{s},\vec{s}\,')] \tag{36}$$

This approach greatly simplifies the treatment of solvation effects, since it includes the response of the environment without having the need for an explicit description. Hence, calculations of molecules in a solvent can be conducted with a comparable effort to gas phase calculations.

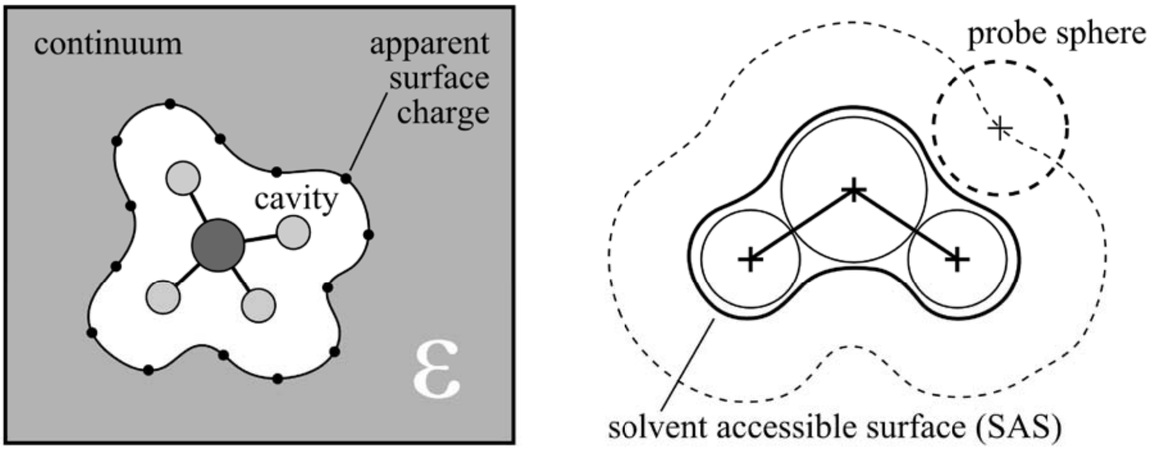


Fig. 4-3: Concept of continuum solvent models that describe the solvent in terms of a molecular cavity that is embedded in a uniform polarizable continuum (left hand side) with a dielectric constant ε. The continuum is often expressed as point charges on the cavity surface. Cavity construction can be conducted through tracing out the solvent accessible surface by use of a probe sphere (right hand side).

A central concept within continuum solvent models is the cavity of a molecule. How to define its shape and size can be quite different in the various models, but they all share the same feature that the solute molecule is considered as being put in a void cavity, surrounded by a dielectric medium. The latter one conclusively mimics the solvent (Fig. 4-3, left hand side). An easy way to set up a molecular cavity would be to simply construct it from overlapping van der Waals spheres. Nevertheless, this does not necessarily represent a good surface that is thoroughly accessible to the solvent molecules. Therefore, the molecular cavity is often constructed as a solvent accesible surface (SAS), which can be obtained by tracing out the SAS with a probe sphere that accounts for the size of the solvent molecules (Fig. 4-3, right hand side).[235]

With the help of the molecular cavity, the solvent reaction can be expressed by apparent charges on the surface, which reduces the source of the response to a charge distribution on top of the closed cavity. Since the integration over such a complex shape is quite challenging, the surface is usually simplified to a discrete number of surface elements, the so-called *tessarae*.[236]

The first method that employed this elaborated methodology was the polarizable continuum model (PCM) that was introduced in 1981,[237] although the concept of a reaction field is considerably older having its origin in the nineteen thirties.[238,239] PCM models are still in use today and actively developed.[240–243] A central quantity within these models is the polarizability of the continuum that itself depends on the dielectric constant of the solvent. The dielectric constant is a unique property for each solvent that has high values for very polar solvents like water or low values for unpolar solvents like alkanes. The relationship between polarization $\vec{P}_i$ and dielectric constant ε_i is given by the following equation that includes the gradient of the total electrostatic potential $\nabla \vec{V}(\vec{r})$ and is important for determining an apparent charge *i* on the cavity surface.

$$\vec{P}_i(\vec{r}) = -\frac{\varepsilon_i - 1}{4\pi} \nabla \vec{V}(\vec{r}) \tag{37}$$

A very popular improvement over these methods is the conductor like screening model COSMO that simplifies the polarizable continuum to a perfect conductor, where $\varepsilon = \infty$.[244] As a consequence, the surface charges are perfectly screened and the boundary condition for the electrostatic problem are strongly simplified, which leads to more efficient calculations. To recover solvent effects for finite dielectric constants, the apparent charge densities $\sigma(\mathbf{s})$ are simply scaled from the ideal screened charge density $\sigma^{is}(\mathbf{s})$ by an empirical function $f(\epsilon)$ of the dielectric constant ε and a parameter k.

$$\sigma(\mathbf{s}) = f(\epsilon)\sigma^{is}(\mathbf{s}) = \frac{\epsilon - 1}{\epsilon + k}\sigma^{ps}(\mathbf{s}) \tag{38}$$

The concept of COSMO was further developed to a more sophisticated version COSMO-RS, [245] a successful tool for the prediction of thermodynamic properties of solvents and mixtures of them. Applications for this are mainly found in the chemical engineering field,[246] but are also rather useful for drug development purposes, like prediction of octanol-water partition coefficients.

4.3 State of the Art for Proteins: The QM/MM Hybrid Method

As stated at the beginning of the chapter, atomistic treatments of biomolecular reactions suffer from the problem of large system sizes with lots of atoms. Since enzymatic reactions mostly occur at the surface of the protein, continuum solvent models are not an appropriate choice to overcome this, because they treat the surrounding in a uniform manner. This becomes apparent by considering the water and protein environment in terms of their dielectric constants that correspond to very polar media (ε=78)[247] and unpolar media (ε<10).[248,249] This conclusively implies problems for implicit models that use the dielectric constant as the central descriptor of the environment. Nevertheless, if the atomistic view should be kept, this naturally leads to the idea of combining the benefits (and hopefully not the drawbacks) of quantum chemical and molecular mechanical treatment.

Such QM/MM approaches have evolved to the state-of-the-art method for treating biomolecular reactions.[183] Albeit their stellar ascent in academic science throughout the last decade, they are far from being black box methods, since they require a lot of methodological decisions before a study can be conducted. Typical questions are:

- Which combination of methods should be employed?
- What should be included in the QM, and what in the MM part?
- Is it necessary to cut bonds across the QM/MM boundary and where should this be conducted?

All of these questions cannot be answered systematically and are specific for every case. They are further complicated by technical concerns like availability of proven implementations or error-prone codes. Therefore, most of the expertise in this field comes from practical experience and exploration of whether the main effects are captured for a given problem or not. Since the pioneering days of QM/MM are over its wide spread use as well as critical assessments of their accuracies, pitfalls, and work flows are constantly being addressed.[250–253]

The key idea of QM/MM methods was first presented in 1976,[254] but its benefits were recognized much later at the beginning of the 1990s.[255] The pivotal point of the hybrid approach is to divide the atomistic system into the chemically important part and a spectator region that accounts for the environment. The respective regions are then treated by quantum mechanics (QM) and molecular mechanics (MM) where the QM part is somehow embedded into the MM part (Fig. 4-4).

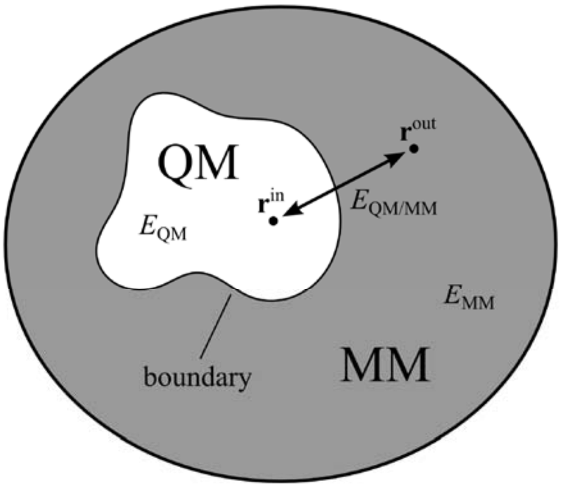


Fig. 4-4: QM/MM scheme, set up from an inner QM part and an outer MM part.

The QM/MM methodology mainly differs in the inner part by how QM and MM regions are coupled, embedded, or how the boundary between them is described. Special concerns arise when covalent bonds need to be cut, since they leave unsaturated valencies.

The general energy terms that arise in such a system are respectively contributions from the QM part E_{QM}, the MM part E_{MM}, and a coupling term $E_{QM/MM}$ that accounts for the interactions between them. The coupling term is only necessary when the QM/MM scheme is set up in an additive scheme, which means that both parts are calculated separately, but include somehow an interaction between each other. Hence, this depends on coordinates for the inner QM part $\mathbf{r}^{\mathrm{in}}$, and the outer MM part $\mathbf{r}^{\mathrm{out}}$.

$$E^{add}_{\mathrm{QM/MM}}(\mathbf{r}^{\mathbf{in}}, \mathbf{r}^{\mathbf{out}}) = E_{QM}(\mathbf{r}^{\mathbf{in}}) + E_{MM}(\mathbf{r}^{\mathbf{out}}) + E_{\mathrm{QM/MM}}(\mathbf{r}^{\mathbf{in}}, \mathbf{r}^{\mathbf{out}}) \tag{39}$$

There is also a second possibility, the subtractive QM/MM scheme, where the whole system $\mathbf{r}^{\mathrm{all}}$ is calculated as the MM part, and the inner part $\mathbf{r}^{\mathrm{in}}$ as the QM description. Simply adding both terms together obviously leads to double counting of interactions, which is compensated by subtracting the MM energy of the inner part $\mathbf{r}^{\mathrm{in}}$, which is obtained by a separate calculation.[183]

$$E^{\mathrm{sub}}_{\mathrm{QM/MM}}(\mathbf{r}^{\mathbf{all}}) = E_{MM}(\mathbf{r}^{\mathbf{all}}) + E_{QM}(\mathbf{r}^{\mathbf{in}}) - E_{MM}(\mathbf{r}^{\mathbf{in}}) \tag{40}$$

In practical use, the additive scheme can be observed as the one mostly applied, in conjunction with an electrostatic embedding of the QM part in the MM environment.[256,257] The latter one means that the QM part “sees” the MM part in terms of its point charges that are explicitly included in the electronic structure calculation and, therefore, account for the polarization of the wave function (or electron density) by the environment.

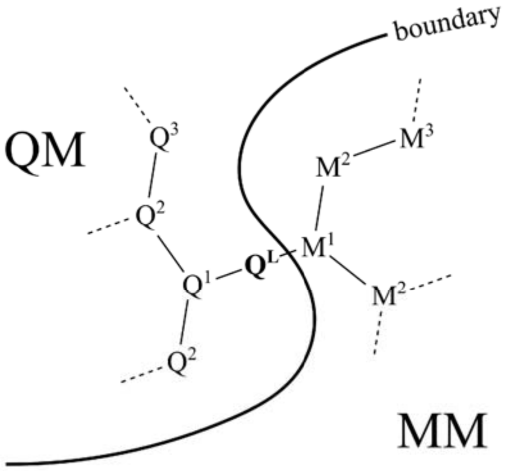


Fig. 4-5: QM/MM boundary across a chemical bond. The link atom scheme saturates the open valence on the QM side through introduction of a link atom Q^L.

Practically, these charges are taken directly from the MM parameter set. An alternative way to couple the QM and MM description is mechanical embedding, where the QM and MM part are simply coupled by the position of their atoms. This scheme lacks the description of electronic coupling, but is much easier to implement and “retrofit” with existing quantum chemical codes.

If QM/MM methodology is applied to biochemical reactions, it is often unavoidable to choose the QM part in a way that it crosses a chemical bond. This is for example necessary when a catalytic residue is studied that sticks inside a long peptide chain of the enzyme. In these cases, the boundary of the QM part contains one or more unsaturated valencies that would impose unwanted effects on the electronic structure that are rather artificial. To circumvent that problem in a chemical sense, the open valencies can simply filled with hydrogen atoms. Although this is somehow more a workaround than an elegant theoretical treatment of the problem, like introducing frozen orbitals, it provides a good and efficient way to fix that issue. The herein introduced hydrogen atoms are called link atoms Q^L and are an often applied practice in QM/MM methods (Fig. 4-5).[255]

The concept of a QM/MM boundary causes a further problem by cutting through covalent bonds, which is related to the point charge description of the MM part. In terms of electrostatic embedding, the cut off atom M^1 that resides on the MM side of the divided bond would be represented as a point charge in very close proximity to the quantum chemical link atom Q^L. Such a close point charge would impose an unphysical strong polarization of the nearby link atom. To account for that over-polarization, a pragmatic way is to shift the point charge further away from the QM part. Since the charge itself should be conserved, it makes sense to redistribute this charge as dipoles on the neighboring MM atoms M^2.[257]

By overcoming the discussed deficiencies and several others, QM/MM methods present a valuable tool for the modeling of very challenging questions, like enzymatic catalysis,[258] charge-relay mechanism inside of enzymes,[259] or X-ray structure refinement of enzyme-inhibitor complexes.[260]

The scaling behavior with respect to the system size can be expected, like for the employed QM part alone, since the MM description usually requires much less computational time than the quantum chemical calculations. For the practical use, QM/MM methods can therefore be considered as QM calculations that have been augmented by a highly sophisticated environment description that does not require significantly increased computational costs.

5 Results and Discussion

5.1 Accuracy of Theoretical Methods for Enzymatic Proton Transfer Reactions

The first step towards a reliable modeling on the atomistic scale is to estimate the accuracy for the available methods. For the given case of an enzymatic reaction, special attention must be paid for the transfer of protons that account for the majority of reactions in biocatalysis. Hence, the determination of protonation states or proton transfers is relevant for most of theoretical studies dealing with enzymes and their actions in biological systems. The prediction of pK_a values for protein residues is one example for a routinely applied procedure. Moreover, proton transfers are involved in the crucial steps of peptide cleavage reactions or inhibition processes. Whenever these reactions are studied, the accurate description of the involved proton transfer reaction becomes important what can be an extremely demanding task.

Whereas the determination of protein residue pK_a values can be efficiently tackled by using empirical pK_a prediction algorithms[261–263] or semi-macroscopic protein dipole/Langevin dipole approaches,[264,265] theoretical studies of biochemical reactions mostly demand for quantum chemical calculations. For whole enzymes that include thousands of atoms the computational effort of calculations based on quantum mechanics is usually beyond the feasible limit. As a consequence, the system size treated with quantum chemical approaches needs to be restricted to the essential part of the investigated reaction while the environment has to be described on a more simple level of theory. Two very common approaches using this strategy are implicit solvent models[236,266–268] or combined quantum mechanical/molecular mechanical calculations (QM/MM).[254,255] For both concepts, recent reviews are available. [183,235,269–271]

Although the original intention of implicit solvent models is the mimicry of a solvent surrounding in an averaged manner, which makes it suitable for the prediction of pK_a values[272] or other thermodynamic properties,[245] their use has also been extended to biochemical reactions. Hereby, the character of the environment is mainly determined by the dielectric constant ε that can also be evaluated for a single case.[273] The spectrum of investigations reaches from the determination of protonation states of catalytic dyads[102] to investigation of hydrolysis reactions[274] as well as mechanistic studies of inhibition mechanisms.[275,276] Nowadays, the majority of studies that deal with proton transfers in a biochemical context employ QM/MM methods and rely on density functional theory in most cases.

A recent example for QM/MM investigation of this kind is the water mediated proton transfer between histidine and a hydrogen peroxide, which plays an important role in the catalytic cycle of the horseradish peroxidase.[277] Another reaction, where proton transfer occurs in a relayed fashion, is studied by comparing the direct and the hydroxyl group mediated hydrogen atom transfer in peptide bond formation of ribosomes.[278] Beside enzyme specific studies, the fundamental understanding of neutral and zwitterionic resting state of cysteine/histidine catalytic dyads is a frequently attended topic during the last years. Insights into this field have been provided by the work of Mladenovic et al. from the cysteine protease cathepsin B,[279,280] by Kaukonen et al. from [Fe,Ni] hydrogenase,[281] or Ke et al. who investigated the catalytic mechanism of an arginine deiminase.[101,282] The KasA enzyme, a drug target against tuberculosis, gives further interesting insights since it contains one cysteine and two histidine residues that have been studied by Lee et al.[283] Nevertheless, such kind of studies are not only conducted for cysteine/histidine residues but also for other types of catalytic protein sites, like demonstrated for the Pilin enzyme, which employs asparagin, lysin, and glutamate as catalytic residues.[284] Furthermore, proton transfer reactions are often involved in the inhibition process of an enzyme. Here, knowledge of the underlying mechanisms provides important hints for the rational drug design of inhibitor molecules. An example where a proton transfer is directly involved in an inhibition reaction is given by Cheng and coworkers[285] who study the covalent inhibition of a serine protease that possesses a catalytic triad. Like most of the studies mentioned above, they use the often applied density functional B3LYP within a QM/MM approach. With an analogue methodology, Mladenovic et al. give insight to the importance of proton shifts for the inhibition of cysteine proteases by aziridines and epoxides as irreversible warheads.[286]

Recent work of Xie et al.[287] demonstrates a complementary use of implicit solvation model on the CCSD(T) level of theory and QM/MM computations employing density functional theory to gain insight into the formation of carbamates from carbon dioxide. Whereas this study uses the two concepts of solvent models separately, Ryde and coworkers[288] apply a combination of explicit and implicit treatment of the environment in terms of the QM/MM-PBSA method, which is suited to estimate free energy differences of proton transfer reactions in a promising manner. Further interesting applications, where proton transfer reactions are involved, are the theoretical interpretation of kinetic isotope effects within a protein by Olsson et al.[289] or the investigation of the intramolecular proton transfer within malonaldehyde,[290] which both represent examples for the calculation of nuclear quantum effects.

As mentioned above, the major part of studies employ density functional theory due to its good ratio between accuracy and computational cost, but there are also many examples where higher or lower levels of theory are used. Especially in case of simulations with *ab initio* methods, the thereby associated computational costs require the use of computationally inexpensive methods,[291] like semi-empirical hamiltonians.[279,292] Although a lot of benchmarking studies are available throughout the literature, that probe the accuracy of quantum chemical methods in the context of proton transfer reactions,[293–299] most focus on accuracy under gas phase conditions. There are only a few examples in the literature[300,301] that provide a comparison throughout different descriptions of the environment, like the work of Sharma et al.[302] who assess the accuracy of different quantum chemistry models in the context of proton transfer tunneling and kinetic isotope effects.

Since implicit solvent models or QM/MM methods can have a large impact on the calculated properties, it is questionable how, or to which extend, benchmarks made under gas phase conditions are valid for these approaches.

Taking these things together, there is actually no study available that systematically investigates the behavior of commonly used quantum chemical methods with respect to different modeling of the environment, like implicit solvation or the QM/MM approach. Therefore it is highly desirable to probe the "robustness" of quantum chemistry methods with regard to this aspect and to reveal a possible bias of parametrized methods towards gas phase models.

Hereby, the focus is set more on the direct comparison of the quantum chemical hamiltonians rather than on methodologies for calculating pK_a values or free energies. For the latter topics, extensive information is available in the literature. A good overview about calculating pK_a values in proteins is given by the review of Gunner and coworkers,[303] whereas an overview about free energy methods has been recently addressed by two recommendable reviews that cover this field in a more general way[291] or with emphasis on the QM/MM approach.[304] Although the accuracy of free energy calculations will not be evaluated here, it should be emphasized that the reliability of methods, which derive Gibbs or Helmholtz free energies from QM or QM/MM potentials, strongly depends on the accuracy of the chosen quantum mechanical hamiltonian. Therefore, the choice of the quantum chemical method can have significant impact on the free energy results, which makes a critical assessment of its reliability necessary.

The results can also help to reveal whether certain methods possibly possess a systematic error behavior with respect to the character of the environment. This offers the possibility to correct such kind of errors by an empirical shifting factor.

For the selection of the probed methods, three important facts are considered. The first one is how commonly used a method is. Therefore, popular approaches like the Møller Plesset perturbation theory MP2, density functional theory using B3LYP, or the classical semi-empirical hamiltonians MNDO, PM3, and AM1 are included. As a second important issue, more recent developments should be included like improvements due to spin component scaling techniques (SCS), meta-hybrid exchange-correlation functionals (e.g. M06-2X) or higher sophisticated parametrizations as found in the PM6 or RM1 method. The third and most decisive fact is, whether a method is implemented for the use in conjunction with an implicit solvent model or the QM/MM approach. Fortunately this holds for the majority of methods that are applied for proton transfer reactions.

Approximated coupled cluster approach CC2 is also included, which is indeed rarely used for the modeling of proton transfer potentials, but well suited for the calculation of excited states and their potential energy surfaces. For problems of this kind, often multireference methods are necessary, although DFT approaches can also deliver satisfying accuracies in some cases.[305–308] Thus the results should give an estimate, which impact can be expected for the CC2 method when other environments than gas phase are applied.

To evaluate the robustness of different quantum chemical methods with respect to the environment modeling, gas phase calculations with an implicit solvent model and an explicit description of the environment in terms of a QM/MM approach are probed.

Since a consistent comparison of a broad spectrum of hamiltonians is aimed, the major bottleneck of this study is the availability of all methods within the different environment models in the used program packages. To provide a complete and consistent comparison of all methods and schemes, 6 program packages are necessary to employ. Especially the use of QM/MM methodology limits the choice of hamiltonians, since an electrostatic embedding scheme, which is the most popular embedding scheme for biomolecular applications,[183] is not possible with all quantum chemical programs. Due to its flexible design and the possibility to interface multiple quantum chemistry codes, the ChemShell program package represents the best choice for this challenge.[309]

As implicit solvent model, the COSMO approach is an approved and often applied method,[244] which is implemented in most program packages. An important parameter for the use of continuum models is the selection of the dielectric constant. Here, it makes sense to consider two values, a low one of ε=4 that represents an environment of low polarity and refers to a protein bulk,[248,249] whereas the second one of ε=78 is usually chosen for very polar environment, comparable to water. The proton transfer reaction between the catalytic residues cysteine 145 and histidine 41 at the active-site of SARS coronavirus main protease represents something in between these two situations, for which all methods are evaluated. The proton transfer represents the transition from the neutral charged resting state, in the following abbreviated with **N**, to the zwitterionic resting state, abbreviated with **ZW**, of the catalytic dyad. The transition state between the two minima is denoted as **TS**.

Experimental studies on the active-site residues of SARS-CoV main protease estimate pK_a values for cystein 145 and histidine 41 to 7.7-8.3 and 6.2-6.4, thus finding the neutral charged state **N** as the predominant situation for the free enzyme.[58,71] The pK_a differences between the two groups can be employed to estimate the free solvation energy differences according to equation 41, which is adopted from the work of Warshel et al.[252,264,310] and discussed there in detail. For the given case, the free energy difference between **N** and **ZW** can expected to be in the range of 7-12 kJ/mol at standard condition for temperature *T*.

$$\Delta G_{sol}^{N \rightarrow ZW} \simeq 2.3\, RT \left(pK_{\mathrm{a}}^{Cys} - pK_{\mathrm{a}}^{His} \right) \tag{41}$$

Geometric parameters from X-ray crystallography,[46,51,52] as well as from molecular modeling studies,[67,78] further reveal that proton transfer is most likely to occur in a water-mediated fashion. The main indicative factor for this finding is the measured distance between the cysteine sulfur and the histidine nitrogen. It is found to be in a range between 3.5 and 4 Å[46,51,52] and is therefore too large for a direct proton transfer that would require a direct hydrogen bond with a sulfur-nitrogen distance about 3 Å. The setup of the model system and the relevant reaction coordinates are shown in Fig. 5-1.

Three properties have been evaluated for the proton transfer path:

1. Thermodynamic and kinetic data in terms of relative energies for minima and transition states
2. Mean unsigned errors of the whole proton transfer potential
3. Location of minima and transition states along the reaction coordinate

The assessment of accuracy for all methods used is done relative to LCCSD(T) | QZVP level of theory for gas phase and QM/MM computations. Due to its advantageous scaling behavior, accompanied by an insignificant loss of accuracy, density fitting approximation (also known as resolution of identity approximation) has been employed.[194,311] Whereas canonical CCSD(T) approaches are expected to have chemical accuracy for at least quadruple zeta basis sets (mean error below 1 kcal/mol) with respect to experimental reaction energies,[191] local approximations can introduce an additional average error of 0.6 kcal/mol.[195]

For computations combining LCCSD(T) and COSMO no code was available, so the SCS-CC2 | QZVP level of theory is taken as reference method, since it yields least errors in comparison to LCCSD(T) results for this case. Nevertheless, its performance is quite similar to SCS-MP2, which should possess a mean absolute error below 1.8 kcal/mol for reaction energies in gas phase.[200,312]

Two basis sets are used for each method, one of triple zeta quality and one of quadruple zeta quality, in order to roughly estimate the magnitude of basis set effect for each method. It is notable that the description of anionic species can be further improved by introducing augmented basis sets, but since convergence problems are experienced quite often with these, their use has been omitted here.

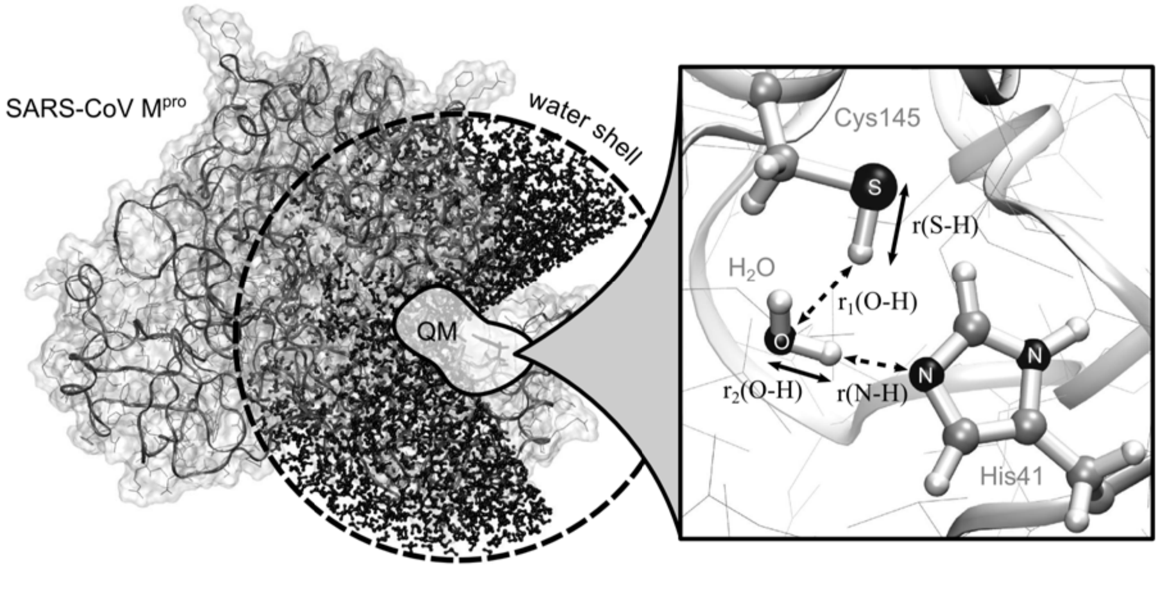


Fig. 5-1: QM/MM scheme for the active-site of SARS coronavirus main protease (left hand side). Cysteine 145, histidine 41, and a bridging water molecule were taken as QM part (right hand side). Proton transfer reaction occurs along the coordinates r(S-H), r_1(OH), r_2(OH), and r(N-H). The r(S-H) coordinate (solid arrow) is used as the "driving" reaction coordinate, all other coordinates (dashed arrows) follow smoothly in forward and backward direction during the proton transfer reaction.

The lowest minimum of the studied proton transfer path is given by cysteine 145 and histidine 41 in their neutral charged state **N**, which is taken as reference state. Therefore, energy differences ΔU of transition states (**TS**) and zwitterionic states (**ZW**) are always taken relative to the respective neutral minimum state **N** on each potential energy curve. A detailed description of the evaluated energies is given in Fig. 5-2.

Errors for the relative energies of the minimum state of the zwitterion ΔΔU(ZW) and the transition state ΔΔU(TS) are calculated as differences between the obtained energy difference ΔU and the reference energy difference $\Delta U_{ref.}$ of the respective single point structure. The groups of wave-function based methods, density functional theory and semi-empirical methods are discussed in an analogue manner: First, unsigned errors and their behaviour with respect to the environment model are discussed. Second, tendencies of over- or underestimation are considered. Third, interesting trends are highlighted.

Calculations of mean unsigned errors (MUE) are done over the sum of all 14 data points *N* of the proton transfer path according to equation 42. Energy differences ΔU of all points are taken relative to the neutral minimum state **N** on the respective potential energy curve, as described above.

$$MUE = \frac{1}{N} \sum_{1}^{N} \left| \Delta U - \Delta U_{ref} \right| \tag{42}$$

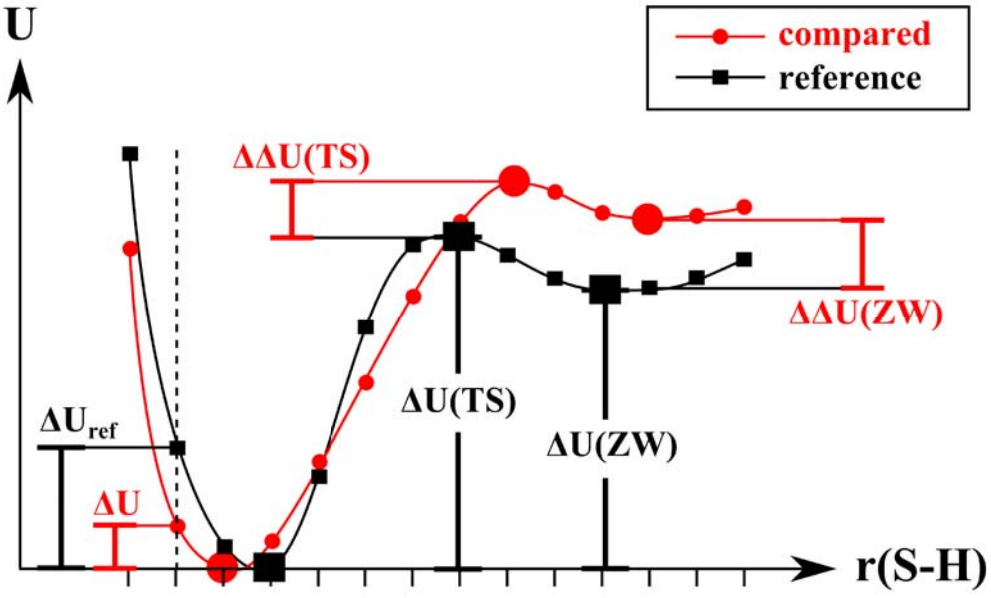


Fig. 5-2: Explanation of the relative internal energies ΔU and their differences ΔΔU, which are calculated from the respective grid points (solid horizontal lines) along the reaction coordinate r(S-H).

A similar approach has been applied for deviations of minima and transition state locations along the reaction coordinate as shown in Fig. 5-3. Errors for the distances of neutral states Δr(N), zwitterionic states Δr(ZW), and transition states Δr(TS) along the reaction coordinate r(S-H) are calculated as differences between the obtained distance r(N,TS,ZW) of the respective minimum/maximum point of the probed method and the reference distance value r_{ref} (N,TS,ZW).

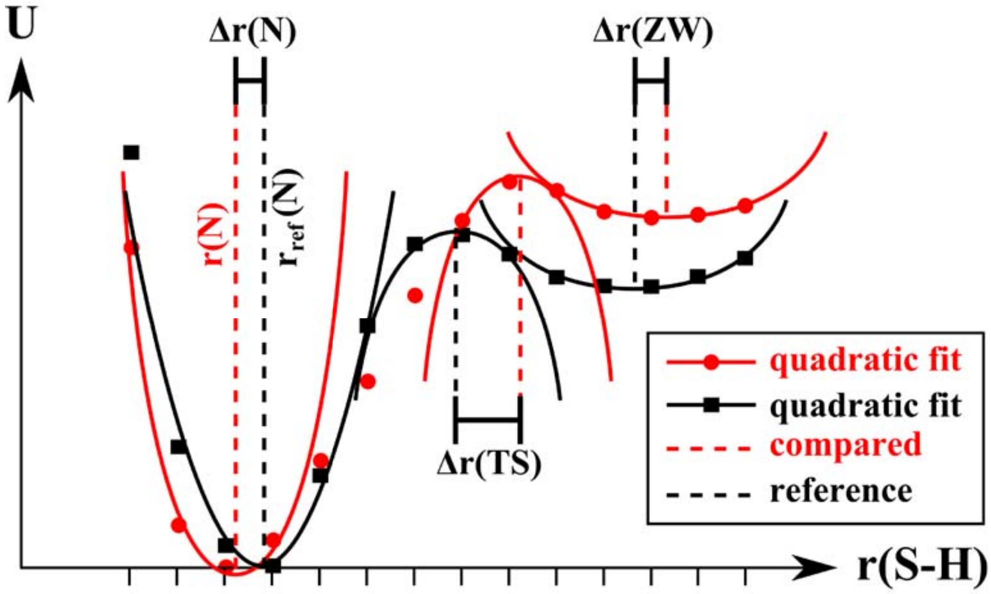


Fig. 5-3: Quadratic fits of extremal values **N**, **TS**, and **ZW**. Fitted functions of the reference (black) and compared potentials (red) are depicted as solid lines. Deviations Δr (dashed vertical lines) are taken relative to fitted extremal values of the reference potential.

The minima and maxima points along the reaction coordinate have been approximated by fits of harmonic functions (polynomial of second degree) that include two neighboring points in each direction around the closest incremental minimum point on the potential curve. The fits are schematically drawn as dashed lines in Fig. 5-3. The location of the minimum or maximum point is hence obtained by determination of the extremal value of the fitted function. For TS, only one neighboring point in each direction has been used around the maximum value, since its location on the reaction coordinate is often very close to ZW. This fitting procedure should give more accurate estimates of the stationary points than simply taking the highest or lowest incremental value on the potential energy curve.

5.1.1 Comparison of Different Environment Models

The proton transfer between cysteine and histidine connects the neutral charged state **N** with the zwitterionic state **ZW**. The kinetics of the reaction is determined by the height of the transition state **TS**. The proton transfer potentials in Fig. 5-4 compare the gas phase results with those obtained for the implicit environment model COSMO using two different dielectric constants, and those computed with the QM/MM approach. The potentials are computed on the LCCSD(T) | QZVP (gas phase, QM/MM) and SCS-CC2 | QZVP level of theory (COSMO). As apparent from Fig. 5-4, all environment models reproduce the experimental finding that **N** is more stable than **ZW**. However, the predictions show significant quantitative differences.

Proton transfer potential in gas phase and QM/MM are in a similar range, predicting an energy difference of 95 and 97 kJ/mol for **ZW** over **N** by using LCCSD(T). A more distinctive difference can be observed for the barrier heights **TS**, where the QM/MM model shows a slightly higher value of 103 kJ/mol for TS with respect to 96 kJ/mol in gas phase. The use of the COSMO model has a large impact on the proton transfer potential in terms of stabilizing **ZW**. Even with a low dielectric constant of ε=4, which is thought to reflect the situation in a protein environment, the zwitterionic state is lowered by 41 kJ/mol with respect to the QM/MM description and moreover by 63 kJ/mol if a higher dielectric constant of ε=78, comparable to water, is applied. This result is counter intuitive at the first sight, since the catalytic residues are located at the surface of the protein and thus the dielectric constant of the surrounding should therefore be somewhere between 4 and 78, but even the COSMO description of a low ε=4 delivers a significant larger stabilization of **ZW** than the QM/MM model. On closer inspection, this difference is not surprising, since the implicit solvent model mimics an ideal and perfect solvent screening for the reaction in contrast to the QM/MM description. The latter one models the environment in an explicit way, thus taking molecular effects into account that might arise at the water protein interface. This is further underlined by the comparison to the experimentally derived free energy difference between 7 and 12 kJ/mol, as calculated from equation 41, which are much better in agreement with the COSMO results. In summary, the qualitative result of having **N** as the lower state is correctly predicted by all proton transfer potentials in Fig. 5-4, but only due to the fact that both states **N** and **ZW** are sufficient distinctive in terms of their relative energies. In cases, where the two states are very close together, a proper relaxation of the environment and further free energy calculations become important, as it has been demonstrated by the work of Lee et al.[283] who investigated similar proton transfers for another enzyme.

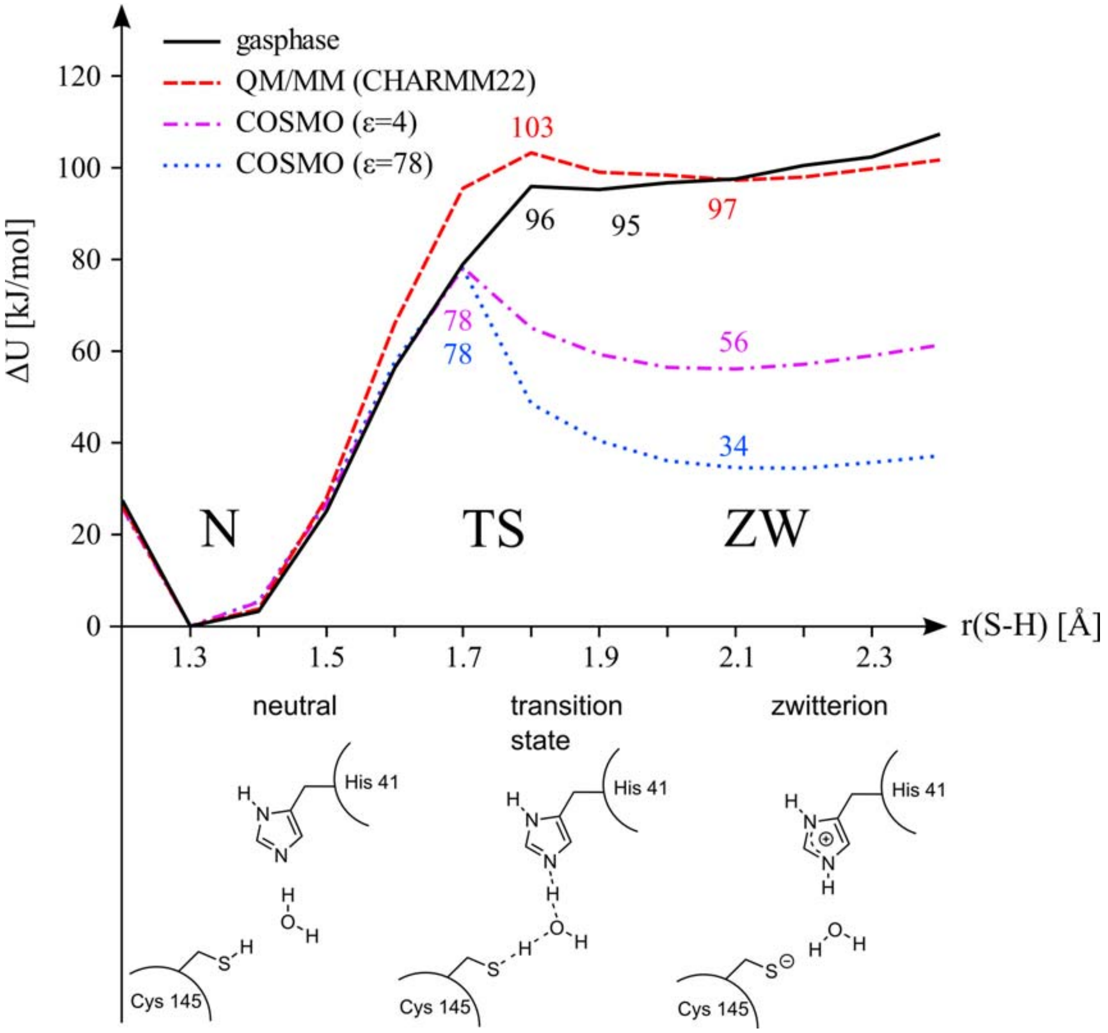


Fig. 5-4: Comparison of proton transfer potential in gasphase, continuum solvent model COSMO, and QM/MM scheme. Gasphase and QM/MM potentials have been obtained on the LCCSD(T) | QZVP level of theory, COSMO potentials employing SCS-CC2 | QZVP. Energies of transition states **TS** and zwitterionic states **ZW** are taken relative to the neutral charged states **N** of each potential curve in kJ/mol. The structures of the minimum energy path have been optimized within the QM/MM scheme on the RI-BLYP | TZVP level of theory.

5.1.2 Probing Thermodynamic, Kinetic, and Structural Properties

Tab. 5-1 probes the accuracy of the wave-function based ab initio methods, the density functional theory methods, and the semi-empirical methods with respect to gas phase calculations, COSMO approach, and QM/MM scheme by a comparison of the relative energies ΔU(ZW) between the neutral charged state **N** and the zwitterionic state **ZW**, as well as the transition state energy ΔU(TS). Tab. 5-2 lists the deviation of each method with respect to this reference and is obtained according to the energy differences defined in Fig. 5-2. In the following, only deviations > 8 kJ/mol are considered as significant. This error bar takes the remaining uncertainties of non gas phase computations into account and are near the expected accuracy limit of the reference methods with respect to experimental gas phase values (1.6 kcal/mol for LCCSD(T) | QZVP[191,195] and 1.8 kcal/mol SCS-CC2 | QZVP[200,312]), as discussed in the previous sub chapter.

Since the performance of a certain method to predict relative energies must not necessarily be reflected in the quality of structure predictions,[299] the BLYP | TZVP geometries were probed against B3LYP | TZVP and MP2 | TZVP structure optimizations, as well as optimizations on the AM1 and PM3 semi-empirical level. The comparison in Tab. 5-1 of MP2 | TZVP (indicated by *) and MP2 | TZVP // BLYP | TZVP and furthermore B3LYP | TZVP (indicated by *) and B3LYP | TZVP // BLYP | TZVP reveals that the proton transfer potentials are quite similar and differ in, or below, the range of significance. Interestingly, structure optimizations with AM1 and PM3 were not able to reproduce the minimum energy path properly and delivered quite different structures as apparent from Fig. 5-5. Detailed plots of direct comparisons between relaxed and unrelaxed geometries can be found in the computational details section in Fig. 7-2 for MP2, Fig. 7-3 for B3LYP, and Fig. 7-4 for the semi-empirical methods.

The inspection of the wave-function based group of *ab initio* methods reveals a very consistent error behavior across the gas phase, COSMO, and the QM/MM approach (Tab. 5-2). In detail, the unsigned errors of the relative energy ΔΔU(ZW) for the post-HF methods (CCSD(T), LCCSD(T), SCS-CC2, SCS-MC2, MP2, CC2) range in general below an absolute value of 14 kJ/mol and change only insignificantly when the environment model is switched (< 5 kJ/mol).

