Comparative Computational Study of Serine Peptidase and Cysteine Peptidase Mechanisms

DISSERTATION

zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften (Dr. rer. nat.)

an der Fakultät für Biologie, Chemie und Geowissenschaften

der Universität Bayreuth

vorgelegt von

Florian Johannes Gisdon

geb. in Augsburg

Bayreuth, 2022

Die vorliegende Arbeit wurde in der Zeit von Januar 2014 bis November 2021 in Bayreuth am Lehrstuhl für Bioinformatik und Strukturbiologie unter Betreuung von Herrn Professor Dr. G. Matthias Ullmann angefertigt.

Vollständiger Abdruck der von der Fakultät für Biologie, Chemie und Geowissenschaften der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.).

Dissertation eingereicht am: 16.11.2021

Zulassung durch die Promotionskommission: 08.12.2021

Wissenschaftliches Kolloquium: 14.06.2022

Amtierender Dekan: Prof. Dr. Benedikt Westermann

Prüfungsausschuss:

Prof. Dr. G. Matthias Ullmann	(Gutachter)
Prof. Dr. Birte Höcker	(Gutachterin)
Prof. Dr. Carlo Unverzagt	(Vorsitz)
Prof. Dr. Stephan Clemens	

"Nil tam difficilest quin quaerendo investigari possiet. – Nichts ist so schwierig, dass es nicht erforscht werden könnte."

Publius Terentius Afer, *Heautontimoroumenos* 4,2

Danksagung

Im Zusammenhang mit meiner Doktorarbeit gibt es Personen, denen ich an dieser Stelle meinen Dank aussprechen möchte.

Meinem Betreuer Prof. Dr. G. Matthias Ullmann danke ich für seine Unterstützung während meiner Promotion. Dies bezieht sich auf allerlei Schwierigkeiten für die Lösungen gefunden wurden. Beeinflusst hat mich besonders sein kritisches Hinterfragen von existierenden Annahmen in den vielen Diskussionen.

Dr. Elisa Bombarda danke ich für die Zusammenarbeit und die Diskussionen im Zusammenhang mit den Arbeiten an der Phytochelatin Synthase.

In diesem Zusammenhang möchte ich auch Prof. Dr. Wulf Blankenfeldt und Dr. Christian Feiler danken für die Betreuung bei der experimentellen Arbeit an der Phytochelatin Synthase vor meiner Promotion. Ebenfalls Prof. Dr. Birte Höcker, die mir ermöglicht hat weitere experimentelle Arbeiten an der Phytochelatin Synthase während der Promotion durchzuführen und ihrem Team, für die schöne Stimmung und die enorme Hilfe.

Meinem damaligen Bürokollegen Martin Culka danke ich für die vielen Diskussionen, die gute Zusammenarbeit und die schöne Atmosphäre. Ebenfalls möchte ich meinen Kollegen aus der B14 danken, für die gute Stimmung, den Spaß und die vielen wissenschaftlichen und oft auch nicht so wissenschaftlichen Diskussionen, besonders mit Jan, Johannes und Lars, aber auch mit Steffen, Simon und Rajeev.

Und besonders möchte ich meiner Familie und meinen Freunden und Verwandten danken, die mich alle unterstützten und mir oft Verständnis entgegenbrachten.

Inhaltsverzeichnis

Zusammenfassung

In der Biochemie wird das Verhalten und die Funktion eines untersuchten biologischen Systems charakterisiert, wobei genaue Kenntnis über dessen Eigenschaften entscheidend ist. Viele Details derartiger Eigenschaften können aus experimentellen Ergebnissen oft nicht ermittelt werden. Das Teilgebiet der computergestützten Biochemie kann dabei Laborexperimente mit mikroskopischen Details ergänzen und mehr Informationen über den makroskopisch messbaren Bereich hinaus beitragen. Die computergestützte Forschung wird auch genutzt, um experimentelle Messungen daraus abzuleiten. Beide Gebiete stehen dabei nicht in Konkurrenz zueinander, sondern ergänzen sich.

Zur Untersuchung wissenschaftlicher Fragestellungen verwendet man in der computergestützten Biochemie dafür geeignete Modelle zur Beschreibung biologischer Systeme. Passende Modelle sind je nach Fragestellung entweder explizit und beschreiben das Verhalten einzelner Atome mit ihren Bestandteilen, oder implizit und simulieren physikalische Eigenschaften ganzer Bereiche des untersuchten Systems. Modelle unterschiedlicher Abstraktionsebenen können auch kombiniert werden, um große Systeme mit hoher Genauigkeit in relevanten Bereichen zu untersuchen. Im Rahmen dieser Arbeit verwendete ich kontinuumelektrostatische und kombinierte quantenmechanische/molekularmechanische Modelle. Die Werkzeuge in der computergestützten Biochemie sind die Algorithmen, die man auf die erstellten Modelle anwendet.

Für die Analyse der Protonierungseigenschaften verwendete ich kontinuumelektrostatische Modelle mit Monte Carlo basierten Abtastalgorithmen, die innerhalb der Arbeitsgruppe entwickelt wurden. Für die Untersuchung der genauen katalytischen Mechanismen und der Enzymkinetik wurden die Reaktionsmechanismen analysiert. Ein wesentlicher Bestandteil der Reaktionspfadanalyse von Peptidasen in der vorliegenden Arbeit war der *conjugate peak refinement* (CPR; deutsch: konjugierte Höchstpunktverfeinerung) Algorithmus. Dieser wurde als PyCPR in Zusammenarbeit mit einem Kollegen

für das Python-basierte Framework pDynamo implementiert. PyCPR ist eine *chain-of-states* (COS; deutsch: Kette-aus-Zuständen) Methode, die einen Reaktionspfad mit diskretisierten aber verbundenen Strukturen darstellt. Der Algorithmus basiert auf Eigenschaften eines Sattelpunkts, mit Fokus darauf sich iterativ einer Sattelpunktregion anzunähern, um den Übergangszustand von Reaktionsschritten zu finden. Die Funktionalität und die Zuverlässigkeit bei der Suche von Übergangszuständen von PyCPR wurde in unseren aufgeführten Beispielen über eine konformationelle Änderung von Butan und über den Reaktionsmechanismus des Glycylradikal Enzyms 4-Hydroxyphenylacetat Decarboxylase gezeigt.

In meiner Arbeit verwendete ich PyCPR, um den ähnlichen, aber dennoch unterschiedlichen Mechanismus von Cystein- und Serinpeptidasen zu analysieren. Peptidasen sind Enzyme mit vielfältiger Funktionalität und sind wichtig in Bereichen wie Peptidabbau, Pathogenabwehr oder Regulation zellulärer Pfade. Es ist bekannt, dass Katalyse durch Cysteinpeptidasen schrittweise über ein Ionenpaarintermediat abläuft, während die Schritte in der Katalyse durch Serinpeptidasen gekoppelt ablaufen. Allerdings sind die Ursachen für den gekoppelten Mechanismus in Serinpeptidasen kaum verstanden. Die Analyse des elektrostatischen Potentials bestätigte die geläufige Meinung, dass um das Cystein ein positives Potential vorliegt, welches das entstehende Thiolat in Cysteinpeptidasen stabilisiert. Im Gegensatz dazu habe ich ein negatives Potential in Serinpeptidasen gefunden. Die Untersuchung der Protonierungseigenschaften zeigte, dass solch ein negatives Potential in Serinpeptidasen wesentich ist, um die Basizität des katalytischen Histidins zu verstärken, um Protonenaufnahme vom Serin zu ermöglichen. Als Folge kann aber kein Ionenpaarintermediat stabilisiert werden. Doch das negative Potential unterstützt auch einen gekoppelten Mechanismus, indem es das nukleophile Potential des katalytischen Serins verstärkt. Zusätzlich muss das aktive Zentrum in Serinpeptidasen kompakter gebaut sein, damit der Protonentransfer zum Histidin und der nukleophile Angriff gekoppelt ablaufen können. Diese Ergebnisse stimmen mit der experimentell nachgewiesenen Inaktivität von Cysteinpeptidasen überein, deren katalytisches Cystein in ein Serin mutiert wurde. Ich untersuchte zunächst den Mechanismus einer Cysteinpeptidase, die von uns ebenfalls im Labor charakterisiert wurde. Die Mutation des katalytischen Cysteins zu einem Serin führte erwartungsgemäß zur Inaktivierung, wodurch das Substrat nicht prozessiert in der

Bindetasche vorgefunden wurde. Diese Struktur wurde als Basis für weitere computergestützte Untersuchungen verwendet. Die Analyse des Reaktionspfads dieser mutierten Peptidase ergab eine viel zu hohe Energie für die Serin-basierte Katalyse innerhalb des aktiven Zentrums einer Cysteinpeptidase. Die hier gezeigte computergestützte Analyse ergab, dass eine zusätzliche Übertragung der gefundenen relevanten Eigenschaften auf die Umgebung des aktiven Zentrums für die Aktivität der Serinpeptidase erforderlich ist.

Im Rahmen dieser Arbeit wurde ein Suchalgorithmus für Reaktionspfade implementiert und ein Verfahren zu deren Charakterisierung etabliert. Unter Anwendung dieses Verfahrens zusammen mit anderen Methoden habe ich eine Cysteinpeptidase untersucht und mit deren inaktiver Serinvariante und einer natürlichen Serinpeptidase verglichen. Dabei wurden weitere Details zur Katalyse von Serinpeptidasen erforscht und zu deren Erfordernis an einen gekoppelten Mechanismus, der durch die Umgebung der katalytischen Triade ermöglicht wird.

Abstract

In biochemistry, detailed knowledge about the properties of an investigated biological system is fundamental to understand its behavior and its function. Experimental results are often limited to describe such properties in every detail. The branch of computational biochemistry can contribute more information to macroscopic measurements, and reveal microscopic details, which go beyond the scope of laboratory experiments. Computational research can also guide experimental measurements. Thereby both fields are not in competition but complement each other.

The research of computational biochemists is based on the description of biological systems with an appropriate computational model regarding the scientific issue. Suitable models can be explicit with detailed descriptions of atoms and their components, or implicit to simulate physical properties of whole parts of the investigated system. Models of different abstraction levels can also be combined, which allows to investigate large systems with high accuracy of the relevant parts. In this thesis I used continuum electrostatic, and hybrid quantum mechanical/molecular mechanical models. The tools of computational biochemists are the algorithms, which are applied on the constructed models.

For analysis of protonation characteristics, I applied continuum electrostatic models with Monte Carlo based sampling algorithms, which have been developed within the group. To investigate mechanistic details and kinetics of enzyme catalysis, reaction paths were analyzed. An essential element in my reaction path investigation of peptidase mechanisms was the conjugate peak refinement (CPR) algorithm. The algorithm was modified, and implemented as PyCPR by me together with a colleague for the Python-based framework pDynamo. PyCPR is a chain-of-states method, which represents a reaction path as discretized but linked structures. The algorithm is based on the characteristics of a saddle point, with focus to approach a saddle point region within an iterative procedure to gradually find the transition state of

reaction steps. The reliable performance of PyCPR was confirmed by our provided examples on the conformational change of butane, and on the mechanism of the glycyl radical enzyme 4-hydroxyphenylacetate decarboxylase.