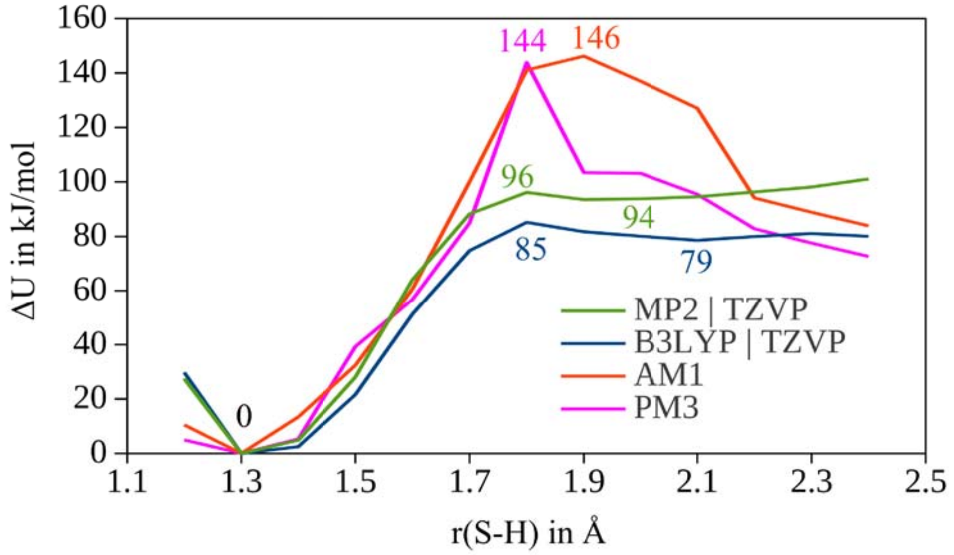
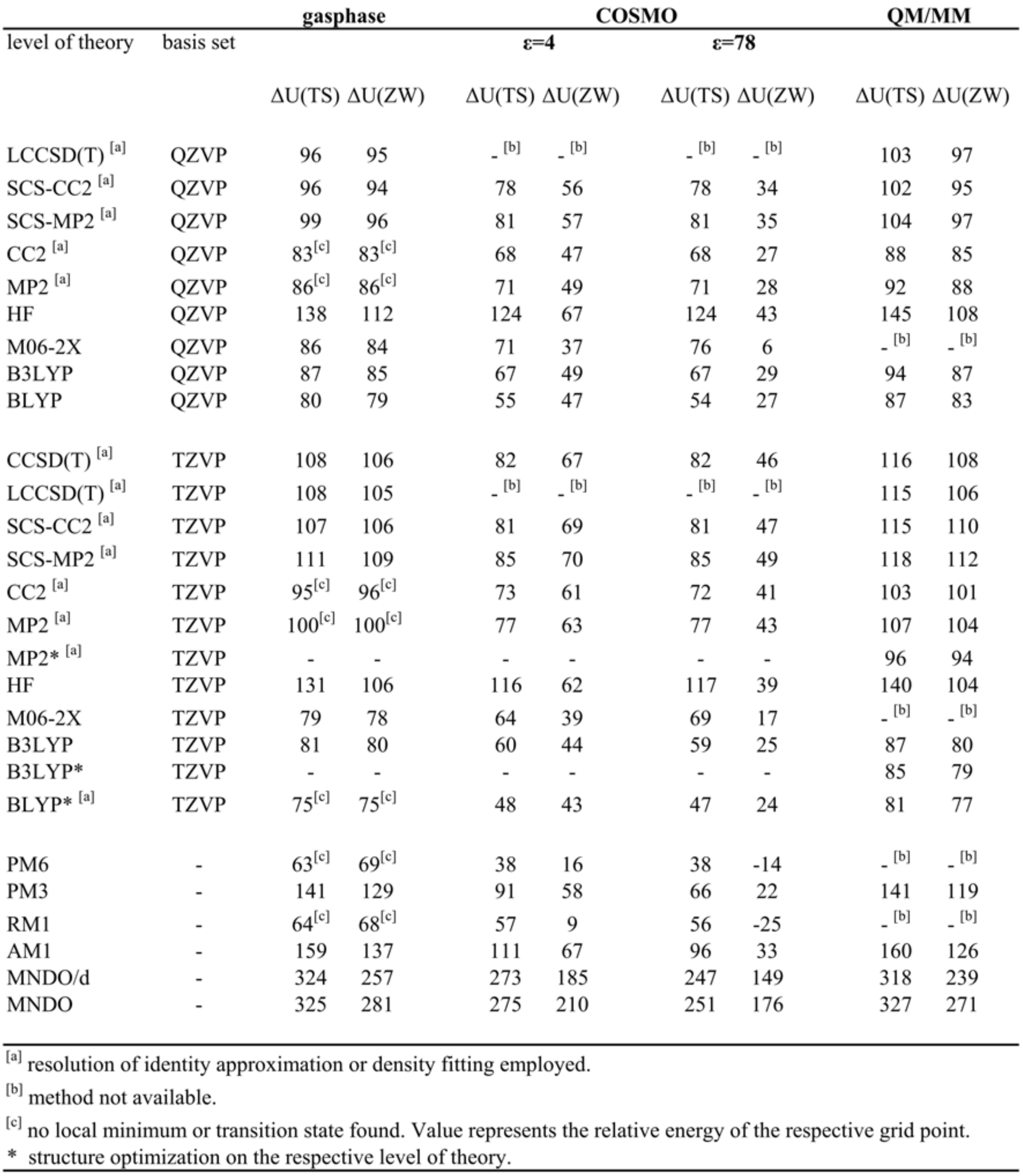


Fig. 5-5: Relaxed QM/MM minimum energy paths of the proton transfer potential, according to the model system described in Fig. 5-1. The MP2 | TZVP and B3LYP | TZVP methods deliver smooth potential energy surfaces for the r(S-H) coordinate, comparable to the unrelaxed BLYP | TZVP geometries. The semi-empirical methods AM1 and PM3 possess significant problems for this case and deliver no reasonable minimum paths without applying further geometrical constraints.

The errors of the relative transition state energy ΔΔU(TS) behave similar to this, since the largest unsigned deviation is 15 kJ/mol in case of the CC2 | QZVP method within QM/MM scheme. The error behavior with respect to the environment model is also in line with the ΔΔU(ZW) values and range at the limit of significance with a maximum absolute difference of 9 kJ/mol (SCS-CC2 | TZVP with COSMO versus QM/MM). The higher level methods CCSD(T), LCCSD(T), SCS-CC2, and SCS-MP2 reveal a consistent tendency of 12-15 kJ/mol overestimation of the energy difference between **N** and **ZW**, when the TZVP basis sets are used. By comparing same basis set quality with each other, MP2 and CC2 show very similar performance and differ in their results only marginally (3-5 kJ/mol). Accuracy might be improved by spin component scaling, but not consistently. Interestingly, spin component scaled methods tend to deliver better results with QZVP basis sets (ΔΔU < 3 kJ/mol vs. ΔΔU < 15 kJ/mol for TZVP), whereas the unscaled CC2 and MP2 counterparts seem to give slightly but not significantly better results in conjunction with the smaller TZVP basis set (ΔΔU < 9 kJ/mol vs. ΔΔU < 15 kJ/mol for TZVP). On the other hand, gas phase calculations employing MP2 and CC2 without spin component scaling do incorrectly reproduce the shape of the potential curve, since they predict a complete repulsive potential, without minimum for

ZW and transition state.



		gasphase		COSMO				QM/MM	
level of theory	basis set			ε=4		ε=78			
		ΔU(TS)	ΔU(ZW)	ΔU(TS)	ΔU(ZW)	ΔU(TS)	ΔU(ZW)	ΔU(TS)	ΔU(ZW)
LCCSD(T) [a]	QZVP	96	95	- [b]	- [b]	- [b]	- [b]	103	97
SCS-CC2 [a]	QZVP	96	94	78	56	78	34	102	95
SCS-MP2 [a]	QZVP	99	96	81	57	81	35	104	97
CC2 [a]	QZVP	83[c]	83[c]	68	47	68	27	88	85
MP2 [a]	QZVP	86[c]	86[c]	71	49	71	28	92	88
HF	QZVP	138	112	124	67	124	43	145	108
M06-2X	QZVP	86	84	71	37	76	6	- [b]	- [b]
B3LYP	QZVP	87	85	67	49	67	29	94	87
BLYP	QZVP	80	79	55	47	54	27	87	83
CCSD(T) [a]	TZVP	108	106	82	67	82	46	116	108
LCCSD(T) [a]	TZVP	108	105	- [b]	- [b]	- [b]	- [b]	115	106
SCS-CC2 [a]	TZVP	107	106	81	69	81	47	115	110
SCS-MP2 [a]	TZVP	111	109	85	70	85	49	118	112
CC2 [a]	TZVP	95[c]	96[c]	73	61	72	41	103	101
MP2 [a]	TZVP	100[c]	100[c]	77	63	77	43	107	104
MP2* [a]	TZVP	-	-	-	-	-	-	96	94
HF	TZVP	131	106	116	62	117	39	140	104
M06-2X	TZVP	79	78	64	39	69	17	- [b]	- [b]
B3LYP	TZVP	81	80	60	44	59	25	87	80
B3LYP*	TZVP	-	-	-	-	-	-	85	79
BLYP* [a]	TZVP	75[c]	75[c]	48	43	47	24	81	77
PM6	-	63[c]	69[c]	38	16	38	-14	- [b]	- [b]
PM3	-	141	129	91	58	66	22	141	119
RM1	-	64[c]	68[c]	57	9	56	-25	- [b]	- [b]
AM1	-	159	137	111	67	96	33	160	126
MNDO/d	-	324	257	273	185	247	149	318	239
MNDO	-	325	281	275	210	251	176	327	271

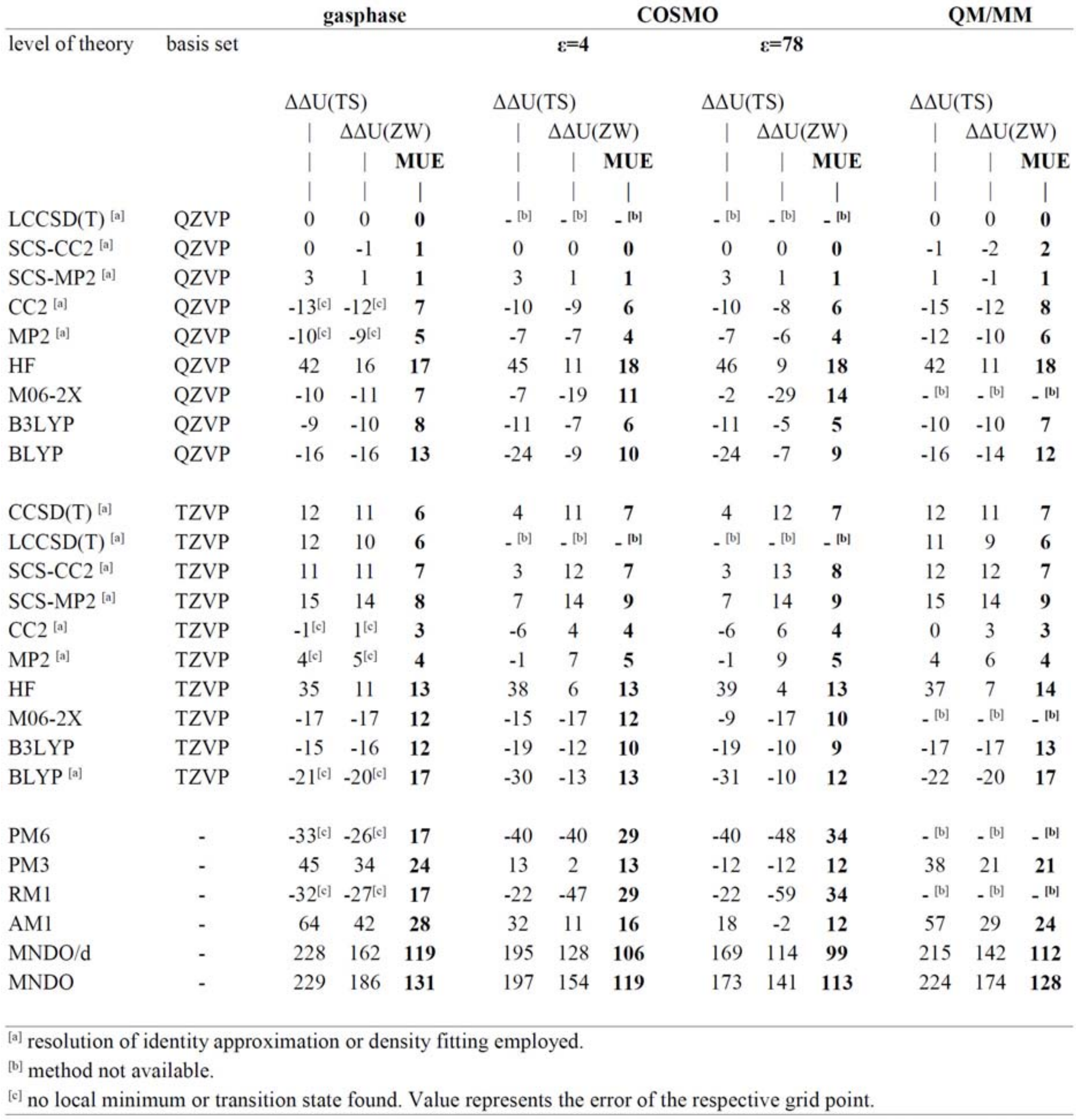
[a] resolution of identity approximation or density fitting employed.

[b] method not available.

[c] no local minimum or transition state found. Value represents the relative energy of the respective grid point.

* structure optimization on the respective level of theory.

Tab. 5-1: Evaluation of thermodynamic and kinetic data from different quantum chemical methods in kJ/mol. Predicted values ΔU(TS) and ΔU(ZW) are taken relative to neutral charged state **N** on the respective level of theory. A detailed description of the relative energies is given in Fig. 5-2.



		gasphase			**COSMO**						**QM/MM**		
level of theory	basis set				**ε=4**			**ε=78**					
		ΔΔU(TS)	ΔΔU(ZW)	**MUE**	ΔΔU(TS)	ΔΔU(ZW)	**MUE**	ΔΔU(TS)	ΔΔU(ZW)	**MUE**	ΔΔU(TS)	ΔΔU(ZW)	**MUE**
LCCSD(T) [a]	QZVP	0	0	**0**	- [b]	- [b]	**- [b]**	- [b]	- [b]	**- [b]**	0	0	**0**
SCS-CC2 [a]	QZVP	0	-1	**1**	0	0	**0**	0	0	**0**	-1	-2	**2**
SCS-MP2 [a]	QZVP	3	1	**1**	3	1	**1**	3	1	**1**	1	-1	**1**
CC2 [a]	QZVP	-13[c]	-12[c]	**7**	-10	-9	**6**	-10	-8	**6**	-15	-12	**8**
MP2 [a]	QZVP	-10[c]	-9[c]	**5**	-7	-7	**4**	-7	-6	**4**	-12	-10	**6**
HF	QZVP	42	16	**17**	45	11	**18**	46	9	**18**	42	11	**18**
M06-2X	QZVP	-10	-11	**7**	-7	-19	**11**	-2	-29	**14**	- [b]	- [b]	**- [b]**
B3LYP	QZVP	-9	-10	**8**	-11	-7	**6**	-11	-5	**5**	-10	-10	**7**
BLYP	QZVP	-16	-16	**13**	-24	-9	**10**	-24	-7	**9**	-16	-14	**12**
CCSD(T) [a]	TZVP	12	11	**6**	4	11	**7**	4	12	**7**	12	11	**7**
LCCSD(T) [a]	TZVP	12	10	**6**	- [b]	- [b]	**- [b]**	- [b]	- [b]	**- [b]**	11	9	**6**
SCS-CC2 [a]	TZVP	11	11	**7**	3	12	**7**	3	13	**8**	12	12	**7**
SCS-MP2 [a]	TZVP	15	14	**8**	7	14	**9**	7	14	**9**	15	14	**9**
CC2 [a]	TZVP	-1[c]	1[c]	**3**	-6	4	**4**	-6	6	**4**	0	3	**3**
MP2 [a]	TZVP	4[c]	5[c]	**4**	-1	7	**5**	-1	9	**5**	4	6	**4**
HF	TZVP	35	11	**13**	38	6	**13**	39	4	**13**	37	7	**14**
M06-2X	TZVP	-17	-17	**12**	-15	-17	**12**	-9	-17	**10**	- [b]	- [b]	**- [b]**
B3LYP	TZVP	-15	-16	**12**	-19	-12	**10**	-19	-10	**9**	-17	-17	**13**
BLYP [a]	TZVP	-21[c]	-20[c]	**17**	-30	-13	**13**	-31	-10	**12**	-22	-20	**17**
PM6	-	-33[c]	-26[c]	**17**	-40	-40	**29**	-40	-48	**34**	- [b]	- [b]	**- [b]**
PM3	-	45	34	**24**	13	2	**13**	-12	-12	**12**	38	21	**21**
RM1	-	-32[c]	-27[c]	**17**	-22	-47	**29**	-22	-59	**34**	- [b]	- [b]	**- [b]**
AM1	-	64	42	**28**	32	11	**16**	18	-2	**12**	57	29	**24**
MNDO/d	-	228	162	**119**	195	128	**106**	169	114	**99**	215	142	**112**
MNDO	-	229	186	**131**	197	154	**119**	173	141	**113**	224	174	**128**

[a] resolution of identity approximation or density fitting employed.

[b] method not available.

[c] no local minimum or transition state found. Value represents the error of the respective grid point.

Tab. 5-2: Errors for relative transition state energies ΔΔU(TS), relative minimum energies ΔΔU(ZW), and mean unsigned errors MUE for the whole set of 14 single point energies along the minimum energy path in kJ/mol. The local coupled cluster approach LCCSD(T) in conjunction with QZVP basis set is taken as reference for gas phase and QM/MM values. Errors for the COSMO approach are taken relative to SCS-CC2 | QZVP level of theory. A detailed description of the relative energies is given in Fig. 5-2.

Nevertheless, basis set effects remain significant and systematically as indicated by the comparison of TZVP and QZVP for each of the wave function based *ab initio* method. Thus applying a bigger basis set results in lowering the zwitterionic state by about 12-16 kJ/mol.

One exception from this exhibits the Hartree Fock method, where a bigger basis set leads to an insignificantly higher lying transition state energy and minimum in the range of 4-7 kJ/mol. This special case can also be found in the comprehensive quantum chemistry textbook of Helgaker et al. who calibrate electronic structure methods systematically.[191] Anyway, the use of "pure" HF method cannot be recommended as apparent from the calculated errors, which rise up to 46 kJ/mol for the predicted transition state energy ΔΔU(TS). As a common trend of all wave-function based methods, overestimation of the relative energy between **N** and **ZW** can be observed for the use of TZVP basis set.

The comparison of thermodynamic and kinetic data predicted by density functional theory based methods reveals the often appreciated cost/benefit ratio provided by DFT methods. Only a few exchange correlation functionals are selected, since extensive benchmarks are available elsewhere,[218,312,313] in particular with focus on proton affinities in amino acids.[298] The selection is done according to the sophistication level on the "Jacobs Ladder" of density functional theory[217] and includes three of the most commonly used functionals, like BLYP based on generalized gradient approximation (GGA), the popular hybrid-GGA functional B3LYP, and the more recent M06-2X functional as a meta-hybrid-GGA type. The use of empirical dispersion correction terms, like that from Grimme et al.,[231,314] is not necessary for this type of reaction, since proton transfers are primary of electrostatic nature and therefore marginal affected by errors due to dispersion. However, the impact of dispersion correction on the given case was checked by B3LYP-D3 test calculations in gas phase, delivering differences below 1 kJ/mol for the results with respect to the uncorrected B3LYP functional. Considering the robustness of DFT methods with respect to the environment model, it becomes apparent that they do not behave as consistent as the wave-function based methods, but still in an acceptable range. Whereas changes in the unsigned errors of B3LYP are below 6 kJ/mol, and therefore insignificant when the environment model is switched, BLYP and M06-2X can change up to 10 kJ/mol and 18 kJ/mol respectively. Especially the M06-2X meta-hybrid functional shows an apparent error bias in conjunction with QZVP when COSMO is employed, which seems to increase as a function of the dielectric constant. The order of absolute errors ΔΔU(ZW) is given by 11 kJ/mol < 19 kJ/mol < 29 kJ/mol for gas phase, COSMO with ε=4, and COSMO with ε=78, respectively. Interestingly this bias totally disappears when the smaller TZVP basis set is used, yielding a perfectly consistent error behavior. The analysis of errors of the relative energies ΔΔU(ZW) shows that results for ZW in case of M06-2X differ by 11-29 kJ/mol, B3LYP by 5-17 kJ/mol, and BLYP by 9-20 kJ/mol, depending on basis set size and environment model. The last one does not predict a minimum for the zwitterionic state in gas phase, when TZVP basis set is used. The deviations for the transition state energies ΔΔU(TS) tend to be in the same range and comprise unsigned errors of 2-17 kJ/mol, 9-16 kJ/mol, and 23-48 respective for M06-2X, B3LYP, and BLYP. The most apparent trend is the fact that all functionals tend to consistently underestimate barriers (2-31 kJ/mol) and relative energies (5-29 kJ/mol). This behavior has recently been discussed by Cohen et al.[229] and particularly by Sharma et al.[302] for proton transfer reactions in gas phase. As shown in Tab. 5-2, this typical drawback of DFT methods also holds for the herein applied environment models COSMO and QM/MM. The effect of the basis set size is generally lower for the DFT methods than for the wave-function based ab initio methods and ranges typically at the limit of significance (2-7 kJ/mol), when triple and quadruple zeta quality basis sets are compared for each functional. Due to their consistent trends to overestimate and underestimate the energy differences when triple zeta basis sets are used, a combination of wave-function based method and DFT method should provide a good estimate of the error bar that can be expected for the proton transfer reaction. By taking SCS-MP2 and B3LYP as an example (Tab. 5-2), the expected error moves within a range of 30 kJ/mol for gas phase, 26-27 kJ/mol for COSMO, and 31-32 kJ/mol for QM/MM when the respective results of the two methods are taken as the upper and lower limit for ΔΔU(TS) and ΔΔU(ZW).

In contrast to the more or less consistent trends of *ab initio* approaches, semi-empirical methods show a very erratic behavior for the accuracy of evaluated thermodynamic and kinetic parameters. The comparison of the proton transfer potentials in gas phase predict only in the case of the "classical" semi-empiric hamiltonians MNDO(/d), AM1, and PM3 a transition state and a zwitterionic minimum along the reaction path. The more recent PM6 and further the reparametrized version of the Austin model 1, RM1, deliver a continuous repulsive energy curve along the minimum path without a local minimum for ZW. Considering robustness with respect to the environment model, PM6 and RM1 seem to be slightly superior, since maximum changes in energy are observed in the range of 7-32 kJ/mol, whereas all older models comprise maximum changes by 44-59 kJ/mol when gas phase, COSMO, and QM/MM are compared for the same method. Unsigned errors for the relative energies ΔΔU(ZW) and ΔΔU(TS) predicted by PM6, PM3, AM1, and RM1 are settled between 2 and 64 kJ/mol and further exceed 112 kJ/mol for the two MNDO variants. Except for this difference, no clear accuracy order can be derived for the semi-empirical methods, if all environment models are taken into account. Considering MNDO and its extended implementation MNDO/d, including d-functions, **ZW** is overestimated by more than 112 kJ/mol, and the barrier heights ΔΔU(TS) are beyond 169 kJ/mol, which make it practical unusable for the proton transfer reaction. This underlines the deficiencies of these methods with respect to hydrogen bond description. The more often employed AM1 and PM3 hamiltonians range in a more acceptable region for ΔΔU(ZW) between -12 and 42 kJ/mol and a slightly broader range for ΔΔU(TS) between underestimation of 12 kJ/mol and overestimation of 64 kJ/mol. With respect to the environment model, they provide surprisingly good results in conjunction with the COSMO model, where the unsigned errors for ΔΔU(ZW) are below 12 kJ/mol. Most interestingly are the results obtained with the recent semi-empirical methods PM6 and RM1. Although no minima for gas phase are found and QM/MM results are not available, the results obtained from the two COSMO cases are rather disappointing. Particularly, when a high dielectric constant of ε=78 is applied, RM1 and PM6 are the only methods throughout this benchmarking that predict the energy of **ZW** below that of **N**. In detail, the error of the relative energy ΔΔU(ZW) reaches -48 kJ/mol for PM6, and -59 kJ/mol for RM1, featuring a massive overstabilization of the zwitterion. In the case of PM6, this problem has also been mentioned by Stewart et al.[180] in terms of the applicability of their semi-empirical hamiltonian to protein modeling. The overstabilization seems further to be a general trend for the Recife Model RM1, since findings from Lee et al.[283] indicate similar problems, when zwitterions are involved.

The mean absolute errors (MUE, defined in equation 42), given in Tab. 5-2, are well suited to get an overall estimate of the accuracies for the whole proton transfer potential and further compare the different environment models. Considering the wave-function based methods HF, CC2, MP2, SCS-MP2, SCS-CC2, CCSD(T), and LCCSD(T), the respective MUEs obtained for the different environment models do not change much (~1-2 kJ/mol). This reflects the robust behavior of the wave-function based methods, already found for the thermodynamic properties. The basis set size (TZVP vs. QZVP) in average changes the computed potential systematically by 1-8 kJ/mol, whereas the spin-component scaled counterparts are slightly more affected (5-8 kJ/mol) than the other wave-function based methods (1-6 kJ/mol). Therefore, SCS-CC2 and SCS-MP2 seem to benefit most from a bigger basis set.

The density functional theories show also a robust behavior towards the different environment models in most cases, and do not exceed 7 kJ/mol difference in the MUEs by going from gas phase over COSMO to the QM/MM approach. Differences between the accuracy of the three functionals are too small to be significant, but nevertheless there is a consistent improvement in accuracy of 1-6 kJ/mol by going from triple zeta basis set to quadruple zeta quality.

The MUE values for the spectrum of semi-empirical methods emphasize the error behavior, found for the thermodynamic and kinetic data. By comparing the trends, it becomes apparent that the older semi-empirical methods MNDO(/d), AM1, PM3 show increased accuracy by going from gas phase towards the COSMO model whereas the most recent hamiltonians PM6 and RM1 decent from their superior performance in gas phase. In gas phase, the latter ones predict the proton transfer potential in a quality (MUE =17 kJ/mol) comparable to the density functional BLYP. However, if the environment is modeled by the COSMO approach (ε=78), the mean unsigned error rises up to 34 kJ/mol. MUEs of the older semi-empirical ones decrease to appealing values of 12 kJ/mol respective for PM3 and AM1. Taking also into account that no zwitterionic minimum is predicted by PM6 and RM1, both approaches are unsuitable for calculating the proton transfer potential for this study. The most robust behavior with respect to the environment model provides PM3, where the maximum difference of MUEs is 12 kJ/mol, followed by AM1 with 16 kJ/mol and 17 kJ/mol for RM1 and PM6. The detailed results of this comparison might also be applied in terms of error scaling, to reduce the systematic errors by going from gas phase models to COSMO or QM/MM approach. The direct comparison of AM1 and RM1 gives further interesting insights to the bias of the semi-empirical methods. Since RM1 represents only a reparametrization of AM1 and therefore does not differ in its functional form or implementation, it shows to which extend the choice of parameters influences the robustness of the semi-empirical method with respect to the environment model. For the given reaction, best performance in gas phase is achieved by RM1 with MUE of 17 kJ/mol and worse values are obtained by AM1 (28 kJ/mol). In conjunction with COSMO, mimicking a very polar environment of ε=78, the AM1 method can even exceed these accuracies with a mean unsigned error of 12 kJ/mol in contrast to a significant dropped performance of 34 kJ/mol for RM1. This change of trend represents a new aspect in terms of method accuracy. It underlines previous conclusions,[296] that despite of the recent developments of RM1 and PM6, semi-empirical methods are still unreliable for proton transfer predictions. It further indicates that the progress of achieving higher accuracy, like recently demonstrated by RM1 or PM6, seems to be bought on cost of robustness, thus introducing a stronger bias towards gas phase data, which are typically used for parameter optimizations.

The proton transfer potential between cysteine and histidine reveals the neutral charged state **N**, the zwitterionic state **ZW** and the transition state **TS**. The r(S-H) distance shown in Fig. 5-1 serves as reaction coordinate for the conducted proton transfer reaction. It represents a covalent sulfur-hydrogen bond for low values and a non-covalent hydrogen bond between the anionic sulfur atom and the oxygen-bound hydrogen atom for larger distances. Tab. 5-3 gives the offsets from the locations, when other methods are compared, which were computed as described in the setup section and illustrated in Fig. 5-3. Further, the reference values (SCS-CC2 | QZVP) are given in Tab. 5-3.

		gasphase			COSMO						QM/MM		
level of theory					ε=4			ε=78					
		Δr(N)	Δr(TS)	Δr(ZW)	Δr(N)	Δr(TS)	Δr(ZW)	Δr(N)	Δr(TS)	Δr(ZW)	Δr(N)	Δr(TS)	Δr(ZW)
LCCSD(T) [a]	QZVP	0	0	-7	[b]	[b]	[b]	[b]	[b]	[b]	0	0	2
SCS-CC2 [a]	**QZVP**	**135**	**184**	**195**	**135**	**171**	**210**	**135**	**169**	**216**	**134**	**182**	**207**
SCS-MP2 [a]	QZVP	0	0	2	0	0	1	0	0	1	0	0	2
CC2 [a]	QZVP	0	[c]	[c]	0	-4	-8	0	-1	-7	0	0	-7
MP2 [a]	QZVP	0	[c]	[c]	0	-1	-6	0	-1	-7	0	0	-5
HF	QZVP	-2	-4	37	-2	1	21	-2	1	16	-2	-3	32
M06-2X	QZVP	1	0	-1	1	-2	1	1	-2	18	[b]	[b]	[b]
B3LYP	QZVP	1	0	1	1	0	-2	1	0	0	1	1	2
BLYP	QZVP	3	1	-6	3	-1	-6	3	1	-6	3	2	-2
CCSD(T) [a]	TZVP	1	0	0	1	2	-2	1	1	-4	31	1	2
LCCSD(T) [a]	TZVP	1	0	3	[b]	[b]	[b]	[b]	[b]	[b]	1	1	5
SCS-CC2 [a]	TZVP	1	1	-3	1	2	-3	1	1	-5	1	2	-1
SCS-MP2 [a]	TZVP	0	0	-1	0	2	-2	0	1	-4	1	1	1
CC2 [a]	TZVP	1	[c]	[c]	1	1	-9	1	0	-10	1	2	-8
MP2 [a]	TZVP	1	[c]	[c]	1	1	-8	1	0	-8	1	2	-6
HF	TZVP	-1	-4	35	-1	1	20	-1	1	15	-1	-3	32
M06-2X	TZVP	2	0	-4	2	-2	0	1	-2	12	[b]	[b]	[b]
B3LYP	TZVP	3	0	1	2	1	-2	2	0	0	2	0	2
BLYP [a]	TZVP	-5	[c]	[c]	4	3	-7	4	1	-6	4	2	-2
PM6	-	3	[c]	[c]	3	-6	-14	4	-6	3	[b]	[b]	[b]
PM3	-	-2	0	26	-1	11	26	-1	11	22	-3	2	25
RM1	-	-1	[c]	[c]	-1	-8	25	-1	-6	47	[b]	[b]	[b]
AM1	-	-2	-1	33	-1	8	29	-1	4	19	-1	1	30
MNDO/d	-	-2	0	76	-2	11	50	-2	12	40	-2	2	64
MNDO	-	-6	0	64	-5	11	41	-5	11	32	-5	2	56

[a] resolution of identity approximation or density fitting employed.

[b] method not available.

[c] no minima or maxima found.

Tab. 5-3: Deviations Δr of minima **N**, transition states **TS**, and zwitterionic minima **ZW** along the reaction coordinate r(S-H) in picometers (pm). Extremal values have been obtained from quadratic fits around the incremental points closest to minima or transition state. Absolute values of minima/maxima obtained from the fitted potential curve on the SCS-CC2 | QZVP level of theory are given as bold numbers and are taken as reference. A detailed description of the relative values Δr is given in Fig. 5-3.

The analysis of the reference values in gas phase and COSMO reveal the tendency to predict a zwitterionic state **ZW** that is shifted further away for increased polarity of the environment, whereas the predicted **TS** moves slightly closer to **N**. The QM/MM approach delivers **TS** at 182 pm, which is similar to the gas phase description, and **ZW** at 207 pm, comparable to the COSMO prediction for a low dielectric constant of ε=4. The lowest state **N** of the proton transfer potential is well described by nearly all methods. Only BLYP and MNDO without d-functions overestimate or underestimate the location of the minimum by up to 6 picometers. The rest of the methods show absolute deviations below 4 picometers, whereas the DFT approaches possess a consistent tendency of marginal overestimations. By taking a look towards **TS** and **ZW** in the gas phase, it becomes obvious that only half of the tested methods predict a transition state, and thus a local minimum for the zwitterionic state on the proton transfer potential. However it should be noted that this case of a proton transfer reaction is challenging, since the shape of the potential is very flat in that region. The comparison of location of the TS among the various methods reveals differences between the accuracy of *ab initio* methods, that in general do not deviate by more than 4 pm from the reference, and the semi-empirical hamiltonians. For the latter ones, an interesting observation with respect to the error behavior within gas phase, COSMO, and QM/MM can be made. Closer inspection of the semi-empirical methods shows that the recent methods RM1 and PM6 behave differently than the older hamiltonians. Although the absolute deviations of the transition state locations range in the same magnitude, namely up to 12 picometers beyond the reference, PM6 and RM1 both tend to underestimate r(S-H) of **TS**, whereas the "classical" methods PM3, AM1, and MNDO(/d) consistently overestimate it. The location of the zwitterionic state ZW shows in general larger deviations than both of the other stationary points. Beside the shortcomings of the different methods, this is also due to the fact that the potential is much flatter in this region (Fig. 5-4). Comparing gas phase, COSMO, and QM/MM, no systematic behavior can be observed. Nevertheless, the relative comparison of all hamiltonians reveals some trends that mainly confirm the observations made for the MUEs and the thermodynamic data. Accordingly, HF shows large deviations from the reference values, but also the unscaled versions of MP2 and CC2 can deviate by more than 10 pm. The DFT methods deliver quite reasonable predictions, in particular B3LYP, what is in line with previous findings from Adamo and coworkers.[315] They comprise errors below 7 pm with the exception of the M06-2X functional. The lineup of the gas phase offset of -1 pm and the COSMO (ε=78) offset of

18 pm results a maximum difference of 19 picometers for the zwitterionic state along the same proton transfer potential when a QZVP basis set is used. Compared to the B3LYP functional that differs by less than 2 picometers from the reference value across all environment models, M06-2X seems to be even less robust than BLYP (changes < 5 pm) which ranks two rungs below on the "Jacobs ladder" of density functional theory. In contrast to the observations made for the thermodynamic properties, where this "COSMO problem" was only observed for the QZVP basis set, this deficiency does not vanish here for the smaller basis set. For TZVP, the error range still spans across 16 pm which is apparently more than the deviations of the other functionals.

Largest deviations and mostly unsystematic error behavior are observed for the semi-empirical methods. Beside the expected bad performance of MNDO(/d) for the location of **ZW**, the recently developed RM1 method delivers a rather disappointing result with a maximum offset of 47 above the reference in case of the COSMO description with ε=78. In contrast to this, PM6 makes a surprisingly good prediction that deviates only 3 pm from the reference. Unfortunately, this encouraging result is accompanied by considerable errors in the relative energy of **ZW**, as already discussed. At the end, PM3 and AM1 hamiltonians remain with deviations of 22-24 pm and 19-33 pm respectively, depending on the environment model.

In summary, only the B3LYP functional provides excellent predictions of all stationary points, by keeping sufficiently consistent throughout the different environment models and the two basis sets. This finding is also supported by reports from Clyaessens et al. who compared QM/MM transition state geometries predicted by B3LYP with LCCSD(T) results.[316] Furthermore, the wave-function based approach SCS-MP2 behaves similar robust.

5.1.3 Conclusions about the Accuracy of the Applied Methods

In summary, some critical aspects for the description of the cysteine-histidine proton transfer have been evaluated that might also hold for other proton transfer reactions in protein environment. The results show the impact of implicit and explicit environment modeling on the proton transfer potential and demonstrates that the application of a continuum solvent model with low epsilon value of 4, corresponding to protein bulk, still tends to overestimate stabilization of ionic state and transition state. Therefore its use for the modeling of biochemical proton transfer reactions seems to be limited. By evaluating the robustness of the probed quantum chemical methods with respect to the environment model, a quite good correlation with the applied level of theory can be observed. Hence, the wave function based *ab initio* methods show consistent error behavior and do not change their accuracy significantly by switching between gas phase, COSMO, and QM/MM approach. For the TZVP basis set, a consistent trend of overestimation (12-16 kJ/mol) of **ZW** can be observed for the post-HF methods. As expected, the CC2 approach, which is usually applied for excited states, delivers similar results as MP2 for our ground state problem.

The density functional theory methods show marginally higher deviations for thermodynamic and kinetic predictions and MUEs, depending on the functional. The recent M06-2X functional reveals weaknesses in combination with the COSMO model, which seems to introduce larger errors (up to 29 kJ/mol) for the relative energies and location of **ZW**, when higher dielectric constants are applied. Therefore it can be assumed that its superior accuracy in gas phase is not necessarily represented within the COSMO description and possibly in QM/MM approaches. By taking performance of thermodynamic data, MUEs, and location of stationary points into account, B3LYP appeals to be the most robust functional with respect to the change of the environment model and can, in agreement with other studies,[295] deliver proton transfer predictions comparable to wave-function-based approaches. As a general trend, consistent underestimation of **TS** and **ZW** (2-31 kJ/mol) is found for all DFT methods.

The semi-empirical methods PM6, PM3, RM1, AM1, and MNDO(/d), which are often used for the computation of free energies, show a very erratic and unsystematic error behavior throughout all evaluated properties. Although the recent methods PM6 and RM1 provide significant advances in gas phase accuracy, in line with assessments in other studies,[180,209–211,317] their application within the COSMO model reveals a dramatic change in performance. For the given case they wrongly predict the stability order of **N** and **ZW**. Whereas this "zwitterion problem" is already described for PM6,[180] it seems also to hold for RM1. Beside this deficiency, none of the semi-empirical methods delivers a satisfactory robustness, since they possess changes up to 59 kJ/mol for the thermodynamic data by going from one environment model to another. Taken these facts together, least problems can only be expected for PM3 and AM1. However, PM3 has the drawback of absolute errors in the maximum range of 45 kJ/mol for the thermodynamic and kinetic predictions. AM1 also behaves similar and comprises a maximum error of 64 kJ/mol for the investigated proton transfer.

In summary, all methods possess strength and weaknesses for this challenging case. Therefore it appears reasonable to apply a combination of two methods to obtain a trustworthy range for the prediction of proton transfer potentials. Due to their consistent trends to overestimate and underestimate when triple zeta basis sets are used, a combination of wave-function based method and DFT method should deliver a good estimate of the error bar that can be expected for the proton transfer reaction, by keeping the computational costs feasible.

5.2 Insights into the Character of SARS-CoV M^{pro}'s Active-Site

The active-site of SARS-CoV M^{pro} shows some distinct features compared to the archetypical cysteine proteases. Conducted pK_a measurements led to the conclusion that the catalytic residues Cys145 and His41 possess an uncharged resting state[71] instead of the expected thiolate/imidazolium ion pair, which is a characteristic feature of papain[94,96,97,102,301,318–320] or cathepsin.[279] A further interesting observation from the experimental side concerns the substrate turnover rates of SARS-CoV M^{pro} for esters and amides. The proteolytic processing seems not to correlate with the chemical reactivity of the two substrates that differ by a factor of 2000 in aqueous media in absence of the enzyme.[58] These findings rise the expectation that also inhibition of SARS-CoV M^{pro} could proceed in a different way, when compared to the typical cysteine proteases. The missing breakthrough with respect to inhibition constants that never significantly have entered the nanomolar range so far, as shown in Tab. 2-2, is an indicator for this assumption. Due to this, the catalytic mechanism of SARS-CoV M^{pro} deserves a detailed investigation on the atomistic level.

A first step towards a deeper understanding is the set up of theoretical model systems, which can reproduce the experimental findings, namely the preference of SARS-CoV M^{pro} to comprise a neutral cysteine/histidine catalytic dyad rather than forming a thiolate/imidazolium ion pair. In the next step, these model systems can be employed to identify the main factors that are responsible for the distinct behaviour of SARS-CoV M^{pro} active-site.

The objective of this chapter is therefore to reach these goals by different theoretical approaches and to gain deeper insight into the character of the catalytic dyad. The starting point for the investigations represent the experimentally obtained X-ray structures of SARS-CoV M^{pro} that represent a profound and reasonable basis for the conducted computational studies. Since X-ray structures are not only available for the free SARS-CoV M^{pro}, but also for inhibitor complexes, and further for substrate complexes, possible effects that might arise from substrate or inhibitor binding can be evaluated in a consistent way. The theoretical investigations are organized in several steps and proceed as follows.

Within the first sub chapter, molecular dynamic simulations are conducted and compared to the recent available X-ray structures. This leads to the identification of the most likely protonation state of the active-site residues. In the second sub chapter, the identified protonation state is checked for possible proton transfer reactions within the active-site, in order to evaluate possible charge-relay mechanisms. Since there are obviously bond breakages involved in the investigations, QM/MM calculations are employed for this purpose. The third sub chapter gives insights to the different effects that promote or inhibit the ion pair formation. This is achieved by a detailed analysis of the electrostatic contributions around the active-site sphere in terms of a charge deletion analysis (CDA). In the last sub chapter, free energy calculations, based on QM/MM/MD simulations, complete the theoretical investigations and provide thermodynamic and kinetic information about the proton transfer reaction between cysteine and histidine that include entropic contributions. The combination of the results from the different theoretical perspectives will finally deliver a comprehensive picture of the catalytic dyad that forms a basis for further investigations on inhibition reactions occurring at the active-site of SARS-CoV M^{pro}.

5.2.1 Comparative Molecular Dynamic Simulations

A topological analysis of the active-site reveals an interesting network of amino acids that exhibit different possibilities for the individual protonation states of the titratable residues. The schematic map of the active-site of SARS-CoV M^{pro} and its neighboring residues (Fig. 2-9) offers 4 titratable residues that could hypothetically exist in plenty of different protonation state combinations. The considered residues compass Cys145, His41, His164, and Asp187. The permutation of all possibilities would formally lead to 2^4 and therefore 16 possible combinations. Asp187 forms a salt bridge with Arg40, therefore it is likely to assume that this residue has an unprotonated carboxylate group, since a protonation of the COO^- moiety of Asp187 would lead to the disruption of the salt bridge and therefore needs to overcome a stabilizing energetic contribution in the magnitude of 15 kJ/mol.[321] For the two histidine residues His41 and His164, the chemical intuition furthermore forbids to place two neighboring positive charges residues next to each other. Finally, there remain 6 hypothetical combinations of protonation states for the residues Cys145, His41, and His164 that are depicted in Fig. 5-7. Although some combinations appear more likely than others, it is finally not clear which one is the predominant situation.

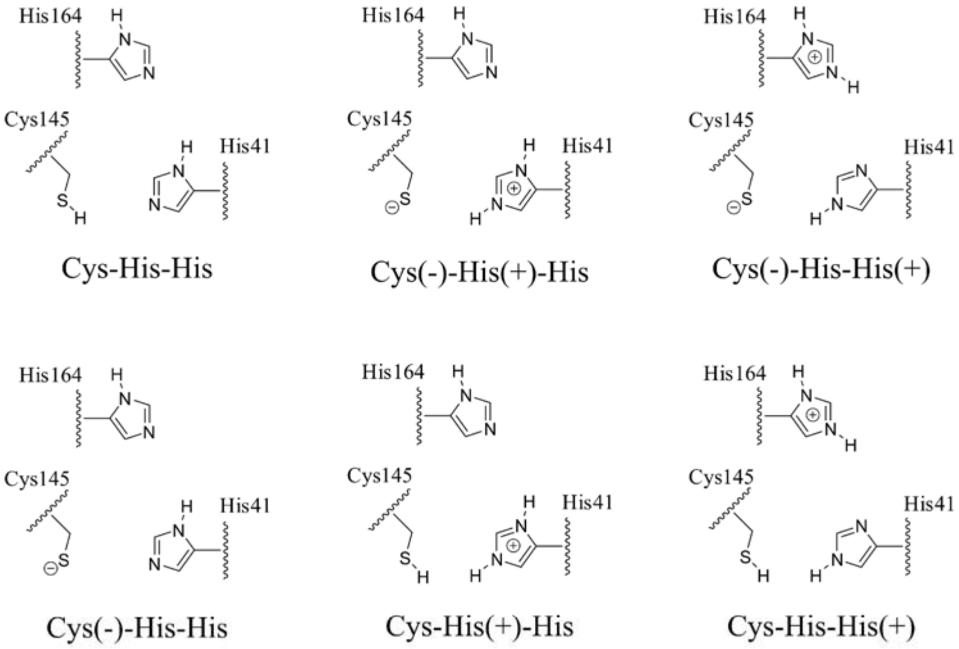


Fig. 5-6: Six possible protonation states of the active-site residues Cys145, His41, and His164. Water molecules are neglected in this scheme for the sake of clarity. A complete scheme with the topology of the active-site of SARS-CoV M^{pro} is given in Fig. 2-9.

The consideration of available experimental X-ray structures unfortunately does not give direct information about the protonation states, since only heavy atoms are detectable within the routinely applied procedures of X-ray crystallography. Nevertheless, this information is partly hidden in the structural arrangement of the heavy atoms, since the atom positions are occasionally determined by the underlying hydrogen bond networks. Hence, the structural comparison of different theoretical protonation state models, such as obtained by molecular dynamic simulations, with the experimentally determined X-ray structure can give valuable insights, which one of the protonation states is likely to exist and which is not. The decisive criteria here is the accordance between simulated and experimental structure that can be ranked in terms of root mean square deviation values (RMSD) or important geometric parameters.

To proceed such a comparison, the six possible protonation states were simulated with identical work flows for 10 ns, as described in the computational details section, and analyzed by their structural properties. The quality of the MD simulations were assessed by their root mean square deviations (RMSD) with respect to the first frame of the simulation. Detailed plots of RMSD values as a function of simulation time are shown in Fig. 7-6 that can be found in the computational details section. All simulations remain stable after a period of about 1 ns and possess average RMSD values between 1.7 and 2.1 Å over the whole simulation time of 10 ns. Although there are small differences between the individual simulations, these are not significant enough to draw conclusions about them.

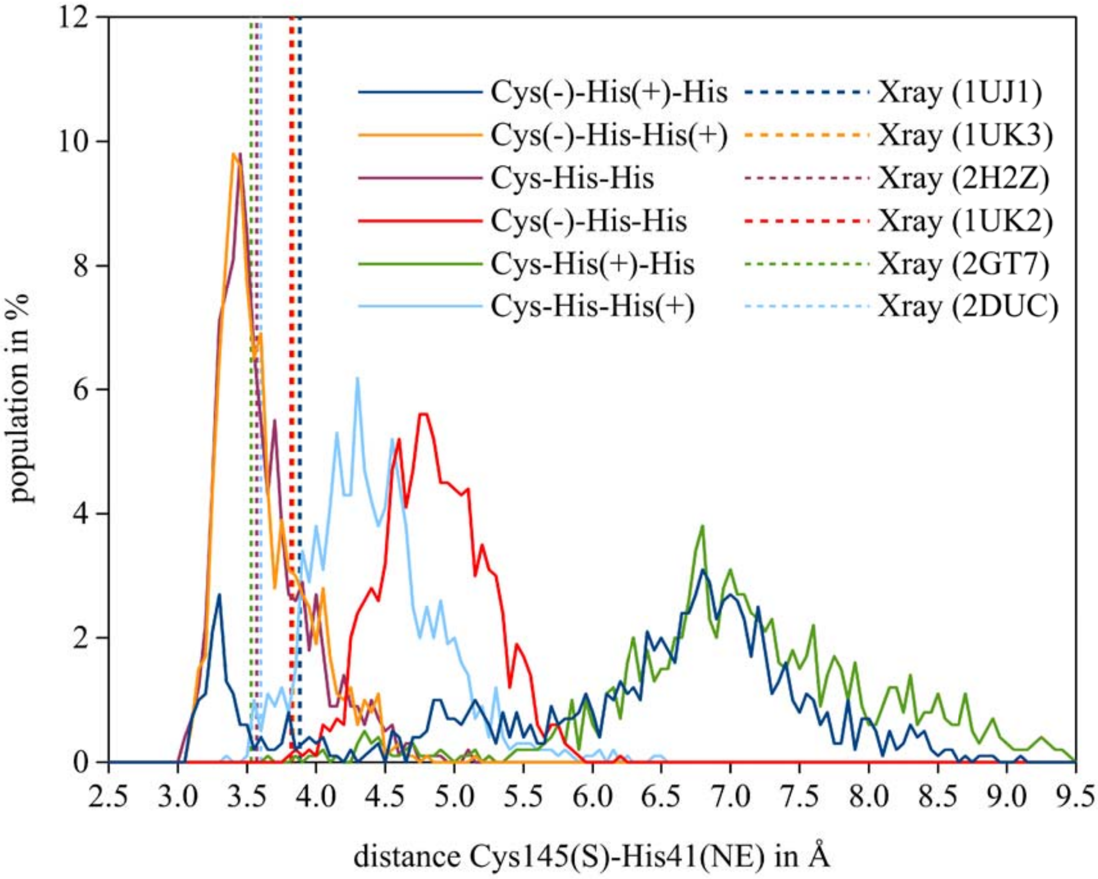


Fig. 5-7: Histogram analysis of the Cys145-His41 distances that are obtained from the six MD simulations (solid lines), referring to the protonation states that are defined in Fig. 5-6. Populations of observed distance values are obtained through statistical binning with a resolution of 0.05 Å and are taken relative in percentages to the overall number of measured points (1000). Experimentally determined values from the six X-ray structures 1UJ1,[46] 1UK3,[46] 2H2Z,[52] 1UK2,[46] 2GT7,[51] 2DUC[53] are indicated by vertical dashed lines.

A more suitable criterion to compare, provides the distance between the catalytic residues Cys145 and His41, which has also been addressed in earlier MD studies of SARS-CoV M^{pro}. [67,322] Fig. 5-7 gives a detailed comparison of the six simulations in terms of a histogram analysis for the observed Cys145-His41 distances that have been recorded throughout the individual trajectories. The histograms reveal which distance values are most likely observed during the simulation and how they are statistically distributed and spread along the distance axis. The distribution profiles are furthermore directly compared to the experimentally available Cys145-His41 distances from six published X-ray structures. The comparison of the individual histograms shows very broad distance distributions for the **Cys(-)-His(+)-His** and **Cys-His(+)-His** protonation states, slightly sharper distributions for the **Cys(-)-His-His** and **Cys-His-His(+)** configurations, and apparent sharp peaks for the distributions of the two protonation states **Cys-His-His** and **Cys(-)-His-His(+)**. Beside these narrow distance distributions, the latter ones also reveal the best agreement with the experimental values from X-ray crystallography. Conclusively, **Cys-His-His** and **Cys(-)-His-His(+)** appear to be the most likely proton configurations for the considered active-site residues. Nevertheless, the histogram data give no clue about, which one of the two configurations is the predominant case, since the distance profiles are very similar.

	Cys145(S)-His41(NE)			His41(ND)-H2O(O)			His164(NE)-H2O(O)			Asp187(OA)-H2O(O)		
	d	Δd		d	Δd		d	Δd		d	Δd	
X-rays*	3.71	± 0.15		2.83	± 0.24		3.06	± 0.10		2.88	± 0.16	
Cys-His-His	3.59	-0.12	(+)	2.90	0.07	(+)	2.96	-0.10	(+)	2.75	-0.13	(+)
Cys(-)-His(+)-His	6.11	2.40	(-)	5.09	2.26	(-)	3.02	-0.04	(+)	2.80	-0.08	(+)
Cys(-)-His-His(+)	3.59	-0.12	(+)	4.84	2.00	(-)	3.77	0.71	(-)	5.57	2.69	(-)
Cys(-)His-His	4.83	1.13	(-)	2.93	0.10	(+)	2.94	-0.12	(-)	2.79	-0.09	(+)
Cys-His(+)-His	7.10	3.39	(-)	5.07	2.23	(-)	3.05	-0.01	(+)	3.77	0.89	(-)
Cys-His-His(+)	4.41	0.71	(-)	6.57	3.73	(-)	3.21	0.15	(-)	2.91	0.03	(+)

* Mean values and standard deviations (±) from the set of six X-Ray structures 1UJ1,[46] 1UK3,[46] 2H2Z,[52] 1UK2,[46] 2GT7,[51] 2DUC.[53]

Tab. 5-4: Comparison of MD simulation results with available X-ray structures. Compared are the mean values of selected geometry parameters of the active-site in Å. Distance values d are obtained from molecular dynamic simulations and are calculated as mean values over the whole simulation trajectory of 10 ns. The deviation Δd indicates, how this value deviates from the experimental value (mean of six X-ray measurements) that is given in the first row. The signs (+) and (-) indicate, whether the deviation of the simulated mean value Δd ranges within the standard deviation (±) of the X-ray reference value, or exceeds it.

To gain some more differentiated information about this, the comparison between the MD simulation results and the X-ray structures was extended by a set of three additional geometry parameters that cover the structural constitution of the active-site over a more widened area. For this purpose, the hydrogen bond network between Cys145, His41, His164, Asp187, and the buried water molecule H_2O@His41 was considered more closely. Averaged values for the four important geometry parameters are summarized in Tab. 5-4. To get an estimate about the significance of the structural agreements or disagreements with the experimental values, average values and standard deviations were calculated for the set of six measurements from the X-ray structures. The averaged distance values from MD simulations were then compared and checked against the experimental standard deviations to decide whether they significantly agree (+) with the X-ray structures or disagree with them (-). Considering all geometry parameters, there is only the **Cys-His-His** state that agrees well with every single value from the set of the four distances.

Hence, the catalytic dyad is most likely to reside in the neutral and uncharged resting state **Cys-His-His**, what is qualitatively in direct agreement with the experimental pK_a values. Furthermore, the MD simulations reveal that His164 is also found to be uncharged, since a protonation of this residue leads to a strong disruption of the rigid hydrogen network that embeds the buried water molecule H_2O@His41. A further interesting finding states the observed Cys145-His41 distance that is determined for 3.6-3.7 Å. This is obviously too large for a direct hydrogen bond between the donating cysteine thiol group and the accepting histidine nitrogen atom. This result therefore indicates that the catalytic dyad possesses a bridging water molecule between the two entities, which is also confirmed by the visual inspection of the simulation trajectory.

5.2.2 Proton Transfer Potentials

Finding the catalytic dyad in an uncharged thiol/imidazol state has led to the conclusion that SARS-CoV M^{pro} follows a general base catalyzed mechanism.[71] This usually includes deprotonation of the thiol group, assisted by the neighboring histidine residue, and subsequent attack of the formed thiolate at the electrophilic substrate's amide moiety or alternatively at an inhibitor warhead. Theoretical studies have shown that this is only possible in a two step mechanism, consisting of deprotonation, and nucleophilic attack, rather than proceeding in a concerted manner.[275] Therefore the kinetic and thermodynamic characteristic of the initial deprotonation step is of general interest, not only for the substrate cleavage mechanism, but also for the inhibition mechanism.

Since the considered proton transfer states the breakage of a bond, and formation of a new bond, molecular mechanics cannot be used here and moreover QM/MM methods have to be employed. Although this hybrid approach can still today not be considered to be a routinely applied procedure, it has emerged throughout the last decades to the state-of-the-art method for treating biomolecular problems of this kind on the atomistic level.[183,270,271]

For the given case, a QM/MM model system was set up to calculate minimum energy paths of possible proton transfer reactions within the active-site of SARS-CoV M^{pro}. The topology of the active-site (Fig. 2-9) shows that beside the obvious proton transfer between Cys145 and His41, there is also the possibility of a second proton transfer between His41 and His164, that would be mediated by the buried water molecule H_2O@His41. In order to evaluate such a proton relay mechanism, the QM part of the system is set up from the side chains of Cys145, His41, His164, and the two bridging water molecules, as shown in Fig. 5-8 (right hand side, top). The MM part includes an aqueous solvation sphere and both protomers of the dimeric SARS-CoV M^{pro} structure. Especially, the latter one is important, since only the dimeric structure possesses an active conformation that is capable to catalyze.[55–57,67] The employed QM/MM set up is illustrated in Fig. 5-9 (a). In order to “drive” the desired proton transfer reaction and calculate the respective potential energy surface, suitable reaction coordinates have to be determined. The high number of involved bond length changes, that have to be considered, indicates that this is no trivial task for the possible charge-relay mechanism which is furthermore mediated by two water molecules. The chain of directly involved bond length changes stretches out over eight bonds (four chemical bonds, and four hydrogen bonds) for the given case, as depicted in Fig. 5-8 (right hand side, top). Nevertheless, most of the bond length changes can be expected to be correlated with each other. To evaluate this, the two transitions between the **Cys-His-His** to **Cys(-)-His(+)-His** states (Fig. 5-8, left-hand side, bottom) and the **Cys(-)-His(+)-His** to **Cys(-)-His-His(+)** state (Fig. 5-8, right-hand side, bottom) have been calculated separately in individual potential energy surface (PES) calculations. They comprise two reaction coordinates respectively, one for the deprotonation of the amino acid moiety, and one for the deprotonation of the bridging water molecule.

The shape of the two PES (Fig. 5-8, bottom) shows that both minimum energy paths proceed in a concerted way, without the occurrence of (meta)stable H_3O^+ intermediates. Therefore the

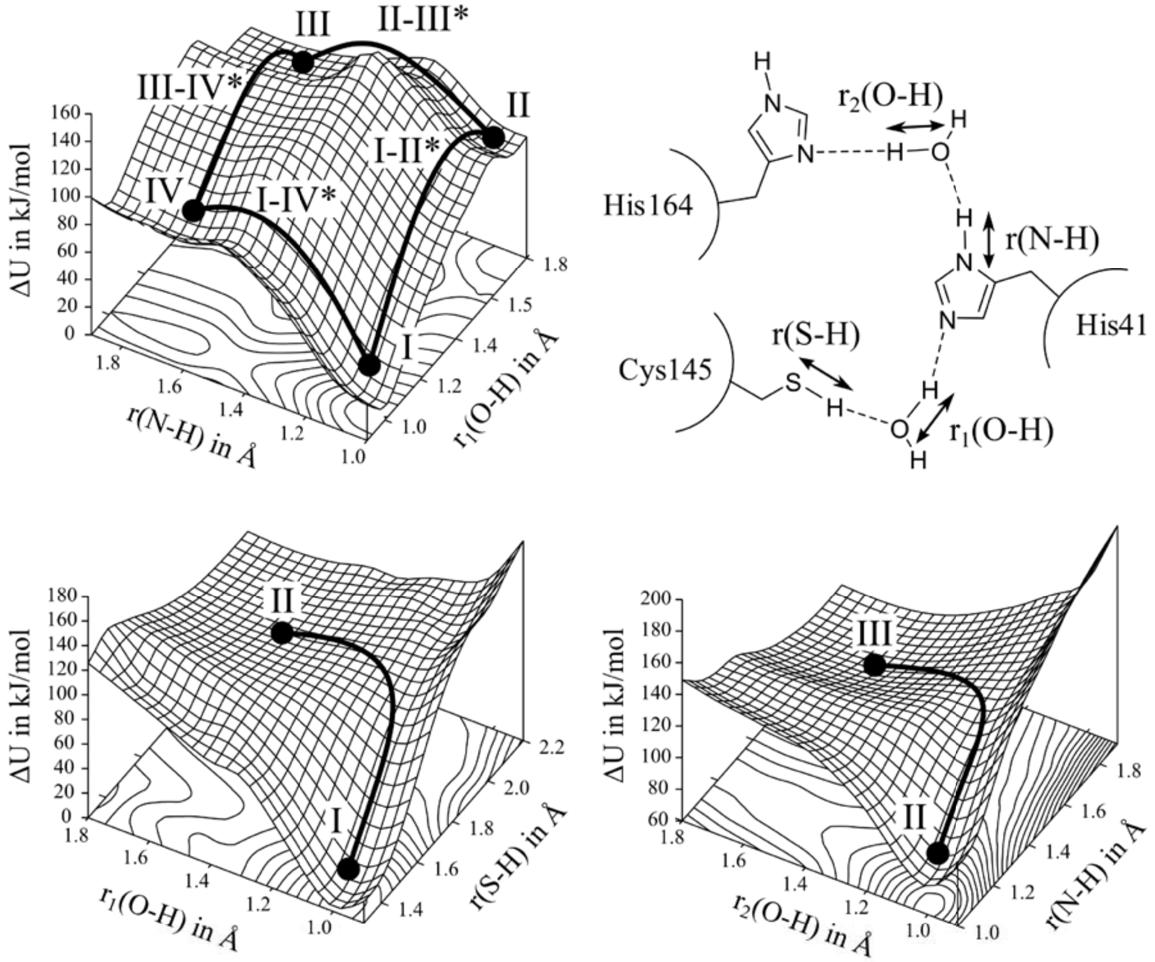


Fig. 5-8: Calculated QM/MM potential energy surfaces of the different protonation states of SARS-CoV M^{pro} active-site residues Cys145, His41, and His164. Proton transfers are mediated by two bridging water molecules. The potential energy surfaces refer to the free enzyme QM/MM system (a) in Fig. 5-9 that possesses no inhibitor or substrate bound to the active-site. The employed QM subsystem and reaction coordinates, that were used for "driving" the proton transfer reactions, are given on the right-hand top side. The reaction coordinates r_1(O-H) and r(S-H) (left-hand side, bottom) and further r_2(O-H) and r(N-H) (right-hand side, bottom) are smoothly coupled to each other, therefore it is sufficient to use only one of these coordinates for the proton transfers between Cys145/His41 and His41/His164. The potential energy surface on the left-hand top side combines the Cys145/His41 proton transfer by using the r_1(O-H) coordinate on the first axis and the His41/His164 proton transfer by using the r(N-H) coordinate on the second axis, hence giving relative energies of the four different protonation states **I-IV** and their respective transition states. All energies are given in kJ/mol.

proton transfer reactions for the **Cys-His-His** to **Cys(-)-His(+)-His** transition and the **Cys(-)-His(+)-His** to **Cys(-)-His-His(+)** state can be described by using only one coordinate for each transition, since all other bond length follow smoothly according to this coordinate and are therefore directly correlated to each other.

For the description of the whole charge-relay mechanism (Fig. 5-8, left-hand side, top) the r_1(O-H) and the r(N-H) coordinates have been selected. From the PES, four local minimum states **I-IV** can be found, which are connected to each other by definite transition state passes. The relative energies ΔU of the located minima and maxima with respect to the state of lowest energy **Cys-His-His** (**I**) are summarized in Tab. 5-5. The thermodynamic equilibrium between the **Cys(-)-His(+)-His** zwitterion **II** and the relayed **Cys(-)-His-His(+)** zwitterion **III** is nearly balanced with relative energies ΔU of 87 kJ/mol and 90 kJ/mol. However, the kinetic parameters for the transition between the two zwitterionic states differ, since the minimum energy pass **I-II*** is located at 101 kJ/mol and **II-III*** at 117 kJ/mol, and makes the formation of the unrelayed zwitterion **Cys(-)-His(+)-His** kinetically more preferable.

minima			transition states		
point	refers to state	ΔU	point	refers to transition	ΔU
I	**Cys-His-His**	0	I-II*	**Cys-His-His to Cys(-)-His(+)-His**	101
II	**Cys(-)-His(+)-His**	87	II-III*	**Cys(-)-His(+)-His to Cys(-)-His-His(+)**	117
III	**Cys(-)-His-His(+)**	90	III-IV*	**Cys(-)-His-His(+) to Cys-His(-)-His(+)**	123
IV	**Cys-His(-)-His(+)**	75	I-IV*	**Cys-His-His to Cys-His(-)-His(+)**	84

Tab. 5-5:Located minima and maxima on the potential energy surface (Fig. 5-8, left-hand side, top) and their relative energies ΔU with respect to the state neutral state **Cys-His-His** (I) in kJ/mol.

Interestingly, the possibility of a further, and rather unexpected, zwitterionic state can be observed that comprises the formation of a histidine anion and histidine cation. Although the **Cys-His(-)-His(+)** state **IV** seems to be an appealing route with quite preferable kinetic (ΔU of 84 kJ/mol) and thermodynamic (ΔU of 75 kJ/mol) parameters, it is a dead end road, since the formation of the Cys145 thiolate has to overcome a barrier of 123 kJ/mol (**III-IV***) to become anionic. Taking these things together, the formation of the non-relayed zwitterion **Cys(-)-His(+)-His** appears to be the most likely situation, although the transition state of 101 kJ/mol (**I-II***) for the minimum energy path to state **I** appears quite high, compared to related enzymatic proton transfer reactions.[277,281,283,285,323] The route to the charge-relayed zwitterion **Cys(-)-His-His(+)** is therefore even more unlikely. This finding is furthermore supported by comparative sequence analysis that reveals that His164 is not well conserved at other coronavirus main proteases.

Due to the fact that the so far conducted investigations possess a prohibitive high barrier for the formation of the zwitterion state **Cys(-)-His(+)-His**, it is desirable to get more information about the influence of substrate or inhibitor binding on the proton transfer potential. Fortunately, extensive structural data of SARS-CoV M^{pro} is available today from X-ray crystallography that include complexes with inhibitors[46,50–52,89,106,116,117,119,130–133,147] and recently also with a substrate.[72] Based on this data, a comparison of the proton transfer potential was made for the free enzyme, the inhibitor-bound enzyme, and the substrate-bound enzyme, by using identical computational QM/MM workflows. For the complexed inhibitor, a high resolution structure with an α,β-unsaturated ester compound was selected (pdb code 2DUC[53]), since this warhead type has been addressed by many studies over the last years. The substrate complex (pdb code 2Q6G[72]) includes the amino aid sequence Ser-Ala-Val-Leu-Gln-↓-Ser-Gly-Phe that can be naturally processed by SARS-CoV M^{pro}.[80] For the free SARS-CoV M^{pro}, without any substrate or inhibitor, the same structure was employed as already described for the PES calculations. Prior to that, 10 ns MD simulations have been conducted for each of the three systems, to generate an extensive set of starting structures for the QM/MM minimum path calculations. Concerning the meaning of single minimum energy paths, as obtained from typically applied QM/MM workflows, it has become clear that starting structures from MD simulations might be somehow arbitrary and that the selection of different frames can lead to significant different results.[250,278,324,325] This accounts especially for the explicit solvent molecule configuration.[302] To avoid that problem, an averaging of the proton transfer potentials was done, consisting of 11 snapshots for each of the three model systems. To help identify meaningful connections between the two minima, proton transfer potentials were furthermore checked by calculations in forward and backward direction.

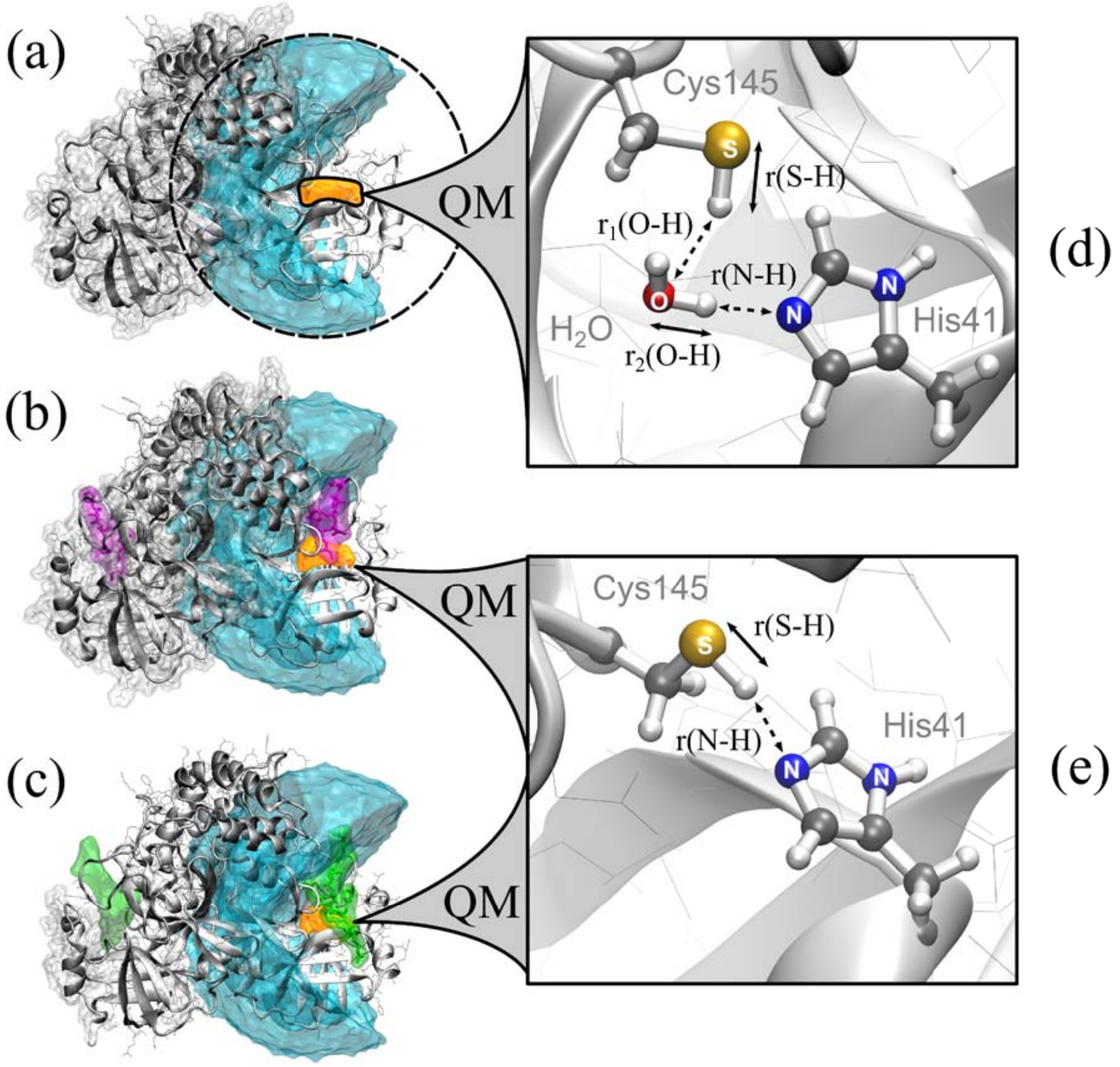


Fig. 5-9: Employed QM/MM model systems. The MM subsystems (left-hand side) include the dimeric protein structure of SARS-CoV M^{pro} (grey) and the solvation spheres around the active-site (cyan). Three different models are employed that describe the free enzyme (a) based on X-Ray structure 2DUC,[53] the enzyme with a Michael acceptor inhibitor (magenta) bound to the active-site (b) based on X-ray 2AMD,[106] and the enzyme with a substrate (green) bound to the active-site (c) based on X-ray 2Q6G.[72] QM subsystems (orange) include the active-site residues Cys145 and His41, without backbone atoms, and are either bridged with a water molecule (d) or directly hydrogen-bonded to each other (e). Reaction coordinates are indicated with arrows and show the coordinates that are used to "drive" the proton transfer (solid lines) and the coordinates that smoothly follow the driving reaction coordinates (dashed lines) on the potential energy surface.

The SARS-CoV M^{pro} X-ray structures of the inhibitor complex and the substrate complex, reveal that the Cys145/His41 distance is slightly closer than in the free enzyme, what indicates that the catalytic dyad is unlikely to be bridged by a water molecule in these cases. In detail, this can be confirmed by visual inspection from the thereof derived MD simulation trajectories, where exclusively a direct hydrogen bond between the thiol group and the accepting nitrogen atom can be observed. An overview about the three QM/MM model systems is illustrated in Fig. 5-9. Respectively, the QM parts include the water bridged catalytic residues Cys145/His41 for the free enzyme (Fig. 5-9, a) and the directly hydrogen bonded counterparts for the complexes (Fig. 5-9, b, c).

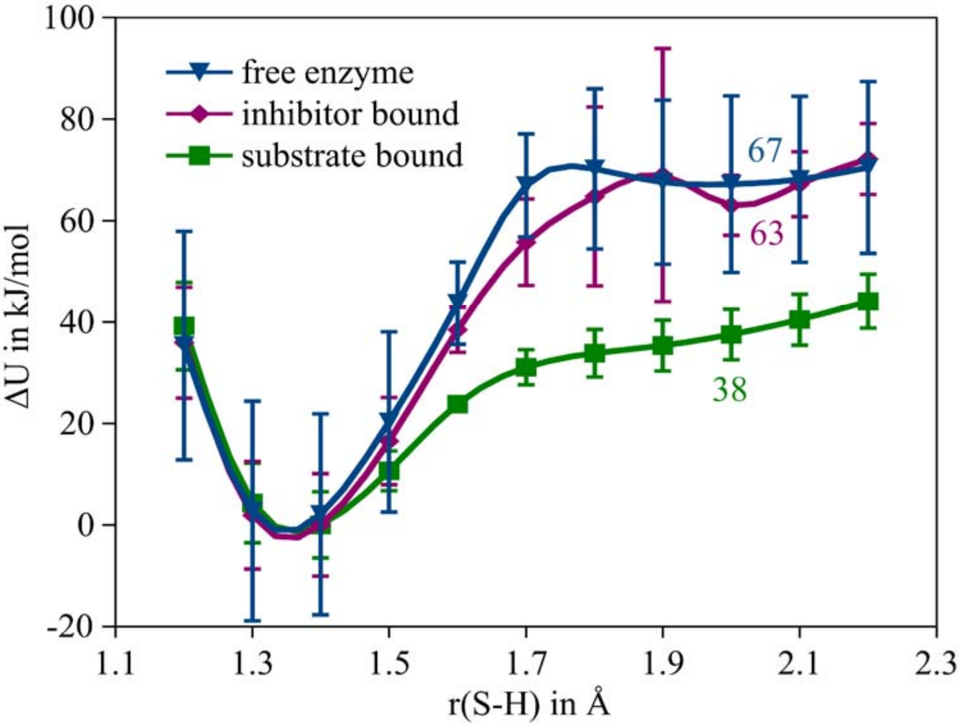


Fig. 5-10:Averaged QM/MM potential energy curves of the proton transfer between Cys145 and His41 for the three model systems (a), (b), (c). The free enzyme (blue line), inhibitor-bound enzyme (magenta line) and substrate-bound enzyme (green line) are compared, as shown in Fig. 5-9. Energy values are given in kJ/mol and are taken relative to the neutral state **Cys-His-His**, as shown in Fig. 5-7. The averaging is done for 20 potentials aligned by least square fitting and is based on minimized snapshots obtained from 10 ns MD simulations of each model system. Standard deviations for each point are indicated by vertical bars. The reaction coordinate r(S-H), as shown in Fig. 5-9 (d) and (e), was used to attain the proton transfer between the **Cys-His-His** and the **Cys(-)-His(+)-His** state.

By comparing the averaged QM/MM potential energy curves of the proton transfer, as given in Fig. 5-10, an interesting effect is observed. For the free enzyme, the averaged proton transfer potential results in a barrier of 70 kJ/mol and a relative energy of 67 kJ/mol by going from **Cys-His-His** to **Cys(-)-His(+)-His** which is 20-31 kJ/mol lower than the values obtained from the single potential energy surface in Fig. 5-8. This results underlines the benefits of averaging over several minimum energy paths that give a better estimate than single paths.

The calculated proton transfer potential in case of the inhibitor-bound structure ranges in the same magnitude and comprises a barrier of 63 kJ/mol and a relative energy of 69 kJ/mol for **Cys(-)-His(+)**. By going to the substrate-bound model system, a significant drop for the relative energies is observed, showing a decreased potential in the range of ~38 kJ/mol for ΔU in the zwitterionic region. Furthermore, the separating barrier between **Cys-His** and **Cys(-)His(+)** does not appear when the substrate is bound.

Obviously, the substrate binding seems to have a modulating effect on the proton transfer potential. This characteristic is moreover not mimicked in case of the inhibitor binding. A view on the respective standard deviation bars in Fig. 5-10, that give an estimate on how strong the potentials fluctuate within the set of calculated proton transfer paths, underlines this conclusion, since the substrate-bound model system comprises a very clear defined average path, expressing little fluctuation compared to the broad deviation bars of the free and inhibitor-bound enzymes. Conclusively, the substrate has a facilitating impact on the formation of the zwitterion.

5.2.3 Charge Deletion Analysis

The results obtained by the QM/MM potentials have given an estimate of 25 to 30 kJ/mol for the stabilization of the zwitterion, whenever a substrate is bound to the active-site of SARS-CoV M^{pro}. So far, this effect is solely derived from the comparison of the proton transfer potentials that give no additional information about the underlying reasons for this energetic difference. It is therefore highly desirable to get a more differentiated picture of this effect, since the stabilizing contributions can stem from various parts of the environment, e.g. from the protein or individual protomers themselves through conformational changes (SARS-CoV M^{pro} is supposed to follow an induced fit mechanism[51]), from the substrate, or maybe from the water shell.

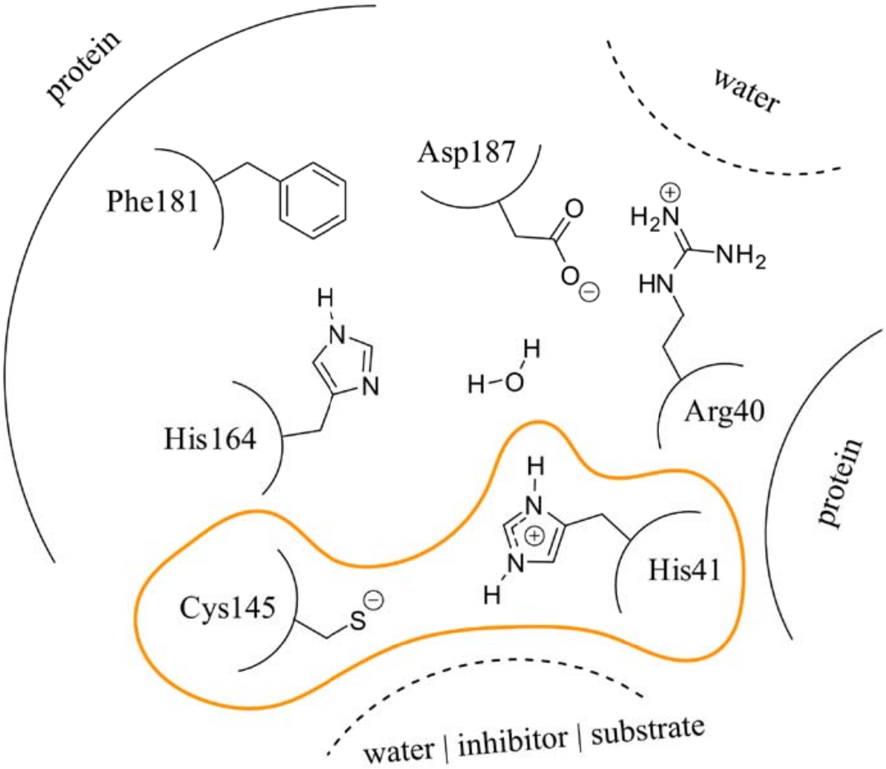


Fig. 5-11: Topology of the SARS-CoV M^{pro} active-site with catalytic dyad in the zwitterionic state **Cys(-)-His(+)-His**. The charges of different neighboring structure elements in the MM subsystem are switched off individually and their energetic impact on the QM subsystem (orange) is estimated in terms of a charge deletion analysis, as given in Tab. 5-6.

In order to dissect the stabilizing contributions, charge deletion analysis (CDA) were performed for the three QM/MM model systems. The CDA allows to selectively switch off specific parts of the MM environment and gives therefore a measurement for the stabilizing or destabilizing contribution of this part. It is notable that the charge deletion analysis only accounts for the electrostatic interactions and not for other intermolecular interactions, like exchange or dispersion interactions, which might play a role, e.g. for phenyl rings. Nevertheless, electrostatic contributions possess typically the greatest impact among the intermolecular interactions due to their $1/r^2$ distance depending behavior. Since possible pitfalls from single minimum energy path calculations, as described above, conclusively also hold for the CDA, an analogue averaging has been performed over 11 snapshots for each case.

residue	free enzyme[a]	inhibitor-bound[b]	substrate-bound[c]
switched off	ΔΔU(ZW)	ΔΔU(ZW)	ΔΔU(ZW)
protein	31 ± 16.5	22 ± 10.6	18 ± 7.0
protomer A	27 ± 16.0	18 ± 11.6	20 ± 6.2
protomer B	4 ± 1.5	1 ± 0.8	-4 ± 3.6
His164	1 ± 3.8	7 ± 2.3	5 ± 3.6
Arg40	-17 ± 2.7	-19 ± 1.2	-18 ± 2.2
Arg40-Asp187	8 ± 5.0	9 ± 1.5	3 ± 2.0
Asp187	23 ± 5.7	26 ± 0.9	19 ± 1.4
Phe181	0 ± 0.1	0 ± 0.2	0 ± 0.2
H2O@His41	18 ± 4.4	14 ± 1.4	0 ± 0.1
water	-5 ± 63.4	18 ± 5.3	25 ± 11.0
substrate	- ± -	- ± -	8 ± 3.1
inhibitor	- ± -	-1 ± 3.9	- ± -

[a] Data set consists of 11 minimized snapshots taken from MD simulation based on 2DUC.

[b]Data set consists of 11 minimized snapshots taken from MD simulation based on 2AMD.

[c]Data set consists of 11 minimized snapshots taken from MD simulation based on 2Q6G.

Tab. 5-6: Charge deletion analysis in kJ/mol. Energy values ΔΔU refer to the change in energy that occurs due to switching off certain parts of the MM system with respect to the unaffected QM/MM system, as shown in Fig. 5-12. Values ΔU(ZW) refer to the relative energy of the zwitterionic state **Cys(-)-His(+)-His** with respect to the neutral state **Cys-His-His** of the catalytic dyad.

The CDA was performed for the local minimum states **Cys-His-His** and **Cys(-)-His(+)-His** and allows to calculate ΔΔU values by subtraction, which yield the energetic impact on the zwitterionic state, as depicted in Fig. 5-11. A more detailed description of the procedure is given in the computational details section. The obtained results from the CDA are summarized in Tab. 5-6 that contains the electrostatic contribution of various parts of the environment and includes furthermore standard deviations of the values as obtained from the respective data set of 11 averaged snapshots.

Most apparent contribution comes from the protein environment that stabilizes the **Cys(-)-**

His(+)-His state by 18-31 kJ/mol, depending on the model system. The major part of this contribution stems from protomer A in which the considered active-site is located. Protomer B has practically no impact on the proton transfer reaction. The more detailed examination of the single residues around the active-site reveals which amino acids contribute most for the first step in catalysis.

Starting with His164, the calculated impact of 1-7 kJ/mol shows that this histidine residue plays no important role for the formation of **Cys(-)-His(+)-His**. Switching off Arg40 leads to a significant benefit for the zwitterionic state, since a stabilizing effect of 17-19 kJ/mol is gained in the absence of this residue. The reason for this can be seen by regarding the relative orientation of Arg40 with respect to the dipole formed by the zwitterionic catalytic dyad, as depicted in orange in Fig. 5-11. The positive charge of the arginine residue is located in close proximity to the positive charged His41 residue, therefore withdrawing one of the positive charges leads to decrease in electrostatic repulsion and thus to a fostering of the **Cys(-)-His(+)-His** state. The opposite effect happens when Asp187 is switched off. The negative charge of the carboxylate group has obviously an important stabilizing effect on the zwitterion formation. By removing this moiety, the favourable interaction between the negative charge of Asp187 and the positive charge of His41 is disturbed, and would lead to an increase by the relative energy of 19-26 kJ/mol for **Cys(-)-His(+)-His** with respect to **Cys-His-His**. Considering the Arg40 and Asp178 residue together as a salt bridge, it becomes apparent that this entity taken as a whole has only a marginal stabilizing effect of 3-9 kJ/mol on the zwitterion formation. Considering the salt bridge and the zwitterion as single dipoles, this marginally effect becomes clear by visual inspection of the three dimensional structure of the active-site. The relative orientation of the two dipoles are arranged not in line and stand nearly perpendicular to each other, with Asp187 being slightly closer to the catalytic dyad. Therefore the impact of the whole salt bridge partly cancels out the stabilizing effect of Asp187 and the destabilizing effect of Arg40. Nevertheless, Asp187 plays the most important role in the closer neighborhood of the active-site and represents an important entity for the first step in general base catalysis of SARS-CoV M^{pro}.

The Phe181 can be expected to have no impact in terms of the CDA, since its phenyl ring side chain possesses only insignificant electrostatic influence. Although there is no electrostatic contribution to the zwitterion formation, it can not be completely excluded that this residue might contribute in this context. Interactions between cationic histidines and aromatic systems are well known[326] and can have stabilizing effects in the range of 4 kJ/mol,[327] and further depends on the relative orientation.[328] Nevertheless, since this type of interaction is primary of non-electrostatic nature, it can not be addressed by the CDA.

So far, all of the protein related contributions are more or less consistently conserved for the three model systems and changes are rarely observed above the limit of significance, as visible from the standard deviations.

Regarding the influence of water, a different trend can be observed. Water plays in a twofold way an important role for the catalysis by the SARS-CoV M^{pro}, since it is present as a solvent shell and further as an inherent and conserved part of the active-site in all X-ray structures. In case of the buried water molecule H_2O@His41, a decreasing importance of this moiety for the stabilization of the zwitterion is observed, with contributions of 18 kJ/mol, 14 kJ/mol, and 0 kJ/mol by going from the free enzyme to the inhibitor-bound enzyme and the substrate-bound enzyme. Since the water molecule is present in all MDs of the three model systems and does not drift out of the rigid hydrogen bond network, as hypothesized at some points in the literature,[67,92] the cause of this effect is moreover due to substrate induced changes of the active-site conformation, and further a changed balance between the various stabilizing effects.

The maybe most interesting effect, concerning the CDA, is observed for the influence of the water shell. By comparing the three model systems it is found that the contribution of the solvation sphere to the formation of the zwitterion behaves quite different for the free enzyme and the two complexes. In the first case, the averaged view reveals a marginally destabilising effect on the formation of **Cys(-)-His(+)-His**, reflecting that the proton transfer path has to proceed through a rigid hydrogen bond network formed by the first solvation shell. This first solvation shell, which can also be denoted as biological water,[329–331] has a constitution that corresponds to the **Cys-His-His** state and can hardly adopt to the **Cys(-)-His(+)-His** state that is formed instantaneously by following the minimum energy path. This situation is also well reflected by the extraordinary high standard deviation of 63 kJ/mol, underlining that only small changes in the first solvation sphere have a strong impact on the proton transfer potential. The situation is quite different for the complexed systems, when a substrate or inhibitor is bound to the active-site. Here, the first solvation sphere is well shielded through the complexed entity and therefore only the stabilizing effect of the water bulk remains on the zwittionic state. In detail, the formation of **Cys(-)-His(+)-His** is fostered by 18-25 kJ/mol through the water bulk, whereas significantly less fluctuation from this contribution is observed (standard deviation < 11 kJ/mol) than in the free enzyme case.

The last subjects of the charge deletion analysis are the complexed entities, namely the inhibitor and the substrate. As visible from Tab. 5-6, the inhibitor does practically not influence the **Cys-His-His/Cys(-)-His(+)-His** equilibrium, whereas the substrate induces a preference of 8 kJ/mol towards the zwitterion. Although this difference appears to be not very decisive, it could represent an inherent property of the substrate that is obviously not featured by the inhibitor. This observation highlights an interesting and novel aspect for inhibitor design, since rational and specific changes regarding this aspect could lead to more efficient inhibitors.

5.2.4 Free Energy Calculations

Although the conducted QM/MM minimum path calculations allow deep insights into the proton transfer reaction, they represent a static picture of the model systems and do not account for dynamics or entropic effects. Potential of mean force based free energy calculations are a commonly used approach to overcome this deficiency,[291,332,333] that have also become popular in combination with the QM/MM methodology.[269,271,304,334] Furthermore, the calculation of free energies can give a complementary view to minimum energy path calculations from the theoretical perspective.

To get access to the free energy profiles of the proton transfer, the model systems of the free enzyme and the substrate-bound complex were set up for QM/MM MD simulations, in an identical manner to the previous QM/MM calculations, as shown in Fig. 5-9 a) and c). Since extensive sampling is necessary to obtain a potential of mean force (PMF), a semi-empirical method (PM3/PDDG) instead of density functional theory had to be employed, in order to keep the computational efforts feasible. Nevertheless, the errors of the employed method can be estimated from the benchmarking in chapter 5.1.2. It should range by an overestimation of 21 kJ/mol for the zwitterionic state, comprising a difference of 38 kJ/mol with respect to the employed DFT approach (B3LYP/TZVP), which was employed for the QM/MM minimum path calculations.