Within my thesis, PyCPR was used for detailed analysis of the similar but different cysteine and serine peptidase mechanisms. Peptidases are enzymes with diverse functionality and perform important tasks in peptide degradation, pathogenic defense, or regulation in cellular pathways. It is known, that cysteine peptidase catalysis proceeds stepwise with an ion-pair intermediate, and serine peptidase catalysis proceeds concerted. However, the reason for a concerted mechanism in serine peptidases is poorly understood. The analysis of the electrostatic potential confirmed the common opinion of a positive potential around the cysteine, which stabilizes the emerging thiolate in cyteine peptidases. Contrary to this, I found a negative potential within serine peptidases. Analysis of protonation characteristics showed, that such a negative potential in serine peptidases is essential to enhance the basicity of the catalytic histidine, which facilitates proton acceptance from serine. As a consequence no ion-pair state intermediate is stabilized. But the negative potential further supports the concerted mechanism by increasing the nucleophilic potential of the catalytic serine. In addition to that, the active site geometry of serine peptidases has to be compact to allow for the simultaneous events of proton transfer to histidine and nucleophilic attack. These findings are in line with the experimentally measured inactivity of cysteine peptidases with their catalytic cysteine mutated into serine. At first I investigated the mechanism of a cysteine peptidase, which we also characterized in the laboratory. As expected, mutation of the catalytic cysteine to a serine led to inactivation, by which the substrate was bound non-processed within the binding pocket. This structure was used for further computational research. Reaction path analysis of this mutant serine peptidase revealed a much too high energy required for serine-based catalysis within the catalytic site of a cysteine peptidase. Computational analysis presented here showed, that additional transfer of features of the environment of the active site is required for serine peptidase activity.

In the context of this thesis, a reaction path search algorithm was implemented, and a procedure was established to investigate reaction path characteristics. With the application of the reaction path search procedure, together with other methods, I investigated a cysteine peptidase, and compared it to

its inactive serine variant and a natural serine peptidase. By that, I revealed further details about serine peptidase catalysis, and the necessity for serine peptidases to have a concerted mechanism, which is facilitated by the surrounding of the catalytic triad.

List of Abbreviations

CI-NEB	Climbing image nudged elastic band
COS	Chain-of-states
CPR	Conjugate peak refinement
DFT	Density functional theory
EM	Electron microscopy
γEC	γ-glutamyl-cysteine
GSH	Glutathione
MD	Molecular dynamics
MM	Molecular mechanics
NEB	Nudged elastic band
NMR	Nuclear magnetic resonance
NsPCS	Phytochelatin synthase of *Nostoc sp.*
PC	Phytochelatin
PCS	Phytochelatin synthase
PES	Potential energy surface
PDB	Protein data bank
QM	Quantum mechanics

Chapter 1

Computational Investigation of Enzyme Mechanisms

1.1 The Field of Computational Biochemistry

Biochemistry is a discipline for the study of the structure or the function of biological systems on the basis of chemistry. The combination of biology with different fields of chemistry allows for the description and explanation of the properties of living organisms.

To maintain life, it is essential for living organisms to obtain energy, to transform it into chemical energy such as adenosine triphosphate, and to use it again. One energy source are nutrients, which have to be broken down to applicable units, for instance by peptidase enzymes. Further processing and transformation is performed to utilize the obtained energy for other processes. To explain the function of involved proteins, their actions and interactions need to be investigated. However, a direct observation of molecular events such as enzymatic mechanisms is difficult. But conclusions can be drawn from indirect approaches, such as the analysis of heat production in binding studies, or the time-dependent measurement of reactant or product concentrations to obtain reaction rates. Further, molecular structures can be determined, which provide information about certain states of a system. Such studies are essential to construct hypotheses, which are then utilized to explain observations, and build scientific theories. [Chapter 1 in ref. 1]

1.1.1 Scientific Models

Scientific theories are a basis to explain investigated systems, processes or phenomena. The descriptions of scientific findings thereby rely on the application of models, which can be interpreted as applicable translations of scientific theories. A biochemical mechanism for instance, by which an enzyme catalyzes a chemical reaction, is represented as a model, which reflects the real events on a certain level of abstraction, and allows for a more universal description. As basic elements in science, models can simplify events or characterizations and make them transferrable to different systems or fields.

In fact, a description of an enzymatic reaction is an abstracted illustration of that biochemical event with focus on relevant steps. These steps can be derived from experimental measurements or computational investigations, and in return also help to interpret obtained results. Consequently, the investigated system itself is represented as a model of interacting atoms, which can be modeled computationally. For further understanding of investigated systems or for predicitons, such computational models can then be used to simulate system properties and behaviors. In my thesis, I focus on the computational investigation of specific biological systems, namely peptidases.

1.1.2 Computer Simulations and Laboratory Experiments

Laboratory experiments are performed in biochemistry to investigate real biological systems by experimental methods. Computer simulations, however, are based on representations of biological systems as computational models. Algorithms are applied on such computational models, analog to the experimental methods in laboratory experiments. These algorithms are the computational methods, which are specific mathematical procedures to obtain results. Common to both, the laboratory field and the computational field, is the requirement of scientific models to interpret obtained results.

Synergy of the Laboratory and the Computational Field. Computational research can be seen as idealized, since it is based on a constructed representation of reality. It is even dabated, if computer simulations are real experiments. Within this debate, the field of computational research is abstracted by Peter Galison to an artificial world, in which experiments take place [2]. This artificial world allows to perform tasks within an idealized

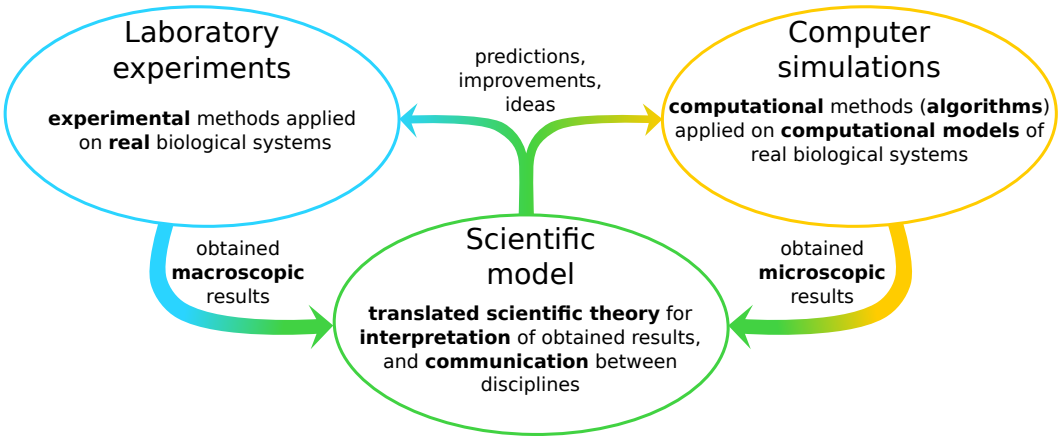


FIGURE 1.1: Illustrated correlation between the laboratory field, which acts on real systems and the computational field, which simulates model representations of real systems. Both fields rely on scientific models for interpretation of obtained results, and supplement each other with predictions, improvements and ideas.

reality. Within this idealized reality the behavior of a biological system is determined by the principles of the applied conceptual model. However, computational analysis can reveal more details than experiments, and is used to understand and explain experimental results or to predict them. But not only computational investigations are limited to the scope of the applied scientific theory. Also laboratory experiments rely on conceptual models for the interpretation of obtained results. When the interpretation of results is no longer possible with existing models, then new scientific theories have to be developed, which can be tested by computational research. Further, predictions on the basis of newly developed scientific theories can be made, which are useful for the design and the interpretation of experiments. So the computational and the laboratory field complement each other based on the overlap of a scientific model. That correlation is illustrated in Figure 1.1.

Macroscopic Results and Microscopic Results. Computational investigation is often essential to explain processes and phenomena, since simulations can go beyond experimental resolution. Laboratory experiments reveal macroscopic characterizations of biochemical events. Obtained results can arise from mixtures of states or sequences of steps, which are not distinguishable. The macroscopic result is a representation of an average of all microscopic details. Computer simulations are used to reveal more details

beyond laboratory experiments. Macroscopic results from laboratory experiments, such as reaction rates for enzymatic catalysis reveal the overall performance of an enzyme. For the microscopic picture, with all single steps of the catalytic event, a probable reaction path has to be simulated. However, the simulation is not straightforward, and further information is required. Crystal structures of initial states prior to catalysis may be available, which are optimal starting points for the reaction path simulation. Further, crystal structures of stable intermediates or transition state analogs can yield information about the probable catalytic pathway. The simulated reaction path and the calculated energy profile can be correlated to the measured reaction rate, and supplements the laboratory characterization. For the biochemical characterization of biological systems or events it is essential to take advantage of the synergy of the laboratory and the computational field. The combination of both fields can improve the interpretations of results and the underlying investigation methods.

1.2 Concepts and Approaches to Model Biological Systems

For biochemical characterization of biological systems or events, a variety of computational methods exist, which can be seen as algorithmic descriptions of computational experiments. Such computational methods are indeed comparable to laboratory experiments. However, computational methods are not applied to real biological systems, but to their modeled *in silico* analogs. The basis of all computational investigations is the computational model. It is an abstraction of the real system, and provides a description of the biological system appropriate for the application. The scientific question determines the level of abstraction. Physical properties of whole system parts can be simulated with an implicit representation by continuum models. The simulation of detailed behaviors and properties of molecules or atoms is performed with explicit models, such as molecular mechanical models or quantum mechanical models. The following subsections outline computational models and computational methods for biochemical investigation. An extensive review is provided in Manuscript A.

1.2.1 Computational Models of Biological Systems

All computational models of a biological system are based on its atomic coordinates, which are required to place the atoms or the corresponding charges in space. The structure of atoms within simple molecules can be generated based on chemical rules. Despite improving modeling techniques, structure generation from scratch is difficult for more complex molecules up to proteins or even larger systems, since the degrees of freedom enormously increase. Recently, AlphaFold [3], an artificial-intelligance-based system was able to predict many protein structures with high accuracy and thus made the use of such predicted structures more reliable. However, computational biochemists often use structures, which are determined by laboratory experiments. Common laboratory methods for structure determination are nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, or electron microscopy (EM).

With NMR spectroscopy structures can be directly investigated in solution, which can reveal dynamic aspects under physiological conditions [4, 5]. However, it is still difficult to determine structures of large proteins or complexes. An alternative approch is X-ray crystallography, where structures of crystallized proteins are determined by their diffraction pattern of an X-ray beam. The major drawback is to obtain suitable crystals for analysis which is even more difficult if proteins do not naturally occur in solution, but for instance are anchored to membranes. An emerging technique for the determination of biomolecular structures is EM, in particular cryo-EM. Instead of protein crystals, whole protein solutions can be frozen and analyzed by an electron beam, analog to light microscopy. Also dynamics can be investigated, since the frozen solution contains several different conformations, while variety in well-orderd crystals is significantly reduced. Determined biomolecular structures are provided for the community in the protein data bank (PDB), which contains more than 150000 entries, mainly crystallographic data [6].

On the basis of biomolecular structures, a computational biochemist models biological systems *in silico* by application of specific model types, such as mechanical models, or continuum models. Thereby it is essential to describe the interactions between the particles, and their properties in a way, that the modeled system appropriately represents the properties of the real biological system. Such a description is obtained by a mathematical equation, the

energy function, which is used to energetically evaluate states of the investigated system. On the basis of an energy function, various algorithms can be designed, for instance, to explore enzymatic reaction paths.

Particle Models. The most explicit computational models in the field of computational biochemistry represent systems on the level of atomic nuclei and electrons. Most models rely on the Born-Oppenheimer approximation, which separates the motion of atomic nuclei and electrons. This is possible because of the mass difference, which effects the different time scales of their motions. To obtain the wavefunction, which is a mathematical description of a system state, *ab initio* approaches are applied, such as Hartree-Fock [7], which solve an approximate Schrödinger equation. However, a major weakness of the Hartree-Fock approach is the mean-field approximation of the Coulomb correlation of electrons. An alternative *ab initio* approach, which is widely applied in computational biochemistry, is density functional theory (DFT) [8]. DFT relies on the total electron density with the wavefunction as its functional. DFT approaches include approximate electron correlation, and a possible use of hybrid functionals as a combination with Hartree-Fock admixes exact electron exchange [9, 10]. In comparison to Hartree-Fock approaches, DFT approaches show a favorable ratio of computational cost and performance [11]. With rising computer power larger systems can be accessed, even whole proteins [12]. However, a computational model of a biological system can consist of hundreds of thousands of atoms, which becomes unfeasible to treat with quantum mechanics (QM) in reasonable time.

In contrast to QM approaches, molecular mechanics (MM) models describe biological systems with classical physics, where atoms are simulated as spherical particles with partial charges. Covalent interactions are approximated with spring potentials, and nonbonded interactions are described with a Lennard-Jones potential or a Coulomb potential. For a realistic simulation of molecules force fields are used, which contain parameters for all interactions, such as bond constants, or charges. The description of complex biological systems with MM is feasable, especially with modern computer power, but a reasonable description is limited to well-parametrized regions. However, it can be difficult to derive proper parameters for cofactors or specific substrates. Further, the simulation of bond breakage or bond formation is restricted, which is necessary for the simulation of enzymatic reactions.

With the combination of QM and MM to hybrid QM/MM models complex biological systems can be simulated with high accuracy in reasonable time. Thereby a small relevant part of the whole system is simulated with QM, and embedded into the remaining MM treated part to allow for consistent behavior of the whole system. This approach was first described by Warshel and Levitt [13] in 1976 and is widely used to investigate enzymes [14, 15].

Continuum Models. For an implicit description of features of a system or its parts continuum models can be applied, which assign average properties to space. For enzymatic behaviors electrostatic effects are prevalent and often sufficient to describe molecular features. A common model is the Poisson-Boltzmann continuum-electrostatics model. Protein regions are represented as a low dielectric with a fixed charge distribution, and aqueous regions as a high dielectric, where ion charges take a Boltzmann distribution [16]. As solution of the Poisson-Boltzmann equation, a position-dependent electrostatic potential is obtained. This potential comprises a Colomb term caused by the distributed charges within a certain permittivity, and a reaction field potential associated with the solvent. Integration of the potential distribution yields the energy for the simulated state of the system. Such models can be applied to visualize the electrostatic potential of molecules [17]. Further, solvation energies or ligand binding energies can be estimated, as well as protonation probabilities for titratable groups [18].

1.2.2 Simulation of Properties of Biological Systems

Analog to laboratory experiments, computational biochemistry investigates biological systems, which are, however, represented as computational models. Application of computational experiments, the computational methods, yield specific system properties. For the investigation of enzymatic mechanisms thermodynamic and kinetic properties are required. Thermodynamic properties characterize different system states, for instance, an educt state, intermediates, or the product states. Kinetic properties describe the behavior of the system by assignment of reaction rates to enzymatic processes, which characterize the likelihood of transitions between stable states. Such computational data can be compared to laboratory findings or can predict them, which supports the design of laboratory experiments. More importantly, the

macroscopic view of laboratory experiments can be supplemented with computed microscopic details to further explain or interpret results.

Thermodynamic Properties. The basis for the computational investigation of enzymatic processes are thermodynamic equilibrium states. These states are stable because they are energetically trapped in a minimum, with no forces acting on them. Therefore, they can be structurally determined by the laboratory methods discussed in subsection 1.2.1. However, not all stable states along a reaction path are available. The initial state of an enzyme with bound substrate is often obtained by inactivated enzyme mutants, which are unable to process the native substrate. Alternatively, substrate analogs can be used, which are not processed by the enzyme, and thus can be structurally determined in the active pocket. Similarly, intermediate states can be determined by appropriate intermediate analogs, which will not be processed by the enzyme. From these structures computational biochemists model respective states. If states are not directly accessible, they can be generated out of available states by chemical intuition, and with use of path exploration methods. Methods such as adiabatic surface mapping or growing string methods allow constraint path searches to find possible stable states. Possible states are then energetically minimized within the respective model. Relative energies between these states provides information about the reaction process, if it is endergonic or exergonic. However, the probability for a transition between stable states is based on the investigation of enzyme kinetics.

Kinetic Properties. In enzyme kinetics, a reaction rate is related to the reaction energy barrier between two states. The rate constant can be seen as a probability factor to get over the reaction energy barrier. It can be calculated from the energy barrier of computationally obtained reaction paths by Eyring-Polanyi equation. The application of computational methods on computational models allows to simulate reaction paths, and to obtain transitions between two states with corresponding energy barrier. An allrounder method to obtain the best simulation of a reaction path does not exist, rather a combination of different methods with different strengths is advisable to succeed. Unfortunately, that combination is often difficult, since available software packages do not contain all existing methods or their variants at once, and are usually not compatible among each other.

The studies within my thesis were performed within the free and versatile pDynamo framework [19]. It contains various approaches to explore and simulate reaction paths. However, a specialized method for reliable transition state search was missing in our opinion. We therefore implemented PyCPR, which is further described as part of this thesis in Masuscript B. PyCPR is an implementation of the conjugate peak refinement (CPR) algorithm, which was developed by Fischer and Karplus [20] based on the work of Sinclair and Fletcher [21]. It was shown to perform effectively to find transition states [22]. Surprisingly, it was just implemented within one software package, which was not appropriate for QM/MM models. The algorithm of CPR exploits the characteristics of a first order saddle point. There, the Hessian matrix has exactly one negative eigenvalue. To locate a saddle point, the algorithm approaches to its vicinity and follows the vector associated with the negative eigenvalue.

In principle, all found saddle points structurally and energetically describe the transition states along the reaction path. By that the rate can be calculated, and compared to results of laboratory experiments. In addition, microscopic structural details from the computational analysis supplement the laboratory results.

1.2.3 Biological Relevance of Performed Computational Investigations

In this thesis, I investigated catalytic mechanisms of proteolytic enzymes with computational approaches. The computational models were constructed on the basis of structures determined by X-ray crystallography, which is a well-established technique to produces high resolution crystal stuctures of enzymes. Structurally flexible parts can sometimes be resolved in several conformations but often it is difficult to locate the positions of the atoms, especially for low quality data. Regions, which are not resolved properly can be modeled in combination with additional information and biochemical understanding. Despite good-quality data and high resolution structures, hydrogen atom positions are normally not well accessible, and have to be modeled. As a consequence, protonation states are not determined. I first prepared MM representations of the biological systems based on the crystal structures. With Poisson-Boltzmann continuum-electrostatics, protonation probabilities of titratable groups were determined, to adjust the MM

models. Electrostatic interactions are important in protein systems and contribute in large part to the catalytic potential of enzymes [16]. Therefore, continuum electrostatic approaches are suitable for the determination of protonation states, and were then applied on the updated models, to estimate ligand binding energies. I further compared the electrostatic potential of proteins to characterize and compare relevant regions of similar proteins. For the explicit simulation of the catalytic mechanisms, I prepared QM/MM models on the basis of the MM models with an explicit water shell to allow for mechanistic interactions. QM/MM is an appropriate state-of-the-art approach to treat biological systems [14, 15]. For the QM models I used hybrid DFT functionals, which have proven to achieve good results [23]. For reaction path exploration and reaction path simulation I used approaches within the pDynamo framework, a versatile and extendable set of computational models and methods, which is free of charge, and constantly developed [19]. Commonly used computational methods and models are described in Manuscript A. For reaction path simulations, mainly PyCPR was applied, which was implemented for the pDynamo framework as part of this theses, and is described in Manuscript B. We critically discuss the results of our PyCPR implementation, and show that they perform well for transition state search [22]. Together, the used approaches are reliable for the representation of biological systems, and for the study of their properties. I applied the approaches to study the catalytic characteristics of cysteine and serine peptidases, which are introduced in Chapter 2.

Chapter 2

Cysteine and Serine Peptidases

2.1 Classification of Peptidases

Peptidases are specific proteolytic enzymes, which facilitate the biochemical cleavage of peptide bonds. Such enzymes are termed proteases, proteinases, or peptidases, whereof the latter will be used throughout this thesis. Initiated by Hartley [24], peptidases can be characterized by their catalytic type, cysteine, serine, threonine, aspartic, glutamic, asparagine, and metallo. This characterization describes the portion, which acts as the nucleophile in the catalytic mechanism. However, peptidases with the same catalytic type have very different molecular structures, and are not all homologues, which is why another classification evolved. The analysis of sequence data and molecular structures enabled the assignment of peptidases to families, which were grouped in clans [25]. The MEROPS database is build upon this classification, and provides a comprehensive set of more than 4000 peptidases [26, 27]. A peptidase family, such as the papain family C1, is usually termed according to the most studied member. The papain family then contains all peptidases with homologous peptidase unit, where homology is shown to papain or another member of that family. Families can be divided into subfamilies, due to a very ancient divergence. Families, which originated from a common ancestor but have already diverged too far to belong to the same family, are combined in a clan. This is the case for papain (family C1) and the phytochelatin synthase (PCS) of *Nostoc sp.* (NsPCS, family C83). NsPCS is structurally similar to papain, and is ascribed to the papain superfamily [28, 29], which paraphrases the cysteine peptidase clan CA. Manuscript C shows

a structural and biophysical investigation of NsPCS. The MEROPS classification further defines species, which are all peptidases within a family, such as PCS of the gamma-glutamylcysteine dipeptidyltranspeptidase family or trypsin of the chymotrypsin family. Each peptidase species can be present in various organisms, and is assumed to have approximately the same biological function.

The MEROPS database currently contains more than one million peptidase sequences, of which about more than 50 % are listed as cysteine or serine peptidases. Within this thesis, I will focus on these two prevalent catalytic types of peptidases. In Manuscript D, I used papain and NsPCS as representative models for cysteine peptidases, and trypsin as a representative model for serine peptidases.

2.2 Relevance of Cysteine and Serine Peptidases

Biological Relevance of Cysteine and Serine Peptidases. Peptide bonds are difficult to cleave. However, proteolysis is required for life [30]. Thus, proteolytic enzymes are supposed to have emerged early with life on earth, and are present in every living organism [31]. The necessity of a biological catalyst for peptide bond cleavage lies in the rather high stability of these chemical bonds. This stability is fundamental to build functional polypeptides, but has to be overcome to break them down. The simple digestion of polypeptides is an important recycle process to reuse amino acids as building blocks, which saves energy for their biosynthesis [32]. It is further important to digest nutrient proteins to acquire essential amino acids. But the potential of peptidases can be seen in various other functions, which have emerged during the evolution of life from a small number of ancestral forms [31]. Diversity of peptidases thereby increases notably from prokaryotes to eukaryotes to multicellular organisms [31, 30]. As described in section 2.1, a large number of all peptidases are of the cysteine or serine catalytic type. These peptidases are involved in various highly regulated biological processes, such as the immune system response [33], blood coagulation [34], or apoptosis [35], where they have to act very specifically. Also pathogens utilize peptidases in a specific way, for instance, to invade the host system [36], or to affect host immunity [37]. These diverse biological processes are carefully regulated, and any alteration can be pathogenic [33, 38]. In turn,

peptidases are of great medical or pharmaceutical importance [39]. But they are also used as highly specific tools within technical applications.

Application-Technical Relevance of Cysteine and Serine Peptidases. The diverse functionality of peptidases, and their high specificity is extensively applied in molecular biology [38]. In this field biological processes are studied on a molecular basis, which requires *in vitro* investigation of involved proteins. The purification of such recombinant proteins is often performed with affinity chromatography, where attached affinity tags have to be cleaved off specifically [40]. In contrast to the analysis of isolated proteins, the field of proteomics investigates the full protein composition within cells under specific conditions. For that purpose highly sequence-specific cleavage is required to analyze obtained data [41]. Furthermore, peptidases are applied in industry, such as food industry [39, 42], where research is, for instance, carried out on rennet substitutes in cheese production [43]. Peptidases are also applied for the reduction of food protein allergy [44, 45]. But also non-food applications are available, such as cleaning of surgical instruments, contact lenses or laundry [42]. All these different applications require a diverse set of natural or engineered peptidases to cover the technical requirements.