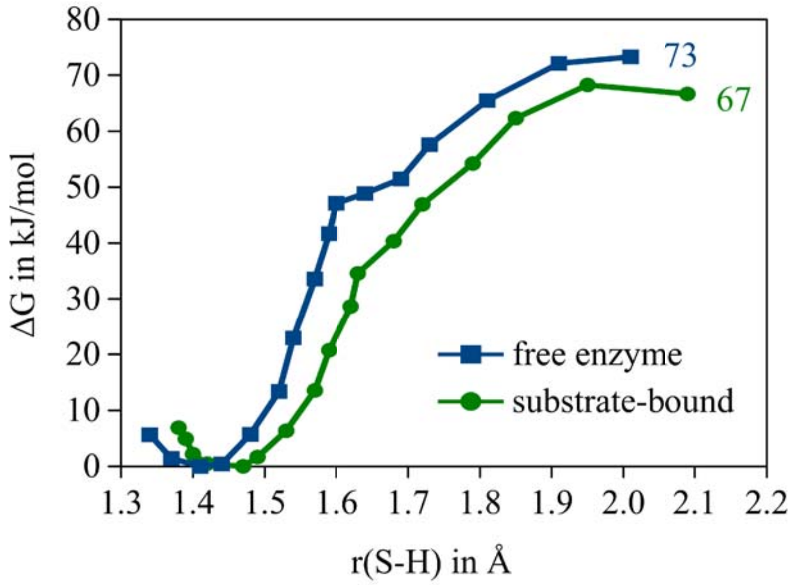


Fig. 5-12: Free energy potentials of the proton transfer between the **Cys-His-His** and the **Cys(+)-His(+)-His** state in kJ/mol. The free enzyme (blue line) as shown in the QM/MM model Fig. 5-9 (a) and the substrate-bound (green line) QM/MM model as shown in Fig. 5-9 (c) are compared. The r(S-H) distance given in Fig. 5-9 (d) and (e) serves as reaction coordinate. Values ΔG are obtained by the potential of mean force method,[333] employing PM3-PDDG,[208,415,417] and using samplings of 50 picoseconds for each point with a subsequent weighted histogram analysis.[420,421]

Fig. 5-12 shows the free energy potentials for the free enzyme model (blue) and the substrate-bound model (green), as obtained by the PMF. The relative free energies ΔG of **Cys(-)-His(+)-His** with respect to the neutral state **Cys-His-His** are quite similar for the two cases, comprising values of 73 kJ/mol and 63 kJ/mol. The latter one refers to the substrate-bound model and yields a significant higher lying **Cys(-)-His(+)-His** state than obtained from the QM/MM minimum path calculations. By taking the intrinsic energetic shift of 38 kJ/mol between the underestimating DFT method (ΔU ~38 kJ/mol), employed for the averaged minimum energy path calculations, and the herein used overestimating PM3 method (ΔG ~67 kJ/mol) into account, the free energy PM3/PDDG results for the substrate-bound PMF agree quite good with the DFT minimum path calculations. The remaining difference of 9 kJ/mol between the free energies ΔG and the internal energies ΔU can be considered as insignificant.

The situation changes for the free enzyme model. Whereas the QM/MM minimum path calculations predict an energy difference ΔU of 67 kJ/mol between the zwitterionic and the neutral state (Fig. 5-10). The ΔG value from PM3/PDDG free energy, corrected by the method intrinsic shift to B3LYP, is only about 35 kJ/mol. Therefore, ΔG is found to be lower than ΔU. There could be three reasons for this stabilizing contribution. The first one refers to the different error bars of the two methods. However the difference of 32 kJ/mol appears too large to represent an artifact. The second reason refers to entropic effects, that are only accounted for in the free energy calculations. However this is counter-intuitive, because a zwitterionic state should cause a higher ordered water shell than the neutral state and lead to a positive entropy contribution where ΔG > ΔU. Since the opposite is found here, this explanation is rather unlikely for the given case. Moreover, entropic contributions can be expected to be rather small for proton transfers inside of the protein.[288,325] The third reason refers to the relaxation of the first solvation sphere. As obvious from the charge deletion analysis, there is a significant difference found for the proton transfer that is due to whether the zwitterion is formed in the direct vicinity of the solvent or screened by an inhibitor or substrate. In the first case, QM/MM minimum path calculations nicely illustrate that the first solvation sphere behaves rather rigid and can only partly relax when the ion pair is formed instantaneously. Regarding this aspect, the QM/MM MD sampling allows a much better relaxation of the first solvation sphere, since it provides 50 ps of dynamic simulation.

The question, which one of the applied approaches can give a better answer, is therefore a matter of the considered time scale, since proton transfer reactions in water are settled below the pico second scale.[335,336] This issue leads inevitably to the discussion about "biological water" that is supposed to possess slightly different properties than bulk water.[330,331] The proposed concept assumes that water at the surface of proteins needs much more time to reorient, in the range of 8 ps to 80 ps,[329] than bulk water with relaxation times < 10 ps,[337] for the dielectric response of the solvent. Although it is uncertain whether this effect plays a role here, it is clear that this can only exclusively be observed by the QM/MM MD simulations.

Conclusively, both effects, entropy and solvent relaxation, could play a role, but unfortunately this cannot be resolved more exactly with the given methods. However, it can be assumed that solvent relaxation accounts for the largest part of the observed difference between the herein performed QM/MM and QM/MM MD calculations. This assumption is supported by test calculations with a simulation time of 25 ps, 50 ps, and 100 ps that reveal apparent decreasing energy differences between the **Cys-His-His** and **Cys(-)-His(+)-His** state. A plot of the results for the three different simulation length can be found in the computational details section.

However, for the substrate-bound case, the relative energies of the initial proton transfer for the averaged QM/MM agree with the QM/MM MD results, under consideration of the method intrinsic inaccuracy. By taking the errors of the applied methods with respect to LCCSD(T)/QZVP calculations into account (see chapter 5.1.2), the energy differences can be estimated. They are respectively an underestimation of 17 kJ/mol for B3LYP/TZVP, and an overstimation of 21 kJ/mol for PM3. With these corrections, a relative energy ΔU of 45 kJ/mol and a relative free energy ΔG of 46 kJ/mol is obtained for the initial proton transfer that both proceed without a significant barrier. For the free enzyme, this additive correction results in a ΔU of 84 kJ/mol and a ΔG about 52 kJ/mol. Whereas the ΔG values differ only insignificantly, the ΔU values are reduced by 39 kJ/mol through the binding of the substrate. This effect might be due to a shielding of the active-site from the surrounding water. Fig. 5-13 shows the active-site (a) partly covered by the inhibitor (b) and fully covered by the substrate (c).

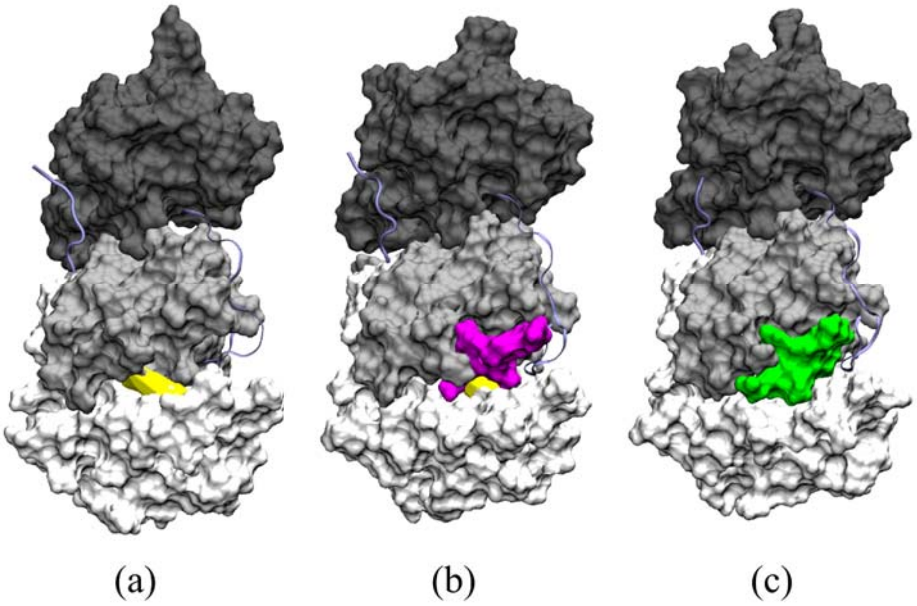


Fig. 5-13: Coverage of the active-site (a, yellow) by the inhibitor (b, magenta) and the substrate (gc, reen). Only one protomer is respectively shown for the sake of clarity.

Taking these things together, the assumption of an "electrostatic trigger" that initiates the catalytic reaction, as reported in the work of Solowiej et al.,[58] can be confirmed by the herein performed calculations. The experimentally available pK_a measurements of SARS-CoV M^{pro} allow to estimate the relative energy ΔG by 7-12 kJ/mol, which is somehow lower than the theoretically predicted values and might result from deficiencies of the applied QM/MM models. However, since the model systems have been kept consistent for the comparisons, the "electrostatic trigger" and the modulation effect of substrate binding on the proton transfer should be rather significant.

5.2.5 Conclusions about the Active-Site of SARS-CoV M^{pro}

In summary, the active-site of SARS-CoV M^{pro} was characterized from different theoretical perspectives. The comparative analysis of several MD simulations with available structural data from X-ray crystallography reveals that the catalytic dyad of SARS-CoV M^{pro} resides in a neutral charge state and that the neighboring His164 residue is also found to be uncharged. Albeit His164 could act as a proton acceptor through a hypothetical charge-relay mechanism, the elaborated results from QM/MM potential energy surface calculations show that this is rather unlikely.

The comparison of free enzyme, inhibitor-bound enzyme, and substrate-bound enzyme by averaged QM/MM potentials has shown that the necessary proton transfer reaction towards the formation of an ion pair is remarkably modulated by substrate binding, which facilitates by 39 kJ/mol. The remaining energy difference ΔU of 45 kJ/mol corresponds to the assumed "electrostatic trigger" reported by Solowiej et al.[58] that initiates the catalytic reaction. However, the modulating effect could not be observed for inhibitor binding (Fig. 5-14).

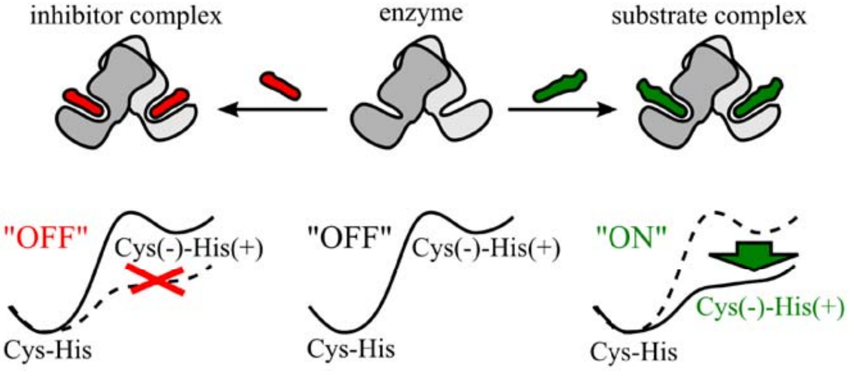


Fig. 5-14: Substrate binding (right hand side) fosters the zwitterion formation. Inhibitor binding lacks this effect (left hand side).

The applied charge deletion analysis furthermore gives insights to the factors that are causative for the modulating effect of the substrate binding on the proton transfer potential.

This difference is mainly explained by the shielding of the active-site by the water shell. This explanation is further supported by the different coverages of the active-site by the inhibitor and the substrate.

Since the main factors stem from the water shell, there is actually no reason to believe that this is an exclusive effect for SARS-CoV M^{pro}. Therefore this might also hold for other enzymes that follow a general base catalyzed mechanism. However, further investigations on similar enzymes have to be made to evaluate the importance of this effect for other cases.

The comparison of the electrostatic impact of substrate and inhibitor rises an interesting and novel approach for the rational design of inhibitors that target enzymes with general base catalyzed mechanism. Inhibition potency of regularly designed inhibitors will typically come to a lower limit, due to the rate determining nature of the initial proton transfer reaction, which is moreover an inherent property of the enzyme. Here, the modulation of the proton transfer potential by the inhibitor could have quite positive effects on the inhibition reaction. Specific designed inhibitors that possess such an "electrostatic booster" property, might therefore achieve better inhibition potencies. However, the precondition of this strategy would be that the electrophilic addition step has a barrier that is significantly lower than the initial proton transfer. Since there are several examples of low barrier additions,[275,282] this strategy could be an interesting approach for novel inhibitors, but needs of course more investigation to prove its applicability.

5.3 Screening of Inhibitor Warheads based on Michael Acceptors

The last chapter gave an extensive theoretical investigation of the catalytic dyad Cys145/His41 and the factors that determine it to be in a neutral thiol/imidazole state rather than resting in an thiolate/imidazolium ion pair state. This chapter tries to extend this knowledge towards the inhibition mechanism. As stated in chapter 2.3 about SARS-CoV M^{pro} inhibitors, warheads based on the Michael addition reaction are an often applied strategy to achieve irreversible or possibly covalent reversible inhibition of SARS-CoV M^{pro} or other cysteine proteases.

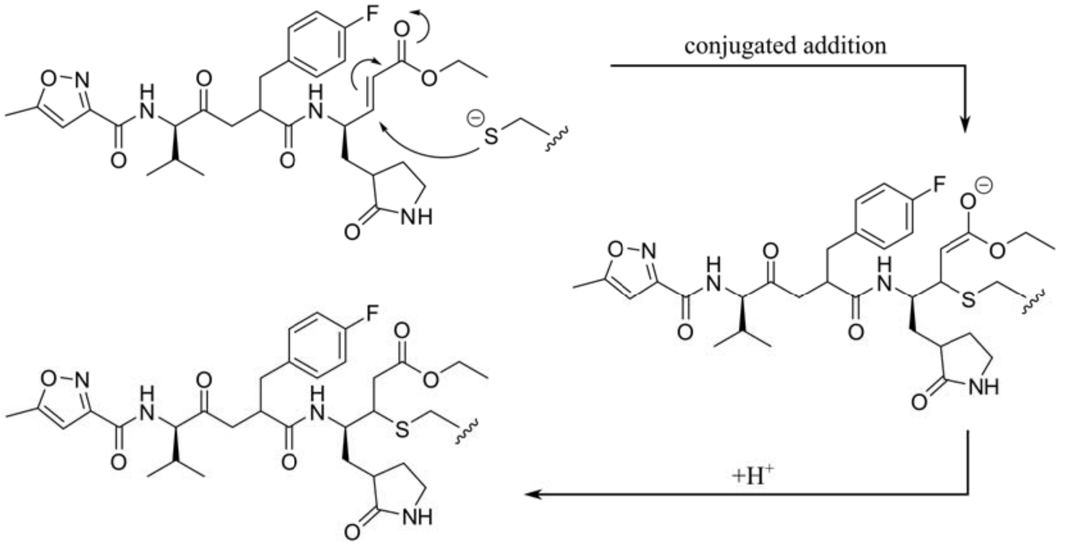


Fig. 5-15: Inhibition of the cysteine moiety by conjugated addition of the thiolate to the α,β-unsaturated carbonyl warhead of AG7088 with subsequent protonation of the anionic intermediate.

Therefore it is desirable to have profound knowledge of the factors that determine inhibition potency of this inhibitor class. One of these factors is the chemical substitution pattern around the α,β,-unsaturated carbonyl moiety, which controls the electrophilicity of the warhead and thus its reactivity with respect to the thiol or thiolate moiety of the enzyme.[338]

Information about this can be experimentally determined through chemoassay screenings that employ commercially available model compounds, like glutathione,[339–341] as an active-site mimicry. Instead of using the purified enzymes that are often difficult to prepare or obtain, this is much more efficient for larger screenings.

Another possibility is the screening of compounds by using a theoretical approach. Such *in silico* screenings can be employed for rational drug design[342–344] but are also useful for the prediction of toxicological properties.[345–347] Especially the latter one is a highly eligible goal, since computational methods with reliable predictive power can help to reduce animal experiments[348] that are undesirable for ethical and economic reasons.

The following chapter employs an *in silico* screening for the reactions between substituted α,β-unsaturated carbonyl compounds and thiols. The results should provide better insights into variations in the underlying mechanisms as a function of the substitution pattern. They are furthermore of interest for understanding the inhibition mechanisms for the SARS-CoV M^{pro} inhibition as well as for the excess toxicity of substituted α,β-unsaturated carbonyl compounds. Since the Michael addition reaction consists of more than one step (Fig. 5-15), the various possible reaction courses including the 1,4-addition followed by a ketonization step have to be considered. The influence of a base-catalyzed step for the reactivity of thiol groups, as observed for the enzymes, must also be considered.

The majority of Michael acceptor compounds that have been employed so far (Tab. 2-2), are derivatives of AG7088 with α,β-unsaturated ester warheads that react in an irreversible fashion. Another approach starts from substituted etacrynic acid[141,349] which possesses an α,β-unsaturated ketone moiety as electrophilic building block. While the vinylogous esters lead to irreversible inhibition, the etacrynic acid derivatives show only a reversible inhibition of proteases.[350] The underlying reason for the difference is not clear at all.

The rational design of improved inhibitors deserves detailed knowledge about the inhibition process and especially about the interplay between the various effects. Of special interest is a distinction between effects which influence the reversible and the irreversible step. The former is influenced by non-covalent interactions between inhibitor and enzyme. The latter may also be influenced by such effects since they may orient the inhibitor in the right arrangement. However, in addition it relies on the chemical reactivity of the electrophilic warhead which is influenced by the substituents.

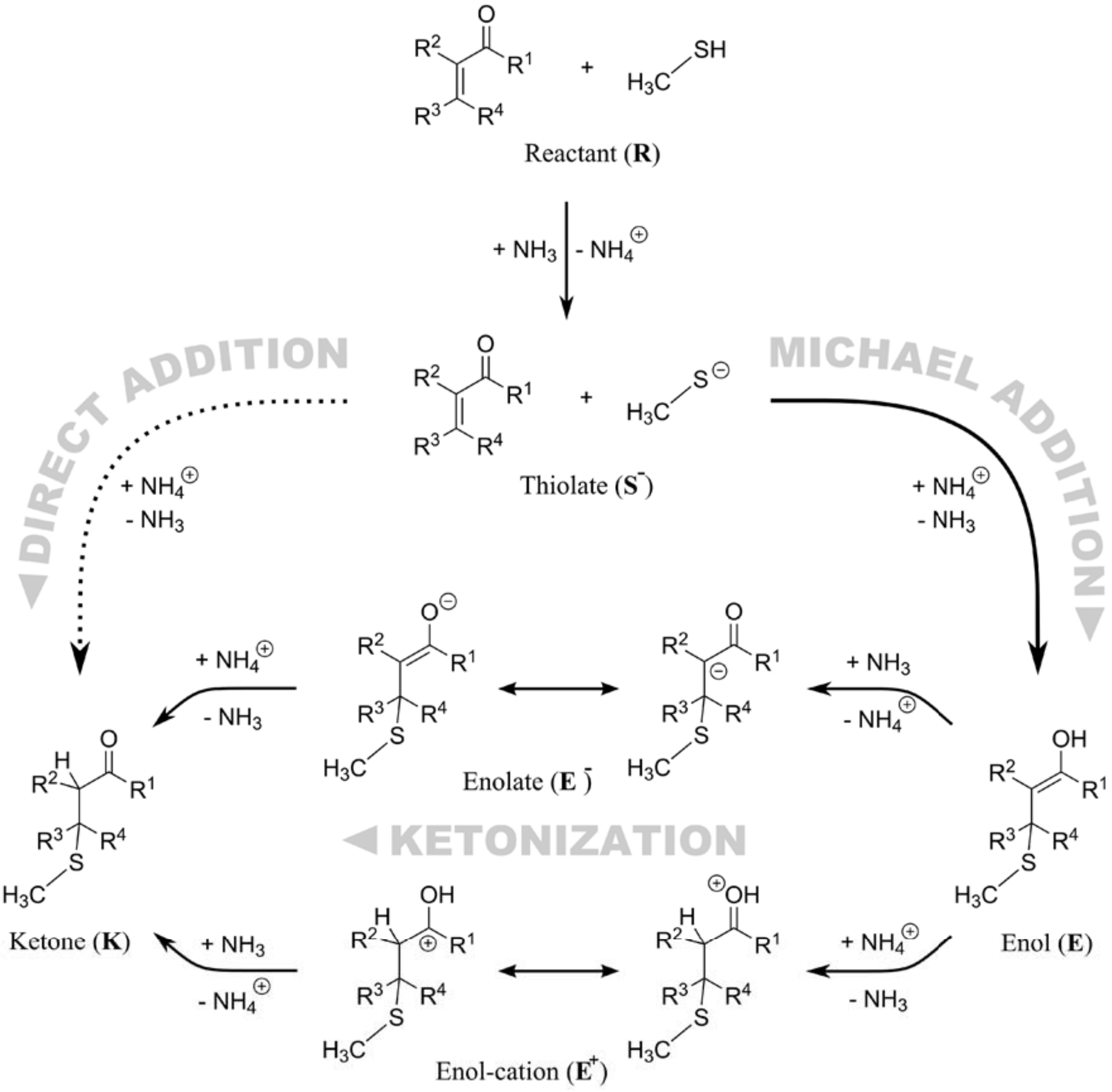


Fig. 5-16: Conceivable reaction mechanisms for the addition of methylthiol to α,β-unsaturated carbonyl compounds.

To disentangle the various effects that determine the reactivity of Michael systems, a systematic screening of different substitution patterns is employed here. Earlier theoretical investigations[351,352] predict that the reaction of a thiolate with a Michael system leads to quite stable covalent bonds between the thiolate and the α,β-unsaturated carbonyl moiety. This corresponds to predicting an irreversible inhibition of thiol containing enzymes by inhibitors containing such electrophiles as warheads. However, a survey of the reported studies reveals, that the used theoretical model systems completely neglected solvent effects and did not consider substituent effects. Additionally, the authors assumed the attack of a negatively charged thiolate moiety, while the catalytic active cysteine residue within the SARS-CoV M$^{\text{pro}}$[71] represents a neutral thiol, as shown in chapter 5.2. The challenge for the herein performed computations is therefore to develop a model system that is simple enough to perform screenings, but also possesses a suitable modeling of the environment, since various reaction paths are conceivable and a strong influence of the environment can be expected. The developed model systems can then be used to extract the influence of the substitution pattern on the inhibition mechanism. As stated above, the addition step of the thiol to the Michael system proceeds via a base-catalyzed mechanism. Possible addition mechanisms include reaction courses which run across enol intermediates, but can alos undergo direct addition pathways (Fig. 5-16). For the former the subsequent enol-keto tautomerization must also be taken into account since the corresponding activation barriers are expected to be comparable or even higher than the barriers of the addition step.

The chapter is organized as follows. After a brief description of the different possible reaction courses for Michael acceptors, the set up of the theoretical model system is explained. This model system is then applied for the screening of model reactions with different substitution patterns. The theoretical results are discussed and afterwards also compared to experimental results to achieve insights how the substitution pattern influences the potency of inhibitors possessing substituted Michael systems as electrophilic warhead. Beside the conclusions that can be drawn about inhibition issues, the resulting information also provides insights into the excess toxicity of the unsaturated carbonyl compounds acting as Michael acceptors.

5.3.1 Mechanism of the Inhibition Reaction

Inhibition mechanisms, in which covalent bonds are formed, can formally be divided into two steps (equation 6).[115] The first step consists of the formation of a non-covalent enzyme-inhibitor complex (E···I). Within this complex, the attacking enzyme moiety is oriented in such a way that the covalent bond (E-I) can be formed, e.g. by an attack of a nucleophilic center of the enzyme at the electrophilic center of the inhibitor. The resulting chemical reaction represents the second step of the overall inhibition. Within this step the bond formation is mainly influenced by enthalpic effects since all entropic contributions which are connected to the formation of the enzyme-inhibitor complex are already included in step 1.[307] Furthermore, entropy effects can be expected quite uniform for similar reactions and hence do not change the relative trends.

Since the first step is mainly determined by intermolecular forces of the whole inhibitor, and not so much from the warhead, it is out of scope here. Moreover the second step is determined mainly by the warhead and therefore the most interesting step for the present investigation. In the present context it involves the formation of a covalent bond between a thiol residue (e.g. of the cysteine residue of the SARS-CoV M^{pro}) and the electrophilic building block of the inhibitor. Models, which aim to describe this step in all details, have to account for the influence of the protein and the solvent environment. This can be achieved with QM/MM approaches which take enzyme and solvent explicitly into account. They are even able to describe effects like regio- or stereoselectivity of the inhibition process.[279,280,286,353] Nevertheless, QM/MM methods would have significant drawbacks by employing them for *in silico* screenings, since they are far away from being used in a black box manner and need considerable effort for the set up of individual substitution patterns. However, as shown recently in investigations about the inhibition of papain like proteases,[276,307,342–344] valuable insights into the influence of the substitution pattern of the warhead are already possible from simpler models. These mimic all important parts of the considered system by truncated model compounds in combination with a continuum model like COSMO.[244] These models approximate the cysteine residue by a methyl thiolate while the environment is modelled either by two water molecules (pKs≈15) to mimic a low proton-donating ability of the environment or by two ammonium molecules (pKs≈9.3) to simulate a higher proton-donating ability. These explicit solvent molecules are employed in combination with the continuum model COSMO which accounts for the overall polarizability of the solvent and the protein environment. From the inhibitor, only the electrophilic warhead and its directly attached substituents are taken into account. Information about the kinetics and thermodynamics of the irreversible step of the inhibition are obtained through potential energy surfaces of the inhibition process which are computed employing the B3LYP/TZVP//BLYP/TZVP level of theory. Since such approaches are considerably less expensive than QM/MM computations they allowed a scan of a large number of substituents to test changes in the inhibition potencies of epoxides and aziridines. Such model calculations are also important to distinguish between influences of the enzyme environment and inherent effects resulting from the electronic structure of the inhibitor itself. That the prediction of the computations could be proven experimentally underlines the reliability of the simple approach.[276,307,342–344]

To transfer such models to the situation in thiol containing enzymes reacting with α,β-unsaturated carbonyl derivatives, the following approximations are adopted and developed into a simplified model system (Fig. 5-17).

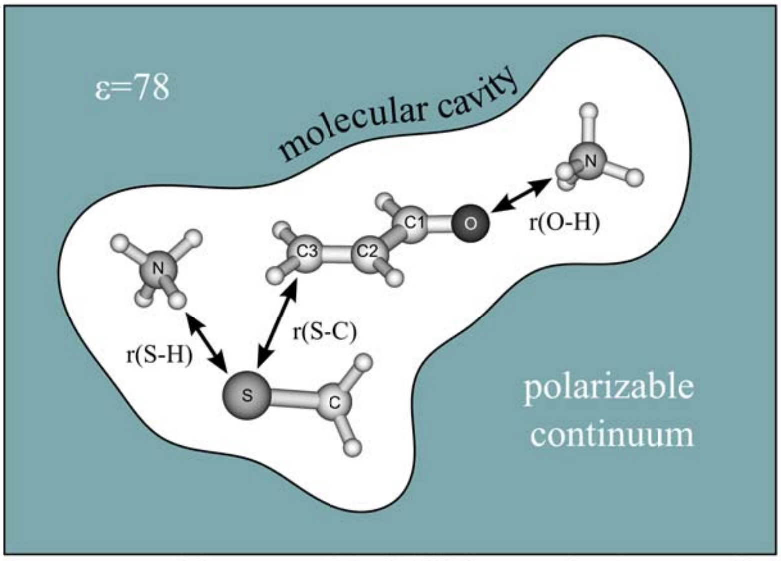


Fig. 5-17: Model system of the inhibition reaction between thiol and Michael acceptor. The employed reaction coordinates that are used to compute the potential energy surfaces are depicted as arrows.

The thiol containing residue of the enzyme is mimicked by methyl thiol. Possible proton acceptor groups, (e.g. the histidine residue of the SARS-CoV M^{pro}) are assumed to play the role of the proton acceptor in the base-catalyzed process. To evaluate the importance of the base catalysis such groups are modeled by an ammonia molecule which has been shown to be a good approximation. This is shown by comparison between simpler QM[307,342] and more elaborated QM/MM[286] approaches. The former used ammonium while the latter employed a histidine group. The success of this approach may result since its pK_a value (9.3) is not too different from that of histidine residues in thiol proteases (pK_a 8-9).

To investigate the influence of the substitution pattern of the α,β-unsaturated carbonyl derivatives on the reaction mechanisms, a screening across the reactions of the model compounds **1**-**11**, as depicted in Fig. 5-18, is performed. The possible reaction paths of the overall addition of thiols to Michael systems are depicted in Fig. 5-16. For the reactions under physiological conditions, thiols exists in protonated (R-SH) and deprotonated (R-S$^-$) form but, as discussed later, only the reactivity of the thiolate seems to be high enough for an efficient attack at the α,β-unsaturated carbonyl moiety. Hence, the proton has to be transferred to a proton acceptor, what corresponds to a base-catalyzed reaction.

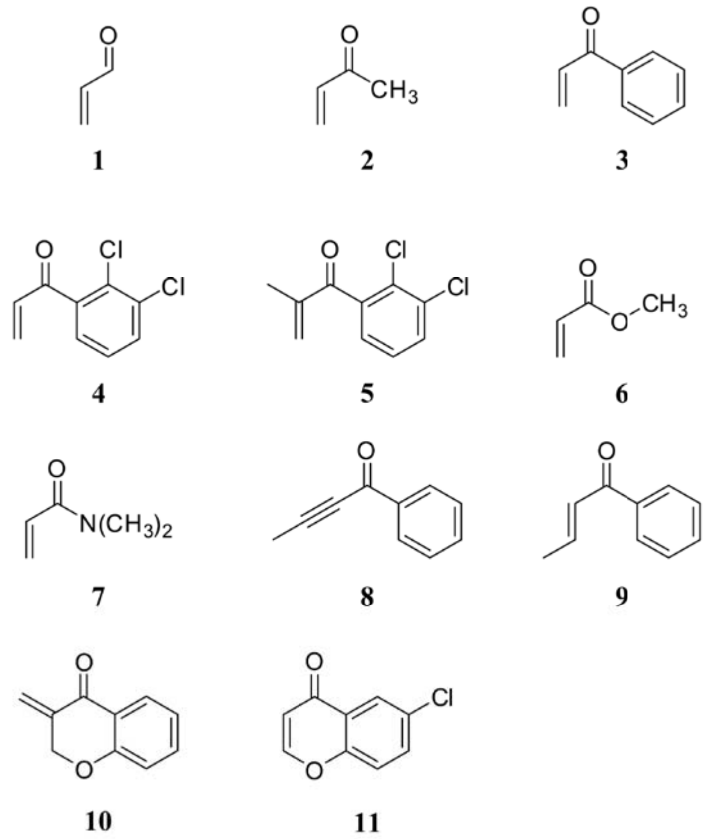


Fig. 5-18: Model compounds that represent different derivatives of α,β-unsaturated carbonyl compounds.

For the resulting thiolate, the reaction can proceed along two pathways. In the direct addition mechanism the thiolate attacks C3 (Fig. 5-17) and at the same time the proton is transferred from the base to C2 (direct E2-like addition). If the proton is instead transferred to the carbonyl oxygen, the enol form of the 1,4-addition is obtained. A comparison between both pathways was computationally performed by Weinstein and coworker,[354] who studied the addition of ammonia to α,β-unsaturated carbonyl compounds. They computed a direct addition as well as the 1,4-addition. However, they did not consider the base-catalyzed reactions discussed in Fig. 5-16. In contrast for the direct addition they assumed an intermolecular proton transfer from the ammonium to the C2. For the 1,4-addition an intramolecular proton transfer from the nitrogen to the carbonyl oxygen was considered. They found that in such cases the 1,4-addition is favored with respect to the direct addition. Both reactions are considerably catalyzed by one water molecule, but the 1,4-addition remains the favorable reactions course. For acrolein the barrier of the 1,4-addition is about 70 kJ/mol lower than the corresponding barrier of the direct addition (96 kJ/mol vs. 27 kJ/mol). For acrylic acid the difference is only about 5 kJ/mol (59 kJ/mol-1 versus 54 kJ/mol), but the 1,4-addition is still favored. Tezer and Ozkan[355] came to similar conclusions.

If the reaction proceeds along the 1,4-addition, the final product is only reached after a ketonization of the enol form. In the studies mentioned above the influence of possible barriers of this step was not considered. Experimentally, the barrier of the acid-catalyzed ketonization of acetone enol in water was determined to 40 kJ/mol. The energy difference between the enol and the keto form was found to be -43 kJ/mol.[356] Comparison to gas phase data show, that this energy difference depends little on the surrounding, but the actual course of the reaction and the kinetics is strongly influenced.[357–360] Recent computations reveal that the barrier of the ketonization strongly depends on the nature of a water bridge, which facilitates the proton transfer. If the bridge consists of three water molecules the computed barrier height is about 73 kJ/mol while a value of more than 210 kJ/mol is computed for the direct transfer. A similar value was also computed by Lien and co-workers.[361,362] If more than three water molecules are added, the water bridge is solvated itself and the barrier increases again. It is noteable that the computations did not consider the proton catalyzation. Further work also showed that in some cases the energy differences between keto- and enol forms depend on the molecular structure of the solvent.[363,364]

Due to the strong dependence on the actual number of water molecules, the computation of the activation energy for this complicated process is too tedious for the goal of the present work which wants to study the influence of various substitution patterns. Nevertheless, since the corresponding barriers are similar or perhaps even higher than the barriers of the addition step, the ketonization can be expected to influence the inhibition potency of a given inhibitor strongly.

The influence of this part of the reaction path is estimated here by the enolate intermediate E^- and the enol cation E^+ (Fig. 5-16). The enolate E^- represents the most important intermediate if the reaction takes place in a quite basic solution (first deprotonation, then addition). The cation E^+ is important for a quite acid environment in which the deprotonation and the protonation step will interchange.

5.3.2 Screening of Substitution Patterns

The determination of relative energies, which are not biased due to hydrogen bonding networks is quite complicated. Since the focus of this *in silico* screening is to identify trends arising from the substitution pattern, it is necessary to compute relative energies without the influence of the network. A suitable reference for all energy considerations is E_0 which represents the energy of the model system in which all reactants and all explicitly accounted solvent molecules are infinitely separated so that no hydrogen bonding network exists. Since the *s-trans* form is usually slightly lower in energy, E_0 is defined as:

$$E_0 = E_{s\text{-}trans}^{\text{reactant}} + E^{\text{methylthiol}} + E^{\text{NH3}} + E^{\text{NH4+}} \tag{43}$$

The energies of the ammonia and of the ammonium ion are necessary for the comparison with the computed PES. Relative energies with respect to E_0 which also do not account for the influence of the hydrogen bonding network arising due to the explicit solvent molecules are abbreviated with $E_0(X)$. As example the relative energy of the enol form E (Fig. 5-16) without considering the influence of the hydrogen bonding network is given as:

$$\Delta E_0(E) = [E(E) + E^{NH3} + E^{NH4}] - E_0 \tag{44}$$

The square brackets in equation 44 indicate that the respective parts of the model system were calculated as a whole complex and not as separated parts. Corresponding computations were

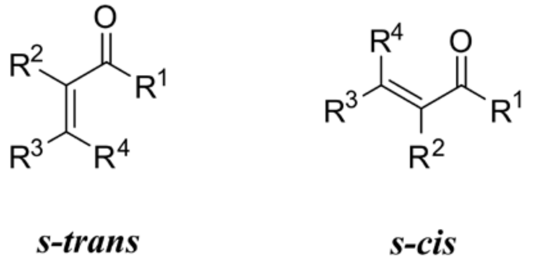


Fig. 5-19: Definition of *s-trans* and *s-cis* isomers and substitution pattern as used in Tab. 5-7.

performed for all other intermediates depicted in Fig. 5-16. The resulting energies computed for the compounds **1-11** are summarized in Tab. 5-7. The *s-trans* and the *s-cis* isomers are defined in Fig. 5-19.

In the first model, the inhibitor is represented by acrolein **1** which represents the simplest Michael system. The potential energy surface for the reaction branch up to the enol form E is depicted in Fig. 5-20 (left hand side). It is computed as a function of the distances between the centre of methyl thiol and the C3 center of the inhibitor (r(S-C), Fig. 5-17), and between the sulfur and the proton of the ammonium (r(S-H), Fig. 5-17).

compound	**R^1**	**R^2**	**R^3**	**R^4**	**conformer**	**R...**	**S$^-$.**	**TS2**	**E..**	**E$^-$.**	**E$^+$**	**K**
1	H	H	H	H	*s-trans*	0	43	-[a]	-20	30	103	-47
					s-cis	11	54	-[a]	-19	33		
2	Me	H	H	H	*s-trans*	0	43	59	-8	50	82	-51
					s-cis	2	45	58	-13	44		
3	Ph	H	H	H	*s-trans*	0	43	59	-10	43	72	-57
					s-cis	-4	39	-[a]	-19	31		
4	PhCl$_2$	H	H	H	*s-trans*	0	43	-[a]	-18	25	78	-52
					s-cis	3	46	-[a]	-19	20		
5	PhCl$_2$	Me	H	H	*s-trans*	0	43	58	-6	41	97	-31
					s-cis	11	54	-[a]	-8	42		
6	OMe	H	H	H	*s-trans*	0	43	-[a]	54	-[b]	97	-58
					s-cis	-2	42	-[a]	53	-[b]		
7	NMe2	H	H	H	*s-trans*	0	43	-[a]	55	-[b]	16	-70
					s-cis	-7	36	-[a]	39	-[b]		
8	Ph	-	Me	-	-	0	43	67	-5	35	-34	-118
9	Ph	H	Me	H	*s-trans*	0	43	76	10	64	94	-35
					s-cis	-4	39	66	0	52		
10	Ph	CH$_2$OR	H	H	*s-cis*	0	43	-[a]	-7	33	76	-51
11	PhCl	H	H	OPh	*s-trans*	0	43	-[a]	45	-[b]	138	1

[a] reaction path from S$^-$ to enolate E proceeds without additional barrier

[b] no stationary point found

Tab. 5-7: Summary of the relative energies, computed for the various intermediates discussed in Fig. 5-16 without considering hydrogen bonding networks. Results are partly taken from earlier work[365] and complemented by additional calculations. All values are given in kJ/mol.

Both coordinates are varied independently since the relationship between them is unknown. The S-H distance (r(S-H), Fig. 5-17) ranges from 1.0 to 3.0 Ǻ while the C3-S distance (r(S-C), Fig. 5-17) is varied between 1.7 and 3.7 Ǻ. For each point of the PES r(S-C) and r(S-H) are kept fixed at certain values while all other internal degrees of freedom are optimized. Transition states are identified as maximum points on the potential curves. Frequency analysis of transition states confirm their nature.

In this PES, the r(O-H) distance (Fig. 5-20, left hand side) is not treated as an independent variable. It represents the main coordinate of the protonation of the emerging intermediate, but during the optimization of the remaining internal degrees of freedom this coordinate smoothly adjusts for varying distances r(S-C) and r(S-H). To investigate the influence of r(O-H) in more detail, a two-dimensional PES for which r(O-H) is varied together with the r(S-C) is computed. The corresponding PES (Fig. 5-20 right hand side) indeed shows that r(O-H) smoothly adjusts. More information is given in the computational details section.

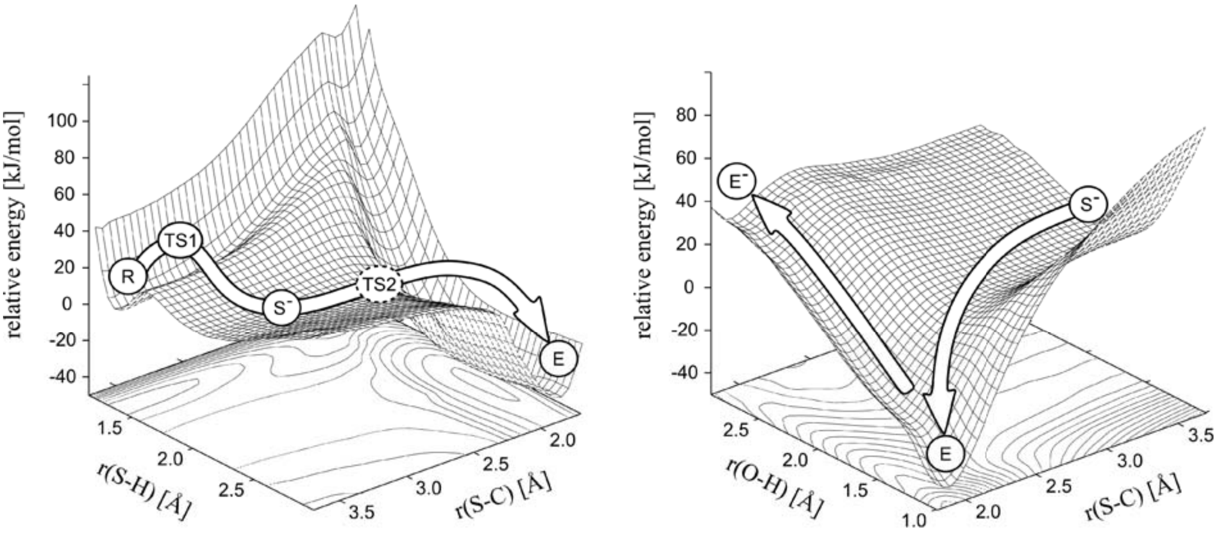


Fig. 5-20: Potential energy surfaces (PES) computed to characterize the reaction branch up to the enol formation for the acrolein model system **1**. Data is taken from earlier work.[365] The internal coordinates are defined in Fig. 5-17. Left hand side: PES obtained as a function of r(S-H) and r(S-C). Right hand side: PES computed as a function of r(S-H) and the r(O-H) distance.

The potential energy surface of the addition reaction (Fig. 5-20, left hand side) gives the qualitative picture of a two step mechanism. As first step, the deprotonation of the thiol reactant R takes place, leading to a zwitterionic state consisting of a thiolate anion CH_3-S^- which interacts with an ammonium ion (NH_4+···$^-$S–CH_3). In the following this state will be abbreviated as S^-. It lays 12 kJ/mol higher than the reactants. The deprotonation process possesses a barrier (TS1) which is computed to 18 kJ/mol. Fig. 5-20 (left hand side) shows that the shape of the potential energy curve describing the deprotonation step is nearly independent from the second reaction coordinate r(S-C) for r(S-C) > 3 Å. For smaller distances the barrier flattens. Nevertheless, such paths remain energetically more unfavorable than deprotonation paths with r(S-C) > 3 Å since they start from considerably higher energies. The approach of the protonated thiol to the Michael system is repulsive (Fig. 5-20, left hand side). The subsequent conjugated addition reaction at the C3 atom of the inhibitor proceeds to the enol intermediate E without passing a second transition state TS2, showing that the prior deprotonation step is the rate determining step. During this reaction step, the carbonyl oxygen atom of acrolein **1** is simultaneously protonated by the proton donor ammonium. This is also shown by the PES which is obtained, if r(S-H) and r(O-H) are varied (Fig. 5-20, right hand side). The whole reaction to the enol possesses an exothermicity of only -45 kJ/mol. For the addition of the deprotonated thiolate to the Michael system a reaction energy of -57 kJ/mol is predicted. The backward reaction is in both cases easily possible.

bond	R	TS1	S^-	E
S-C	3.70	3.70	3.70	1.90
S-H	1.40	1.60	2.20	2.20
C1-O	1.24	1.25	1.25	1.36
C1-C2	1.46	1.45	1.45	1.35
C2-C3	1.36	1.35	1.35	1.49

Tab. 5-8: Geometric parameters of the model system acrolein **1** for selected points on the potential energy surface,[365] as given in Fig. 5-20. Data is taken from earlier work.[365] The atom numbers are defined in Fig. 5-17. All bond lengths are given in Å.

An analysis of selected bond parameters within the Michael system at the characteristic points of the potential energy surface is given in Tab. 5-8. The bond lengths within the nearly undisturbed Michael system, referring to a S-C distance of 3.7 Å, agree well with the experimentally determined counterparts,[203] reported as 1.22 Å for the C1-O carbonyl bond, 1.49 Å for the C1-C2 single bond, and 1.35 Å for the C2-C3 conjugated double bond. The product of the conjugated addition shows the typically observed bond lengths referring to a sulfur-carbon single bond C1-S, a C2-C3 single bond, a C1-C2 double bond, and a C1-O single bond.

The exothermicity of the reaction strongly depends on the proton donor molecule in the vicinity of the inhibitor´s carbonyl oxygen atom.[365] If the ammonia molecule is replaced by a water molecule the inhibitor is not protonated during the addition reaction and the exothermicity lowers by about 40 kJ/mol. Furthermore, the conjugated addition step possesses a barrier (TS2) of about 3 kJ/mol, therefore the deprotonation of the thiol remains the rate-determining step. While the energies change considerably, the corresponding geometrical parameters resemble those found before. Without any proton donor, the conjugated addition becomes the rate-determining step with a barrier height of 24 kJ/mol for the second transition state TS2. The reaction energy of the complete reaction is +5 kJ/mol, therefore the reaction becomes endothermic.[365] The geometrical parameters again resemble those found for the previous systems. The strong influence of the proton donor underlines the importance of a proper orientation of the carbonyl oxygen atom of the Michael system towards a stabilizing group, like the oxyanion hole.

The influence of the hydrogen bonding network is reflected by the relative energies from the PES scan (Fig. 5-20) compared with those calculated by equation 44. If the influence of the hydrogen bonding network is switched off, the energy difference between the reactant and the intermediate S^{-} increases to about 40 kJ/mol. The enol form E is predicted to lie about -20 kJ/mol below the reactants. These values can be compared with the relative energies for the ketonization step (enolate E^{-} or enol cation E^{+}) to get information about the complete 1,4-addition mechanism. The enolate E^{-} is found to lie about 10 kJ/mol below the S^{-} intermediate. Hence, the computations predict that for acrolein **1** the proton transfer leading to the S^{-} intermediate represents the rate-determining step. This is in line with previous findings. [354,355,361,362] For the addition of ammonium to acrolein **1**, previous computations also indicate that the 1,4-addition mechanism takes place and that the addition step is rate-determining. This also supports that the herein chosen approach is sufficiently accurate.

By going from acrolein **1** to other α,β-unsaturated carbonyl compounds, changes in the relative energies and mechanisms can be expected, since substituents are known to influence the inhibition potency of Michael system based inhibitors. In order to study such effects, the influence of different substituents of the Michael system on the relative energies of the intermediates are investigated, as depicted in Fig. 5-16. The results obtained for the compounds depicted in Fig. 5-18 are summarized in Tab. 5-7. The *s-cis* form of acrolein **1** (Fig. 5-19) is about 10 kJ/mol higher in energy than the *s-trans* form. The corresponding barriers are very similar so that acrolein **1** is expected to react in the *s-trans* form. The addition of thiolate S^{-} to the α,β-unsaturated carbonyl compound with R^1 = Me possesses an additional barrier of 10 to 15 kJ/mol which is not present for acrolein **1**. Together with the energy which is necessary to form S^{-}, the total barrier for reaching the enol form E increases to about 60 kJ/mol. The relative energy of the enolate is computed to about 50 kJ/mol. Hence, the ketonization is also predicted to be slower than for acrolein **1**, but the addition of the thiolate remains the rate-determining step. The increase with respect to the unsubstituted acrolein can be explained by the +I-effect of the methyl group. This effect also explains why the enol cation E^+ becomes about the same degree more stable, but it is still considerably higher in energy than the enolate E^-. The barriers computed for the *s-trans* and the *s-cis* isomers are virtually identical which represents an additional difference to the unsubstituted counterpart. In summary, the methyl substituent decreases the reactivity of the α,β-unsaturated moiety, but the overall mechanism is not expected to change.

For compound **3** (R^1 = Ph) similar effects are seen, but in difference to compound **2** the *s-cis* form becomes slightly more stable than the *s-trans* form. Furthermore, while the addition step for the *s-trans* form possesses a barrier (TS2 = 59 kJ/mol) no barrier for the *s-cis* form is computed. This could result from mesomeric effects between the carbonyl group and the phenyl group, but the latter one is rotated with respect to the α,β-unsaturated moiety so that mainly the inductive effect of the phenyl group remains. The size seems to be comparable to the effects seen for R^1 = Me. The differences to compound **2** result from steric effects which are mainly present in the *s-trans* form. Going from compound **3** to **4** the +I effects should be diminished due to the electron withdrawing effects of the chlorine substituents. Such effects are mainly seen for the *s-trans* form for which the barrier of the addition step disappears. Furthermore, enol E and enolate form E^- become more stable. The steric effects strongly increase if one goes from compound **4** to **5**. The *s-cis* form lies even higher in energy and the barriers of the addition step increases as well as the enolate intermediate. In summary **5** is expected to be similar reactive as **3**, thus the advantages gained through the chlorine substituents are lost due to the additional Me group. This indicates that many of the effects are additive.

Despite these variations the overall mechanism remains the same for compounds **1**-**5**. In all cases the formation of the thiolate in combination with the addition to the Michael system represents the rate-determining step. The ketonization process possesses lower lying intermediates and seems to proceed across the enolate E^- which is considerably lower in energy than the enol cation E^+. This does not seem to be the case for the remaining compounds. For compound **6** the barrier for the addition step is similar to the other systems, but the necessary ketonization process seems to be impossible. The enol is already considerably less stable than the reactants (+ 55 kJ/mol) and the enolate is predicted to be not stable. The enol cation lies at about + 100 kJ/mol so that a possible ketonization would represent the rate-determining step. However, test computations performed for the direct E2-like addition in which the proton donor was placed in the vicinity of C2 indicated considerably lower barriers at about 60 to 70 kJ/mol. Hence, for R^1 = OMe, the computations indicate a direct addition. This switch in the mechanism is in line with previous experimental findings of Miyata et al.[366] who explained stereospecific nucleophilic additions of thiols to derivatives of α,β-unsaturated carboxylic acids by a fast protonation of the arising enolate at C2.

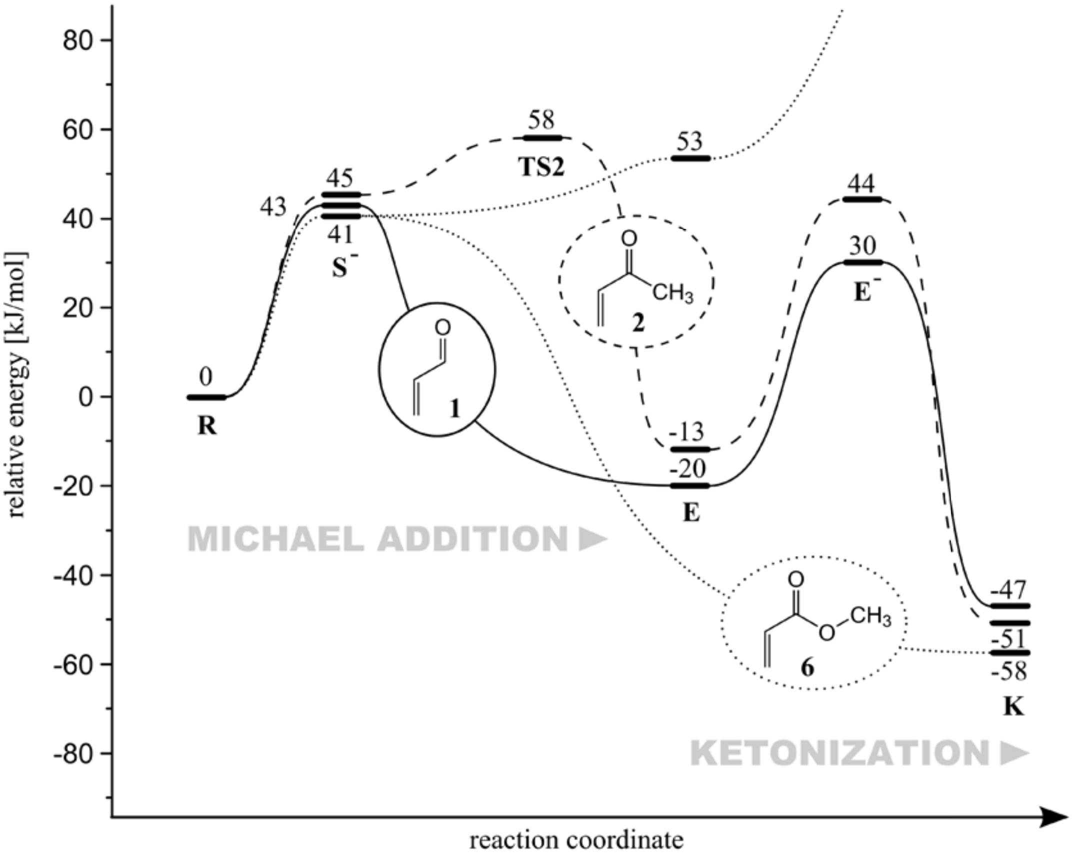


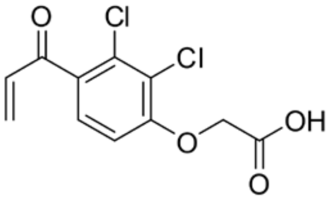
Fig. 5-21: Sketch of the three different mechanisms found for the reaction of α,β-unsaturated carbonyl compounds with thiols. Prototype compounds for the different situations are also given.

A ketonization across the enolate form is also quite unfavorable for compound **7**. In this case, however, the electron pushing properties of the NMe_2 group stabilizes E^+ in such a way that the ketonization should be easily possible. Also for compound **8**, the addition step is predicted to represent the rate-determining step and the ketonization will proceed across the enol cation. Due to its strong stability, the enol cation is expected to be directly formed. In general, compound **8** represents an outlier due to its different electronic structure, since it contains a conjugated triple bond instead of a double bond. Compound **9** is characterized in order to investigate in combination with compound **3** how R^3 = Me influences the reactivity. According to the computations, this substitution should lower the reactivity since the barriers associated with the 1,4-addition and the ketonization step become higher. Compound **10** and **11** enable to experimentally distinguish between the reactivity of *s-cis* and *s-trans* forms, since the two structure are sterically hindered in rotation around the *s*-bond between carbonyl function and double bond and hence fixed in one form. The computations predict that the *s-cis* form (compound **10**) reacts considerably faster than the *s-trans* form. However, the difference results mainly from the ketonization step. While it should proceed quite fast for **10**, it should not take place for **11**.

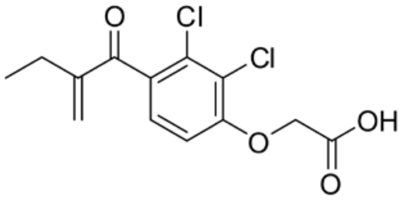
Taking the results together, the computations point to three different mechanisms. The prototypes of these mechanisms are sketched in Fig. 5-21. For acrolein **1** the formation of the thiolate S^- is the rate-determining step, while for its addition to the α,β-unsaturated carbonyl moiety, no additional energy barrier has to be overcome. Compound **2** represents a prototype for a mechanism for which the barrier of the conjugated addition step represents the highest point on the total reaction path. This reaction type is found for various alkyl substituted compounds. Finally, for compounds like the α,β-unsaturated carboxylic ester **6**, the ketonization step is so unfavorable that a direct addition is assumed to take place.

5.3.3 Comparison to Experimental Results

The theoretical insights gained from the *in silico* screening of the different α,β-unsaturated compounds point towards three different mechanism. Although these are quite reasonable and conclusive, they should be evaluated by comparing them to experimental observations. Some data from the literature are collected in Tab. 5-9 that correspond to most of the herein investigated compounds. The etacrynic acid compounds **12** and **13** (Fig. 5-22) are augmented versions of the compounds **4** and **5** (Fig. 5-18) and include the respective warheads.



12



13

Fig. 5-22: Etacrynic acid derivatives **12** and **13** that contain the electrophilic warheads of compound **4** and **5**.

The comparison of the experimental data with the calculated results provides information whether the chosen theoretical model captures the main effects which determine the reactivity of the α,β-unsaturated carbonyl compounds.

compound	turnover in % (time)	k_2 in $M^{-1}min^{-1}$	RC_{50} in mM	reference
1	-	-	0.086	Böhme et al.[340]
2	-	-	0.090	Böhme et al.[340]
6	-	-	0.55	Böhme et al.[340]
9	88% (12 min) >99% (42 min)	-	-	Schiller[367]
10	>99 (12 min)	-	-	Schiller[367]
11	<5% (2 days)	-	-	Schiller[367]
12	>99% (12 min)	-	-	Schiller[367]
13	63% (12 min) 90% (42 min) >99% (90 min)	25.6	-	Schiller[367]

Tab. 5-9: Experimental results for the reactivity of the considered Michael acceptor model compounds as obtained from UV-vis and NMR spectroscopy. The turnover rates and rate constants of second order k_2 (equation 5) measure the reactivity with respect to the model nucleophile thiophenol. RC_{50} values denote the concentration that produce 50% reaction with glutathione in 120 minutes.

According to the computations in Tab. 5-7, an alkyl substitution at C2 (Fig. 5-19) increases the reaction barrier of the addition step to the enol form (+15 kJ/mol) and also shifts the enolate intermediate to higher energies (+16 kJ/mol), as the comparison of compound **4** and **5** shows. This is reflected in the decrease of experimentally determined turnover that is found if one goes from compound **13** to **12** (Tab. 5-9). Further experimental results from the literature confirm this trend, as shown by the work of Schüürmann and co-worker in two examples.[340] In the first example, the comparison of 2-cyclopentene-1-one and 2-methyl-2-cyclopentene-1-one corresponds to a similar alkyl substitution at the C2 position. The respective RC_{50} values increase from 0.58 to 8.40 and going from methyl acrylate to methyl methacrylate the value increases from 0.42 to 74.1. These trends are moreover reported by Schultz et al.[368,369]

As a second trend, the computations predict that an alkyl substituent at the C3 position (Tab. 5-7, **3** versus **9**) will further decrease the reactivity of the α,β-unsaturated carbonyl compound, since both, the barrier of the addition (+ 27 kJ/mol) and the relative energy of the enolate (+ 21 kJ/mol) increase. This prediction is also reflected in the decrease of the RC_{50} values by going from 1-pentene-3-one to 4-hexene-3-one.[340] A similar effect is also described by Schultz et al.[368] who came to the conclusion that a methyl substitution at the olefin moiety reduces the activity.

The well known experience that aldehydes are more reactive than the corresponding ketones is reflected in the difference computed between **1** and **2** and can be furthermore found in experimental data.[368,369]

A key finding of the *in silico* screening is the prediction that derivatives of α,β-unsaturated carboxylic esters are less reactive than the other counterparts. The reason for this can be found in the ketonization step that is strongly disfavored and steers the reaction path towards a direct addition. Coincidence for this is given by the explication from Miyata et al.[366] for stereospecific nucleophilic additions of thiols to derivatives of α,β-unsaturated carboxylic acids. An experimental confirmation of the predicted decrease in the reactivity is again given by Schüürmann et al.[340] who found decreasing RC_{50} values for the model compounds 1-pentene-3-one (0.09) and methyl acrylate (0.42) that represent a similar comparison.

A further issue from the theoretical predictions is the comparison between *s-trans* and *s-cis* which is an inherent conformational feature for most of the α,β-unsaturated carbonyl compounds. In many cases the computations predict that the *s-cis* forms possess a higher reactivity than their *s-trans* counterparts. Compounds **10** and **11** are exceptional and valuable variants of Michael acceptors, since they possess fixed conformations with respect to the relative orientation of carbonyl group and double bond, due to their cyclic structure. This enables an experimentally observation of the postulated *s-trans* and *s-cis* reactivity relationship. The results in Tab. 5-9 show that **10** is indeed very reactive. However, for **11** no reaction takes place. This observation is nicely reflected by the computed results that explain the astonishing stability of **11** by very high barriers for the ketonization step. Since the addition that leads to the enol form is strongly endothermic, no reaction will take place. That the herein employed simplified model can explain even this unexpected behavior, indicates that it indeed captures the main aspects that are important for the reactivity of α,β-unsaturated carbonyl compounds with respect to thiol containing enzymes.

Experimental inhibition assays for SARS-CoV M^{pro} and related enzymes show that some Michael acceptors inhibit the thiol-containing enzymes irreversibly, while for some others only reversible inhibition is found. Since all agents possess an α,β-unsaturated carbonyl derivatives as electrophilic warheads, the question remains why this is the case. For the herein performed investigations, two possible explanations can be concluded. In some cases, e.g. for compound **5** or **9**, the reaction energy is estimated to be exothermic in the range of 30 kJ/mol. In such cases a back reaction is much slower than the forward reaction but may still occur. In other cases the reaction energies are sufficiently high to impede any back reaction. Even in such cases a reversible inhibition may occur since, complicated multi-step inhibition mechanisms have to take place in order to reach the final keto form. For all compounds, a base catalyzation by a neighboring proton acceptor is necessary, since only the thiolate form is sufficiently reactive for the addition reaction, as shown in Fig. 5-20. The second step consists in most cases of a 1,4-addition leading to the enol intermediate. The final keto form is then obtained through a ketonization of the enol form. This ketonization, as shown in previous investigations, is only possible if the necessary proton transfer is catalyzed somehow, e.g. by an efficient water bridge. Since such catalysis cannot take place in the active-site of many enzymes, the reaction might get stuck at the enol intermediate. Another possibility, if no proton donor is available, would be that the reaction might become stuck in an enolate intermediate, which is strongly stabilized by an available oxyanion hole. In such cases the computations predict considerably smaller reactions energies. Hence, the back reaction can take place easily and only a reversible inhibition is observed after all. Especially for derivatives of α,β-unsaturated carboxylic acids, the enol form is so high in energy that a direct addition to the vinylogue double bond becomes more favorable than the 1,4-addition. For the direct addition, an efficient proton donor is necessary in the vicinity of the double bond, since the reaction proceeds E2-like. However, as obvious from many enzyme X-ray structures, this role could be taken by an assisting histidine residue, which would also serve as the base in the first step of the overall reaction. If one assumes that a histidine residue can play the double role of an acceptor and donor, and that the electronic structure of the warhead furthermore steers the reaction into the direction of a direct addition, the final keto form can be reached without a ketonization process from the enol form. Only in such cases, an irreversible inhibition would be found. In other cases, the reaction can only proceed until the enolate is formed and consequentially a reversible inhibition is found.

According to the computations, such a direct addition is more favorable for α,β-unsaturated carboxylic esters, as provided by the inhibitor AG7088 that shows irreversible inhibition of the rhinovirus 3C protease.[149] Whenever alkyl or phenyl substituents are present at the Michael acceptor moiety, the 1,4-addition is predicted to be more favorable. The experimental confirmation for this finding represents the inhibition behavior of etacrynic acid derivatives that inhibit SARS-CoV M^{pro} in a reversible fashion.[350]

5.3.4 Conclusions about the Michael Addition Reaction as Inhibition Strategy

The performed work in this chapter develops a simplified model system to describe the reaction of α,β-unsaturated carbonyl derivatives with the catalytic cysteine thiol of an enzyme. As a key finding, the computed potential energy surface reveals that the addition of thiols to acrolein is only possible with a previously occurred base-catalyzed process. This previous step is found to be unavoidable for an efficient addition.

The applied model system is moreover suitable to perform *in silico* screenings that are hardly to achieve with QM/MM approaches. The performed calculations can explain the various trends found in chemoassays, which are a commonly employed tool to estimate the excess toxicity of this important class of compounds.

Finally, the *in silico* screening provides an explanation why, depending on warhead and enzyme, a reversible or an irreversible inhibition of the enzymatic active-site thiol is found for the different substitution patterns of the considered Michael acceptors.

From the obtained data, the conclusion can be drawn that an irreversible inhibition is only found when the reaction directly steers to the final keto form. The computations predict this behaviour for the unsaturated carbonyl esters. If the reaction proceeds across the 1,4-addition, which is predicted for most of the derivatives, the reaction will get stuck in the enol or in a stabilized enolate form. In this case, the necessary ketonization step can be expected as strongly hindered in active-site of the protein. Thus, the inhibition is observed as a reversible one.

Conclusively, the different substituent effects and their impact on the inhibition mechanism can be used to steer the Michael addition reaction into a desired direction. A possible application would be the “fine tuning” of the reactivity of a covalent reversible inhibitor. As demonstrated here, the theoretical calculations performed with the applied model system are a successful tool for such a rational drug development process. They give not only investigative insights, but also provide predictive power to some extend.

5.4 Revealing the Binding Mode of a SARS-CoV M^{pro} Inhibitor

The so far conducted model calculations in chapter 5.1 and 5.2 relied mostly on initial structure information that are obtained by X-ray crystallography. The latter is a routinely applied tool for structural chemists since the nineteen sixties.[370] However, in cases of protein structure determination, X-ray data has to be refined and interpreted with computational methods to obtain a comprehensive structure. The basis for the X-ray scattering experiment is a crystallized sample of the protein or protein-inhibitor complex from which an electron density map is determined.

Reliable electron density maps of good quality are obtained by the subsequent structure refinement. This is typically a straight forward process today, but requires clear defined electron densities from the scattering experiment. In cases where the electron density map exhibits unclear or “smeared” regions, a successful refinement of the structure strongly depends on the interpreting experimentalists and their chemical intuition. The following chapter develops a workflow that can support the structure refinement process in such difficult cases by atomistic molecular modeling techniques.

The problem to be solved refers to a non-covalent protein-inhibitor complex of SARS-CoV M^{pro}, where the inhibitor (TS174) is only represented by a smeared electron density distribution in the density map.[371,372] Moreover, TS174 is known to have a low binding affinity. Due to the lack of further structure information, the inhibitor poses cannot be uniquely determined and have to be more or less guessed. In contrast to the inhibitor, the positions of the protein atoms are well defined in the electron density maps, therefore the problem is expected to be due to the binding properties of the inhibitor rather than due to deficiencies in the X-ray experiment. This is further supported by several revisits of the experiment with similar outcomes.

Hence, the question has to be answered why the inhibitor is smeared in the electron density map, and how the low binding affinity can be explained. In the following, a theoretical investigation is conducted that should give insights into this problem and provide ideas for improvements in the structure of TS174.

5.4.1 Development of a Workflow to Support X-Ray Structure Refinement

It has been demonstrated that QM/MM computational modeling approaches can be a valuable tool for the refinement of X-ray structures and the understanding of ligand binding.[373–376] MD simulation studies have also been applied for SARS-CoV M^{pro} inhibitors.[128] The interplay between experiment and theory leads hereby to a better understanding of the protein-ligand interactions and helps to refine the obtained structures. However, these approaches are inappropriate to create and propose totally new poses.

The approach that is developed and applied herein has therefore a special emphasis on generating new poses. It furthermore probes the reasonability of the proposed and generated poses by their relative energies, their deviation between initial and simulated structure, as well as their fluctuation during simulation.

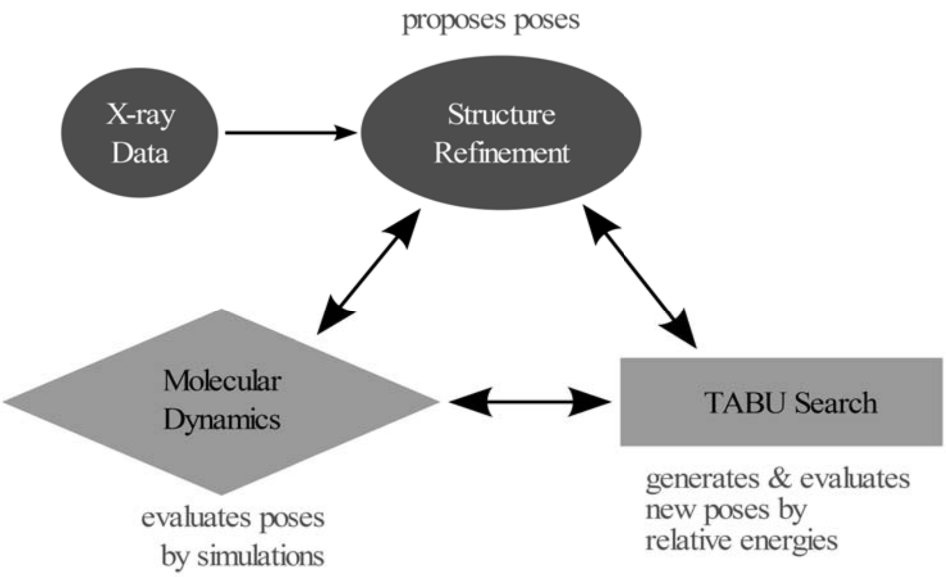

proposes poses
X-ray Data
Structure Refinement
Molecular Dynamics
TABU Search
evaluates poses by simulations
generates & evaluates new poses by relative energies

Fig. 5-23: Workflow scheme for developing an improved inhibitor pose structure refinement. An initial set of inhibitor-complex structures is proposed from experimental X-ray data. The set is augmented by further proposals generated by a TABU search based conformational search algorithm. The whole set of inhibitor poses is evaluated by molecular dynamic simulations that yield a ranking of the best structures. The whole cycle can be repeated in both directions, as an iterative process.

As a precondition, the workflow must be capable to handle a bunch of structure proposals that are furthermore simulated individually. Hence, molecular mechanics is chosen as theoretical basis (CHARMM force field) and used for generating and simulating the structure proposals. An overview about the workflow is shown in Fig. 5-23.

The workflow starts from an educated guess for possible inhibitor poses, which were given by the experimentalist. Based on this set of structures, further poses are proposed by a conformational search algorithm. The rapid screening of different inhibitor poses is typically a domain of docking methods that rely on simplified scoring functions and often suffer from bad performance.[377] However, it has been shown that molecular mechanics can in many cases significantly improve this deficiency.[378] The tabu search algorithm is a very efficient method for conformational searches with molecular mechanical methods and is used in the workflow to generate the new inhibitor poses.[379] The evaluation of the different inhibitor poses is done by relative energy calculations and analysis of molecular dynamic simulation results. This allows finally a pragmatic rating of the quality of a pose.

5.4.2 Theoretical Evaluation of 22 Inhibitor Pose Proposals

The chemical structure of TS174 is shown in Fig. 5-24. It possesses 4 important chemical substituents: a methyl-cylclohexyl group (CHEX), an isopropyl group (ISO), a pyridine group (PYR), and a 2,3-butene group (BUT). All are attached to a peptide backbone (BACK). The inhibitor has a chiral center at the carbon atom connecting PYR, BUT, and BACK. The experimental investigations were performed with a diastereomeric mixture, i.e. both stereoisomers have to be considered.

Some selected inhibitor poses are summarized in Tab. 5-10. Most of the proposed structures are educated guesses from initial X-ray structure refinements, suggested by Uwe Dietzel from AK Kisker.[372] They consist of 17 poses that are generated as different inhibitor substituents-binding pocket combinations. The predominantly occupied binding pockets of SARS-CoV M^{pro} are S1`, S1, S2, and S3 that can be permuted with the different inhibitor substituents CHEX, ISO, PYR, and BUT. The structure series **or1-X** and **or2-X** differ in the orientation of TS174 with respect to the binding site, and denote respectively twisted orientations of 180 degrees. They were developed for the *S*-configuration. The **orsi-X** series of structures denotes stereo isomers with *R*-configuration of TS174 for the indicated chiral center in Fig. 5-24.

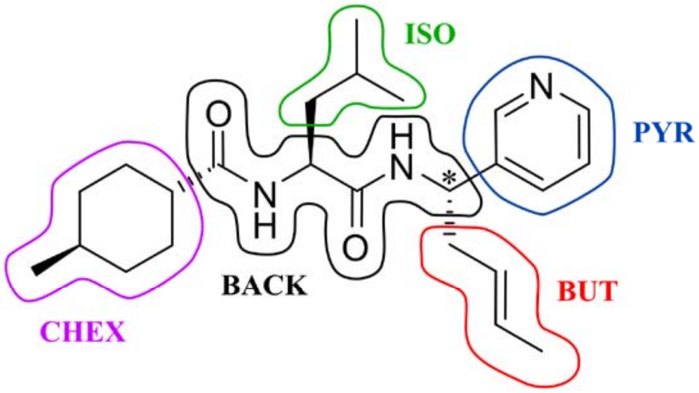


Fig. 5-24: Structure of SARS-CoV M^{pro} inhibitor TS174 and definition of five inhibitor fragments, namely cyclohexyl moiety (CHEX), isopropyl groupsubstituent (ISO), pyridine ring (PYR), 2,3-butene (BUT), and inhibitor backbone (BACK). Note that the inhibitor is available as diastereomeric mixture with a chiral carbon atom (*).

Structures abbreviated as **ts-X** were generated by a tabu search based conformational search algorithm, i.e. these represent *in silico* developed pose proposals. The starting points for the five poses stem from the initial X-ray structure refinements and are indicated by the name for **X** in the **ts-X** nomenclature. The tabu search employs the global minimum force field energy as a criteria to find the best structure proposal along a conformational search run. All tabu search pose proposals were taken from the work of Christoph Grebner.[380]

All 22 pose proposal were compared for the following three properties:

1) Relative energies of the solvated and minimized protein-inhibitor complex
2) Averaged root mean square deviations (RMSD) of the TS174 inhibitor as a whole and of the single substituents taken from a 5 ns MD trajectory
3) Root mean square fluctuations (RMSF) of individual inhibitor substituent-binding pocket distances

In the first step, all 22 pose proposals were set up and minimized as a solvated model system of the protein-inhibitor complex with a diameter of 110 Å. The detailed procedure is given in the computational details section. Before running the molecular dynamic simulations, an analysis of the relative energies for the minimized model systems was performed. The results are all shown in Tab. 5-10.

pose	E(tot)	E(prot-wat)	E(inh)	E(tot)-E(wat+prot)	**ΔE(inh)**	**ΔE[(tot)-(wat+prot)]**
	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol	kcal/mol
or1-1	-281749	-281692	26	-57	**2**	**18**
or1-2	-281700	-281658	39	-42	**15**	**33**
or1-3	-281931	-281898	38	-33	**13**	**42**
or1-4	-281770	-281713	43	-57	**19**	**18**
or1-5	-281633	-281606	53	-27	**29**	**48**
or1-6	-282114	-282071	35	-43	**11**	**32**
or2-1	-281909	-281855	38	-55	**14**	**20**
or2-2	-281631	-281611	47	-21	**23**	**55**
or2-3	-281732	-281696	42	-36	**18**	**40**
or2-4	-281831	-281795	46	-36	**22**	**40**
or2-5	-281971	-281933	34	-38	**10**	**37**
or2-6	-281717	-281682	42	-34	**17**	**41**
orsi-1	-281721	-281673	30	-48	**6**	**27**
orsi-2	-281951	-281894	45	-57	**21**	**18**
orsi-3	-281984	-281934	37	-50	**13**	**26**
orsi-4	-281856	-281811	32	-45	**8**	**30**
orsi-5	-281676	-281636	33	-39	**9**	**36**
ts-or1-3	-282575	-282508	24	-67	**0**	**9**
ts-or1-4	-282567	-282496	29	-71	**5**	**4**
ts-or1-5	-282239	-282181	29	-58	**5**	**17**
ts-or2-1	-282576	-282500	25	-75	**1**	**0**
ts-orsi-3	-282552	-282490	31	-62	**7**	**13**

Tab. 5-10: Energies E of all minimized structure proposals as initially obtained from X-ray refinement and tabu search based conformational search. Relative energies ΔE are calculated with respect to pose of lowest energy.

The energy E(tot) represents the force field energy of the whole system. It is not suitable to judge about the quality of a pose, since changes due to the inhibitor are drowned out by the “noise” of protein and water interactions. The molecular strain of the inhibitor itself is expressed by the energy contribution of the isolated inhibitor E(inh) and can be used to calculate relative energies ΔE(inh). Nevertheless, these values have to be interpreted with caution, because the interactions with the environment are missing. The latter might overcompensate the molecular strain and make an unfavorable inhibitor pose more preferable.

The total system can be dissected into two parts that comprise the energy of the protein and water shell E(prot+wat) and the energy of the total system E(tot). This allows to calculate a differential energy value E[(tot)-(prot+wat)] that accounts for strain effect of the inhibitor and the inhibitor interactions with the environment.

The relative energy ΔE[(tot)-(prot+wat)] can be correlated with the thermodynamic stability of a given pose, since it includes the influence of the water shell and the protein interactions with the inhibitor.

A survey of the molecular strains of the 22 inhibitor poses in terms of their relative energies ΔE(inh) reveals the poses **or1-1**, **ts-or1-3**, and **ts-or2-1** as the three most relaxed structures. They differ by less than 2 kcal/mol. The picture changes if interactions with the environment are taken into account. The ΔE[(tot)-(prot+wat)] values reveal **ts-or2-1** > **ts-or1-4** > **ts-or1-3** as the ranking of the most stable poses. The energy variations span a window of about 9 kcal/mol.

The group of tabu search generated structures generally tends to lower energies than the educated guesses from the X-ray structure refinements. This possibly expresses a good performance of the tabu search algorithm, but could also result from a methodological bias effect. A conceivable reason for such an artifact represent the different energy functions that are used for X-ray structure refinement and tabu search based conformational search. Since the last one uses the same molecular mechanical energy expression as employed for calculating the relative energies, it naturally comes closer to the found results than the X-ray structure refinements that use a different energy expression.

In the next step, MD simulation with 5 ns duration of each pose were performed and trajectories determined. The acquired data was analyzed for root mean square deviation RMSD(inh) of the TS174 inhibitor, which gives information about how far the inhibitor drifts away from its initial position during the simulation (equation 17). The computed numbers can also be interpreted as how well the simulated structures match the initially proposed structures from X-ray refinement or tabu search.

In the third step of the property evaluation, individual structural parameters were analyzed. By this approach it becomes possible to get information about which one of the inhibitor substituents addresses which binding pocket and how strong it fluctuates. This information can be rather helpful for the rational drug design, since it obviously discriminates between the different inhibitor substituents, and furthermore estimates how the different substituent-binding pocket combinations behave during the dynamic simulation.

To achieve this, centers of mass were calculated for all substituents of the TS174 inhibitor (CHEX, ISO, PYR, BUT, BACK) and all relevant binding pockets of the SARS-CoV M^{pro} (S1`, S1, S2, S3). The underlying definitions for each center of mass are depicted in Fig. 5-24, and furthermore in Tab. 7-1, which is given in the computational details section. The reduction of the substituents and pockets to single points in space enables to calculate individual distances between them. Fig. 5-25 gives an example for the inhibitor substituent PYR and its individual distances to each of the binding pockets.

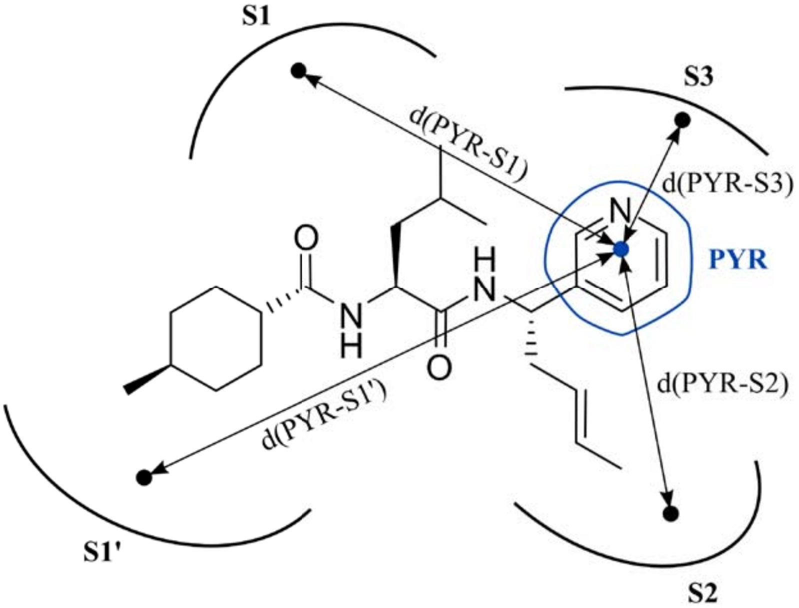


Fig. 5-25: Definition of distance between an inhibitor substituent and the binding pockets of SARS-CoV M^{pro}. Center of masses are determined for each entity and their distances can be measured for each individual permutation. In the example above, the d(PYR-S3) distance is the lowest one of all four binding pockets, hence, the pyridine group binds to the S3 pocket.

For example, the comparison of the four distances shows that the pyridine ring is located in the S3 pocket, since the d(PYR-S3) distance represents the lowest one. The colored matrix in Tab. 5-11 summarizes the conformational configuration for all 22 inhibitor poses.

The configuration of a pose delivers a picture of how the inhibitor binds to the enzyme, but nothing about the fluctuation within an individual inhibitor substituent-binding pocket combination during the MD simulation. This can be expressed by the root mean square fluctuation (RMSF) of the respective distance d(A-B) between inhibitor substituent and binding pocket.

$$RMSF = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(\mathbf{d(A\text{-}B)}^{\boldsymbol{i}} - \langle \mathbf{d(A\text{-}B)} \rangle\right)^2} \tag{45}$$

It measures the intrinsic range of fluctuation around the mean value $\langle$d(A-B)$\rangle$ over all time steps *t a*nd accounts for the spread of distance distributions.[381] Since introducing additional numbers into Tab. 5-11 would make the interpretation of the results quite cumbersome, the RMSF values are simply expressed by colors. The color coding in Tab. 5-11 is therefore related to the calculated RMSF of the monitored substituent-pocket distances d(A-B), as defined in Fig. 5-25. The colors reflect that the RMSF value of a specific substituent-pocket distance d(A-B) is in the lower field (green, <25%), the middle field (yellow, 25%-50%), or the higher field (red, >50%). The overall range covered by these fields is bordered by the maximum and minimum RMSF value observed for this specific distance d(A-B) over the whole set of MD trajectories. In case of the example shown in Fig. 5-25, this would refer to all d(PYR-S3) distances for all poses. It is therefore a relative measurement that discriminates between tight binding (relative RMSF is low), more flexible binding (relative RMSF is mean), and loosely bound (relative RMSF is high). The detailed distance and RMSF values are given in the computational details section in Tab. 7-2, Tab. 7-3, Tab. 7-4, and Tab. 7-5.

The comparison reveals the three most inflexible poses with respect to structure dynamics, namely **orsi-2**, **ts-or1-3**, and **ts-orsi-3.** They show RMSD values of 1.6 Å, 1.3 Å and 1.1 Å. Interestingly, this result does not reflect the poses of initially lowest energies (Tab. 5-10, Tab. 5-11), with exception of **ts-or1-3**. A closer inspection of this finding by correlation analysis (see computational details section, Fig. 7-8) shows that there is indeed no relationship between relative energy of the starting structure and RMSD value averaged over the 5 ns simulation. The same holds for the standard deviations of the RMSD values that reflect the structural fluctuation of the inhibitor during the simulation.

pose	S1'	S1	S2	S3	**RMSD (inh)**	**ΔE(inh)**	**ΔE[(tot)-(wat+prot)]**
						kcal/mol	kcal/mol
or1-1	🟩 BUT	🟥 CHEX	🟨 PYR	🟩 CHEX	**2.5 ± 0.5**	**2**	**18**
or1-2	🟥 BUT	🟥 ISO	🟥 PYR	🟥 CHEX	**4.0 ± 1.1**	**15**	**33**
or1-3	🟩 BUT	🟨 ISO	🟥 PYR	🟩 CHEX	**2.2 ± 0.6**	**13**	**42**
or1-4	🟨 PYR	🟩 CHEX	🟩 BUT	🟩 CHEX	**2.2 ± 0.6**	**19**	**18**
or1-5	🟩 PYR	🟨 CHEX	🟩 BUT	🟩 CHEX	**2.3 ± 0.3**	**29**	**48**
or1-6	🟨 PYR	🟩 ISO	🟩 BUT	🟨 ISO	**2.2 ± 0.3**	**11**	**32**
or2-1	🟩 CHEX	🟨 ISO	🟨 CHEX	🟩 PYR	**2.3 ± 0.4**	**14**	**20**
or2-2	🟥 ISO	🟥 PYR	🟥 CHEX	🟨 BUT	**4.2 ± 0.8**	**23**	**55**
or2-3	🟨 ISO	🟩 ISO	🟨 CHEX	🟥 BUT	**4.8 ± 0.5**	**18**	**40**
or2-4	🟥 ISO	🟨 PYR	🟨 CHEX	🟥 PYR	**5.2 ± 0.6**	**22**	**40**
or2-5	🟩 CHEX	🟨 BUT	🟩 ISO	🟨 BUT	**3.0 ± 0.5**	**10**	**37**
or2-6	🟥 CHEX	🟥 PYR	🟩 ISO	🟥 PYR	**2.8 ± 0.7**	**17**	**41**
orsi-1	🟥 BUT	🟩 ISO	🟥 PYR	🟩 CHEX	**1.7 ± 0.6**	**6**	**27**
orsi-2	🟩 CHEX	🟩 PYR	🟩 ISO	🟩 BUT	**1.6 ± 0.4**	**21**	**18**
orsi-3	🟥 PYR	🟨 ISO	🟩 BUT	🟩 ISO	**2.8 ± 0.6**	**13**	**26**
orsi-4	🟨 BUT	🟩 ISO	🟨 PYR	🟩 ISO	**2.1 ± 0.5**	**8**	**30**
orsi-5	🟩 BUT	🟨 CHEX	🟨 PYR	🟥 CHEX	**2.6 ± 0.8**	**9**	**36**
ts-or1-3	🟩 PYR	🟩 ISO	🟩 BUT	🟩 CHEX	**1.3 ± 0.3**	**0**	**9**
ts-or1-4	🟩 ISO	🟩 ISO	🟥 CHEX	🟥 BUT	**2.6 ± 0.9**	**5**	**4**
ts-or1-5	🟥 PYR	🟥 ISO	🟨 BUT	🟩 CHEX	**5.6 ± 1.5**	**5**	**17**
ts-or2-1	🟥 CHEX	🟥 PYR	🟨 CHEX	🟥 BUT	**4.9 ± 1.3**	**1**	**0**
ts-orsi-3	🟩 CHEX	🟩 ISO	🟨 CHEX	🟩 BUT	**1.1 ± 0.3**	**7**	**13**

Tab. 5-11: Initial inhibitor pose configurations with respect to the binding pockets of the SARS-CoV M^{pro}. The binding motifs are given in the colored matrix. The colors indicate the relative RMSF values for distances between the individual inhibitor substituents (CHEX, PYR, ISO, BUT) and binding pockets (S1', S1, S2, S3). Green indicates a low RMSF (tight binding), yellow a mean RMSF (flexible inhibitor substituent), and red a high RMSF (loosely bound) for the respective d(A-B).Averaged RMSD values are calculated from MD simulations and refer to heavy atoms of the inhibitor. The first simulation frame structure is taken as reference. Relative energies are taken from Tab. 5-10.

It is important to note that a correlation between RMSD or RMSF values and relative energies of the starting structure cannot be expected. High values result from very flat shaped potentials irrespective of its binding energy. Small values, in contrast result from very narrow minima, which might be quite high in energy. Furthermore, the RMSF values are calculated for the individual inhibitor substituents, while the binding energy sums up over all interactions. For example, for a high binding energy a strong binding of one or two inhibitor substituents may be sufficient. For a good ranking in the RMSD(inh) already one very floppy inhibitor substituent-binding pocket combination leads to a bad scoring. Calculated RMSD values for each individual inhibitor substituent do not even improve this, since they also show no obvious correlation with the RMSF values (see computational details, Fig. 7-9). Further investigations seem to be necessary to disentangle the various effects.

Nevertheless, a survey across the substituent based RMSF values in Tab. 5-11 reveals that there is a qualitative relationship between the RMSFs and the RMSD(inh) value of an individual inhibitor pose. This means that a low RMSD(inh) value is reflected by more “green” substituent contributions while higher RMSD(inh) values have more “yellow” or more “red” contributions, indicating that the substituent binding pocket distance fluctuates stronger during simulation. Conclusively, the three best poses with respect to the measured RMSD(inh) value, namely **orsi-2**, **ts-or1-3**, and **ts-orsi-3**, possess exclusively low RMSF values for all of their respective inhibitor substituent-binding pocket distances. Thus, qualitatively spoken, they bind tightly to the active-site of SARS-CoV M^{pro} over the whole simulation period. The only exception represents the CHEX group of **ts-orsi-3** that occupies a position in between the S1' and the S2 pocket and fluctuates in the middle range. It is important to note that the occupation of a position in between two binding pockets does not necessarily lead to higher fluctuation. The poses **or1-4**, **orsi-4**, or **ts-or1-4** are examples where the cyclohexyl or isopropyl group address an intermediate position between two binding pockets and possess low fluctuations.

By taking these results together, the **orsi-2**, **ts-or1-3**, and **ts-orsi-3** poses are the most likely of the whole set of 22 inhibitor-complexes, although this result is not fully reflected in the relative energy of the initial starting structures. Nevertheless, the pose **ts-or1-3** appears to be very reasonable, because it ranks among the most stable poses in terms of relative energy, as well as RMSD, and RMSF values obtained from MD simulations. However, a distinctive rating between the three best structures is not possible, since the differences between the averaged RMSD(inh) values are insignificant. This is particularly illustrated by the RMSD(inh) standard deviations in Tab. 5-11.