2.3 Diverse Functionality but Identical Principle

As outlined before, nature created a wide variety of cysteine and serine peptidases with diverse functionality. However, all this diversity rests on the same simple principle. A nucleophilic amino acid, the nucleophile, is embedded in a supporting environment, which is defined by the protein. In an acylation step, the substrate is cleaved by the formation of an ester bond between the nucleophile and the N-terminal part of the substrate (Figure 2.1, Peptidase Mechanism: Acylation). The C-terminal part of the peptide substrate is cleaved off, and leaves the binding pocket. In a deacylation step, the native enzyme is restored by hydrolytic cleavage of the enzyme-substrate ester (Figure 2.1, Peptidase Mechanism: Deacylation). In this thesis I focus on the acylation reaction, since there the different characteristics of a cysteine or a serine nucleophile require a different active site environment to facilitate substrate cleavage. The succeeding deacylation, however, is performed by water as a nucleophile, independent of the peptidase catalytic type. In general, the

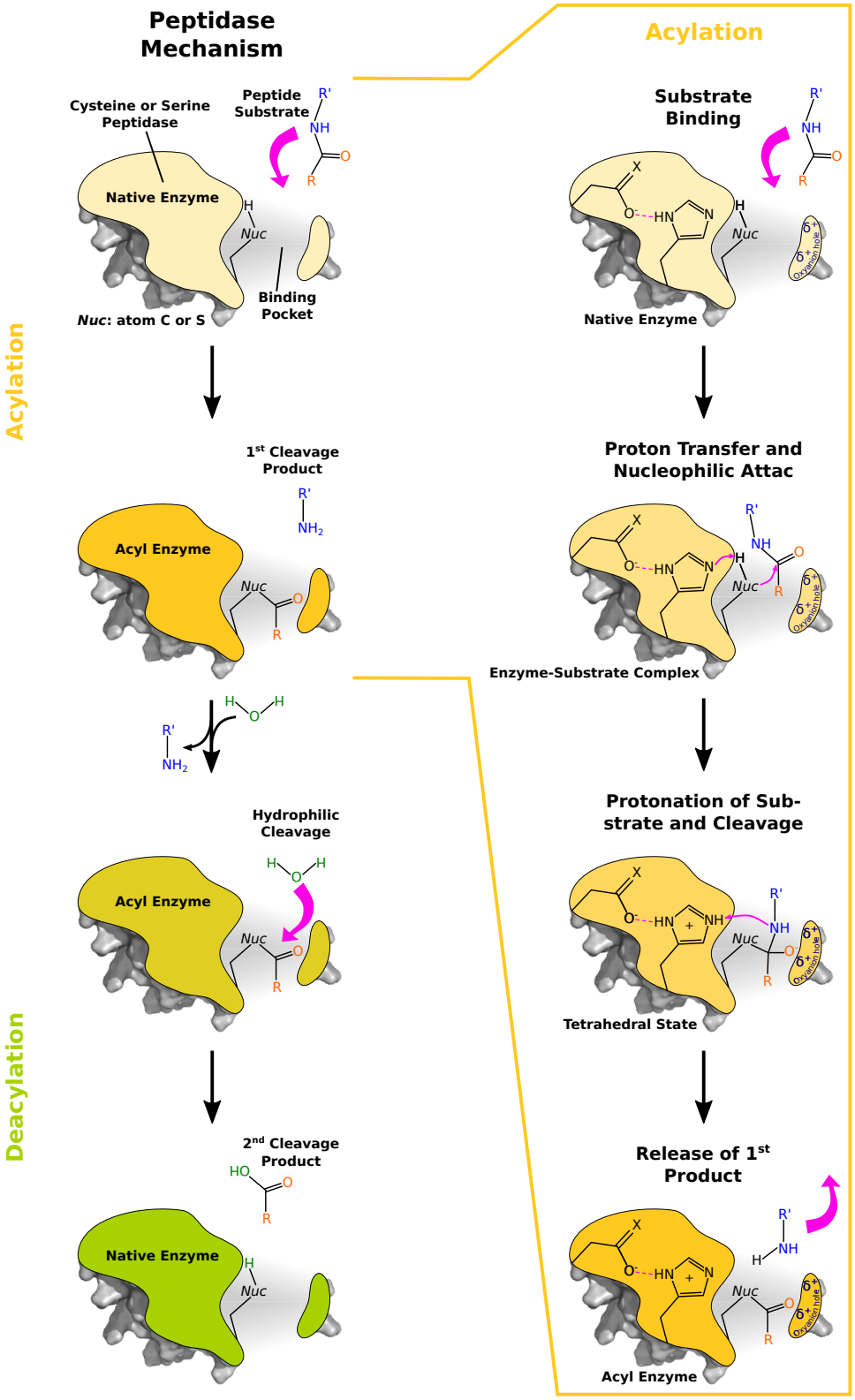


FIGURE 2.1: Illustration of the general peptidase mechanism for cysteine and serine peptidases. The peptidase mechanism can be divided into an acylation reaction and a deacylation reaction. The acylation reaction is displayed in detail. The peptidases employ a catalytic dyad or triad, which consists of the respective nucleophile, a histidine, and optionally an asparagine or aspartate. Another catalytic requirement is the oxyanion hole, which supports the generation of a tetrahedral state.

acylation reaction is performed by a catalytic dyad or triad, which consists of the respective nucleophile, a histidine, and optionally an asparagine or aspartate (Figure 2.1, Acylation). The third residue coordinates the histidine, which accepts the proton of the nucleophile in a first step and donates it to the substrate in a second step. In the first step, the bound substrate is attacked by the nucleophile, which forms a tetrahedral state with the substrate. The proton of the nucleophile is transfered to the histidine. In the tetrahedral state, the carbonyl group of the peptide bond of the substrate formally carries a negative charge. The formation of a tetrahedral state is supported by an essential element of peptidases, the oxyanion hole, which is a region with positive partial charges. Once the abstracted proton at the histidine is donated to the substrate in a second step, cleavage of the substrate occurs. The N-terminal part of the substrate forms an ester bond with the enzyme, and the C-terminal part of the substrate leaves the binding pocket.

This identical principle is utilized in cysteine and serine peptidases to perform various tasks. Diverse utilization of peptidases is basically the cause of different environments, under which peptidases act, in combination with their various substrate specificity [46]. Thereby an identical substrate can also be cleaved at different positions, which makes peptidases highly sequence specific scissors. The specificity can vary from cleavage of a broad range of peptides to highly specific cleavage of just one target peptide at a certain position. Further, the activities of peptidases depend on the conditions within different cellular compartments or extracellular environments. Oxidizing conditions in general inactivate cysteine peptidases, which is utilized to regulate cysteine peptidases within the cellular redox environment [47, 48]. However, oxidation inhibits the peptidase activity, which makes cysteine peptidases less generally applicable. But a cysteine nucleophile can be effectively activated by deprotonation, in contrast to a serine nucleophile, with the respective pK_a values of around 9 [49, 50] and 16 [51]. As a consequence the nucleophilic attack in serine peptidases and the proton transfer to the histidine occurs concerted [52], while for cysteine catalysis it occurs stepwise [53]. These different characteristics of serine and cysteine based catalysis require a different set of amino acids to create a proper surrounding of the catalytic residues and support the catalytic action of the nucleophile. In the following subsections, I will describe the current state of research about the

common properties of cysteine and serine peptidases, and different requirements for their mechanisms.

2.3.1 Specific Mechanisms of Cysteine and Serine Peptidases

Cysteine and serine peptidases utilize the same principle for cleavage of peptide bonds (see Figure 2.1). However, the nucleophilic attack occurs concerted [52] for serine peptidases, and stepwise for cysteine peptidases [53], which utilize an activated ion-pair intermediate state. The difference relies on the deprotonation of cysteine, which results in an activated ion-pair intermediate state. This intermediate state is stabilized by hydrogen bonds, and a positive electrostatic potential around the cysteine [53, 54, 55]. However, requirements on the active site environment for efficient serine peptidase catalysis are not known. The following paragraphs describe the mechanistics of three peptidases, which were used in Manuscript D as model peptidases for computational investigation of cysteine and serine peptidase mechanisms.

Trypsin Trypsin is one of the most studied serine peptidase. According to MEROPS classification [26, 27], trypsin is a member of the chymotrypsin family, which is the largest classified family. Trypsin is an intestinal digestion enzyme, and contains the catalytic triad serine, histidine and aspartate. It is an endopeptidase, and cleaves substrates exclusively C-terminal to aginine, or lysine [56]. Peptide bond cleavage is catalyzed with a concerted mechanism, which means, that the nucleophilic attack of the nucleophile at the substrate occurs concerted with the proton transfer from the nucleophile to the catalytic histidine (Figure 2.2). The resulting tetrahedral state formally carries a negative charge at the carbonyl oxygen, which is stabilized by an oxyanion hole. The occurence of a meta-stable tetrahedral intermediate is still under debate [57], which is addressed in Manuscript D. With the proton transfer from the histidine to the substrate the peptide bond gets cleaved, and the enzyme-substrate ester is formed.

Papain and Phytochelatin Synthase NsPCS Papain is one of the most studied cysteine peptidases. According to MEROPS classification [26, 27], papain is a member of the papain family. Papain has a broad substrate range and is naturally available in high concentration in the papaya latex, where it is used for defense. Papain contains the catalytic triad cysteine, histidine

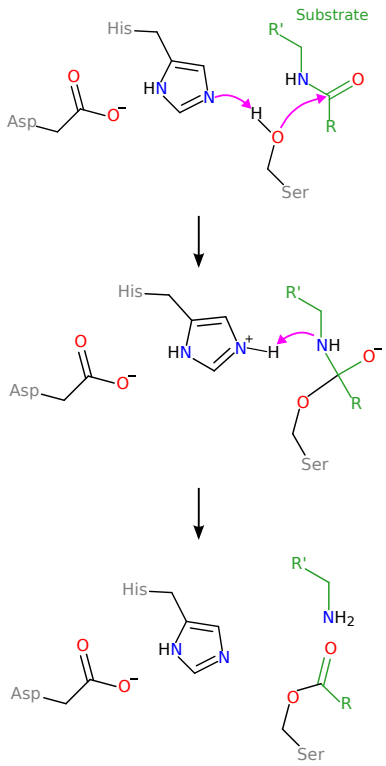


FIGURE 2.2: Illustration of the catalytic mechanism of the serine peptidase trypsin. In the first step, the initial nucleophilic attack and the proton transfer to the histidine occur concerted. In a second step, the proton is transfered to the substrate, which initiates cleavage of the substrate.

and asparagine. The phytochelatin synthase NsPCS is a cyanobacterial enzyme, which has peptidase activity and cleaves off glycine from glutathione. NsPCS is ascribed to the papain superfamily and is supposed to have a comparable enzymatic mechanism [28, 29]. NsPCS contains the catalytic triad cysteine, histidine and aspartate, which is directly comparable to the catalytic triad of trypsin. Peptide bond cleavage is catalyzed with a stepwise mechanism, where in the first step a reactive ion-pair intermediate is formed by hydrogen transfer from the cysteine nucleophile to the catalytic histidine (Figure 2.3). For the hydrogen transfer a low energy barrier was shown in computational studies [58, 59]. In the second step, nucleophilic attack of the nucleophile at the substrate occurs. Analog to serine peptidases, also for cysteine peptidases the occurrence of a stable tetrahedral intermediate is debated [59, 60, 61]. This tetrahedral state carries a formal charge at the carbonyl oxygen of the cleavage bond of the substrate, which is stabilized by an oxyanion hole. Analog to serine peptidases, cleavage of the substrate is performed with the proton transfer from the histidine to the substrate.

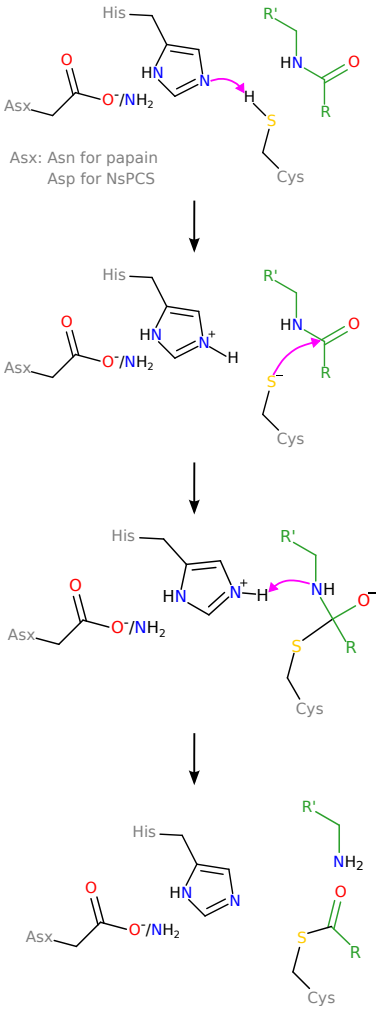


FIGURE 2.3: Illustration of the catalytic mechanism of the cysteine peptidase papain. In the first step, the formation of a reactive ion-pair state is performed by proton transfer from cysteine to histidine. In a second step, nucleophilic attack of the nucleophile at the substrate occurs. In a third step, the proton is transfered to the substrate, which is then cleaved.

2.3.2 Characteristics of Cysteine and Serine-Based Catalysis

In Manuscript D, I address the details of the concerted serine based catalysis and the stepwise cysteine based catalysis of peptidases. Cysteine and serine peptidases apply the same basic mechanism for the cleavage of peptide bonds but the different nature of their nucleophiles require a different environment for the active site, which has to be provided by the surrounding enzyme. This can be seen in nucleophile replacement studies. Serine peptidases, in which the nucleophile is mutated to a cysteine normally show just very little activity [62, 63]. Cysteine peptidases, in which the nucleophile is mutated to a serine are rather inactive [64]. It is assumed, that cysteine peptidases have a positive electrostatic potential around the nucleophile to

stabilize the emerging thiolate ion [54]. Thus, reduced activity of cysteine to serine mutations is attributed to the absence of such a positive potential in wildtype serine peptidases. However, not much is known about requirements for serine peptidase catalysis. In Manuscript D, the requirements for a serine based and a cysteine based catalysis are investigated computationally, and reveal details about the chemical environment for serine peptidases, which are necessary for their proper function.

Chapter 3

Synopsis of the Manuscripts

The focus of this thesis is the comparative computational investigation of cysteine and serine peptidase catalysis. The catalytic mechanism of both peptidases appears similar but proceeds essentially different with distinct requirements for the environment of the active site. My research was based on computational approaches to analyze biophysical properties of selected peptidases and their biochemistry. This thesis comprises a general description of concepts in the field of compuational biochemistry, with a focus on the investigation of enzyme mechanisms, and specifically the application of these concepts on the description and analysis of cysteine and serine peptidase mechanisms (see Figure 3.1). The Manuscript A presents principles of computational models and computational methods in the field of computational biochemistry, which comprise the basic theory of this thesis. The Manuscript B focuses on specific reaction path search methods, of which one was used to characterize the mechanisms of the enzymes, which are analyzed within this thesis in the Manuscripts C and D. The Manuscript C provides a first detailed characterization of one cysteine peptidase and its serine peptidase mutant, a phytochelatin synthase, NsPCS. Both variants were used to investigate characteristics of cysteine and serine peptidases, which is elaborated in the Manuscript D. The Manuscript D discusses investigated characteristics of cysteine and serine peptidases and compares the two well characterized model peptidases, the serine peptidase trypsin and the cysteine peptidase papain. The cysteine peptidase NsPCS with its serine variant, which were characterized in the Manuscript D, replaces papain as a model, since it allows for the direct comparison of both peptidase types. This manuscript

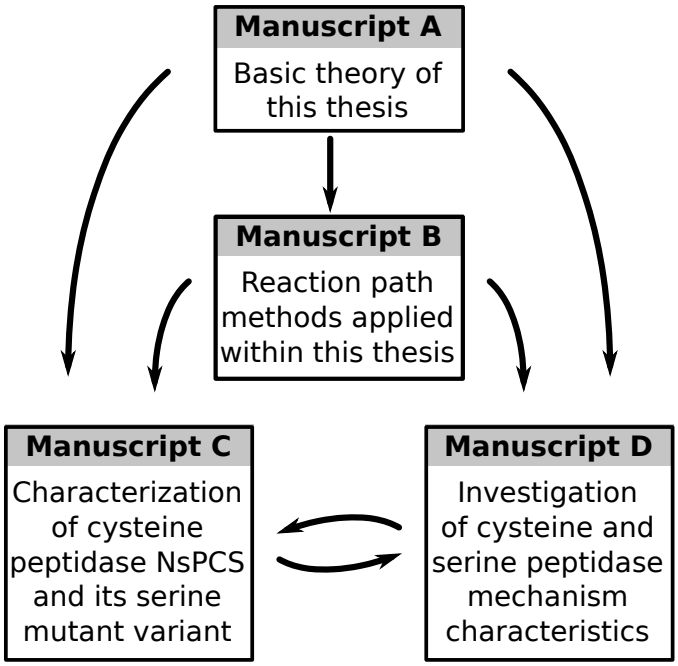


FIGURE 3.1: Schematic representation of the synopsis of the manuscripts, which are part of this thesis. Arrows indicate the context of the manuscripts within this thesis, while the arrow head shows the information flow.

treats the main topic of this thesis, and provides new insights into the requirements for cysteine and serine peptidase catalysis.

Computational Models and Computational Methods. A variety of computational methods exist to simulate enzyme behavior. Common to all is, that the investigated biological system is represented as a computational model, which describes relevant characteristics on a certain level of abstraction. In principle all calculations could be performed on a very precise replica of the investigated biological system. However, this would extremely increase the computational cost, and at the same time the enormous amount of details would distract from substantial findings. Therefore the choice of a computational model with proper abstraction level has to match the leading question of the research. For analysis of a biological system, computational methods have to be applied on computational models. A computational method describes the algorithm, which reveals the state or the behavior of the modeled system. A detailed review of a variety of computational models and possible algorithms for the computational investigation of biological systems is provided in Mauscript A.

Computational Investigation of Enzyme Mechanisms. For computational investigation of catalytic mechanisms, I applied a continuum electrostatic model with a sampling algorithm to obtain information about probable states of the biological system. This information is used to generate a model of the initial state to simulate the catalytic reaction. For enzymes, QM/MM models have emerged as a favorable description of a biological system, which combines accuracy and performance. Initially I applied the popular nudged elastic band (NEB) method and its climbing image variant (CI-NEB). These algorithms belong to the chain-of-states (COS) methods and optimize a chain of coupled structures, which connect two stable states. In addition to constrained energy minimization of all structures, CI-NEB tries to optimize the highest structure towards a saddle point. By this, the chain of structures discretizes a path between two stable states, which passes through a saddle point. However, the found transition state within the obtained path sometimes underestimates the energy barrier. A more reliable method to find transition states is conjugate peak refinement (CPR). It was only implemented within the program CHARMM, which was not well capable of hybrid QM/MM calculations. The QM/MM models were instead prepared with pDynamo, which allows great adjustability, and which is highly compatible and extensible due to its python based nature. Compared to pDynamo, CHARMM and its QM/MM interface did not feature that. Together with a colleague I implemented a modified version of the CPR algorithm with improved features (PyCPR) for the pDynamo framework, which is described in Manuscript B. Compared to a previously published NEB-derived reaction path of 4-hydroxyphenylacetate, we have found a structurally more reliable transition state with PyCPR. With continuum electrostatics, and with QM/MM reaction path search approaches, which we partially implemented, I investigated the central topic of my thesis, the similar but different catalysis of cysteine and serine peptidases.

Comparison of Cysteine and Serine Peptidase Mechanisms. Computational models and methods are important tools in computational biochemistry to describe biophysical properties and behaviors of biological systems. Such tools are necessary to reveal details, which are not easily accessible through laboratory experiments. The mechanism of cysteine and serine peptidases proceeds differently. Cysteine peptidases apply a stepwise mechanism with an ion-pair intermediate, while serine peptidase catalysis proceeds

concerted. Investigation of cysteine peptidases in the literature revealed requirements for the active site environment for the catalytic residues. A suitable cysteine peptidase active site environment supports catalysis by providing a stabilizing potential for the ion-pair intermediate, which enhances the nucleophilicity of the cysteine. However, environmental requirements for serine peptidase catalysis are poorly understood. In a detailed comparative study, I investigated requirements for serine peptidase catalysis and present their correlation to the concerted mechanism. I performed continuum electrostatic comparisons and analysis of active site geometries, together with QM/MM reaction path calculations. By that, I was able to explain, how the environment of the active site contributes to serine peptidase catalysis, and that the concerted mechanism is a consequence of that required environment. Thus it will not be possible to obtain a normally active peptidase by exchange of the catalytic residues serine and cysteine unless the environment of the active site is adapted as well. In Manuscript C, the bacterial cysteine peptidase NsPCS was investigated, which looses activity with change of its catalytic cysteine to serine (NsPCS-C70S). Fortunately, this mutant did not process the substrate glutathione but kept it bound non-covalently within the binding pocket. This provided new insights for the interaction prior to catalysis and was the structural basis for my QM/MM calculations on that peptidase. The inactivity of NsPCS-C70S was experimentally shown, and could be seen computationally by a very high energy barrier for catalysis. The actual reasons for that inactivity, and possible requirements for the activation of the mutant enzyme needed to be researched. In my further studies in Manuscript D, I comparatively investigated requirements for cysteine and serine peptidase catalysis. I used the serine peptidase trypsin, the cysteine peptidase NsPCS, and its inactive serine mutant NsPCS-C70S. This research revealed essential properties of an active site surrounding to allow for proper catalytic function of a catalytic triad with serine or cysteine as nucleophile. I further applied my observations to mimic an active site environmment for a serine peptidase within the inactive serine mutant NsPCS-C70S. From the computational perspective, this procedure successfully generated a suitable active site environment for a serine. Manuscript C and Manuscript D comprise a detailed analysis of cysteine and serine peptidases, which leads to a deeper understanding of essential features for the catalytic mechanism of both enzyme families.

3.1 Manuscript A: Computational Biochemistry – Enzyme Mechanisms Explored

The field of computational biochemistry approaches the properties of biological systems by computational methods. One essential component within biochemical processes are proteins, whose states, interactions, or enzymatic behaviors can be investigated in detail by computational approaches.

In this review article, we discuss the complementarity between computational and experimental results. We first review the use and the limitations of several computational models on different abstraction levels, such as continuum electrostatic models or QM/MM models. We further describe methods, which can be applied on such models to derive thermodynamic or kinetic properties, which can be compared to experimental results. However, experimental measurements provide macroscopic observations, which are supplemented by computational analysis with microscopic details. Thermodynamic properties provide information about probable states of investigated sytems, which characterize stable states on the potential energy surface. Kinetic properties supplement information about transitions, which energetically evaluate the rate of chemical or physical processes in proteins. These transitions can be described by transition state theory, which requires explicit analysis of intermediate structures along the reaction coordinate.

To explore the potential energy surface, methods are applied on computational models to obtain biochemical properties of the investigated system. We review several reaction path search methods, such as the adiabatic surface mapping, which is useful to obtain basic information about the investigated process. The search starts from one stable state and is driven into specific directions, which are derived by chemical intuition and experimental data. With basic information about the reaction coordinate, more specific path search methods can be applied, which often utilize a chain of states, that discretizes the reaction path between states. Such chain-of-states methods optimize the whole reaction path until convergence criteria are fulfilled.

One chain-of-states method is conjugate peak refinement, which I used in this theses for specific reaction path search. The conjugate peak refinement algorithm is subject of Manuscript B.