Interestingly, the comparison of the configuration of the three poses **orsi-2**, **ts-or1-3**, and **ts-orsi-3** with respect to the binding pockets (Tab. 5-11) shows that they address the binding site in a completely different way. For example, the PYR group can be found in the S1 pocket (**orsi-2**), as well as in the S1' pocket (**ts-or1-3**), or outside the binding site in a more solvent exposed position (**ts-orsi-3**). This is furthermore illustrated by the visual inspection of the three binding motifs, as depicted in Fig. 5-26.

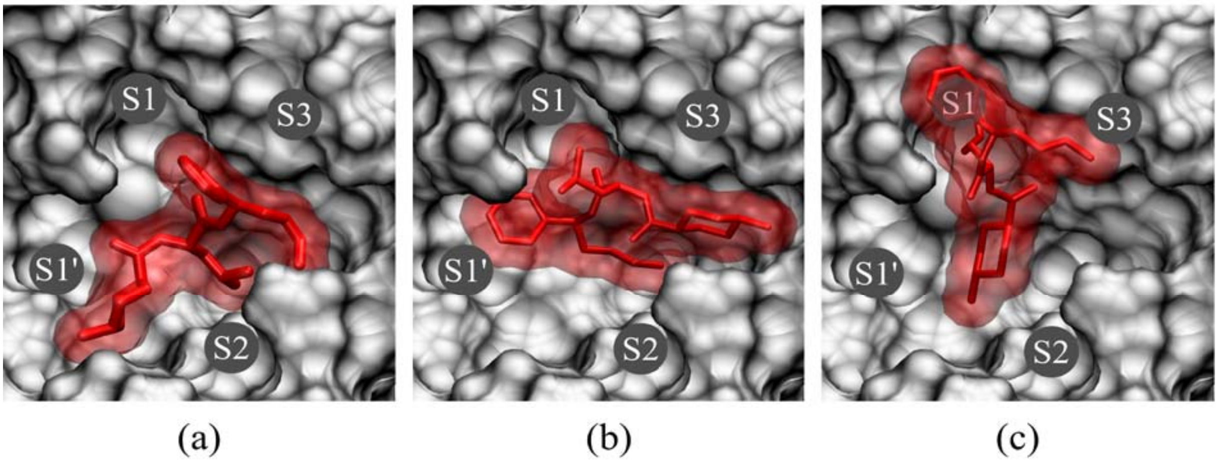


Fig. 5-26: The three most likely inhibitor poses of TS174 in complex with SARS-CoV M^{pro}, as ranked from comparative MD simulation results. They include the poses **orsi-2** (a), **ts-or1-3** (b), and **ts-orsi-3** (c).

Two major conclusions can be drawn from this finding. The first one is that the TS174 inhibitor is capable to bind to the active-site in multiple ways and conformations. The second one is that both stereoisomers are found in the active-site of SARS-CoV M^{pro}. Both results are in line with the experimentally determined electron density map that shows a clearly defined density for the protein, but smeared regions for the TS174 inhibitor. This reflects the situation of an inhibitor pose mix that causes an averaged electron density in the X-ray scattering experiment. Hence, unique atom position cannot be determined. This finding also explains the low inhibition potency of TS174, since in case of a strong binding, only one or a few preferred binding motifs would be present.

A qualitative analysis of the RMSF results in Tab. 5-11 yields some ideas for improvements of the substitution pattern of TS174. The CHEX group seems to fluctuate least in the S3 pocket, the ISO group in the S1 pocket, and the BUT group in the S2 pocket. This binding motif is fulfilled e.g. by the pose **ts-or1-3** that also possesses the lowest molecular strain. Hence, this pose seems to be a good starting point for optimizing the substituent-binding pocket interactions. Moreover, the PYR group tends to bind more floppy in general than the other inhibitor substituents. Thus, replacing this substituent by another one might improve the overall binding affinity.

5.4.3 Conclusions about the Binding Mode of the Inhibitor

The developed workflow has been applied to support and interpret the X-ray structure refinement of SARS-CoV M^{pro} in complex with the TS174 inhibitor. It applies a combination of educated guesses and *in silico* generated structure proposals that are evaluated by subsequent MD simulations and relative energies. In summary, the inhibitor is found to bind in multiple poses to the active-site. Hence, a unique X-ray structure of the enzyme-inhibitor complex cannot be determined. This problem is moreover amplified by the fact that both stereoisomers of TS174 seem capable to bind.

The methodological approach of analyzing individual structure parameters over the whole simulation trajectory allows sophisticated insights into the binding performance of individual inhibitor substituents. The results indicate that the pose proposal **ts-or1-3** seems to be a good starting point for a rational improvement of TS174 and that the pyridine substituent should be replaced by another one. These findings underline the additional benefits of MD simulations in the support of X-ray structure refinement and virtual screening over energy and scoring functions.

However, the results indicate that there is no obvious relationship between the relative energies of the initial minimum energy structures and the RMSD (or RMSF) values of an inhibitor pose. There could be several reasons for this finding. One deficiency could stem from an insuficient equilibration of the MD simulations. Conducting short molecular dynamic simulations that start from different points of a trajectory, and a subsequent analysis of the RMSD/RMSF values, could give further insights into this problem. Another approach could start from the thermodynamic point of view. Considering not only the relative energies from initially minimized inhibitor poses, but also the relative energies from several other points of the simulation trajectory, would give a more detailed picture.

Furthermore, entropic effects could contribute to the lack of correlation between the relative energies and the RMSD/RMSF values. They are inherently included in the simulations, but not accounted for by the relative energy calculations, since the latter only measure potential energy differences.

Moreover kinetic effects could play a role. E.g. a low barrier between two stable and low lying conformers causes a high fluctuation due to a frequent change between the structures, but would yield a low relative energy; in the opposite case a high lying structure might be trapped within high and steep energy barriers but would nevertheless result in a high relative energy with low fluctuation.

Considering all possible explanations, further investigations seem to be necessary to disentangle the various effects.

6 Conclusions

The review of all elaborated results from this work delivers an improved understanding of the SARS coronavirus main protease. Some aspects might also hold for other proteases. Considering the four questions defining the aim of this work in chapter 3, the results can be summarized as follows.

As an entry point, the question of how well enzymatic proton transfer reactions can be described by current quantum chemical methods has been raised. An extensive benchmark for the specific case of SARS-CoV M^{pro} revealed the use of QM/MM methodology as well founded, since it solely meets the requirements for this purpose that cannot be achieved by continuum solvent models or gas phase models. Considering the error behavior of the different hamiltonians, a simplified rule of thumb can be concluded that says: as less empirical parameters a method contains, the more systematically it behaves when applied in different environment models, like gas phase, COSMO, or QM/MM. Conclusively, semi-empirical methods suffer from unpredictable performance changes in this case and should be scrutinized before applied. This holds especially for the recent RM1 method, for which a dramatic deficiency in the description of zwitterions has been untapped. However, DFT and wavefunction based methods delivered mostly robust performance, and a combination of them can even yield accuracies of highly correlated methods, by keeping the computational costs feasible.

The second question of this work has been related to the active-site of SARS-CoV M^{pro} and its characteristic features. Extensive theoretical modeling of possible proton transfer reactions between the active-site residues gave a comprehensive explanation for the experimental findings that render SARS-CoV M^{pro} rather distinctive compared to archetypical cysteine proteases. The neutral resting state of the Cys/His dyad could be correctly reproduced and furthermore a possible charge-relay mechanism was excluded. As a striking outcome it was shown that substrate binding of SARS-CoV M^{pro} fosters the zwitterion formation. This effect was not observed for the binding of the Michael acceptor type inhibitor. The origin of this effect seems to stem from the water shell that is completely screened from the active-site in the first case, but only partly in the second case. Since the rigid hydrogen bonding network of the water shell cannot instantaneously relax in the vicinity of the activity-site, it introduces a higher barrier for the proton transfer than in the screened case, where only water bulk effects are accounted for.

The third question defined by the aim of this work requests insights into the inhibition mechanism of SARS-CoV M^{pro} with Michael acceptor warheads. It was shown that an *in silico* screening of model compounds with a continuum solvent approach was a successful tool in comparing various conceivable reaction pathways. The formation of a zwittionic catalytic Cys/His dyad was found to be an inevitable precondition for an inhibitor binding reaction. The subsequent conjugated 1,4-addition of the thiolate to the α,β-unsaturated carbonyl moiety can, moreover, only proceed irreversible, when the subsequent protonation step leads to the keto form. If the ketonization step is hindered, reversible binding is observed. These key findings give a comprehensive explanation of the experimental data available from chemoassays, by predicting different reaction pathways for the investigated substitution patterns.

The last question was related to the problem of explaining the non-covalent inhibitor binding of a SARS-CoV M^{pro} inhibitor. The developed workflow, consisting of an interplay between experimental X-ray structure refinement, tabu search based pose proposals, and molecular dynamic simulations, was demonstrated as a pragmatic way to improve X-ray structures in difficult cases. For the applied example of the TS174 inhibitor, the undefined electron density regions of the experimentally determined electron density map were interpreted by the presence of multiple inhibitor poses. These were found to address different binding pockets at the active-site of SARS-CoV M^{pro}. It was furthermore shown that both stereoisomers of TS174 are likely to bind.

Summarizing the answers to the four asked questions, some conclusions can be drawn about the overall mechanisms of substrate cleavage and inhibition reaction by the SARS coronavirus main protease. Most important one represents the meaning of the initial zwitterion formation of the catalytic dyad. There is evidence from the results that this step can have a rate determining character, which is in line with experimental results.[58] This electrostatic trigger can be expected to have a significant impact on all chemical reactions occurring inside of the active-site.

This mechanistic feature has important consequences for the inhibitor design. As known from papain, the reactivity of the Michael acceptor group has a direct correlation to the inhibition constant.[338] This can be employed to increase or decrease the inhibition potency by modifying the substitution pattern of the α,β-unsaturated carbonyl warhead. However, this can only work if the chemical step of conjugated addition is the rate-determining one. Since the catalytic dyad of SARS-CoV M^{pro} needs a further activation step to form the ion pair, and thus switching on the reactive state, the overall rate depends either partly or solely on this critical step. Nevertheless, this is inherently controlled by the enzyme and not by the inhibitor. Hence, inhibitor potency cannot be accelerated by an increased electrophilicity of the inhibitor warhead.

As demonstrated by the Michael acceptor type inhibitor-complex model of SARS-CoV M^{pro}, the inhibitor additionally lacks the ability to foster the zwitterion formation of the catalytic dyad. This might be one of the reasons, perhaps the most decisive reason, why the inhibitor development against SARS-CoV M^{pro} has not been very successful during the last ten years.

As a final outlook, several ideas and recommendations can be derived from the new insights. The first one deals with the question, how far the herein made discoveries are. Since the drawn conclusions of this work have been elaborated exclusively from SARS coronavirus main protease, it would be of interest whether some features, like the electrostatic trigger, are also found for other enzymes. Comparable theoretical studies for other proteases could show the overall importance of this mechanism. This could represent a general mechanism to increase substrate specificity of enzymes, beyond the concept of molecular recognition or induced fit binding. Since it is well known that the surrounding of enzymes can influence ligand or substrate pK_a values,[382] there is no reason to believe why this should not work in reverse, respectively for an active-site residue.

The results also give novel implications for the rational inhibitor design. On account of the fostering of the zwitterion formation being significant, this effect should be mimicked by the inhibitors. One improvement could start from an increased coverage of the active-site by the inhibitor. In view of the fact that this was found as the main reason for the different behavior of substrate and inhibitor, it could represent a promising way towards higher inhibition potency. Another approach could be based on a rational design of the electrostatic properties of the inhibitor warhead. The simplest way to do this would be to develop a dipol with complementary orientation to the zwitterion. This could conclusively lead to a stabilizing interaction that makes the zwitterion more preferable and thus accelerate the whole inhibition reaction.

7 Computational Details

Chapter 5.1 - Accuracy of Theoretical Methods for Enzymatic Proton Transfer Reactions. In order to get a selection of starting structures, molecular dynamic simulations were conducted with the NAMD 2.7 program package,[383,384] based on crystal structure of SARS-CoV M^{pro} (PDB code 2DUC)[53] taken from RCSB protein data bank.[54] The initial protein structure was embedded in the center of a TIP3P[385] water shell with 110 Å diameter and energy minimized by keeping the protein structure fixed. Protonation states of titratable residues were determined by empirical pK_a prediction with the PROPKA methodology,[386–388] satisifying a pH value of 7. Since no residual charges were left in the whole system, it was not necessary to place counter ions. To ensure a well equilibrated model system, it was slowly heated to 310 Kelvin in gradual steps of 10 Kelvin. After heating of the system, atom position constraints put on protein structure were successive released and simulation was equilibrated for 1 ns, followed by a production run of 10 ns. Heating, equilibration, and simulation procedure were conducted in a NVT ensemble with 1fs time steps using the SHAKE algorithm[389] for all water molecules and Verlet algorithm as integration method.[168] Spherical harmonic boundary conditions were applied beyond a radius of 55 Å around the center of sphere to prevent water molecules from evaporating. The root mean square deviation (RMSD) from starting structure of the protein backbone, plotted as a function of time (**Fig. 4**), indicates a stable simulation system with average RMSD of 1.6 Å.

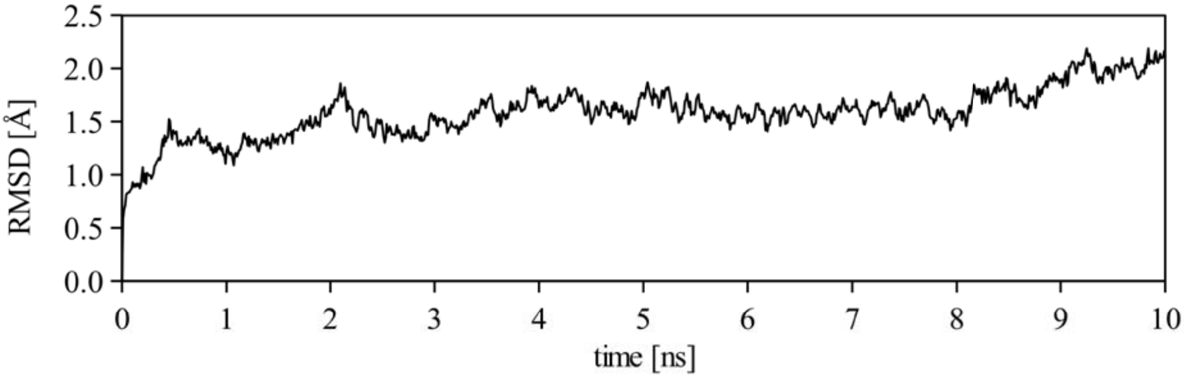


Fig. 7-1: Root mean square deviation (RMSD) from starting structure as function of time for molecular dynamic simulation of SARS coronavirus main protease.

All QM/MM calculations were performed within the ChemShell 3.4 computational chemistry suite,[309] using an electrostatic embedding scheme in conjunction with charge shift method and link atom approach for the QM/MM boundary treatment.[255–257] The QM/MM model system was set up from protein structure of SARS-CoV M^{pro}. As starting structure for all calculations, one suitable snapshot from the previously performed molecular dynamics simulations (MD) was used. The active-site residues Cys145 and His41 and one bridging water molecule were selected as quantum mechanical part (QM) of the model system as depicted in Fig. 5-1. All remaining atoms were treated on the molecular mechanic level of theory (MM). For the MM part the all-atom forcefield CHARMM22 was used.[160]

To save computational resources in QM/MM calculations, the water shell was reduced to a radius of 30 Å around the active-site residues His41 and Cys145. The outer water layer and protein structure at radii greater than 10 Å were kept fixed during structure optimizations. The minimum proton transfer path was obtained from a relaxed constrained minimization series along the reaction coordinate as shown in Fig. 5-1, using increments of 0.1 Å for the r(S-H) distance. All the other relevant reaction coordinates r_1(O-H), r_2(O-H), r(N-H) adjusted smoothly to the stepwise changed r(S-H) distance during the constrained minimizations. The minimization series was proceeded in both directions to ensure that the obtained minimum path comprises a meaningful connection between the neutral charged state and zwitterionic state. QM/MM structure minimizations were performed on the BLYP | TZVP level,[220,390] using resolution of identity approximation.[391,392]

For the benchmarking, single point calculations of the obtained minimum path were performed for all methods. In detail, this included 14 grid points of the proton transfer potential which connect the neutral state **N** and the zwitterionic state **ZW**. It is noteworthy that instead of calculating single point energies, there is also the possibility to perform geometry optimizations on each level of theory, which might deliver slightly different results. Due to the fact that gradients are not available for every method/environment combination and further that higher level single point calculations "on top" of DFT geometries are a rather established and routinely applied approach for QM/MM computations, the first possibility was chosen. Nevertheless, since it cannot be assured that GGA functionals like BLYP deliver reasonable structures for proton transfer reactions, as recently demonstrated by Adamo and coworkers,[299] furthermore structure optimizations on the MP2 | TZVP and B3LYP |TZVP level as well as with the semi-empirical methods AM1 and PM3 were performed. The comparison between relaxed and unrelaxed structures are shown in Fig. 7-2, Fig. 7-3, Fig. 7-4.

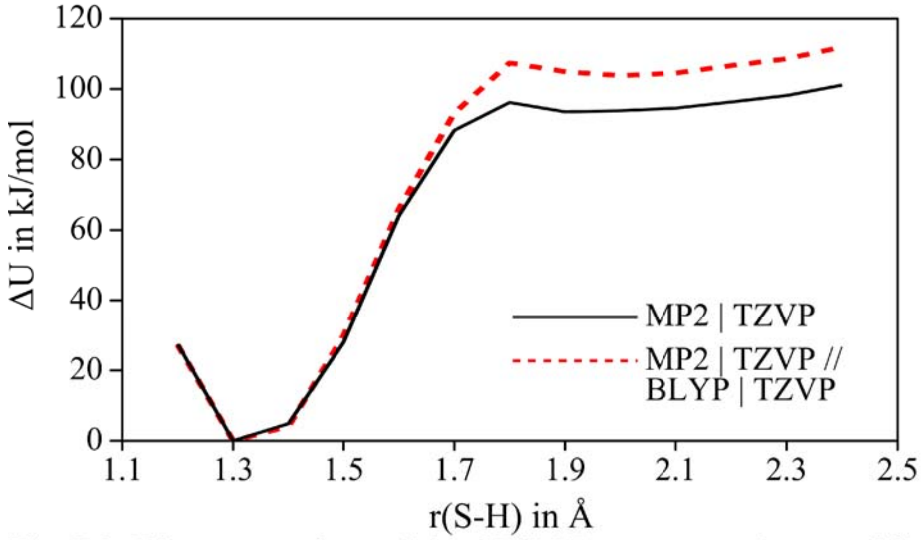


Fig. 7-2: Direct comparison of the QM/MM proton transfer potentials obtained on the MP2 | TZVP level (solid line, black) and the MP2 | TZVP // BLYP | TZVP level (dashed line, red).

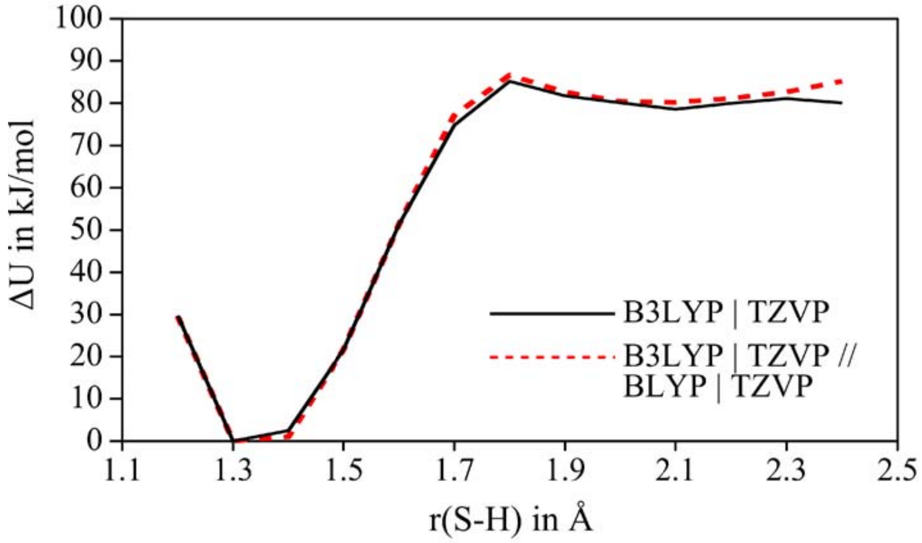


Fig. 7-3: Direct comparison of the QM/MM proton transfer potentials obtained on the B3LYP | TZVP level (solid line, black) and the B3LYP | TZVP // BLYP | TZVP level (dashed line, red).

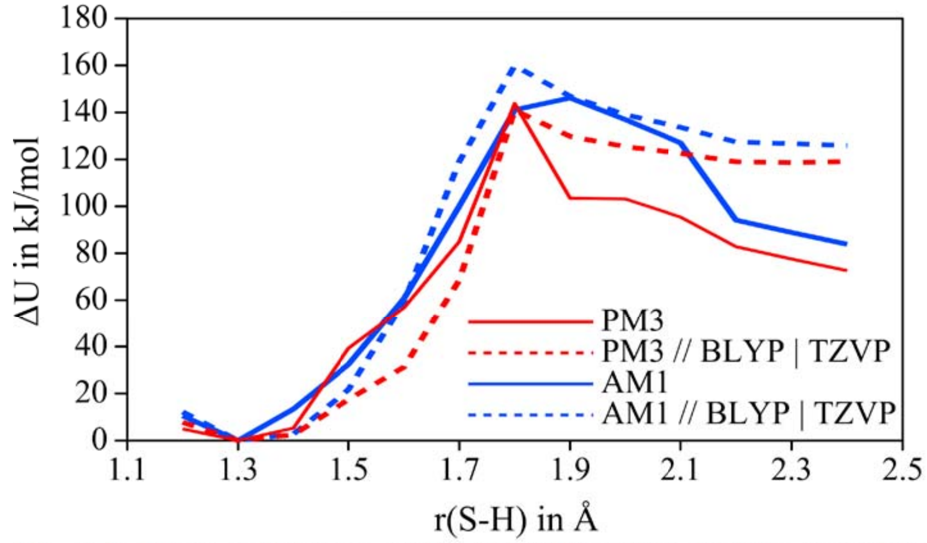


Fig. 7-4: Direct comparison of the QM/MM proton transfer potentials obtained with AM1 (solid line, blue) and AM1 relative energies on top of BLYP | TZVP geometries and further PM3 (solid line, red) and PM3 relative energies on top of BLYP | TZVP geometries (dashed line, red).

As implicit solvent model, the COSMO approach of Klamt and Schüürmann[244] was used. Since default settings partly differ in MOLPRO, GAMESS-US, MOPAC, and TURBOMOLE, they have been set manually to keep results comparable between the different program packages. The value for the dielectric constant have been set to ε=4 and ε=78, which represent estimates of protein environment[248,249] and water.[247] Multiplicative factor for cavity construction (RSOLV) was set to a radius of 1.2 times the respective van der Waals radius as recommended by Tomasi and coworkers.[393,393] Furthermore, the number of surface elements per atomic sphere was set to 92 and distance threshold radius for segment-segment interactions to 10 Å in all COSMO calculations, which represent the most commonly used default values. Nevertheless, since the choice of the multiplicative factor might severely change the resulting proton transfer potentials, the influence of RSOLV on the relative transition state energies and zwitterionic minimum state energies at least for MP2 | TZVP, M06-2X | TZVP, and PM3 was probed by using a dielectric constant ε of 78. Changing RSOLV in the range from 0.5 to 3.0, influences the relative MP2 energies by a maximum deviation of 1 kJ/mol (TS) and 2 kJ/mol (ZW), relative M06-2X energies by 3 kJ/mol (TS) and 6 kJ/mol (ZW), and relative energies predicted by PM3 in the range of 2 kJ/mol (TS) and 6 kJ/mol (ZW). The respective energy plots are shown in Fig. 7-5.

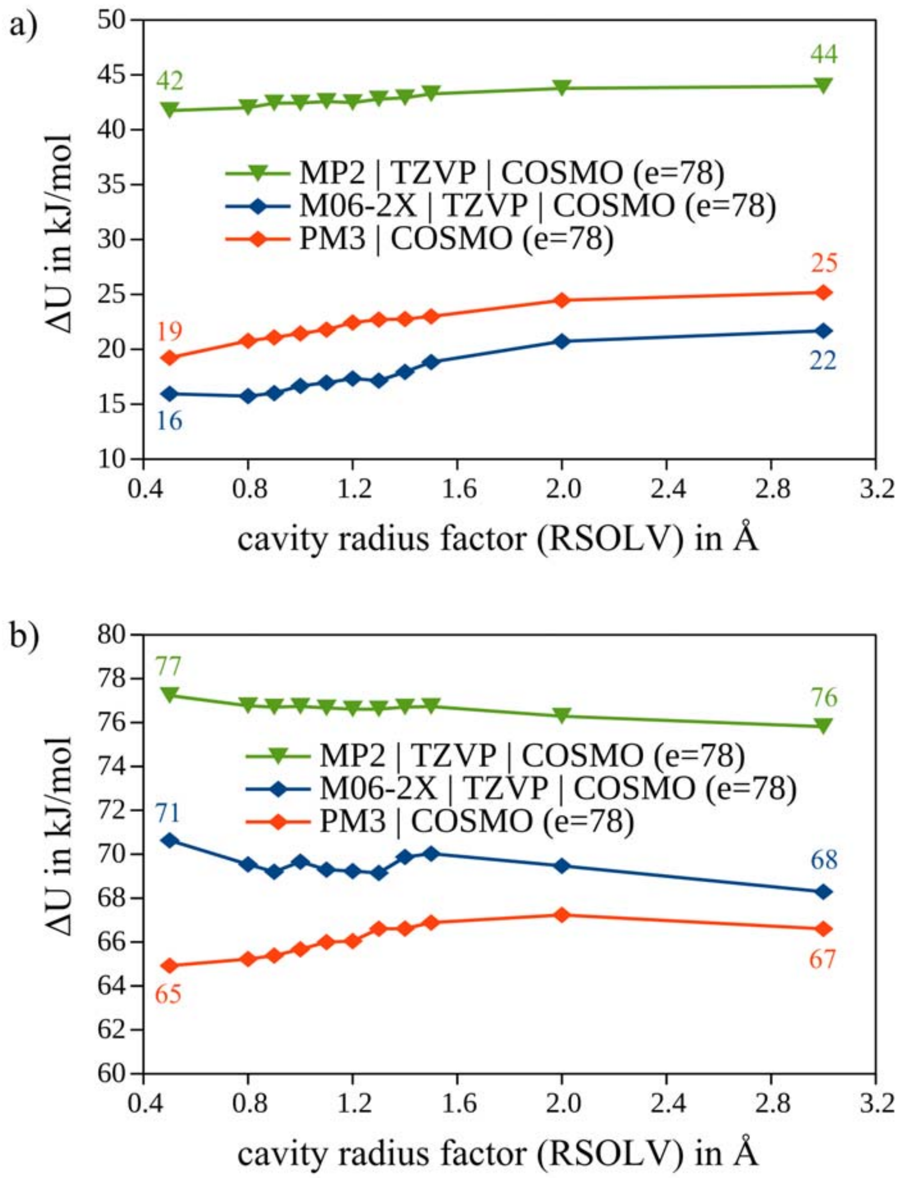


Fig. 7-5: Influence of the multiplicative factor that is used for cavity construction (RSOLV) on the relative energies of the zwitterionic state **ZW** (a) and the transition state **TS** (b). Cavity radii are calculated as RSOLV * vdW radii of the respective atoms.

The QM/MM calculations always included the whole system as described above, whereas COSMO and gas phase calculations were carried out only with the QM part. Reference calculations with the local coupled cluster method LCCSD(T)[193] were conducted with the

MOLPRO 2010 quantum chemistry package,[394] using density fitting approximation.[194,395] Non-local CCSD(T) calculations, as well as Møller Plesset perturbation theory MP2,[396] the approximated coupled cluster theory CC2,[198] Hartree Fock (HF) calculations,[397] and density functional theory computations with the BLYP[220] and B3LYP[221] functional, were performed with the TURBOMOLE 6.2 program package.[398] Since spin component scaling (SCS) improves the accuracy of MP2 and CC2 calculations without causing further computational costs,[200,201] it has been applied for the reference calculation on the MP2 level and the CC2 calculation.

To estimate effects due to the extent of basis functions, polarized basis sets of single valence (SVP),[390] triple zeta (TZVP),[399] and quadruple zeta quality (QZVP)[400] were employed as implemented in TURBOMOLE and MOLPRO. The resolution of identity approximation[392] was used for MP2, CC2, and B-LYP calculations in conjunction with suitable auxiliary basis sets.[391] DFT calculations employing the meta-hybrid exchange-correlation functional M06-2X[223] were carried out with the GAMESS-US program package version OCT 2010 (R1).[401] Here, TZVP and QZVP basis set files were manually converted from TURBOMOLE, since these are not included in GAMESS-US. A cross check with B3LYP results between TURBOMOLE and GAMESS-US results confirmed a faultless conversion. QM/MM calculations with M06-2X could not be performed, since electrostatic embedding scheme is not available for the GAMESS-US interface of ChemShell.

Gas phase and COSMO calculations that employ the semi-empirical methods AM1,[207] RM1, [209] PM3,[208] PM6,[210] MNDO[203] and MNDO including d-orbitals,[205,206] were conducted with the MOPAC2009 program.[402] For QM/MM calculations with semi-empirical methods, the MNDO2005 program was used.[403] Since RM1 and PM6 are not implemented in MNDO2005, they had to be omitted.

Chapter 5.2.1 - Comparative Molecular Dynamic Simulations. All MD simulations were conducted with the NAMD 2.6 parallel molecular dynamics program.[383,384] Each of the six simulations contained an overall number of about 68500 atoms and was simulated for 10 nanoseconds by using timesteps of 1 picosecond. Solvent was explicitly included in the simulations as a droplet approach and comprises a diameter of 110 Å for the employed water shell. Simulations were conducted as NVT ensemble, by using the Verlet integration method[168] and the SHAKE algorithm[404] for the water molecules to reduce internal degrees of freedom. Furthermore, constraints were applied by using a spherical boundary condition beyond a radius of 55 Å from the model system's center of mass, to prevent water molecules from evaporating from the water shell into vacuum during the simulation.

All simulations were prepared in an identical manner, using the following protocol, and are based on the PDB structure 2DUC[53] of the SARS coronavirus main protease dimer that was obtained from the RCSB database.[54]

The protonation state of titratable residues, like histidines, were set according to a pH value of 7 and empirical pK_a value predictions with the PROPKA methodology.[387,388,405] The protonation states of the active-site residues His41, Cys145, and the neighboring histidine residue His164 were set according to one of the six possible protonation states, as shown in Fig. 5-6, in six individual MD simulations. The protein structures were embedded into a TIP3 water shell[385] and water molecules initially minimized in their energy for 1000 steps to remove molecular strains, especially at the protein-water interface. Subsequently, the model systems were heated in gradual steps of 10 Kelvin until reaching the final simulation temperature of 310 Kelvin, while keeping the protein structure fixed. After heating, simulations were equilibrated for 1 ns under successive release of spatial constraints that were put on the protein backbone. Finally, production runs were performed for 10 ns and trajectories recorded in 10.000 step windows, resulting in 1000 frames for each simulation. Subsequent data analysis for root mean square deviation (RMSD) plotting and statistical structure analysis (histograms) were performed with VMD.[406] The respective RMSD values as function of simulation time are shown in Fig. 7-6.

Histogram analysis for Cys145-His41 distance are performed based on atom distances between sulfur (Cys145) and nitrogen (His41) for each trajectory. A binning size of 0.05 Å is employed for the distance range of 2.0 to 10.0 Å and the number of counts per bin are taken relative to the overall number of 1000 recorded distances.

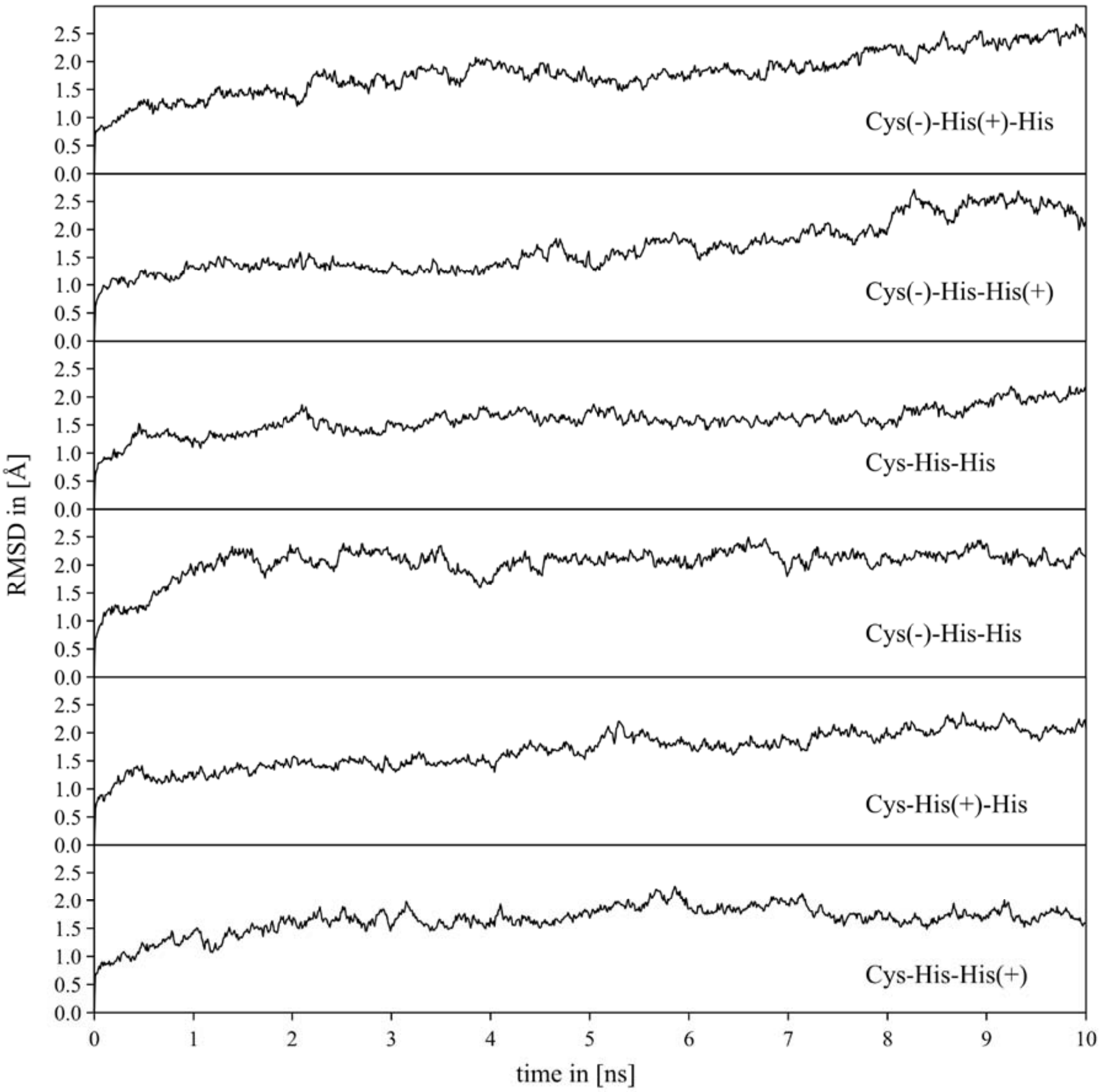


Fig. 7-6: Root mean square deviation (RMSD) of the six MD simulations in Å as a function of time in ns. Deviations refer to Cα atoms of the protein backbone and are taken relative to the first frame of the respective simulation. Average RMSD values are from top to bottom 1.8 Å, 1.7 Å, 1.6 Å, 2.1 Å, 1.7 Å, 1.7 Å.

Chapter 5.2.2 - Proton Transfer Potentials. QM/MM potential energy surfaces have been calculated with the Chemshell 3.2 program package.[309] The QM/MM system was set up from a representative snapshot of the MD simulation according to the **Cys-His-His** state. The water shell was reduced to a radius of 30 Å around the active-site residues and further kept fix for radii greater than 10 Å to save computational resources during energy minimizations. Water molecules within the radius of 10 Å were kept rigid with respect to internal degrees of freedom, with the exception of the bridging water molecules that are involved in the proton transfer reactions. The employed QM part was defined by the side chains of active site residues Cys145 and His41, His164, and the two bridging water molecules as illustrated in Fig. 5-9. Calculations of the QM part were performed with the TURBOMOLE 5.9 suite of programs[398] by using density functional theory. Due to their good ratio between accuracy and computational costs, the B3LYP exchange correlation functional[220–222] in combination with a TZVP basis set[399,407] was used for energy calculations "on top" of RI-BLYP/TZVP structure optimizations,[220,391,392,399,407,408] as successfully applied in several earlier reports.[275,276,279,280,286,344] According to the benchmarking in Tab. 5-1, the error with respect to structure optimizations on the B3LYP/TZVP level of theory can be expected below 2 kJ/mol. For the MM part, the CHARMM22 all atom forcefield[160] was employed.

For QM/MM coupling and boundary treatment across chemical bonds, an electrostatic embedding scheme[256] and furthermore the link atom approach[255] was employed. A comprehensive insight to the applied approaches, is given in the review of Senn and Thiel.[183]

Valid reaction coordinates that connect the different protonation states between the active-site residues Cys145, His41, and His164 in a meaningful way, were carefully identified by probing multiple potential energy surface calculations with different combinations of bond length changes within the complex hydrogen bond network, as shown in Fig. 5-8. The reaction coordinates r_1(O-H) and r(N-H) were assessed to be suitable descriptors to model the proton transfer reactions between Cys145/His41 and His41/His164. Relaxed QM/MM potential energy surfaces were mapped by changing the reaction coordinates in incremental steps of 0.1 Å and subsequent constrained energy minimizations with respect to the two coordinates for each grid point. For energy minimizations, hybrid delocalized internal coordinates (HDLC)[409] were used.

Averaged QM/MM energy potentials for the proton transfer between **Cys-His-His** and **Cys(-)-His(+)-His** state were calculated for three QM/MM set ups that were prepared according to the same procedure as described for the QM/MM energy surfaces above. The set up of the three model systems (free-enzyme, inhibitor-bound enzyme, substrate-bound enzyme) is based on X-ray structures 2DUC,[53] 2AMD,[106] and 2Q6G,[72] as illustrated in Fig. 5-9. For the averaging of proton transfer potentials, 10 snapshots were taken from each MD simulation, respectively one for every 1 ns window over the whole simulation time of 10 ns. The QM part was restricted to the relevant parts, comprising the side chains of Cys145 and His41 residues, which partly includes one bridging water molecule in case of the free enzyme. QM/MM energy curves were calculated in forward and backward direction to ensure meaningful minimum energy paths for the transition between the **Cys-His-His** and **Cys(-)-His(+)-His** states. Therefore, averaging was done in sum over 20 minimum energy paths that were aligned by least square fit technique. Furthermore, standard deviations for the relative energies ΔU were calculated for each point of the proton transfer curve.

Chapter 5.2.3 - Charge Deletion Analysis. The charge deletion analysis, also denoted as decomposition analysis, represents changes in the relative energies ΔΔU that are due to switching off the electrostatic point charges for certain parts of the QM/MM system by simply setting them to zero. This method has been applied successfully in different forms by several other QM/MM studies.[279,410–414] For the given purpose, the two minima **Cys-His-His** and **Cys(-)-His(+)-His** that are located at r(S-H)=1.4 Å and r(S-H)=2.0 Å were taken from the eleven calculated proton transfer potentials, as described above, and calculated for their relative QM/MM energy ΔU(ZW), which represents the thermodynamic difference between the two states. Then, a series of QM/MM energy calculations were repeated with identical structures, but with individual residues that were deleted by their point charges in the MM part. This was respectively done for the side chains of the amino acids Arg40, His164, Phe181, Asp187, and the whole salt bridge formed by the Arg40-Asp187 residues. Other important parts around the active-site, like the buried water molecule H_2O@His41, the inhibitor, or the substrate, were also included in the analysis. Furthermore, the solvent water shell, the whole protein structure, and the individual protomers A and B were switched off respectively. The difference between the unaffected relative energy ΔU(ZW) and the obtained energy $\Delta U(ZW)^{\text{MM charges off}}$, when switching off a certain part of the system, gives an estimate about the stabilizing or destabilizing electrostatic effect of this part on the zwitterionic state **Cys(-)-His(+)-His**.

$$\Delta\Delta U(ZW)=\Delta U(ZW)^{MM\,charges\,off}-\Delta U(ZW) \tag{46}$$

The sign of the ΔΔU value reveals whether the selected part has a stabilizing or destabilizing impact on the formation of the zwitterion. From positive energy values for ΔΔU, a stabilizing contribution can be concluded, whereas negative values indicate a destabilizing effect. Since single minima might be somehow arbitrary in their structural conformation, an averaging has been done over 11 snapshots, similar to the procedure described for the proton transfer potential. This also allowed the calculation of standard deviations in order to get an estimate for the fluctuation of the electrostatic contribution.

Chapter 5.2.4 - Free Energy Calculations. Free energy potentials of the cysteine histidine proton transfer were obtained by QM/MM potential of mean force (PMF) calculations,[415] using the AMBER 11 suite of programs.[416] Simulations were performed for the free enzyme and the substrate-bound model. Reasonable starting structures were already available from the classical MD simulations. The selection of the QM part was done in an analogue manner to the QM/MM minimum energy path calculations and included the side chains of Cys145 and His41, and further one bridging water molecule in case of the free enzyme. The semi-empirical PM3/PDDG method[208,417] was employed for the quantum chemical calculations, since it delivers least errors compared to other available semi-empirical hamiltonians (RM1, AM1, MNDO). This finding was evaluated explicitly for the given case, as apparent from the benchmarking results in chapter 5.1. For the MM part, the ff99SB parameter set from Hornak and Simmerling[163] was employed, as implemented in AMBER. The three model systems were set up identically to the classical MD simulations, comprising spherical boundary conditions, 1 fs time steps, and 310 Kelvin simulation temperature. In order to perform PMF calculations, an umbrella sampling[418,419] was employed by using a bias potential along the reaction coordinate. Each window was simulated for 50 ps. The influence of the simulation length on the PMF results is shown in Fig. 7-7 that gives an estimate about the change in relative energies with respect to the extent of sampling applied. Although longer simulation length lead to better relaxation, the samplings had to be restricted to 50 ps due to the high computational costs. The maximum deviation between relative energies obtained by 50 ps and 100 ps ranges below 16 kJ/mol and are therefore clearly smaller than the intrinsic error of the PM3 hamiltonian. The bias potential was set up for the r(S-H) bond of Cys145 and the r(N-H) distance of His41 in case of the substrate-bound model system, as illustrated in Fig. 5-9 (e). The proton transfer required 16 simulation windows with incremental shifts of 0.1 Å for the bias potential minimum, to achieve the transition from the **Cys-His-His** to the **Cys(-)-His(+)-His** state. In case of the free enzyme, which comprises a water-mediated proton transfer, the calculation of a two dimensional PMF energy surface with 272 simulation windows was necessary to obtain the minimum energy path that connects the **Cys-His-His** and **Cys(-)-His(+)-His** state. Here, the r(S-H)/r_1(O-H) and r_2(O-H)/r(N-H) coordinates were used, as illustrated in Fig. 5-9 (d).

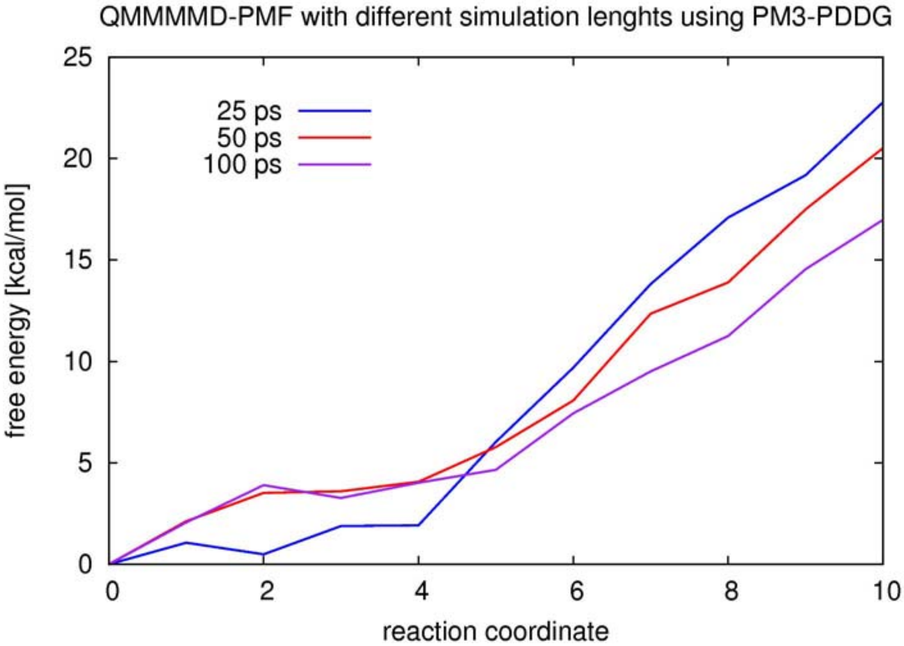


Fig. 7-7: Test calculations to probe the influence of the simulation length on the relative free energy between Cys-His-His state on the left hand side and Cys(-)-His(+)-His state on the right hand side.

A force constant of 800 kcal/mol Å^2 was found to be a suitable value for the bias potential to generate a distribution with reasonable overlap between neighboring windows. The recorded data sets where subsequently processed with the weighted histogram analysis method (WHAM),[333,420,421] to obtain the PMFs.

Chapter 5.3.2 – Screening of Substitution Patterns. Parts of the calculations were adopted from earlier work.[365] The points of the PESs were computed with DFT employing the TURBOMOLE program package,[398] with standard settings. The potential energy surface scans were performed with a resolution of 0.1 Å for the respective grid points. The optimization of the geometrical parameters, except the two characteristic bond distances r(S-C) and r(S-H), that define the potential energy surface, were performed with the BLYP exchange correlation functional,[220–222] the resolution of identity approximation (RI),[392] and the triple zeta basis set TZVP,[399] which includes three sizes of contracted functions and further p-functions to take polarization into account. The electronic energies of all obtained geometries were recalculated by additional single point calculations, employing Becke´s three parameter hybrid functional B3LYP[220–222] and TZVP basis set. Solvent effects were taken into account within the conductor-like screening model (COSMO),[244] which was used for all calculations (geometry optimizations and single point calculations). The dielectric constant of the polarizable environment was set to 78.39, which is the corresponding value for water. For cavity radii, standard settings were used.

Chapter 5.4.2 - Theoretical Evaluation of 22 Inhibitor Pose Proposals. Molecular dynamic simulations (MDs) were conducted with the NAMD 2.8 program package[383,384] using a graphic processor unit (GPU) accelerated implementation.[173] Force-field parameters for the SARS-CoV M^{pro} were taken from CHARMM22 all-atom force field[160] by using the cross-term map correction (CMAP),[159] which improves the description of the peptide backbone. Parameters for the TS174 inhibitor were constructed from the CHARMM General Force Field (CGenFF version 2b6, program version 0.9.1 beta)[422] and checked for their reasonability by penalty scores (similarity index of unknown parameters with respect to known parameters) and visual inspection of several energy minimizations of the inhibitor without enzyme. All MDs of the protein-inhibitor complex were prepared in an identical manner according to the following procedure.

Since no information about hydrogen atoms are available from X-ray structures, the protonation states of all titratable groups were determined by empirical pK_a predictions and set up identical for all MDs. The commonly used PROPKA methodology,[386–388] which should be reasonable accurate for this purpose,[262] was employed. Protein structures of every inhibitor pose were embedded in a water sphere of 110 Å and saturated with one counterion (chlorine) to balance the residual protein charge and achieving an electrostatic neutral system. Internal degrees of freedom of the TIP3P water molecules[385] were constrained by the SHAKE algorithm[389] to save computational time. To prevent water molecules from evaporating into vacuum, spherical boundary conditions were applied beyond the surface of the water droplet, which was set at a radius of 55 Å with respect to the center of mass. Initial structures were minimized to release molecular strains, especially at the protein-water interface. This was first applied solely for the water shell with fixed protein (2000 steps) and after wards for the whole system (500 steps). The minimized system was allowed to heat up in gradual steps of 10 Kelvin to a simulation temperature of 310K. The protein structure was kept fix with spatial constraints and slowly released during the subsequent equilibration period of 1ns. Productive simulation runs were finally conducted for 5 ns at a temperature of 310 Kelvin. Equilibrations and simulations employed the Verlet integration algorithm[168] and 1 fs time step size within a canonical ensemble of fixed particle number, volume, and temperature (NVT). Data analysis of recorded trajectories with respect to root mean square deviation (RMSD), structural parameters, and root mean square fluctuations (RMSF) was done with VMD.[406]

For each simulation trajectory, a data set of 500 frames was acquired and analyzed.

binding pocket	involved residues
S1'	Thr25, Leu27, His41, Val42, Gly143, Cys145
S1	Phe140, Leu141, Asn142, Ser144, His163, Glu166, His172
S2	His41, Met49, Met165, Asp187, Arg188, Gln189
S3	Glu166, Leu167, Pro168

Tab. 7-1: Definition of SARS-CoV M^{pro} binding pockets as used for the inhibitor substituent distance analysis. Nomenclature of binding pockets is taken from Schechter and Berger.[81]

Calculation of RMSD values was done with respect to the first frame structure of the simulation trajectory and included the heavy atoms of the inhibitor TS174. For the analysis of distance values, centers of mass were defined for the inhibitor substituents of TS174 and the binding pockets of SARS-CoV M^{pro}. Tab. 7-1 gives the detailed description, which amino acids were used for the definition of the different binding pocket entities S1', S1, S2, and S3. The definition of inhibitor substituents is given in Fig. 5-24.

All calculated distances and RMSF values are given in Tab. 7-2, Tab. 7-3, Tab. 7-4, and Tab. 7-5. The most probable binding pocket occupations for a given pose were determined from the lowest distance of all inhibitor substituents and are summarized in Tab. 5-11.

structure	S1'		S1		S2		S3	
	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF
or1-1	14.26	0.45	**8.99**	**0.92**	10.78	0.70	**4.19**	**0.40**
or1-2	14.54	0.83	9.66	1.45	10.13	1.17	**3.74**	**1.88**
or1-3	12.49	0.55	10.42	0.50	6.08	0.45	**5.63**	**0.32**
or1-4	14.42	0.55	**9.00**	**0.50**	10.89	0.42	**3.96**	**0.28**
or1-5	13.32	0.40	**8.86**	**0.63**	9.21	0.46	**3.85**	**0.45**
or1-6	12.31	0.44	10.74	0.34	5.48	0.44	6.17	0.31
or2-1	**5.37**	**0.45**	9.34	1.02	**5.79**	**0.70**	11.04	0.45
or2-2	7.99	1.17	10.08	0.97	**2.34**	**0.94**	9.38	0.39
or2-3	9.59	1.26	11.34	1.25	**5.69**	**0.58**	10.09	0.84
or2-4	9.75	1.40	11.49	1.13	**5.61**	**0.62**	10.08	1.06
or2-5	**5.30**	**0.32**	10.65	0.31	6.97	0.32	12.69	0.45
or2-6	**5.30**	**1.06**	10.64	0.63	7.00	0.60	12.69	0.66
orsi-1	14.63	0.69	9.57	0.57	10.74	0.41	**4.29**	**0.37**
orsi-2	**5.54**	**0.28**	11.38	0.61	7.49	0.58	13.45	0.46
orsi-3	13.78	0.67	11.23	0.77	7.12	1.28	5.50	0.60
orsi-4	12.88	0.57	11.12	0.55	5.97	0.56	6.16	0.28
orsi-5	14.17	0.63	9.05	0.78	10.47	0.74	**4.03**	**1.45**
ts-or1-3	12.95	0.36	10.53	0.26	6.68	0.35	**5.35**	**0.27**
ts-or1-4	12.27	1.19	9.79	0.73	**6.59**	**1.16**	5.19	0.61
ts-or1-5	14.29	0.70	7.87	0.52	11.80	0.86	**4.08**	**0.39**
ts-or2-1	**5.71**	**1.13**	8.04	1.06	**6.27**	**0.79**	9.96	1.60
ts-orsi-3	**5.98**	**0.28**	8.46	0.30	**5.57**	**0.75**	9.86	0.26
max		1.40		1.45		1.28		1.88
min		0.28		0.26		0.32		0.26
mean		0.70		0.72		0.68		0.63

Tab. 7-2: Distances between binding pockets and cyclohexyl inhibitor substituent (CHEX), as obtained from X-ray structure proposals. RMSF values were calculated from 500 distances along the 5 ns MD trajectory. Most probable binding pocket occupations are indicated by bold numbers.

structure	S1'		S1		S2		S3	
	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF
or1-1	11.11	0.38	9.40	0.56	5.33	0.86	5.81	0.62
or1-2	10.25	1.63	**6.80**	**2.55**	8.63	1.09	6.17	1.67
or1-3	10.08	0.85	**6.34**	**1.22**	8.49	0.48	5.84	1.18
or1-4	11.06	0.46	9.26	0.51	5.75	0.35	5.81	0.30
or1-5	9.52	0.53	9.42	0.66	6.55	0.76	8.32	0.62
or1-6	9.94	0.58	**6.48**	**0.48**	8.29	0.37	**5.93**	**0.62**
or2-1	9.75	0.74	**6.22**	**0.95**	8.49	0.46	6.18	0.79
or2-2	**5.11**	**1.23**	9.15	2.45	6.19	1.12	11.15	1.45
or2-3	**5.52**	**0.89**	**7.82**	**0.48**	6.02	0.52	9.80	0.62
or2-4	**5.49**	**1.22**	8.82	1.49	6.13	0.90	10.65	1.66
or2-5	8.38	0.27	9.72	0.40	**2.77**	**0.40**	8.65	0.60
or2-6	8.42	0.41	9.70	0.62	**2.82**	**0.35**	8.58	0.37
orsi-1	11.01	0.63	**9.13**	**0.39**	5.50	0.42	5.65	0.42
orsi-2	7.77	0.32	9.48	0.31	**2.98**	**0.35**	8.94	0.38
orsi-3	10.45	0.44	**6.62**	**0.90**	8.32	0.60	**5.38**	**0.43**
orsi-4	11.69	0.76	**6.93**	**0.75**	9.20	0.50	**4.48**	**0.61**
orsi-5	10.62	0.80	9.61	0.89	7.54	1.27	7.92	1.04
ts-or1-3	9.80	0.59	**5.84**	**0.48**	9.68	0.37	7.02	0.55
ts-or1-4	**8.07**	**0.37**	**3.71**	**0.37**	9.52	0.61	7.65	0.32
ts-or1-5	12.31	1.22	**4.42**	**1.93**	13.86	2.31	8.07	0.93
ts-or2-1	11.56	1.70	7.92	0.90	8.92	1.66	5.69	1.15
ts-orsi-3	8.72	0.34	**2.83**	**0.25**	9.94	0.57	7.12	0.26
max	12.31	1.70	9.72	2.55	13.86	2.31	11.15	1.67
min	5.11	0.27	2.83	0.25	2.77	0.35	4.48	0.26
mean	9.39	0.74	7.53	0.89	7.31	0.74	7.31	0.75

Tab. 7-3: Distances between binding pockets and isobutyl inhibitor substituent (ISO), as obtained from X-ray structure proposals. RMSF values were calculated from 500 distances along the 5 ns MD trajectory. Most probable binding pocket occupations were determined from lowest distance of all inhibitor substituents and are indicated by bold numbers.

structure	S1'		S1		S2		S3	
	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF
or1-1	8.39	0.53	10.27	0.62	**2.23**	**0.52**	9.16	0.38
or1-2	8.13	1.18	10.11	0.79	**2.28**	**0.90**	9.27	0.79
or1-3	7.98	0.56	10.21	0.62	**2.21**	**0.47**	9.48	0.54
or1-4	**5.18**	**0.68**	9.14	0.85	6.85	0.75	11.38	0.38
or1-5	**5.45**	**0.41**	9.42	0.49	6.35	0.46	11.27	0.36
or1-6	**5.65**	**0.62**	10.17	0.43	6.38	0.43	11.85	0.40
or2-1	14.31	0.64	9.72	0.71	9.66	0.39	**3.70**	**0.51**
or2-2	10.76	1.09	**6.88**	**1.42**	8.38	0.64	5.17	1.34
or2-3	12.90	0.86	10.89	0.53	6.12	1.03	5.79	0.61
or2-4	11.25	0.73	**7.22**	**1.18**	8.98	0.84	**5.38**	**1.60**
or2-5	12.88	0.71	10.77	0.78	6.23	0.84	5.66	1.54
or2-6	12.06	0.82	**7.63**	**2.05**	9.13	0.47	**4.56**	**1.04**
orsi-1	8.23	0.62	10.33	0.81	**2.12**	**0.94**	9.39	0.35
orsi-2	10.46	0.43	**6.38**	**0.64**	8.98	0.40	5.77	0.62
orsi-3	**5.13**	**0.80**	9.68	0.51	6.63	0.66	11.78	0.37
orsi-4	8.29	0.47	10.24	0.69	**2.22**	**0.66**	9.23	0.33
orsi-5	8.36	0.49	10.17	0.79	**2.26**	**0.65**	9.12	0.55
ts-or1-3	**5.36**	**0.52**	6.49	0.34	9.40	0.31	10.90	0.32
ts-or1-4	12.65	0.96	5.04	0.68	14.20	0.73	8.40	1.04
ts-or1-5	**5.69**	**0.92**	5.58	1.39	9.71	0.84	10.48	0.73
ts-or2-1	12.81	0.93	**5.21**	**2.05**	14.46	0.96	8.62	1.18
ts-orsi-3	12.86	0.53	5.23	0.34	14.27	0.61	8.30	0.36
max		1.18		2.05		1.03		1.60
min		0.41		0.34		0.31		0.32
mean		0.70		0.85		0.66		0.70

Tab. 7-4: Distances between binding pockets and pyridine inhibitor substituent (PYR), as obtained from X-ray structure proposals. RMSF values were calculated from 500 distances along the 5 ns MD trajectory. Most probable binding pocket occupations were determined from lowest distance of all inhibitor substituents and are indicated by bold numbers.

structure	S1'		S1		S2		S3	
	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF	d in Å	RMSF
or1-1	**5.89**	**0.42**	10.28	0.60	5.41	0.33	11.49	0.29
or1-2	**5.09**	**1.89**	9.29	1.09	5.54	1.25	11.07	0.72
or1-3	**5.66**	**0.73**	10.60	0.75	5.65	0.58	11.99	0.72
or1-4	7.42	0.43	9.77	0.54	**2.77**	**0.43**	9.48	0.27
or1-5	7.23	0.36	9.80	0.39	**2.81**	**0.42**	9.75	0.44
or1-6	7.32	0.41	10.06	0.30	**2.54**	**0.41**	9.86	0.29
or2-1	13.36	0.58	11.36	0.39	6.39	0.50	5.98	0.38
or2-2	11.71	1.04	9.21	0.50	6.21	0.96	**4.99**	**0.74**
or2-3	12.60	0.71	8.05	1.11	8.87	0.70	**3.66**	**1.26**
or2-4	11.63	0.84	9.69	0.73	5.83	0.97	5.62	0.98
or2-5	11.96	0.48	**7.82**	**0.66**	8.49	0.48	**4.27**	**0.93**
or2-6	12.62	0.55	10.49	0.87	6.23	0.39	5.58	0.55
orsi-1	**5.96**	**1.32**	10.06	1.12	5.26	1.22	11.20	0.84
orsi-2	11.74	0.38	9.84	0.42	5.70	0.44	**5.68**	**0.52**
orsi-3	7.45	0.40	9.86	0.50	**2.64**	**0.59**	9.58	0.35
orsi-4	**5.86**	**0.85**	9.94	1.07	5.32	0.71	11.16	0.75
orsi-5	**5.99**	**0.46**	10.27	0.74	5.23	0.33	11.36	0.86
ts-or1-3	6.67	0.42	8.59	0.26	**4.21**	**0.28**	9.13	0.32
ts-or1-4	13.69	0.57	7.30	0.63	11.66	0.92	**4.46**	**1.79**
ts-or1-5	6.46	0.49	6.98	0.90	**6.06**	**0.77**	8.57	0.84
ts-or2-1	14.50	1.24	7.66	1.29	12.65	1.59	**4.88**	**1.05**
ts-orsi-3	13.43	0.38	7.20	0.30	11.24	0.45	**4.19**	**0.33**
max		1.89		1.29		1.59		1.79
min		0.36		0.26		0.28		0.27
mean		0.68		0.69		0.67		0.69

Tab. 7-5: Distances between binding pockets and 2-butene inhibitor substituent (BUT), as obtained from X-ray structure proposals. RMSF values were calculated from 500 distances along the 5 ns MD trajectory. Most probable binding pocket occupations were determined from lowest distance of all inhibitor substituents and are indicated by bold numbers.

pose	CHEX	ISO	BUT	PYR
or1-1	2.37 ± 0.67	3.08 ± 0.72	2.13 ± 0.41	2.39 ± 0.59
or1-2	3.04 ± 1.80	5.23 ± 1.97	3.87 ± 1.82	3.19 ± 0.92
or1-3	1.68 ± 0.47	3.50 ± 0.82	1.92 ± 0.86	1.71 ± 0.73
or1-4	1.95 ± 0.42	1.78 ± 0.37	2.15 ± 0.63	3.46 ± 1.04
or1-5	1.55 ± 0.52	3.43 ± 0.74	2.33 ± 0.54	2.83 ± 0.54
or1-6	2.00 ± 0.29	2.19 ± 0.52	1.62 ± 0.35	3.42 ± 0.47
or2-1	3.03 ± 0.69	2.08 ± 0.83	1.6 ± 0.32	2.29 ± 0.42
or2-2	2.88 ± 0.60	5.79 ± 1.12	2.73 ± 0.78	4.97 ± 1.09
or2-3	7.61 ± 1.22	5.08 ± 0.81	2.94 ± 0.94	3.92 ± 0.50
or2-4	5.33 ± 0.91	6.24 ± 1.23	5.06 ± 1.24	4.91 ± 1.30
or2-5	3.72 ± 0.34	2.29 ± 0.37	2.8 ± 0.56	4.15 ± 1.25
or2-6	3.43 ± 0.74	1.88 ± 0.66	2.18 ± 0.76	3.08 ± 1.62
orsi-1	1.50 ± 0.72	1.38 ± 0.62	2.36 ± 1.02	1.71 ± 0.78
orsi-2	1.58 ± 0.74	1.92 ± 0.31	1.29 ± 0.39	1.93 ± 0.70
orsi-3	3.98 ± 0.80	2.00 ± 0.65	2.02 ± 0.83	3.31 ± 0.78
orsi-4	2.11 ± 0.76	2.43 ± 0.72	2.51 ± 0.96	1.95 ± 0.50
orsi-5	2.79 ± 1.31	3.47 ± 1.14	1.78 ± 0.46	2.19 ± 0.51
ts-or1-3	1.14 ± 0.29	1.67 ± 0.74	1.13 ± 0.34	1.20 ± 0.42
ts-or1-4	3.23 ± 1.27	1.73 ± 0.41	2.74 ± 1.72	2.48 ± 0.78
ts-or1-5	4.07 ± 0.93	9.92 ± 2.85	2.67 ± 1.11	4.43 ± 1.39
ts-or2-1	4.66 ± 0.86	3.89 ± 0.84	4.09 ± 1.48	6.43 ± 2.28
ts-orsi-3	1.25 ± 0.36	0.87 ± 0.29	0.99 ± 0.45	1.26 ± 0.41

Tab. 7-6: Averaged RMSD values for individual inhibitor substituents. The substituents are defined in Fig. 5-24. Values are given in Å and are calculated from MD simulations. They refer to heavy atoms of the inhibitor, by taking first simulation frame structure as reference.

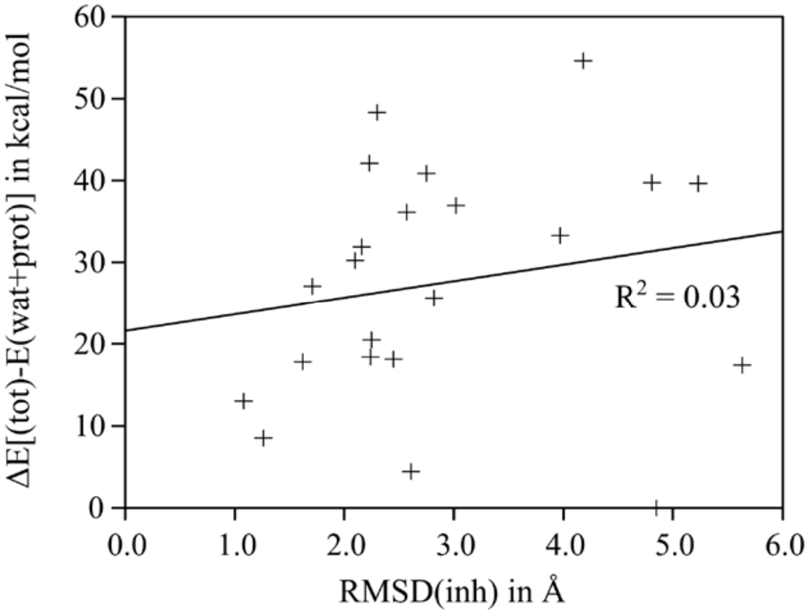


Fig. 7-8: Correlation analysis between RMSD of the inhibitor and its relative energy (Tab. 5-11). No correlation is found as shown by the linear regression (solid line) and coefficient of determination R^2.

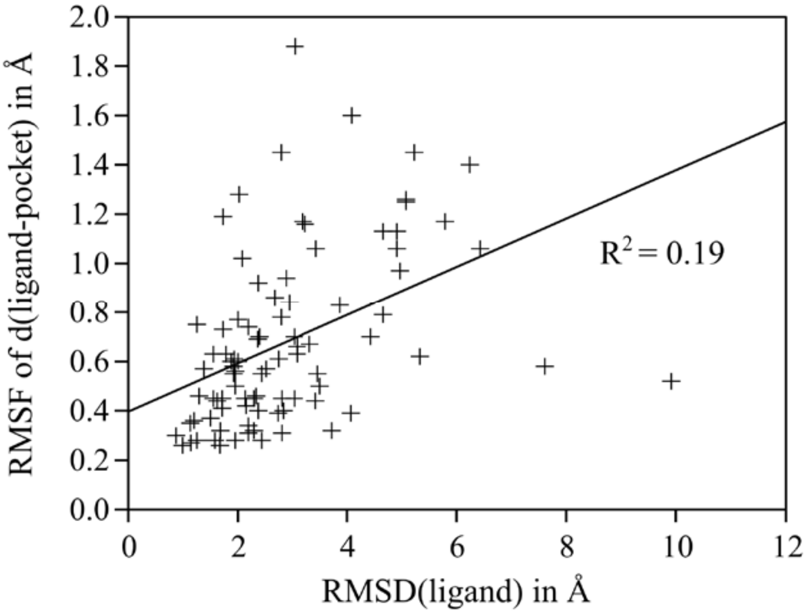


Fig. 7-9: Correlation analysis between RMSD of the individual inhibitor substituents (Tab. 7-6) and the RMSF of the substituent-pocket distance (Tab. 5-11, Tab. 7-2, Tab. 7-3, Tab. 7-4, Tab. 7-5). No obvious correlation is found as shown by the linear regression (solid line) and coefficient of determination R^2.

8 Appendix

8.1 Abbreviations

$3CL^{pro}$	Chymotrypsin-like protease
ACE2	Angiotensin-converting enzyme 2
Ala	Alanine
AM1	Austin model 1
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
B3	Becke`s three parameter functional
CC	Coupled cluster
CDA	Charge deletion analysis
CMK	Chloro methyl ketone
CoV	Coronavirus
COSMO	Continuum solvent model
Cys	Cysteine
DF	Density fitting
DLS	Dynamic ligation screening
DNA	Desoxyribonucleic acid
DSK	Dewar-Sabelli-Klopman approximation
E-protein	Envelope protein
GGA	Generalized gradient approximation
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
GPU	Graphic processor unit
HF	Hartree Fock
HIV	Human immunodeficiency virus
His	Histidine
Ile	Isoleucine
k	Rate constant
k_2	Rate constant of 2nd order
K_D	Dimerization constant
kDa	kilo Dalton
K_i	Dissociation constant
K_M	Michaelis constant
LDA	Local density approximation
Leu	Leucine
LYP	Functional after Lee, Yang, and Parr
Lys	Lysine
MD	Molecular dynamics
Met	Methionine
MM	Molecular mechanical
MNDO	Modified neglect of diatomic overlap
Mn	Minnesota functional in the n-th version, e.g. M06
MPn	Møller Plesset perturbation theory of n-th order
M^{pro}	Main protease
M-protein	Membrane protein
NDDO	Neglect of differential overlaps
N-protein	Nuclecapsid protein
PCM	Polarizable continuum model
PDB	Protein data base
Phe	Phenylalanine
pK_a	Acid dissociation constant
Pl^{pro}	Papain-like protease
PMn	Parametrized method in the n-th version
Pro	Proline
QM	Quantum mechanical
QM/MM	Quantum mechanical/molecular mechanical
RCSB	Research Collaboratory for Structural Bioinformatics
RI	Resolution of identity
RM1	Recife model 1
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid
SARS	Severe acute respiratory syndrome
SAS	Solvent accessible surface
SCC-DFTB	Self-consistent charge density functional tight-binding

SCF	Self consistent field
SCS	Spin component scaling
Ser	Serine
S-protein	Spike protein
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine
ZDO	Zero differential overlap

8.2 Publications

This work was conducted under the direction of Prof. Dr. Bernd Engels at the Institute of Organic Chemistry / Institute of Physical and Theoretical Chemistry at the University Würzburg from September 2007 to January 2013.

Parts of this work have been published in scientific journals and conference contributions.

A. Paasche, T. Schirmeister, B. Engels. "Benchmark Study for the Cysteine–Histidine Proton Transfer Reaction in a Protein Environment: Gas Phase, COSMO, QM/MM Approaches“. *Journal of Chemical Theory and Computation* **2013**, 9, 3, 1765–1777.

T. Schmidt, A.Paasche, C. Grebner, K. Ansorg, J. Becker, W. Lee, B. Engels. "QM/MM Investigations Of Organic Chemistry Oriented Questions“. In *Topics in Current Chemistry*, 1–77. Berlin, Heidelberg: Springer Berlin Heidelberg, **2012**.

M. Breuning, A. Paasche, M. Steiner, S. Dilsky, V. Gessner, C. Strohmann, B. Engels. "Theoretical and spectroscopic studies on the conformational equilibrium of 9-oxabispidines in solution“. *Journal of Molecular Structure* **2011**, 1005, 1–3, 178–185.

A. Paasche, M. Schiller, T. Schirmeister, B. Engels. "Mechanistic Study of the Reaction of Thiol-Containing Enzymes with α,β-Unsaturated Carbonyl Substrates by Computation and Chemoassays“. *ChemMedChem* **2010**, 5, 6, 869–880.

A. Paasche, M. Mladenovic, B. Engels. "Cysteine Proteases and their Inhibition by Michael Acceptor Compounds". Poster presentation (awarded), *46th Symposium on Theoretical Chemistry.* Münster, **2010**.

A. Paasche, M. Mladenovic, B. Engels. "Cysteine Proteases and their Inhibition by Michael Acceptor Compounds". Poster presentation, *International Symposium on Theoretical and Computational Chemistry,* Mülheim an der Ruhr, **2010**.

E. Janke, E. Basílio, S. Schlund, A. Paasche, B. Engels, R. Dede, I. Hussain, P. Langer, M.

Rettig, K. Weisz. "Tautomeric Equilibria of 3-Formylacetylacetone: Low-Temperature NMR Spectroscopy and ab Initio Calculations“. *The Journal of Organic Chemistry* **2009**, 74, 13, 4878–4881.

M. Breuning, M. Steiner, C. Mehler, A. Paasche, D. Hein. "A Flexible Route to Chiral 2-endo-Substituted 9-Oxabispidines and Their Application in the Enantioselective Oxidation of Secondary Alcohols“. *The Journal of Organic Chemistry* **2009**, 74, 3, 1407–1410.

A. Paasche, M. Arnone, R. Fink, T. Schirmeister, B. Engels. "Origin of the Reactivity Differences of Substituted Aziridines: CN vs CC Bond Breakages“. *The Journal of Organic Chemistry* **2009**, 74, 15, 5244–5249.

9 References

[1] *SciFinder*, American Chemical Society, **2012**.

[2] *Web of Knowledge*, Thomas Reuters, **2012**.

[3] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Peñaranda, B. Bankamp, K. Maher, M.-H. Chen, et al., *Science* **2003**, *300*, 1394–1399.

[4] M. A. Marra, S. J. M. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. N. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, et al., *Science* **2003**, *300*, 1399–1404.

[5] J. Peiris, S. Lai, L. Poon, Y. Guan, L. Yam, W. Lim, J. Nicholls, W. Yee, W. Yan, M. Cheung, et al., *Lancet* **2003**, *361*, 1319–1325.

[6] T. G. Ksiazek, D. Erdman, C. S. Goldsmith, S. R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. A. Comer, W. Lim, et al., *N. Engl. J. Med.* **2003**, *348*, 1953–1966.

[7] C. Drosten, S. Günther, W. Preiser, S. van der Werf, H.-R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R. A. M. Fouchier, et al., *N. Engl. J. Med.* **2003**, *348*, 1967–1976.

[8] "World Health Organization. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003.," can be found under http://www.who.int/csr/sars/country/table2004_04_21/en/index.html, **2004**.

[9] N. Lee, D. Hui, A. Wu, P. Chan, P. Cameron, G. M. Joynt, A. Ahuja, M. Y. Yung, C. B. Leung, K. F. To, et al., *N. Engl. J. Med.* **2003**, *348*, 1986–1994.

[10] Y. Guan, B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, P. H. Li, L. J. Zhang, Y. J. Guan, et al., *Science* **2003**, *302*, 276–278.

[11] W. Li, Z. Shi, M. Yu, W. Ren, C. Smith, J. H. Epstein, H. Wang, G. Crameri, Z. Hu, H. Zhang, et al., *Science* **2005**, *310*, 676–679.

[12] W. Ren, X. Qu, W. Li, Z. Han, M. Yu, P. Zhou, S.-Y. Zhang, L.-F. Wang, H. Deng, Z. Shi, *J. Virol.* **2008**, *82*, 1899–1907.

[13] K. V. Holmes, *Science* **2005**, *309*, 1822 –1823.

[14] M. S. Klempner, D. S. Shapiro, *N. Engl. J. Med.* **2004**, *350*, 1171–1172.

[15] T. Wong, C. Lee, W. Tam, J. T. Lau, T. Yu, S. Lui, P. K. S. Chan, Y. Li, J. S. Bresee, J. J. Y. Sung, et al., *Emerging Infect. Dis.* **2004**, *10*, 269–276.

[16] R. M. Anderson, C. Fraser, A. C. Ghani, C. A. Donnelly, S. Riley, N. M. Ferguson, G. M. Leung, T. H. Lam, A. J. Hedley, *Phil. Trans. R. Soc. Lond. B* **2004**, *359*, 1091–1105.

[17] D. L.-M. Goh, B. W. Lee, K. S. Chia, B. H. Heng, M. Chen, S. Ma, C. C. Tan, *Emerging Infect. Dis.* **2004**, *10*, 232–234.

[18] Stacey Knobler, Adel Mahmoud, Stanley Lemon, Alison Mack, Laura Sivitz, and Katherine Oberholtzer, Editors, Forum on Microbial Threats, *Learning from SARS: Preparing for the Next Disease Outbreak - Workshop Summary*, The National Academies Press, Washington, D.C., **2004**.

[19] P. A. Singer, *BMJ* **2003**, *327*, 1342–1344.

[20] J. S. M. Peiris, Y. Guan, K. Y. Yuen, *Nat. Med.* **2004**, *10*, S88–S97.

[21] L. J. Stockman, R. Bellamy, P. Garner, *PLoS Med* **2006**, *3*, e343.

[22] "World Health Organization. Novel coronavirus infection in the United Kingdom," can be found under http://www.who.int/csr/don/2012_09_23/en/index.html, **2012**.

[23] A. M. Zaki, S. van Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, R. A. M. Fouchier, *N. Engl. J. Med.* **2012**, *367*, 1814–1820.

[24] S. van Boheemen, M. de Graaf, C. Lauber, T. M. Bestebroer, V. S. Raj, A. M. Zaki, A. D. M. E. Osterhaus, B. L. Haagmans, A. E. Gorbalenya, E. J. Snijder, et al., *MBio* **2012**, *3*, e00473–00412.

[25] S. K. Lal, Ed., *Molecular Biology of the SARS-Coronavirus*, Springer, **2010**.

[26] P.-K. Hsieh, S. C. Chang, C.-C. Huang, T.-T. Lee, C.-W. Hsiao, Y.-H. Kou, I.-Y. Chen, C.-K. Chang, T.-H. Huang, M.-F. Chang, *J. Virol.* **2005**, *79*, 13848–13855.

[27] H. Hofmann, S. Pöhlmann, *Trends Microbiol.* **2004**, *12*, 466–472.

[28] W. Li, M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough, et al., *Nature* **2003**, *426*, 450–454.

[29] P. Wang, J. Chen, A. Zheng, Y. Nie, X. Shi, W. Wang, G. Wang, M. Luo, H. Liu, L. Tan, et al., *Biochemical and Biophysical Research Communications* **2004**, *315*, 439–444.

[30] D. R. Beniac, A. Andonov, E. Grudeski, T. F. Booth, *Nat. Struct. Mol. Biol.* **2006**, *13*, 751–752.

[31] D. R. Beniac, S. L. deVarennes, A. Andonov, R. He, T. F. Booth, *PLoS ONE* **2007**, *2*, e1082.

[32] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, D. C. Wiley, *Nature* **1997**, *387*, 426–430.

[33] A. M. King, E. Lefkowitz, M. J. Adams, E. B. Carstens, *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, Elsevier, **2011**.

[34] Y. Ruan, C. L. Wei, A. E. Ling, V. B. Vega, H. Thoreau, S. Y. Se Thoe, J.-M. Chia, P. Ng, K. P. Chiu, L. Lim, et al., *The Lancet* **2003**, *361*, 1779–1785.

[35] A. Moya, E. C. Holmes, F. González-Candelas, *Nat. Rev. Microbiol.* **2004**, *2*, 279–288.

[36] E. J. Snijder, P. J. Bredenbeek, J. C. Dobbe, V. Thiel, J. Ziebuhr, L. L. M. Poon, Y. Guan, M. Rozanov, W. J. M. Spaan, A. E. Gorbalenya, *J. Mol. Biol.* **2003**, *331*, 991–1004.

[37] B. H. Harcourt, D. Jukneliene, A. Kanjanahaluethai, J. Bechill, K. M. Severson, C. M. Smith, P. A. Rota, S. C. Baker, *J. Virol.* **2004**, *78*, 13600–13612.

[38] N. D. Rawlings, A. J. Barrett, A. Bateman, *Nucleic Acids Res.* **2011**, *40*, D343–D350.

[39] J. Ziebuhr, J. Herold, S. G. Siddell, *J. Virol.* **1995**, *69*, 4331–4338.

[40] S. Zhang, N. Zhong, F. Xue, X. Kang, X. Ren, J. Chen, C. Jin, Z. Lou, B. Xia, *Protein & Cell* **2010**, *1*, 371–383.

[41] V. Thiel, K. A. Ivanov, Á. Putics, T. Hertzig, B. Schelle, S. Bayer, B. Weißbrich, E. J. Snijder, H. Rabenau, H. W. Doerr, et al., *J. Gen. Virol.* **2003**, *84*, 2305–2315.

[42] A. K. Ghosh, J. Takayama, K. V. Rao, K. Ratia, R. Chaudhuri, D. C. Mulhearn, H. Lee, D. B. Nichols, S. Baliji, S. C. Baker, et al., *J. Med. Chem.* **2010**, *53*, 4968–4979.

[43] A. K. Ghosh, J. Takayama, Y. Aubin, K. Ratia, R. Chaudhuri, Y. Baez, K. Sleeman, M. Coughlin, D. B. Nichols, D. C. Mulhearn, et al., *J. Med. Chem.* **2009**, *52*, 5228–5240.

[44] K. Stadler, V. Masignani, M. Eickmann, S. Becker, S. Abrignani, H.-D. Klenk, R. Rappuoli, *Nat. Rev. Microbiol.* **2003**, *1*, 209–218.

[45] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, R. Hilgenfeld, *Science* **2003**, *300*, 1763–1767.

[46] H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, et al., *PNAS* **2003**, *100*, 13190 –13195.

[47] A. Hegyi, J. Ziebuhr, *J .Gen. Virol.* **2002**, *83*, 595–599.

[48] G. J. Palenik, W. P. Jensen, I.-H. Suh, *J. Chem. Educ.* **2003**, *80*, 753.

[49] M.-F. Hsu, C.-J. Kuo, K.-T. Chang, H.-C. Chang, C.-C. Chou, T.-P. Ko, H.-L. Shr, G.-G. Chang, A. H.-J. Wang, P.-H. Liang, *J. Biol. Chem.* **2005**, *280*, 31257–31266.

[50] I.-L. Lu, N. Mahindroo, P.-H. Liang, Y.-H. Peng, C.-J. Kuo, K.-C. Tsai, H.-P. Hsieh, Y.-S. Chao, S.-Y. Wu, *J. Med. Chem.* **2006**, *49*, 5154–5161.

[51] T.-W. Lee, M. M. Cherney, J. Liu, K. E. James, J. C. Powers, L. D. Eltis, M. N. G. James, *J. Mol. Biol.* **2007**, *366*, 916–932.

[52] X. Xue, H. Yang, W. Shen, Q. Zhao, J. Li, K. Yang, C. Chen, Y. Jin, M. Bartlam, Z. Rao, *J. Mol. Biol.* **2007**, *366*, 965–975.

[53] H. Wang, Y. T. Kim, T. Muramatsu, C. Takemoto, M. Shirouzu, S. Yokoyama, **2007**.

[54] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucl. Acids Res.* **2000**, *28*, 235–242.

[55] J. Shi, J. Sivaraman, J. Song, *J. Virol.* **2008**, *82*, 4620–4629.

[56] K. Fan, P. Wei, Q. Feng, S. Chen, C. Huang, L. Ma, B. Lai, J. Pei, Y. Liu, J. Chen, et al., *J. Biol. Chem.* **2004**, *279*, 1637–1642.

[57] H. Chen, P. Wei, C. Huang, L. Tan, Y. Liu, L. Lai, *J. Biol. Chem.* **2006**, *281*, 13894 –13898.

[58] J. Solowiej, J. A. Thomson, K. Ryan, C. Luo, M. He, J. Lou, B. W. Murray, *Biochemistry* **2008**, *47*, 2617–2630.

[59] F. Jordan, *Science* **2004**, *306*, 818–820.

[60] V. Graziano, W. J. McGrath, L. Yang, W. F. Mangel, *Biochemistry* **2006**, *45*, 14632–14641.

[61] S. Chen, L. Chen, J. Tan, J. Chen, L. Du, T. Sun, J. Shen, K. Chen, H. Jiang, X. Shen, *J. Biol. Chem.* **2005**, *280*, 164–173.

[62] C.-Y. Chou, H.-C. Chang, W.-C. Hsu, T.-Z. Lin, C.-H. Lin, G.-G. Chang, *Biochemistry* **2004**, *43*, 14958–14970.

[63] P.-Y. Lin, C.-Y. Chou, H.-C. Chang, W.-C. Hsu, G.-G. Chang, *Arch. Biochem. Biophys.* **2008**, *472*, 34–42.

[64] B. Xia, X. Kang, *Protein & Cell* **2011**, *2*, 282–290.

[65] W. Appel, *Clin. Biochem.* **1986**, *19*, 317–322.

[66] K.-W. Zheng, Q.-S. Yu, Y.-H. Wang, B. Zhang, G.-X. Hu, *Act. Chim. Sin.* **2004**, *62*, 542–549.

[67] K. Zheng, G. Ma, J. Zhou, M. Zen, W. Zhao, Y. Jiang, Q. Yu, J. Feng, *Proteins* **2007**, *66*, 467–479.

[68] J. Shi, Z. Wei, J. Song, *J. Biol. Chem.* **2004**, *279*, 24765–24773.

[69] J. Barrila, U. Bacha, E. Freire, *Biochemistry* **2006**, *45*, 14908–14916.

[70] N. Zhong, S. Zhang, P. Zou, J. Chen, X. Kang, Z. Li, C. Liang, C. Jin, B. Xia, *J. Virol.* **2008**, *82*, 4227–4234.

[71] C. Huang, P. Wei, K. Fan, Y. Liu, L. Lai, *Biochemistry* **2004**, *43*, 4568–4574.

[72] X. Xue, H. Yu, H. Yang, F. Xue, Z. Wu, W. Shen, J. Li, Z. Zhou, Y. Ding, Q. Zhao, et al., *J. Virol.* **2008**, *82*, 2515–2527.

[73] A. D. McNaught, A. Wilkinson, *Compendium of Chemical Terminology: Iupac Recommendations: Gold Book*, IUPAC International Union Of Pure And Applied Chem, **1997**.

[74] A. Cornish-Bowden, *Fundamentals of Enzyme Kinetics*, Wiley-VCH Verlag GmbH & Co. KGaA, **2012**.

[75] V. Graziano, W. J. McGrath, A. M. DeGruccio, J. J. Dunn, W. F. Mangel, *FEBS Letters* **2006**, *580*, 2577–2583.

[76] U. Bacha, J. Barrila, A. Velazquez-Campoy, S. A. Leavitt, E. Freire, *Biochemistry* **2004**, *43*, 4906–4912.

[77] S. Chen, L. Chen, H. Luo, T. Sun, J. Chen, F. Ye, J. Cai, J. Shen, X. Shen, H. Jiang, *Acta Pharmacol. Sin* **2005**, *26*, 99–106.

[78] J. Tan, K. H. G. Verschueren, K. Anand, J. Shen, M. Yang, Y. Xu, Z. Rao, J. Bigalke, B. Heisen, J. R. Mesters, et al., *J. Mol. Biol.* **2005**, *354*, 25–40.

[79] T. Hu, Y. Zhang, L. Li, K. Wang, S. Chen, J. Chen, J. Ding, H. Jiang, X. Shen, *Virology* **2009**, *388*, 324–334.

[80] K. Fan, L. Ma, X. Han, H. Liang, P. Wei, Y. Liu, L. Lai, *BBRC* **2005**, *329*, 934–940.

[81] I. Schechter, A. Berger, *Biochem. Biophys. Res. Commun.* **1967**, *27*, 157–162.

[82] C.-P. Chuck, L.-T. Chong, C. Chen, H.-F. Chow, D. C.-C. Wan, K.-B. Wong, *PLoS ONE* **2010**, *5*, e13197.

[83] K. Phakthanakanok, K. Ratanakhanokchai, K. Kyu, P. Sompornpisut, A. Watts, S. Pinitglang, *BMC Bioinformatics* **2009**, *10*, S48.

[84] R. Menard, J. Carriere, P. Laflamme, C. Plouffe, H. E. Khouri, T. Vernet, D. C. Tessier, D. Y. Thomas, A. C. Storer, *Biochemistry* **1991**, *30*, 8924–8928.

[85] E. Fischer, *Ber. Dtsch. Chem. Ges.* **1890**, *23*, 2611–2624.

[86] E. Fischer, *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985–2993.

[87] D. E. Koshland, *PNAS* **1958**, *44*, 98–104.

[88] D. E. Koshland, *Angew. Chem. Int. Ed. Engl.* **1995**, *33*, 2375–2378.

[89] J. Yin, C. Niu, M. M. Cherney, J. Zhang, C. Huitema, L. D. Eltis, J. C. Vederas, M. N. G. James, *J. Mol. Biol.* **2007**, *371*, 1060–1074.

[90] J. Kraut, *Annual Review of Biochemistry* **1977**, *46*, 331–358.

[91] P. Carter, J. A. Wells, *Nature* **1988**, *332*, 564–568.