3.2 Manuscript B: PyCPR – a Python-based Implementation of the Conjugate Peak Refinement (CPR) Algorithm for Finding Transition State Structures

Conjugate peak refinement (CPR) belongs to the chain-of-states (COS) methods, which are described in Manuscript A. With COS methods enzymatic mechanisms can be investigated to obtain reaction coordinates, which are specified by stable states and transition states on a potential energy surface (PES). The PES originates from the 3N-dimensional energy function of a molecular system, regarding the coordinates of its N atoms. Thereby, the PES can be imagined as a landscape with stable states as valleys, and transition pathways as mountain passes, which connect the stable states. Information about stable states and transitions are used in transition state theory to attribute a rate to a reaction. The rate of an elementary step in catalysis is calculated from the energy difference of a stable state and the transition state along its reaction coordinate. The largest energy difference of all elementary steps defines the rate-limiting step of the reaction. The calculated overall reaction rate can be correlated with experimentally measured values. In addition to experimental measurements, computational analysis can provide further details about the process of enzyme catalysis, and the individual catalytic effects. Detailed knowledge about enzyme catalysis is relevant for the understanding of catalytic events, and further to adopt the principles, or adapt the process.

In this article we extensively describe the theory behind CPR and its implementation (PyCPR) for the Python-based framework pDynamo. We further provide examples and a strategy for PyCPR application. The algorithm of CPR is based on the characteristics of a first order saddle point, whose Hessian matrix has exactly one negative eigenvalue. Diagonalization of the Hessian with a conjugate basis set yields a matrix with its eigenvalues as its diagonal elements. The idea of this method is to approach to the vicinity of a saddle point and follow the vector associated with the negative eigenvalue to locate the saddle point. Subsequent energy minimization along all remaining conjugate directions, which are associated to the positive eigenvalues, optimizes the structure to the first order saddle point. This procedure is an

adaptation of the conjugate gradient method, but applies an initial search direction for energetic maximization and further provides a term, which ensures to be conjugate to the initial direction. By that, a set of conjugate vectors can be constructed, which define the conjugate basis set. This strategy is possible, since diagonalization of the Hessian matrix is independent of the set of conjugate basis vectors.

The CPR method is an adaptation of this outlined theory to molecular systems. It is a flexible COS method, which dynamically adjusts the number of explicit structures along the reaction coordinate. Transition state regions are approached by a higher density of explicit structures to find the proper transition structure. Explicit structures can be deleted from the path as well, if they describe regions, which are off a suitable reaction coordinate. The whole reaction coordinate is linked for the optimization, whereas just a single structure is optimized at time, with direct influence of its neighboring structures. The CPR method interpolates the segments between the path points to find the structure with the highest energy of each segment. If this structure is unspecified, it will undergo a refinement cycle. The structure is first energetically maximized along the tangential direction of the path, which approximates the eigenvector of the negative eigenvalue. Subsequent minimization is performed along the conjugate directions. This procedure is performed iteratively until all high energy structures are optimized. Some of the optimized structures will approximate first order saddle points and describe the transition states along the reaction coordinate.

We have implemented the CPR algorithm into pDynamo, a Python-based framework, capable of hybrid QM/MM molecular simulations. Our PyCPR implementation is open source, and can be combined with the features of pDynamo. Unlike the original CPR implementation in CHARMM, our PyCPR implementation provides several scaling factors for the influence of the previous search direction, as they can be found for variations of the conjugate gradient method. We further utilize a pure mathematical expression for the conjugacy measure, which can be chosen to be an exit criterion for the minimization steps. Further minimization, although the minimization directions are no longer conjugate, can thus lead to different minima and vary the reaction coordinate. We also implemented a highly efficient procedure to prevent the iterative addition and deletion of the same path points during

the complete PyCPR run. Our PyCPR implementation supplements the existing methods in pDynamo, and profits from the computational models and algorithms available within the pDynamo framework.

As application of PyCPR, we provide two examples, a principle study on the conformational change of butane from anti-periplanar to gauche, and the analysis of the catalytic mechanism of the glycyl radical enzyme 4-hydroxyphenylacetate decarboxylase. Both examples are compared to nudged elastic band (NEB) calculations. For the conformational change of butane, PyCPR and NEB reach the same conclusion, while the different behavior of NEB and CPR can be observed. The focus of NEB lies on the path as a whole, whereas CPR concentrates on transition state regions. However, the comparable results confirm the functionality of PyCPR. The comparison of results from PyCPR and NEB for the radical enzyme catalysis showed similar transitions for all elementary steps except the rate-limiting step. NEB missed a reliable transition state, which could be found by PyCPR.

PyCPR is a useful and reliable alternative to other reaction path search methods, and emerged as an essential tool for my reaction path search procedure. I applied PyCPR for the detailed analysis of peptidase catalysis, which is shown in Manuscript C and Manuscript D.

3.3 Manuscript C: Structural and Biophysical Analysis of the Phytochelatin-Synthase-like Enzyme from *Nostoc* sp. shows that its Protease Activity is Sensitive to the Redox State of the Substrate

NsPCS (phytochelatin synthase of *Nostoc* sp.) is a cysteine peptidase of the papain superfamily. It is a phytochelatin-synthase-like enzyme of the cyanobacterium *Nostoc* sp. and it catalyzes the cleavage of glutathione (GSH) into glycine and γ-glutamyl-cysteine (γEC). This cleavage is the first step of the phytochelatin (PC) synthesis. The final product are PCs that are glutathione derivatives, where the glycine is cleaved off and another GSH molecule is covalently linked via a peptide bond to γEC. By that reaction, also longer PC chains can be formed, where all γEC units possess one thiol group. Because

of the thiol groups PCs can chelate heavy metal ions and are assumed to participate in heavy metal detoxification or homeostasis.

In this article we performed a quantitative biophysical analysis of NsPCS. Although the enzyme is named phytochelatin synthase, NsPCS is not catalyzing the synthesis of PC molecules. Weak activity is reported in some cases but the actual reaction catalyzed by NsPCS is the cleavage of GSH into glycine and γEC. The oxidized form of the substrate GSH, GSSG, with two molecules of GSH covalently connected via a disulfide bond at the cysteine residues is also bound by NsPCS and the glycine of one GSH moiety is cleaved off. However, after cleaving off the glycine the remaining γEC-SG is not released and is inhibiting the enzyme. Thus, we show that GSSG acts as an inhibitor for NsPCS and makes it possible to trap the enzyme-substrate complex in an acyl-enzyme form (acyl-NsPCS).

NsPCS is not a phytochelatin synthase but the actual function is unclear. Since NsPCS processes GSH in its reduced form but is inhibited by its oxidized form GSSG, the enzyme might be associated with signaling of the redox state of the environment. However, the fact that the two GSH moieties in GSSG are close in the active site might have also sporadically led to a transpeptidase reaction. Then, instead of a water molecule the uncleaved GSH moiety of GSSG would attack the thioester bond of the enzyme-substrate complex and could lead to the formation of PCs. The characterization of NsPCS and the understanding of its reaction mechanism might therefore give insights on how the function of phytochelatin synthesis evolved and spread among organisms for which the presence of heavy-metal-chelating molecules became an advantage.

In detail, we analyzed the structures of the wild type enzyme (wt-NsPCS), the acyl enzyme (acyl-NsPCS) with γEC covalently bound to the catalytic cysteine, and the inactive serine variant (C70S-NsPCS), where the catalytic cysteine is mutated to a serine. The serine mutant structure C70S-NsPCS has been used to measure the binding affinity of the substrates GSH and GSSG to NsPCS, since C70S-NsPCS is not active and trapped the uncleaved substrate GSH non-covalently bound inside the binding pocket. The fact that GSH was found in the crystal structure of the inactive enzyme C70S-NsPCS led to the conclusion that GSH was co-purified with the enzyme, which assumes a tight binding. To characterize that binding, we performed binding studies with the inactive variant C70S-NsPCS. For these studies GSH first had to

be removed from the binding pocket of C70S-NsPCS, which was achieved by unfolding and refolding. Proper unfolding and refolding was monitored with spectroscopic methods and in addition also the activity of the native and the refolded wild type enzyme (wt-NsPCS) was measured to evaluate how the refolding affects the function of the enzyme. The spectroscopic results showed properly restored characteristics and also the activity test showed that the refolded enzyme was active, though just to an amount of about 10 %. This amount matched with the active portion of refolded C70S-NsPCS, which was determined by the binding studies. Nevertheless, the concentration of the enzyme is considered in the applied binding model and is therefore not affecting the analysis. The performed binding studies and the evaluation of the electrostatic interaction energies indicated a tight binding of the substrate and a slow exchange, which explains the co-purified GSH in the C70S-NsPCS structure. Although the crystal was grown under aerobic conditions and the substrate GSH is sensitive to oxidation (i.e. GSSG formation), reduced GSH is found in the solved crystal structure. The substrate in the active enzyme was trapped in the acyl-enzyme state, since the cleavage product of the oxidized form GSSG , i. e. γEC-SG, is inhibiting the enzyme. We performed a Monte Carlo simulation to analyze the encounter complex formation with GSH or γEC present in the active pocket. Encounter complex formation was clearly visible when the cleaved off glycine moiety of GSH was not present, which causes different electrostatics. Thus, in the active enzyme, where the glycine can be cleaved off the acyl enzyme can form an encounter complex with another GSH molecule, which then could lead to γEC-SG formation under aerobic conditions. However, when the glycine moiety of GSH is not cleaved off no encounter complex formation was observed. These docking results explain the presence of a reduced GSH in the binding pocket of C70S-NsPCS.

Our research provides a first quantitative biophysical analysis of the PCS-like protein NsPCS and leads us to the assumption that NsPCS may sense the redox state of the solution and the cleaved off glycine serves as a signal for reducing conditions. We performed binding studies and activity measurements and complemented them with continuum electrostatic and QM/MM calculations, whose theories are described in Manuscript A and Manuscript B. The analysis of NsPCS is important for my thesis since the enzyme is a cysteine peptidase of the papain superfamily and has a catalytic triade

proper for comparison with the serine peptidase trypsin. In particular, we have produced C70S-NsPCS, the inactive serine variant of NsPCS, which builds a substantial basis for the comparison of cysteine and serine peptidases, which is discussed in Manuscript D.

3.4 Manuscript D: Serine and Cysteine Peptidases – So Similar, Yet Different. How the Active-Site Electrostatics Facilitates Different Reaction Mechanisms

Cysteine and serine peptidases perform the cleavage of peptide bonds by nucleophilic attack and subsequent hydrolytic release of the product. While cysteine peptidases make use of a stepwise mechanism with an ion-pair intermediate, serine peptidases apply a concerted mechanism. Ion-pair stabilization in cysteine peptidases is enabled by hydrogen bonds and a positive electrostatic potential around the cysteine. Requirements for the active site environment for serine peptidase catalysis are indeed poorly understood. In this article I performed a comparative analysis of cysteine and serine peptidase catalysis by continuum electrostatic approaches, and detailed energetic and geometric analysis of the QM/MM reaction path. For my analysis, I mainly applied computational methods to a continuum electrostatic model, and a QM/MM model, which are explained in more detail in Manuscript A and Manuscript B.

At first, I analyzed the catalytic mechanisms of the serine peptidase trypsin, and the cysteine peptidases papain and NsPCS. NsPCS is a phytochelatin-synthase-like enzyme of *Nostoc* sp., which catalyses the cleavage of GSH into glycine and γEC, in particular the peptide bond between glycine and cysteine of glutathione is cleaved. Detailed information about NsPCS is provided in Manuscript C. The enzyme belongs to the papain superfamily of cysteine peptidases, and shares a similar fold. However, NsPCS utilizes an aspartate as the third catalytic triad residue, which is analog to trypsin, instead of an asparagine, which is found in papain. Hence, NsPCS was used as a model for cysteine peptidases for comparative analysis to trypsin, as a model for serine peptidases. Further, a serine mutant of NsPCS (C70S-NsPCS) exists, which experimentally shows no activity (see Manuscript C).