[92] Y.-P. Pang, *Proteins* **2004**, *57*, 747–757.

[93] M. Bartlam, H. Yang, Z. Rao, *Curr. Opin. Struc. Biol.* **2005**, *15*, 664–672.

[94] S. D. Lewis, F. A. Johnson, J. A. Shafer, *Biochemistry* **1976**, *15*, 5009–5017.

[95] A. J. Barrett, N. D. Rawlings, J. F. Woessner, *Handbook of Proteolytic Enzymes*, Academic Press, San Diego [etc.], **1998**.

[96] A. Storer, R. Ménard, in *Methods in Enzymology* (Ed.: Alan J. Barrett), Academic Press, **1994**, pp. 486–500.

[97] L. G. Theodorou, J. G. Bieth, E. M. Papamichael, *Bioresource Technology* **2007**, *98*, 1931–1939.

[98] Z. Sárkány, Z. Szeltner, L. Polgár, *Biochemistry* **2001**, *40*, 10601–10606.

[99] Z. Sárkány, L. Polgár, *Biochemistry* **2003**, *42*, 516–522.

[100] B. A. Malcolm, *Protein Science* **1995**, *4*, 1439–1445.
[101] Z. Ke, Y. Zhou, P. Hu, S. Wang, D. Xie, Y. Zhang, *J. Phys. Chem. B* **2009**, *113*, 12750–12758.
[102] M. Shokhen, N. Khazanov, A. Albeck, *Proteins* **2009**, *77*, 916–926.
[103] G. Davies, M. L. Sinnott, S. G. Withers, M. L. Sinnott, *Comprehensive Biological Catalysis*, **1998**.
[104] A. Fersht, *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding*, WH Freeman, **1998**.
[105] J. C. Powers, J. L. Asgian, Ö. D. Ekici, K. E. James, *Chem. Rev.* **2002**, *102*, 4639–4750.
[106] H. Yang, W. Xie, X. Xue, K. Yang, J. Ma, W. Liang, Q. Zhao, Z. Zhou, D. Pei, J. Ziebuhr, et al., *Plos Biol* **2005**, *3*, e324.
[107] L. Stryer, J. M. Berg, J. L. Tymoczko, *Biochemistry*, W H Freeman, **2002**.
[108] P. J. Gray, R. G. Duggleby, *Biochem J* **1989**, *257*, 419–424.
[109] M. Dixon, E. C. Webb, *Enzymes*, Longman, **1979**.
[110] R. Leung-Toung, Y. Zhao, W. Li, T. F. Tam, K. Karimian, M. Spino, *Curr. Med. Chem.* **2006**, *13*, 547–581.
[111] T. P. Smyth, *Bioorg. Med. Chem.* **2004**, *12*, 4081–4088.
[112] M. McCall, C. Toso, J. Emamaullee, R. Pawlick, R. Edgar, J. Davis, A. Maciver, T. Kin, R. Arch, A. M. J. Shapiro, *Surgery* **2011**, *150*, 48–55.
[113] S. D. Linton, T. Aja, R. A. Armstrong, X. Bai, L.-S. Chen, N. Chen, B. Ching, P. Contreras, J.-L. Diaz, C. D. Fisher, et al., *J. Med. Chem.* **2005**, *48*, 6779–6782.
[114] Z. Yu, P. Caldera, F. McPhee, J. J. De Voss, P. R. Jones, A. L. Burlingame, I. D. Kuntz, C. S. Craik, P. R. Ortiz de Montellano, *J. Am. Chem. Soc.* **1996**, *118*, 5846–5856.
[115] H.-H. Otto, T. Schirmeister, *Chem. Rev.* **1997**, *97*, 133–172.
[116] S. Yang, S.-J. Chen, M.-F. Hsu, J.-D. Wu, C.-T. K. Tseng, Y.-F. Liu, H.-C. Chen, C.-W. Kuo, C.-S. Wu, L.-W. Chang, et al., *J. Med. Chem* **2006**, *49*, 4971–4980.
[117] K. H. G. Verschueren, K. Pumpor, S. Anemüller, S. Chen, J. R. Mesters, R. Hilgenfeld, *Chem. Biol.* **2008**, *15*, 597–606.
[118] E. Martina, N. Stiefl, B. Degel, F. Schulz, A. Breuning, M. Schiller, R. Vicik, K. Baumann, J. Ziebuhr, T. Schirmeister, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 5365–5369.
[119] C.-P. Chuck, C. Chen, Z. Ke, D. Chi-Cheong Wan, H.-F. Chow, K.-B. Wong, *European Journal of Medicinal Chemistry* **2013**, *59*, 1–6.
[120] P. Wei, K. Fan, H. Chen, L. Ma, C. Huang, L. Tan, D. Xi, C. Li, Y. Liu, A. Cao, et al., *Biochem. Biophys. Res. Commun* **2006**, *339*, 865–872.
[121] C.-C. Lee, C.-J. Kuo, M.-F. Hsu, P.-H. Liang, J.-M. Fang, J.-J. Shie, A. H.-J. Wang, *FEBS Letters* **2007**, *581*, 5454–5458.
[122] C.-N. Chen, C. P. C. Lin, K.-K. Huang, W.-C. Chen, H.-P. Hsieh, P.-H. Liang, J. T.-A. Hsu, *Evid. Based. Complement Alternat. Med.* **2005**, *2*, 209–215.
[123] C. Yung-Chi, W. H. Prusoff, *Biochem. Pharm.* **1973**, *22*, 3099–3108.
[124] P.-H. Liang, *Curr. Top. Med. Chem.* **2006**, *6*, 361–376.
[125] E. De Clercq, *Expert Rev. Anti Infect. Ther.* **2006**, *4*, 291–302.
[126] S. Zhai, W. Liu, B. Yan, *Recent. Pat. Antiinfect. Drug Discov.* **2007**, *2*, 1–10.
[127] K.-C. Tsai, S.-Y. Chen, P.-H. Liang, I.-L. Lu, N. Mahindroo, H.-P. Hsieh, Y.-S. Chao,

L. Liu, D. Liu, W. Lien, et al., *J. Med. Chem* **2006**, *49*, 3485–3495.

[128] P. Mukherjee, F. Shah, P. Desai, M. Avery, *J. Chem. Inf. Model.* **2011**, *51*, 1376–1392.

[129] M. F. Schmidt, A. Isidro-Llobet, M. Lisurek, A. El-Dahshan, J. Tan, R. Hilgenfeld, J. Rademann, *Angew. Chem. Int. Ed.* **2008**, *47*, 3275–3278.

[130] A. K. Ghosh, K. Xi, K. Ratia, B. D. Santarsiero, W. Fu, B. H. Harcourt, P. A. Rota, S. C. Baker, M. E. Johnson, A. D. Mesecar, *J. Med. Chem.* **2005**, *48*, 6767–6771.

[131] T.-W. Lee, M. M. Cherney, C. Huitema, J. Liu, K. E. James, J. C. Powers, L. D. Eltis, M. N. G. James, *J. Mol. Biol.* **2005**, *353*, 1137–1151.

[132] D. H. Goetz, Y. Choe, E. Hansell, Y. T. Chen, M. McDowell, C. B. Jonsson, W. R. Roush, J. McKerrow, C. S. Craik, *Biochemistry* **2007**, *46*, 8744–8752.

[133] K. Akaji, H. Konno, H. Mitsui, K. Teruya, Y. Shimamoto, Y. Hattori, T. Ozaki, M. Kusunoki, A. Sanjoh, *J. Med. Chem.* **2011**, *54*, 7962–7973.

[134] J. E. Blanchard, N. H. Elowe, C. Huitema, P. D. Fortin, J. D. Cechetto, L. D. Eltis, E. D. Brown, *Chemistry & Biology* **2004**, *11*, 1445–1453.

[135] R. P. Jain, H. I. Pettersson, J. Zhang, K. D. Aull, P. D. Fortin, C. Huitema, L. D. Eltis, J. C. Parrish, M. N. G. James, D. S. Wishart, et al., *J. Med. Chem.* **2004**, *47*, 6113–6116.

[136] L. Chen, C. Gui, X. Luo, Q. Yang, S. Günther, E. Scandella, C. Drosten, D. Bai, X. He, B. Ludewig, et al., *J. Virol* **2005**, *79*, 7095–7103.

[137] Q. Yang, L. Chen, X. He, Z. Gao, X. Shen, D. Bai, *Chem. Pharm. Bull.* **2008**, *56*, 1400–1405.

[138] J.-J. Shie, J.-M. Fang, C.-J. Kuo, T.-H. Kuo, P.-H. Liang, H.-J. Huang, W.-B. Yang, C.-H. Lin, J.-L. Chen, Y.-T. Wu, et al., *J. Med. Chem.* **2005**, *48*, 4469–4473.

[139] L.-R. Chen, Y.-C. Wang, Y. W. Lin, S.-Y. Chou, S.-F. Chen, L. T. Liu, Y.-T. Wu, C.-J. Kuo, T. S.-S. Chen, S.-H. Juang, *Bioorganic & Medicinal Chemistry Letters* **2005**, *15*, 3058–3062.

[140] J.-J. Shie, J.-M. Fang, T.-H. Kuo, C.-J. Kuo, P.-H. Liang, H.-J. Huang, Y.-T. Wu, J.-T. Jan, Y.-S. E. Cheng, C.-H. Wong, *Bioorg. Med. Chem* **2005**, *13*, 5240–5252.

[141] U. Kaeppler, N. Stiefl, M. Schiller, R. Vicik, A. Breuning, W. Schmitz, D. Rupprecht, C. Schmuck, K. Baumann, J. Ziebuhr, et al., *J. Med. Chem* **2005**, *48*, 6832–6842.

[142] C.-Y. Wu, K.-Y. King, C.-J. Kuo, J.-M. Fang, Y.-T. Wu, M.-Y. Ho, C.-L. Liao, J.-J. Shie, P.-H. Liang, C.-H. Wong, *Chemistry & Biology* **2006**, *13*, 261–268.

[143] Y.-M. Shao, W.-B. Yang, H.-P. Peng, M.-F. Hsu, K.-C. Tsai, T.-H. Kuo, A. H.-J. Wang, P.-H. Liang, C.-H. Lin, A.-S. Yang, et al., *ChemBioChem* **2007**, *8*, 1654–1657.

[144] K. Akaji, H. Konno, M. Onozuka, A. Makino, H. Saito, K. Nosaka, *Bioorganic & Medicinal Chemistry* **2008**, *16*, 9400–9408.

[145] C.-J. Kuo, H.-G. Liu, Y.-K. Lo, C.-M. Seong, K.-I. Lee, Y.-S. Jung, P.-H. Liang, *FEBS Letters* **2009**, *583*, 549–555.

[146] T. T. Hanh Nguyen, H.-J. Ryu, S.-H. Lee, S. Hwang, V. Breton, J. H. Rhee, D. Kim, *Bioorg. Med. Chem. Lett.* **2011**, *21*, 3088–3091.

[147] L. Zhu, S. George, M. F. Schmidt, S. I. Al-Gharabli, J. Rademann, R. Hilgenfeld, *Antiviral Res.* **2011**, *92*, 204–212.

[148] S. L. Binford, F. Maldonado, M. A. Brothers, P. T. Weady, L. S. Zalman, J. W. Meador, D. A. Matthews, A. K. Patick, *Antimicrob Agents Chemother* **2005**, *49*, 619–626.

[149] D. A. Matthews, P. S. Dragovich, S. E. Webber, S. A. Fuhrman, A. K. Patick, L. S. Zalman, T. F. Hendrickson, R. A. Love, T. J. Prins, J. T. Marakovits, et al., *PNAS* **1999**, *96*, 11000 –11007.

[150] A. G. Taranto, P. Carvalho, M. A. Avery, *J. Mol. Graph. Model* **2008**, *27*, 275–285.

[151] R. Withnall, B. Z. Chowdhry, S. Bell, T. J. Dines, *J. Chem. Educ.* **2007**, *84*, 1364.

[152] W. Thiel, *J. Phys. Chem. A* **2009**, *113*, 11457–11464.

[153] W. Thiel, *Angew. Chem. Int. Ed. Engl.* **2011**, 50, 9216–9217.

[154] J. A. Roberts, M. J. Kuiper, B. R. Thorley, P. M. Smooker, A. Hung, *Journal of Molecular Graphics and Modelling* **2012**, *38*, 165–173.

[155] D. B. Boyd, K. B. Lipkowitz, *J. Chem. Educ.* **1982**, *59*, 269.

[156] F. Jensen, *Introduction to Computational Chemistry*, John Wiley & Sons, **2006**.

[157] N. L. Allinger, in *Advances in Physical Organic Chemistry* (Ed.: V. Gold and D. Bethell), Academic Press, **1976**, pp. 1–82.

[158] J. E. Lennard-Jones, *Proc. R. Soc. London, Ser. A* **1925**, *109*, 584–597.

[159] A. D. Mackerell Jr, M. Feig, C. L. Brooks 3rd, *J. Comput. Chem.* **2004**, *25*, 1400–1415.

[160] MacKerell, D. Bashford, Bellott, Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al., *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[161] M. Buck, S. Bouguet-Bonnet, R. W. Pastor, A. D. MacKerell, *Biophys. J.* **2006**, *90*, L36–L38.

[162] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[163] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **2006**, *65*, 712–725.

[164] L. D. Schuler, X. Daura, W. F. van Gunsteren, *J. Comput. Chem.* **2001**, *22*, 1205–1218.

[165] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, W. F. van Gunsteren, *Eur. Biophys. J.* **2011**, *40*, 843–856.

[166] J. A. McCammon, B. R. Gelin, M. Karplus, *Nature* **1977**, *267*, 585–590.

[167] V. E. Lamberti, L. D. Fosdick, E. R. Jessup, C. J. C. Schauble, *J. Chem. Educ.* **2002**, *79*, 601.

[168] L. Verlet, *Phys. Rev.* **1967**, *159*, 98–103.

[169] K. Lipkowitz, *J. Chem. Educ.* **1995**, *72*, 1070.

[170] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2012**, *8*, 1542-1555.

[171] B. G. Levine, J. E. Stone, A. Kohlmeyer, *J. Comp. Phys.* **2011**, *230*, 3556–3569.

[172] J. C. Phillips, J. E. Stone, *Commun. ACM* **2009**, *52*, 34–41.

[173] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, K. Schulten, *J. Comput. Chem.* **2007**, *28*, 2618–2640.

[174] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, K. Schulten, *Journal of Molecular Graphics and Modelling* **2010**, *29*, 116–125.

[175] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, et al., *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

[176] I. V. Leontyev, A. A. Stuchebrukhov, *J. Chem. Theory Comput.* **2012**, *8*, 3207–3216.

[177] F. R. Salsbury Jr, *Current Opinion in Pharmacology* **2010**, *10*, 738–744.

[178] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, D. E. Shaw, *JACS* **2011**, *133*, 9181–9183.

[179] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, *J. Phys. Chem. A* **2001**, *105*, 9396–9409.

[180] J. Stewart, *J. Mol. Model.* **2009**, *15*, 765–805.

[181] I. S. Ufimtsev, N. Luehr, T. J. Martinez, *J. Phys. Chem. Lett.* **2011**, *2*, 1789–1793.

[182] L. Hu, J. Eliasson, J. Heimdal, U. Ryde, *J. Phys. Chem. A* **2009**, *113*, 11793–11800.

[183] H. M. Senn, W. Thiel, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[184] M. Dal Peraro, P. Ruggerone, S. Raugei, F. L. Gervasio, P. Carloni, *Current Opinion in Structural Biology* **2007**, *17*, 149–156.

[185] J. Thar, W. Reckien, B. Kirchner, in *Atomistic Approaches in Modern Biology* (Ed.: M. Reiher), Springer Berlin / Heidelberg, **2007**, pp. 133–171.

[186] A. Szabo, J. Szabo, Chemistry, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Pubn Inc, **1996**.

[187] R. Eisberg, R. Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, Wiley, **1985**.

[188] E. R. Davidson, D. Feller, *Chem. Rev.* **1986**, *86*, 681–696.

[189] C. Møller, M. S. Plesset, *Phys. Rev.* **1934**, *46*, 618–622.

[190] B. Levy, G. Berthier, *International Journal of Quantum Chemistry* **1968**, *2*, 307–319.

[191] T. Helgaker, P. Jorgensen, J. Olsen, *Molecular Electronic-Structure Theory*, Wiley, **2000**.

[192] G. E. Scuseria, T. J. Lee, *J. Chem. Phys.* **1990**, *93*, 5851–5855.

[193] H.-J. Werner, M. Schütz, *J. Chem. Phys.* **2011**, *135*, 144116.

[194] M. Schütz, F. R. Manby, *Phys. Chem. Chem. Phys.* **2003**, *5*, 3349.

[195] H.-J. Werner, K. Pflüger, in *Annual Reports in Computational Chemistry* (Ed.: David C. Spellmeyer), Elsevier, **2006**, pp. 53–80.

[196] T. B. Adler, G. Knizia, H.-J. Werner, *J. Chem. Phys.* **2007**, *127*, 221106.

[197] T. B. Adler, H.-J. Werner, *J. Chem. Phys.* **2011**, *135*, 144117.

[198] C. Hättig, F. Weigend, *J. Chem. Phys.* **2000**, *113*, 5154–5161.

[199] O. Christiansen, H. Koch, P. Jørgensen, *Chemical Physics Letters* **1995**, *243*, 409–418.

[200] S. Grimme, *J. Chem. Phys.* **2003**, *118*, 9095.

[201] A. Hellweg, S. A. Grün, C. Hättig, *Phys. Chem. Chem. Phys.* **2008**, *10*, 4119.

[202] J. Antony, S. Grimme, *J. Phys. Chem. A* **2007**, *111*, 4862–4868.

[203] M. J. S. Dewar, W. Thiel, *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

[204] G. Klopman, *J. Am. Chem. Soc.* **1964**, *86*, 4550–4557.

[205] W. Thiel, A. A. Voityuk, *Theor. Chim. Acta* **1992**, *81*, 391–404.

[206] W. Thiel, A. A. Voityuk, *Theor. Chim. Acta* **1996**, *93*, 315–315.

[207] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

[208] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 209–220.

[209] G. B. Rocha, R. O. Freire, A. M. Simas, J. J. P. Stewart, *J. Comput. Chem.* **2006**, *27*, 1101–1111.

[210] J. Stewart, *J. Mol. Model.* **2007**, *13*, 1173–1213.

[211] K. W. Sattelmeyer, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. A* **2006**, *110*, 13551–13559.

[212] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, G. Seifert, *Phys. Rev. B* **1998**, *58*, 7260–7268.

[213] N. Otte, M. Scholten, W. Thiel, *J. Phys. Chem. A* **2007**, *111*, 5751–5755.

[214] C. M. Maupin, B. Aradi, G. A. Voth, *J. Phys. Chem. B* **2010**, *114*, 6922–6931.

[215] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864–B871.

[216] K. Burke, L. O. Wagner, *Int. J. Quantum Chem.* **2013**, *113*, 96-101.

[217] J. P. Perdew, K. Schmidt, *AIP Conf. Proc.* **2001**, *577*, 1–20.

[218] M. Korth, S. Grimme, *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

[219] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098–3100.

[220] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 1372.

[221] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[222] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.

[223] Y. Zhao, D. G. Truhlar, *Theor Chem Account* **2007**, *120*, 215–241.

[224] Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* **2006**, *125*, 194101.

[225] Y. Zhao, D. G. Truhlar, *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.

[226] R. Peverati, D. G. Truhlar, *J. Phys. Chem. Lett.* **2011**, *2*, 2810–2817.

[227] R. Peverati, D. G. Truhlar, *J. Phys. Chem. Lett.* **2011**, *3*, 117–124.

[228] P. R. Schreiner, *Angew. Chem.* **2007**, *119*, 4295–4297.

[229] A. J. Cohen, P. Mori-Sanchez, W. Yang, *Science* **2008**, *321*, 792–794.

[230] E. R. Johnson, I. D. Mackie, G. A. DiLabio, *J. Phys. Org. Chem.* **2009**, *22*, 1127–1135.

[231] S. Grimme, *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[232] M. Wong, *Chemical Physics Letters* **1996**, *256*, 391–399.

[233] J. P. Merrick, D. Moran, L. Radom, *J. Phys. Chem. A* **2007**, *111*, 11683–11700.

[234] I. S. Ufimtsev, T. J. Martinez, *Computing in Science Engineering* **2008**, *10*, 26–34.

[235] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999–3094.

[236] B. Roux, T. Simonson, *Biophys. Chem.* **1999**, *78*, 1–20.

[237] S. Miertuš, E. Scrocco, J. Tomasi, *Chemical Physics* **1981**, *55*, 117–129.

[238] L. Onsager, *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.

[239] J. G. Kirkwood, *J. Chem. Phys.* **1939**, *7*, 911–919.

[240] C. L. de O. Mendes, C. O. da Silva, E. C. da Silva, *J. Phys. Chem. A* **2006**, *110*, 4034–4041.

[241] M. F. Iozzi, M. Cossi, R. Improta, N. Rega, V. Barone, *J. Chem. Phys.* **2006**, *124*, 184103–184103–9.

[242] D. G. Fedorov, K. Kitaura, H. Li, J. H. Jensen, M. S. Gordon, *J. Comput. Chem.* **2006**, *27*, 976–985.

[243] C. Steinmann, K. L. Blædel, A. S. Christensen, J. H. Jensen, *submitted* **2013**.

[244] A. Klamt, G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

[245] A. Klamt, F. Eckert, W. Arlt, *Annu. Rev. Chem. Biomol. Eng.* **2010**, *1*, 101–122.

[246] R. de P. Soares, *Ind. Eng. Chem. Res.* **2011**, *50*, 3060–3063.

[247] G. C. Akerlof, H. I. Oshry, *J. Am. Chem. Soc.* **1950**, *72*, 2844–2847.

[248] M. K. Gilson, B. H. Honig, *Biopolymers* **1986**, *25*, 2097–2119.

[249] T. Simonson, C. L. Brooks, *J. Am. Chem. Soc.* **1996**, *118*, 8452–8458.
[250] M. Klähn, S. Braun-Sand, E. Rosta, A. Warshel, *J. Phys. Chem. B* **2005**, *109*, 15645–15650.
[251] J. Zheng, A. Altun, W. Thiel, *J. Comput. Chem.* **2007**, *28*, 2147–2158.
[252] S. C. L. Kamerlin, M. Haranczyk, A. Warshel, *J. Phys. Chem. B* **2009**, *113*, 1253–1272.
[253] S. Kumbhar, F. D. Fischer, M. P. Waller, *J. Chem. Inf. Model.* **2012**, *52*, 93–98.
[254] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227–249.
[255] M. J. Field, P. A. Bash, M. Karplus, *Journal of Computational Chemistry* **1990**, *11*, 700–733.
[256] P. E. Sinclair, A. de Vries, P. Sherwood, C. Richard A. Catlow, R. A. van Santen, *J. Am. Chem. Soc., Faraday Trans.* **1998**, *94*, 3401–3408.
[257] D. Bakowies, W. Thiel, *J. Phys. Chem.* **1996**, *100*, 10580–10594.
[258] O. Acevedo, W. L. Jorgensen, *Acc. Chem. Res.* **2010**, *43*, 142–151.
[259] L. Polgár, P. Halász, *Biochem J.* **1982**, *207*, 1–10.
[260] Z. Fu, X. Li, Y. Miao, K. M. Merz, *J. Chem. Theory Comput.* **2013**, *9*, 1686-1693.
[261] M. N. Davies, C. P. Toseland, D. S. Moss, D. R. Flower, *BMC Biochem.* **2006**, *7*, 18.
[262] C. L. Stanton, K. N. Houk, *J. Chem. Theory Comput.* **2008**, *4*, 951–966.
[263] C. Liao, M. C. Nicklaus, *J. Chem. Inf. Model.* **2009**, *49*, 2801–2812.
[264] Y. Y. Sham, Z. T. Chu, A. Warshel, *J. Phys. Chem. B* **1997**, *101*, 4458–4472.
[265] R. Borštnar, M. Repič, S. C. L. Kamerlin, R. Vianello, J. Mavri, *J. Chem. Theory Comput.* **2012**, *8*, 3864–3870.
[266] J. Tomasi, M. Persico, *Chem. Rev.* **1994**, *94*, 2027–2094.
[267] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99*, 2161–2200.
[268] M. Orozco, F. J. Luque, *Chem. Rev.* **2000**, *100*, 4187–4226.
[269] H. Hu, W. Yang, *Journal of Molecular Structure: THEOCHEM* **2009**, *898*, 17–30.
[270] R. Zhang, B. Lev, J. E. Cuervo, S. Y. Noskov, D. R. Salahub, in *Advances in Quantum Chemistry*, Elsevier, **2010**, pp. 353–400.
[271] T. C. Schmidt, A. Paasche, C. Grebner, K. Ansorg, J. Becker, W. Lee, B. Engels, in *Topics in Current Chemistry*, Springer Berlin Heidelberg, Berlin, Heidelberg, **2012**, pp. 1–77.
[272] S. Zhang, *J. Comput. Chem.* **2012**, *33*, 517–526.
[273] C. N. Schutz, A. Warshel, *Proteins: Structure, Function, and Bioinformatics* **2001**, *44*, 400–417.
[274] M. Štrajbl, J. Florián, A. Warshel, *J. Phys. Chem. B* **2001**, *105*, 4471–4484.
[275] A. Paasche, M. Schiller, T. Schirmeister, B. Engels, *ChemMedChem* **2010**, *5*, 869–880.
[276] A. Paasche, M. Arnone, R. F. Fink, T. Schirmeister, B. Engels, *J. Org. Chem.* **2009**, *74*, 5244–5249.
[277] E. Derat, S. Shaik, C. Rovira, P. Vidossich, M. Alfonso-Prieto, *J. Am. Chem. Soc.* **2007**, *129*, 6346–6347.
[278] J. Kästner, P. Sherwood, *Mol. Phys.* **2010**, *108*, 293.
[279] M. Mladenovic, R. F. Fink, W. Thiel, T. Schirmeister, B. Engels, *J. Am. Chem. Soc.* **2008**, *130*, 8696–8705.
[280] M. Mladenovic, K. Junold, R. F. Fink, W. Thiel, T. Schirmeister, B. Engels, *J. Phys.*

Chem. B **2008**, *112*, 5458–5469.

[281] M. Kaukonen, P. Söderhjelm, J. Heimdal, U. Ryde, *J. Chem. Theory Comput.* **2008**, *4*, 985–1001.

[282] Z. Ke, H. Guo, D. Xie, S. Wang, Y. Zhang, *J. Phys. Chem. B* **2011**, *115*, 3725–3733.

[283] W. Lee, S. R. Luckner, C. Kisker, P. J. Tonge, B. Engels, *Biochemistry* **2011**, *50*, 5743–5756.

[284] X. Hu, H. Hu, J. A. Melvin, K. W. Clancy, D. G. McCafferty, W. Yang, *J. Am. Chem. Soc.* **2011**, *133*, 478–485.

[285] Y. Cheng, X. Cheng, Z. Radić, J. A. McCammon, *J. Am. Chem. Soc.* **2007**, *129*, 6562–6570.

[286] M. Mladenovic, T. Schirmeister, S. Thiel, W. Thiel, B. Engels, *ChemMedChem* **2007**, *2*, 120–128.

[287] H.-B. Xie, Y. Zhou, Y. Zhang, J. K. Johnson, *J. Phys. Chem. A* **2010**, *114*, 11844–11852.

[288] M. Kaukonen, P. Söderhjelm, J. Heimdal, U. Ryde, *J. Phys. Chem. B* **2008**, *112*, 12537–12548.

[289] M. H. M. Olsson, P. E. M. Siegbahn, A. Warshel, *J. Am. Chem. Soc.* **2004**, *126*, 2820–2828.

[290] K. F. Wong, J. L. Sonnenberg, F. Paesani, T. Yamamoto, J. Vaníček, W. Zhang, H. B. Schlegel, D. A. Case, T. E. Cheatham, W. H. Miller, et al., *J. Chem. Theory Comput.* **2010**, *6*, 2566–2580.

[291] C. D. Christ, A. E. Mark, W. F. van Gunsteren, *J. Comput. Chem.* **2010**, *31*, 1569–1582.

[292] G.-S. Li, M. T. C. Martins-Costa, C. Millot, M. F. Ruiz-López, *Chem.Phys. Letters* **1998**, *297*, 38–44.

[293] B. S. Jursic, *J. Mol. Struct.* **1998**, *427*, 137–142.

[294] S. Sadhukhan, D. Muñoz, C. Adamo, G. E. Scuseria, *Chem.Phys. Letters* **1999**, *306*, 83–87.

[295] D. J. Anick, *J. Phys. Chem. A* **2003**, *107*, 1348–1358.

[296] K. Range, D. Riccardi, Q. Cui, M. Elstner, D. M. York, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.

[297] J. A. Frey, A. Müller, M. Losada, S. Leutwyler, *J. Phys. Chem. B* **2007**, *111*, 3534–3542.

[298] N. F. Brás, M. A. S. Perez, P. A. Fernandes, P. J. Silva, M. J. Ramos, *J. Chem. Theory Comput.* **2011**, *7*, 3898–3908.

[299] G. F. Mangiatordi, E. Brémond, C. Adamo, *J. Chem. Theory Comput.* **2012**, *8*, 3082–3088.

[300] G.-S. Li, B. Maigret, D. Rinaldi, M. F. Ruiz-Lopez, *J. Comput. Chem.* **1998**, *19*, 1675–1688.

[301] M. J. Harrison, N. A. Burton, I. H. Hillier, *J. Am. Chem. Soc.* **1997**, *119*, 12285–12291.

[302] R. Sharma, M. Thorley, J. P. McNamara, C. I. F. Watt, N. A. Burton, *Phys. Chem. Chem. Phys.* **2008**, *10*, 2475.

[303] M. R. Gunner, J. Mao, Y. Song, J. Kim, *Biochchim. Biophys. Acta* **2006**, *1757*, 942–968.

[304] H. Hu, W. Yang, *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

[305] B. Engels, S. D. Peyerimhoff, *J. Phys. Chem.* **1989**, *93*, 4462–4470.

[306] H. U. Suter, V. Pleß, M. Ernzerhof, B. Engels, *Chem. Phys. Letters* **1994**, *230*, 398–404.

[307] H. Helten, T. Schirmeister, B. Engels, *J. Phys. Chem. A* **2004**, *108*, 7691–7701.

[308] H.-M. Zhao, J. Pfister, V. Settels, M. Renz, M. Kaupp, V. C. Dehm, F. Würthner, R. F. Fink, B. Engels, *J. Am. Chem. Soc.* **2009**, *131*, 15660–15668.

[309] P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, et al., *J. Mol. Struct.* **2003**, *632*, 1–28.

[310] A. Warshel, *Biochemistry* **1981**, *20*, 3167–3177.

[311] F. Weigend, A. Köhn, C. Hättig, *J. Chem. Phys.* **2002**, *116*, 3175–3183.

[312] L. Goerigk, S. Grimme, *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670.

[313] X. Xu, I. M. Alecu, D. G. Truhlar, *J. Chem. Theory Comput.* **2011**, *7*, 1667–1676.

[314] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.

[315] V. Barone, L. Orlandini, C. Adamo, *Chemical Physics Letters* **1994**, *231*, 295–300.

[316] F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, H.-J. Werner, *Angew. Chem.* **2006**, *118*, 7010–7013.

[317] J. Řezáč, J. Fanfrlík, D. Salahub, P. Hobza, *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

[318] D. J. Creighton, M. S. Gessouroun, J. M. Heapes, *FEBS Letters* **1980**, *110*, 319–322.

[319] F. A. Johnson, S. D. Lewis, J. A. Shafer, *Biochemistry* **1981**, *20*, 44–48.

[320] S. D. Lewis, F. A. Johnson, J. A. Shafer, *Biochemistry* **1981**, *20*, 48–51.

[321] A. C. Tissot, S. Vuilleumier, A. R. Fersht, *Biochemistry* **1996**, *35*, 6786–6794.

[322] S. Chen, T. Hu, J. Zhang, J. Chen, K. Chen, J. Ding, H. Jiang, X. Shen, *J. Biol. Chem* **2008**, *283*, 554–564.

[323] H. Tachikawa, M. Igarashi, J. Nishihira, T. Ishibashi, *J. Photochem. Photobiol., B* **2005**, *79*, 11–23.

[324] S. C. L. Kamerlin, A. Warshel, *Proteins* **2010**, *78*, 1339–1375.

[325] H. M. Senn, J. Kästner, J. Breidung, W. Thiel, *Can. J. Chem.* **2009**, *87*, 1322–1337.

[326] J. C. Ma, D. A. Dougherty, *Chem. Rev* **1997**, *97*, 1303–1324.

[327] R. Loewenthal, J. Sancho, A. R. Fersht, *J. Mol. Biol.* **1992**, *224*, 759–770.

[328] E. Cauët, M. Rooman, R. Wintjens, J. Liévin, C. Biot, *J. Chem. Theory Comput.* **2005**, *1*, 472–483.

[329] D. Zhong, S. K. Pal, A. H. Zewail, *Chemical Physics Letters* **2011**, *503*, 1–11.

[330] S. K. Pal, J. Peon, A. H. Zewail, *PNAS* **2002**, *99*, 1763–1768.

[331] N. Nandi, B. Bagchi, *J. Phys. Chem. B* **1997**, *101*, 10954–10961.

[332] W. L. Jorgensen, *Acc. Chem. Res.* **1989**, *22*, 184–189.

[333] B. Roux, *Comput. Phys. Commun.* **1995**, *91*, 275–282.

[334] Y. Zhang, H. Liu, W. Yang, *J. Chem. Phys.* **2000**, *112*, 3483–3492.

[335] D. Marx, *ChemPhysChem* **2007**, *8*, 209–210.

[336] D. Marx, *ChemPhysChem* **2006**, *7*, 1848–1870.

[337] K. Bhattacharyya, *Acc. Chem. Res.* **2003**, *36*, 95–101.

[338] S. Liu, R. P. Hanzlik, *J. Med. Chem.* **1992**, *35*, 1067–1075.

[339] T. W. Schultz, J. W. Yarbrough, R. S. Hunter, A. O. Aptula, *Chem. Res. Toxicol.* **2007**, *20*, 1359–1363.

[340] A. Böhme, D. Thaens, A. Paschke, G. Schüürmann, *Chem. Res. Toxicol.* **2009**, *22*, 742–750.

[341] J. A. H. Schwöbel, D. Wondrousch, Y. K. Koleva, J. C. Madden, M. T. D. Cronin, G. Schüürmann, *Chem. Res. Toxicol.* **2010**, *23*, 1576–1585.

[342] H. Helten, T. Schirmeister, B. Engels, *J. Org. Chem.* **2005**, *70*, 233–237.

[343] R. Vicik, H. Helten, T. Schirmeister, B. Engels, *ChemMedChem* **2006**, *1*, 1021–1028.

[344] V. Buback, M. Mladenovic, B. Engels, T. Schirmeister, *J. Phys. Chem. B* **2009**, *113*, 5282–5289.

[345] J. C. Dearden, *J. Comput. Aided Mol. Des* **2003**, *17*, 119–127.

[346] D. Mulliner, D. Wondrousch, G. Schüürmann, *Org. Biomol. Chem.* **2011**, *9*, 8400.

[347] D. Wondrousch, A. Böhme, D. Thaens, N. Ost, G. Schüürmann, *J. Phys. Chem. Lett.* **2010**, *1*, 1605–1610.

[348] T. Hartung, C. Rovida, *Nature* **2009**, *460*, 1080–1081.

[349] U. Kaeppler, T. Schirmeister, *Med. Chem.* **2005**, *1*, 361–370.

[350] R. Vicik, M. Busemann, K. Baumann, T. Schirmeister, *Curr. Top. Med. Chem.* **2006**, *6*, 331–353.

[351] B. E. Thomas, P. A. Kollman, *J. Org. Chem.* **1995**, *60*, 8375–8381.

[352] Z. Kunakbaeva, R. Carrasco, I. Rozas, *Theochem. J. Mol. Struct.* **2003**, *626*, 209–216.

[353] M. Mladenovic, K. Ansorg, R. F. Fink, W. Thiel, T. Schirmeister, B. Engels, *J. Phys. Chem. B* **2008**, *112*, 11798–11808.

[354] L. Pardo, R. Osman, H. Weinstein, J. R. Rabinowitz, *JACS* **1993**, *115*, 8263–8269.

[355] N. Tezer, R. Ozkan, *Theochem. J. Mol. Struct.* **2001**, *546*, 79–88.

[356] Y. Chiang, A. J. Kresge, N. P. Schepp, *J. Am. Chem. Soc.* **1989**, *111*, 3977–3980.

[357] D. Lee, C. K. Kim, B.-S. Lee, I. Lee, B. C. Lee, *J. Comput. Chem.* **1997**, *18*, 56–69.

[358] E. D. Raczyńska, W. Kosińska, B. Ośmiałowski, R. Gawinecki, *Chem. Rev.* **2005**, *105*, 3561–3612.

[359] C. S. Cucinotta, A. Ruini, A. Catellani, A. Stirling, *ChemPhysChem* **2006**, *7*, 1229–1234.

[360] M. Zakharov, A. E. Masunov, A. Dreuw, *J. Phys. Chem. A* **2008**, *112*, 10405–10412.

[361] C.-C. Su, C.-K. Lin, C.-C. Wu, M.-H. Lien, *J. Phys. Chem. A* **1999**, *103*, 3289–3293.

[362] C.-C. Wu, M.-H. Lien, *J. Phys. Chem.* **1996**, *100*, 594–600.

[363] S. Schlund, E. M. Basílio Janke, K. Weisz, B. Engels, *J. Comput. Chem.* **2010**, *31*, 665–670.

[364] E. M. B. Janke, S. Schlund, A. Paasche, B. Engels, R. Dede, I. Hussain, P. Langer, M. Rettig, K. Weisz, *J. Org. Chem.* **2009**, *74*, 4878–4881.

[365] A. Paasche, *Diplomathesis,* Würzburg, **2007**.

[366] O. Miyata, T. Shinada, I. Ninomiya, T. Naito, T. Date, K. Okamura, S. Inagaki, *J. Org. Chem.* **1991**, *56*, 6556–6564.

[367] M. Schiller, *Dissertation*, Würzburg, **2009**.

[368] T. W. Schultz, T. I. Netzeva, D. W. Roberts, M. T. D. Cronin, *Chem. Res. Toxicol.* **2005**, *18*, 330–341.

[369] J. W. Yarbrough, T. W. Schultz, *Chem. Res. Toxicol.* **2007**, *20*, 558–562.

[370] W. M. Macintyre, *J. Chem. Educ.* **1964**, *41*, 526.

[371] M. Stempka, *Dissertation*, Würzburg, **2011**.

[372] U. Dietzel, C. Kisker, *unpublished results* **2012**.

[373] J. Kongsted, U. Ryde, J. Wydra, J. H. Jensen, *Biochemistry* **2007**, *46*, 13581–13592.

[374] U. Ryde, *Dalton Trans.* **2007**, 607.

[375] K. Nilsson, H.-P. Hersleth, T. H. Rod, K. K. Andersson, U. Ryde, *Biophys. J.* **2004**, *87*, 3437–3447.

[376] U. Ryde, K. Nilsson, *JACS* **2003**, *125*, 14232–14233.

[377] B. T. Sutch, R. M. Romero, N. Neamati, I. S. Haworth, *J. Chem. Educ.* **2012**, *89*, 45–51.

[378] J.-H. Hsieh, S. Yin, X. S. Wang, S. Liu, N. V. Dokholyan, A. Tropsha, *J. Chem. Inf. Model.* **2012**, *52*, 16–28.

[379] C. Grebner, J. Becker, S. Stepanenko, B. Engels, *J. Comput. Chem.* **2011**, *32*, 2245–2253.

[380] C. Grebner, *Dissertation*, Würzburg, **2013**.

[381] A. Grossfield, D. M. Zuckerman, *Annu. Rep. Comput. Chem.* **2009**, *5*, 23–48.

[382] P. Chenprakhon, B. Panijpan, P. Chaiyen, *J. Chem. Educ.* **2012**, *89*, 791–795.

[383] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten, *J. Comp. Phys.* **1999**, *151*, 283–312.

[384] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[385] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926–935.

[386] H. Li, A. D. Robertson, J. H. Jensen, *Proteins* **2005**, *61*, 704–721.

[387] D. C. Bas, D. M. Rogers, J. H. Jensen, *Proteins* **2008**, *73*, 765–783.

[388] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, J. H. Jensen, *J. Am. Chem. Soc.* **2011**, *7*, 525–537.

[389] J.-P. Ryckaert, G. Ciccotti, H. J. Berendsen, *J. Comp. Phys.* **1977**, *23*, 327–341.

[390] A. Schäfer, H. Horn, R. Ahlrichs, *J. Chem. Phys.* **1992**, *97*, 2571–2577.

[391] K. Eichkorn, F. Weigend, O. Treutler, R. Ahlrichs, *Theor. Chem. Account.* **1997**, *97*, 119–124.

[392] R. A. Kendall, H. A. Früchtl, *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **1997**, *97*, 158–163.

[393] R. Bonaccorsi, P. Palla, J. Tomasi, *J. Am. Chem. Soc.* **1984**, *106*, 1945–1950.

[394] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, et al., *MOLPRO, Version 2010.1, a Package of Ab Initio Programs*, **2010**.

[395] H.-J. Werner, F. R. Manby, P. J. Knowles, *J. Chem. Phys.* **2003**, *118*, 8149–8160.

[396] F. Weigend, M. Häser, *Theor. Chem. Account.* **1997**, *97*, 331–340.

[397] M. Häser, R. Ahlrichs, *J. Comput. Chem.* **2004**, *10*, 104–111.

[398] R. Ahlrichs, M. Bär, M. Häser, H. Horn, C. Kölmel, *Chem. Phys. Letters* **1989**, *162*, 165–169.

[399] A. Schäfer, C. Huber, R. Ahlrichs, *J. Chem. Phys.* **1994**, *100*, 5829.

[400] F. Weigend, F. Furche, R. Ahlrichs, *J. Chem. Phys.* **2003**, *119*, 12753–12762.

[401] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, et al., *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[402] J. J. P. Stewart, *MOPAC2009*, Stewart Computational Chemistry, **2009**.

[403] W. Thiel, *MNDO2005*, Max-Planck-Institut Für Kohlenforschung, Mülheim, **2005**.

[404] H. C. Andersen, *Journal of Computational Physics* **1983**, *52*, 24–34.

[405] M. Rostkowski, M. Olsson, C. Sondergaard, J. Jensen, *BMC Struct. Biol.* **2011**, *11*, 6.

[406] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graph.* **1996**, *14*, 33–38, 27–28.

[407] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

[408] F. Weigend, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

[409] S. R. Billeter, A. J. Turner, W. Thiel, *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177–2186.

[410] P. A. Bash, M. J. Field, R. C. Davenport, G. A. Petsko, D. Ringe, M. Karplus, *Biochemistry* **1991**, *30*, 5826–5832.

[411] A. J. Mulholland, W. G. Richards, *Proteins* **1997**, *27*, 9–25.

[412] K. F. Wong, J. B. Watney, S. Hammes-Schiffer, *J. Phys. Chem. B* **2004**, *108*, 12231–12241.

[413] C. Hensen, J. C. Hermann, K. Nam, S. Ma, J. Gao, H.-D. Höltje, *J. Med. Chem.* **2004**, *47*, 6673–6680.

[414] J. C. Hermann, C. Hensen, L. Ridder, A. J. Mulholland, H.-D. Höltje, *J. Am. Chem. Soc.* **2005**, *127*, 4454–4465.

[415] R. C. Walker, M. F. Crowley, D. A. Case, *J. Comput. Chem.* **2008**, *29*, 1019–1031.

[416] D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, et al., *AMBER 11*, University Of California, San Francisco, **2010**.

[417] M. P. Repasky, J. Chandrasekhar, W. L. Jorgensen, *J. Comput. Chem.* **2002**, *23*, 1601–1622.

[418] G. M. Torrie, J. P. Valleau, *Journal of Computational Physics* **1977**, *23*, 187–199.

[419] M. Souaille, B. Roux, *Comput. Phys. Commun.* **2001**, *135*, 40–57.

[420] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, J. M. Rosenberg, *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[421] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, *J. Comput. Chem.* **1995**, *16*, 1339–1350.

[422] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al., *J. Comput. Chem.* **2009**, 671–690.