The QM/MM reaction path results showed an energetically very high transition state barrier of 31.5 kcal/mol, thus reinforced the inactivity of C70S-NsPCS. The inactivity is also consistent with other studies on the conversion of peptidases. The characterization of the electrostatic potential of the active site environments of NsPCS and trypsin reflected the common picture of cysteine peptidase catalysis, which facilitates ion-pair stabilization by a surrounding positive potential. Electrostatic analysis of the active site of the serine peptidase trypsin revealed a prevalent negative potential. Protonation characteristics of catalytic residues reveales the importance of a strong histidine base for serine catalysis, which is achieved by interaction with the catalytic aspartate, and a negative potential around. Additionally the nucleophilicity of the catalytic serine increases within a negative potential. However, such an environment of the nucleophile is not able to stabilize an ion pair. As a consequence serine peptidase catalysis has to proceed concerted.

To ensure reactivity of the catalytic residues, cysteine and serine peptidases impose different requirements on their active site environments. Therefore the placement of a serine within an active site environment of a cysteine peptidase causes inactivity, as can be seen for C70S-NsPCS. The enzyme evolved for ion-pair stabilization but not for proper activation of a catalytic serine. Comparison of the active site shape of cysteine and serine peptidases further revealed a more compact geometry of the catalytic residues of serine peptidases. A compact active site geometry is a necessity for a concerted mechanism and thus is essential for serine peptidases. The active site environment of the cysteine peptidase NsPCS neither provides a compact active site geometry, nor a negativ electrostatic potential related to a strong histidine base. Further the catalytic aspartate in NsPCS is not mainly interacting with the catalytic histidine, but with an arginine on top of the histidine. The simple change of the cysteine to a serine is therefore not sufficient for the conversion of the peptidases.

Based on the analysis, I mimicked an active site environment of a serine peptidase for C70S-NsPCS. In exchange with another residue, the catalytic aspartate was put at a different position, which increased the interaction with the catalytic histidine. This modification increased the basicity of the histidine, and further made the active site electrostatic potential more negative. Introduction of two more mutations moved the catalytic residues

closer together, and closer to the substrate, which provides the required compact geometry for a concerted mechanism. The combination of the introduced effects in a computationally varied structure, NsPCS-mut4, caused a decrease of the calculated transition barrier compared to C70S-NsPCS by about 11 kcal/mol. The resulting energy barrier of 20.2 kcal/mol for NSPCS-mut4 reaches the range of natural serine peptidases, which indicates a suitable active site environment for serine based catalysis.

This article reveals details about serine peptidase catalysis, which proceeds concerted in contrast to the stepwise cysteine peptidase catalysis. The concerted mechanism appears to be a consequence of the negative electrostatic environment of the serine nucleophile, which is essential for serine activation. A precondition of the concerted mechanism is a compact active site geometry to facilitate the simultaneous events. The detailed analysis of the different requirements for cysteine and serine peptidase catalysis in correlation with their different procedures revealed insights about the catalytic behavior of both enzyme families.

3.5 Outlook

The peptide bond is an essential component for life on earth. In particular its stability is important for the creation of stable polypeptides or proteins. However, cleavage of polypeptide chains for digestion or modification requires reactive enzymes which compensate the stability of a peptide bond. Nature evolved different types of peptide bond cleaving enzymes, so-called peptidases, of which two wide-spread types are cysteine and serine peptidases. The catalytic mechanism of cysteine peptidases and serine peptidases appears similar but is essentially different. This difference is remarkable, since a cysteine and a serine merely differ in one atom, a sulfur and an oxygen, respectively. The surroundings of the active sites are highly adapted to the characteristics of the present amino acid to effectively initiate catalysis, which is performed stepwise in cysteine peptidases and concerted in serine peptidases. The explanation for this phenomenon is described in Manuscript D and basically is: specific electrostatics around the catalytic triad residues in combination with an optimized geometry.

The basic research of my work about the functioning of cysteine and serine peptidases might have an impact in the design of specific peptidases. But

this work is not only restricted to peptidases, since the basic mechanism is applicable for hydrolases in general, which for instance includes also esterases. Esterases gained particular importance in the last years, since enzymes have been found, so-called PETases, which are capable of polyethylene terephthalate (PET) degradation. The durability of such plastics in nature and distribution of microplastics in various habitats poses a problem for living organisms. The detailed understanding of the functioning of hydrolases could support the design of novel enzymes, which are specifically optimized to digest different kids of plastics. PETases are commonly serine hydrolases. In Manuscript D I have shown, that the serine nucleophile is difficult to activate in comparison to a cysteine, and hence specific characteristics are required for the active site and its surrounding to facilitate a concerted mechanism. A concerted mechanism is needed, since a stepwise activation of the nucleophile via an ion-pair intermediate with a deprotonated serine is not feasible due to the high pK value of the serine. Indeed mutation studies showed, that the mutation of the catalytic serine of a serine peptidase into a cysteine results in a slightly active enzyme, which often shows just a minor activity against already activated substrates. However, the mutation of the catalytic cysteine of a cysteine peptidase into a serine causes complete loss of function, which was also observed for the serine mutant C70S-NsPCS in Manuscript C.

It might be interesting to use specific cysteine peptidases, which recognizes specific targets, and mutate the cysteine nucleophile into a serine, especially since a serine is less sensitive to the redox state of the environment. In Manuscript D I performed calculations on the serine mutant of the cysteine peptidase NsPCS. Based on my observations about a serine peptidase I suggested possible mutations (see Manuscript D, NsPCS-mut4) for the inactive serine mutant C70S-NsPCS to adapt the characteristics of a serine peptidase. Reaction path calculations showed an energy barrier for NsPCS-mut4, which is in the range of functioning serine peptidases. To test that suggestion, I expressed and purified NsPCS-mut4. However, the introduced mutations caused instability of the protein, which is why the expression product was not soluble and had to be purified in denaturing conditions to have it unfolded and therefore soluble. After purification NsPCS-mut4 was refolded and the activity was measured with NMR but no activity was observed. No

conclusion can be drawn directly if the introduced mutations will enable peptidase activity in the calculated structure, since the structure itself is destabilized and not folded properly. Despite the insolubility of NsPCS-mut4 after expression, the unfolding and refolding experiments for wild type NsPCS, which are described in Manuscript C also showed just around 10 % of successfully refolded protein. Possible next steps would be to test the mutations individually to account for their effects and to identify destabilizing mutations to avoid the purification in denaturing conditions.

This thesis provides an overview of computational concepts for the analysis of enzyme mechanisms. It further shows the application of some presented computational concepts. First, in a complementing structural and biophysical analysis of a protein combining experimental results and calculations. Second, in a theoretical study on the comparison of specific enzyme mechanisms. In particular, I focus on the catalytic mechanisms of cysteine and serine peptidases, of which many representatives have already been extensively characterized. However, for a meaningful comparison of both peptidase types a rather uncharacterized cysteine peptidase, NsPCS, was investigated as part of this thesis, since it corresponds well to the extensively characterized serine peptidase trypsin. Although the principle mechanisms of both peptidase types are known it was not understood how serine peptidases facilitate a concerted mechanism and why it is required. My studies give detailed insights about the geometrical and electrostatic requirements for cysteine and serine peptidases, and contributes to the basic research of this wide-spread type of catalytic reaction.

Bibliography

[1] Nelson, D. and Cox, M. *Lehninger Biochemie (German Edition)*. 4th ed. Springer, 2009. ISBN: 3540686371.

[2] Galison, P. L. Computer Simulations and the Trading Zone. In: *Peter Galison & David J. Stump (eds.), The Disunity of Science: Boundaries, Contexts, and Power*. Stanford University Press, 1996, 118–157.

[3] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žıdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. 596 (2021), 583–589. DOI: 10.1038/s41586-021-03819-2.

[4] Palmer, A. G. Enzyme Dynamics from NMR Spectroscopy. *Acc. Chem. Res.* 48 (2015), 457–465. DOI: 10.1021/ar500340a.

[5] Frueh, D. P., Goodrich, A. C., Mishra, S. H., and Nichols, S. R. NMR Methods for Structural Studies of Large Monomeric and Multimeric Proteins. *Curr. Opin. Struct. Biol.* 23 (2013), 734–739. DOI: 10.1016/j.sbi.2013.06.016.

[6] Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In: *Methods in Molecular Biology*. Springer New York, 2017, 627–641. DOI: 10.1007/978-1-4939-7000-1_26.

[7] Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*. Wiley, 2004. ISBN: 0470091827.

[8] Hohenberg, P. and Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* 136 (1964), B864–B871. DOI: 10.1103/physrev.136.b864.

[9] Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* 98 (1993), 5648–5652. DOI: 10.1063/1.464913.

[10] Görling, A. and Levy, M. Hybrid Schemes Combining the Hartree–Fock Method and Density-Functional Theory: Underlying Formalism and Properties of Correlation Functionals. *J. Chem. Phys.* 106 (1997), 2675–2680. DOI: 10.1063/1.473369.

[11] Mourik, T. van, Bühl, M., and Gaigeot, M.-P. Density Functional Theory Across Chemistry, Physics and Biology. *Phil. Trans. Math. Phys. Eng. Sci.* 372 (2014), 20120488. DOI: 10.1098/rsta.2012.0488.

[12] Cole, D. J. and Hine, N. D. M. Applications of Large-Scale Density Functional Theory in Biology. *J. Phys. Condens. Matter* 28 (2016), 393001. DOI: 10.1088/0953-8984/28/39/393001.

[13] Warshel, A. and Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* 103 (1976), 227–249. DOI: 10.1016/0022-2836(76)90311-9.

[14] Senn, H. M. and Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* 48 (2009), 1198–1229. DOI: 10.1002/anie.200802019.

[15] Sousa, S. F., Ribeiro, A. J. M., Neves, R. P. P., Brás, N. F., Cerqueira, N. M. F. S. A., Fernandes, P. A., and Ramos, M. J. Application of Quantum Mechanics/Molecular Mechanics Methods in the Study of Enzymatic Reaction Mechanisms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 7 (2016), e1281. DOI: 10.1002/wcms.1281.

[16] Ullmann, G. M. and Bombarda, E. Continuum Electrostatic Analysis of Proteins. In: *Protein Modelling*. Springer International Publishing, 2014, 135–163. DOI: 10.1007/978-3-319-09976-7_6.

[17] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. USA* 98 (2001), 10037–10041. DOI: 10.1073/pnas.181342398.

[18] Ullmann, G. M., Kloppmann, E., Essigke, T., Krammer, E.-M., Klingen, A. R., Becker, T., and Bombarda, E. Investigating the Mechanisms of Photosynthetic Proteins Using Continuum Electrostatics. *Photosynth. Res.* 97 (2008), 33–53. DOI: 10.1007/s11120-008-9306-1.

[19] Field, M. J. The pDynamo Program for Molecular Simulations using Hybrid Quantum Chemical and Molecular Mechanical Potentials. *J. Chem. Theory Comput.* 4 (2008), 1151–1161. DOI: 10.1021/ct800092p.

[20] Fischer, S. and Karplus, M. Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems With Many Degrees of Freedom. *Chem. Phys. Lett.* 194 (1992), 252–261. DOI: 10.1016/0009-2614(92)85543-j.

[21] Sinclair, J. E. and Fletcher, R. A New Method of Saddle-Point Location for the Calculation of Defect Migration Energies. *J. Phys. C: Solid State Phys.* 7 (1974), 864–870. DOI: 10.1088/0022-3719/7/5/009.

[22] Gisdon, F. J., Culka, M., and Ullmann, G. M. PyCPR – A Python-Based Implementation of the Conjugate Peak Refinement (CPR) Algorithm for Finding Transition State Structures. *J. Mol. Model.* 22 (2016). DOI: 10.1007/s00894-016-3116-8.

[23] Himo, F. Quantum Chemical Modeling of Enzyme Active Sites and Reaction Mechanisms. *Theor. Chem. Acc.* 116 (2006), 232–240. DOI: 10.1007/s00214-005-0012-1.

[24] Hartley, B. S. Proteolytic Enzymes. *Annu. Rev. Biochem.* 29 (1960), 45–72. DOI: 10.1146/annurev.bi.29.070160.000401.

[25] Rawlings, N. D. and Barrett, A. J. Evolutionary Families of Peptidases. *Biochem. J.* 290 (1993), 205–218. DOI: 10.1042/bj2900205.

[26] Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. MEROPS: The Database of Proteolytic Enzymes, Their Substrates and Inhibitors. *Nucleic Acids Res.* 42 (2014), D503–D509. DOI: 10.1093/nar/gkt953.

[27] Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., and Finn, R. D. The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database. *Nucleic Acids Res.* 46 (2017), D624–D632. DOI: 10.1093/nar/gkx1134.

[28] Vivares, D., Arnoux, P., and Pignol, D. A Papain-Like Enzyme at Work: Native and Acyl-Enzyme Intermediate Structures in Phytochelatin Synthesis. *Proc. Natl. Acad. Sci. USA* 102 (2005), 18848–18853. DOI: 10.1073/pnas.0505833102.

[29] Rea, P. A. Phytochelatin Pynthase, Papain's Cousin, in Stereo. *Proc. Natl. Acad. Sci. USA* 103 (2006), 507–508. DOI: 10.1073/pnas.0509971102.

[30] Page, M. J. and Cera, E. D. Evolution of Peptidase Diversity. *J. Biol. Chem.* 283 (2008), 30010–30014. DOI: 10.1074/jbc.m804650200.

[31] López-Otın, C. and Bond, J. S. Proteases: Multifunctional Enzymes in Life and Disease. *J. Biol. Chem.* 283 (2008), 30433–30437. DOI: 10.1074/jbc.r800035200.

[32] Bröer, S. and Bröer, A. Amino Acid Homeostasis and Signalling in Mammalian Cells and Organisms. *Biochem. J.* 474 (2017), 1935–1963. DOI: 10.1042/bcj20160822.

[33] Heutinck, K. M., Berge, I. J. ten, Hack, C. E., Hamann, J., and Rowshani, A. T. Serine Proteases of the Human Immune System in Health and Disease. *Mol. Immunol.* 47 (2010), 1943–1955. DOI: 10.1016/j.molimm.2010.04.020.

[34] Walsh, P. N. and Ahmad, S. S. Proteases in Blood Clotting. *Essays Biochem.* 38 (2002), 95–111. DOI: 10.1042/bse0380095.

[35] Solary, E., Eymin, B., Droin, N., and Haugg, M. Proteases, Proteolysis, and Apoptosis. *Cell Biol. Toxicol.* 14 (1998), 121–132. DOI: 10.1023/a:1007481921502.

[36] Blackman, M. J. Proteases in Host Cell Invasion by the Malaria Parasite. *Cell. Microbiol.* 6 (2004), 893–903. DOI: 10.1111/j.1462-5822.2004.00437.x.

[37] Potempa, J. and Pike, R. N. Corruption of Innate Immunity by Bacterial Proteases. *J. Innate Immun.* 1 (2008), 70–87. DOI: 10.1159/000181144.

[38] Mótyán, J., Tóth, F., and Tőzsér, J. Research Applications of Proteolytic Enzymes in Molecular Biology. *Biomolecules* 3 (2013), 923–942. DOI: 10.3390/biom3040923.

[39] Li, Q., Yi, L., Marek, P., and Iverson, B. L. Commercial Proteases: Present and Future. *FEBS Lett.* 587 (2013), 1155–1163. DOI: 10.1016/j.febslet.2012.12.019.

[40] Arnau, J., Lauritzen, C., Petersen, G. E., and Pedersen, J. Current Strategies for the Use of Affinity Tags and Tag Removal for the Purification of Recombinant Proteins. *Protein Expr. Purif.* 48 (2006), 1–13. DOI: 10.1016/j.pep.2005.12.002.

[41] Giansanti, P., Tsiatsiani, L., Low, T. Y., and Heck, A. J. R. Six Alternative Proteases for Mass Spectrometry–Based Proteomics Beyond Trypsin. *Nat. Protoc.* 11 (2016), 993–1006. DOI: 10.1038/nprot.2016.057.

[42] Tavano, O. L., Berenguer-Murcia, A., Secundo, F., and Lafuente, R. F. Biotechnological Applications of Proteases in Food Technology. *Compr. Rev. Food Sci. F.* 17 (2018), 412–436. DOI: 10.1111/1541-4337.12326.

[43] Jacob, M., Jaros, D., and Rohm, H. Recent Advances in Milk Clotting Enzymes. *Int. J. Dairy Technol.* 64 (2010), 14–33. DOI: 10.1111/j.1471-0307.2010.00633.x.

[44] Rizzello, C. G., Angelis, M. D., Cagno, R. D., Camarca, A., Silano, M., Losito, I., Vincenzi, M. D., Bari, M. D. D., Palmisano, F., Maurano, F., Gianfrani, C., and Gobbetti, M. Highly Efficient Gluten Degradation by Lactobacilli and Fungal Proteases during Food Processing: New Perspectives for Celiac Disease. *Appl. Environ. Microbiol.* 73 (2007), 4499–4507. DOI: 10.1128/aem.00260-07.

[45] Osborn, D. A., Sinn, J. K., and Jones, L. J. Infant Formulas Containing Hydrolysed Protein for Prevention of Allergic Disease. *Cochrane Database Syst. Rev.* (2018). DOI: 10.1002/14651858.cd003664.pub6.

[46] Rawlings, N. D. Peptidase Specificity from the Substrate Cleavage Collection in the MEROPS Database and a Tool to Measure Cleavage Site Conservation. *Biochimie* 122 (2016), 5–30. DOI: 10.1016/j.biochi.2015.10.003.

[47] Giles, N. M., Watts, A. B., Giles, G. I., Fry, F. H., Littlechild, J. A., and Jacob, C. Metal and Redox Modulation of Cysteine Protein Function. *Chem. Biol.* 10 (2003), 677–693. DOI: 10.1016/s1074-5521(03)00174-1.

[48] Nagahara, N. Catalytic Site Cysteines of Thiol Enzyme: Sulfurtransferases. *J. Amino Acids* 2011 (2011), 1–7. DOI: 10.4061/2011/709404.

[49] Nozaki, Y. and Tanford, C. [84] Examination of Titration Behavior. In: *Method. Enzymol.* Elsevier, 1967, 715–734. DOI: 10.1016/s0076-6879(67)11088-4.

[50] Thurlkill, R. L., Grimsley, G. R., Scholtz, J. M., and Pace, C. N. pK Values of the Ionizable Groups of Proteins. *Protein Sci.* 15 (2006), 1214–1218. DOI: 10.1110/ps.051840806.

[51] Ballinger, P. and Long, F. A. Acid Ionization Constants of Alcohols. II. Acidities of Some Substituted Methanols and Related Compounds1, 2. *J. Am. Chem. Soc.* 82 (1960), 795–798. DOI: 10.1021/ja01489a008.

[52] Ishida, T. and Kato, S. Theoretical Perspectives on the Reaction Mechanism of Serine Proteases: The Reaction Free Energy Profiles of the Acylation Process. *J. Am. Chem. Soc.* 125 (2003), 12035–12048. DOI: 10.1021/ja021369m.

[53] Mladenovic, M., Fink, R. F., Thiel, W., Schirmeister, T., and Engels, B. On the Origin of the Stabilization of the Zwitterionic Resting State of Cysteine Proteases: A Theoretical Study. *J. Am. Chem. Soc.* 130 (2008), 8696–8705. DOI: 10.1021/ja711043x.

[54] Beveridge, A. J. A Theoretical Study of the Active Sites of Papain and S195C Rat Trypsin: Implications for the Low Reactivity of Mutant Serine Proteinases. *Protein Sci.* 5 (1996), 1355–1365. DOI: 10.1002/pro.5560050714.

[55] Dardenne, L. E., Werneck, A. S., Oliveira Neto, M. de, and Bisch, P. M. Electrostatic Properties in the Catalytic Site of Papain: A Possible Regulatory Mechanism for the Reactivity of the Ion Pair. *Proteins: Struct., Funct., Genet.* 52 (2003), 236–253. DOI: 10.1002/prot.10368.

[56] Olsen, J. V., Ong, S.-E., and Mann, M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Mol. Cell. Proteom.* 3 (2004), 608–614. DOI: 10.1074/mcp.t400003-mcp200.

[57] Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* 102 (2002), 4501–4524. DOI: 10.1021/cr000033x.

[58] Harrison, M. J., Burton, N. A., and Hillier, I. H. Catalytic Mechanism of the Enzyme Papain: Predictions with a Hybrid Quantum Mechanical/Molecular Mechanical Potential. *J. Am. Chem. Soc.* 119 (1997), 12285–12291. DOI: 10.1021/ja9711472.

[59] Wei, D., Huang, X., Liu, J., Tang, M., and Zhan, C.-G. Reaction Pathway and Free Energy Profile for Papain-Catalyzed Hydrolysis of N-Acetyl-Phe-Gly 4-Nitroanilide. *Biochemistry* 52 (2013), 5145–5154. DOI: 10.1021/bi400629r.

[60] Harrison, M. J., Burton, N. A., Hillier, I. H., and Gould, I. R. Mechanism and Transition State Structure for Papain Catalysed Amide Hydrolysis, Using a Hybrid QM/MM Potential. *Chem. Commun.* (1996), 2769. DOI: 10.1039/cc9960002769.

[61] Han, W.-G., Tajkhorshid, E., and Suhai, S. QM/MM Study of the Active Site of Free Papain and of the NMA-Papain Complex. *J. Biomol. Struct. Dyn.* 16 (1999), 1019–1032. DOI: 10.1080/07391102.1999.10508311.

[62] Neet, K. E., Nanci, A., and Koshland, D. E. Properties of Thiol-Subtilisin. The Consequences of Converting the Active Serine Residue to Cysteine in a Serine Protease. *J.Biol. Chem.* 243 (1968), 6392–6401.

[63] Higaki, J. N., Evnin, L. B., and Craik, C. S. Introduction of a Cysteine Protease Active Site Into Trypsin. *Biochemistry* 28 (1989), 9256–9263. DOI: 10.1021/bi00450a004.

[64] Clark, P. I. and Lowe, G. Conversion of the Active-Site Cysteine Residue of Papain into a Dehydro-serine, a Serine and a Glycine Residue. *Eur. J. Biochem.* 84 (1978), 293–299. DOI: 10.1111/j.1432-1033.1978.tb12168.x.

Manuscripts

Contributions to the Joint Publications

Manuscript A

Martin Culka*, **Florian J. Gisdon***, G. Matthias Ullmann (2017): Computational biochemistry – enzyme mechanisms explored. *Adv. Protein Chem. Struct. Biol.*, 109, 77-112
DOI: 10.1016/bs.apcsb.2017.04.004

The core chapters about structural computer models in biological science and techniques to computationally simulate enzyme catalysis were jointly written by me and Martin Culka. The introduction and conclusion was written by Matthias Ullmann, who contributed to discussions and helped to prepare the final version of the manuscript.

Manuscript B

Florian J. Gisdon*, Martin Culka*, G. Matthias Ullmann (2016): PyCPR – a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures. *J. Mol. Model.*, 22, 242
DOI: 10.1007/s00894-016-3116-8

The reverse engineering of the CPR code, further developments, and implementation within the pDynamo environment were performed by me and Martin Culka. Discussions about the theoretical background of the CPR theory and further algorithmic developments were held together with Martin Culka and Matthias Ullmann. The CPR and NEB calculations on butane were performed by me. The calculations on HPD were performed by Martin Culka. Results of calculations were analyzed by me and Martin Culka. All parts of the manuscript were written by me and Martin Culka with support of Matthias Ullmann.

Manuscript C

Florian J. Gisdon, Christian G. Feiler, Oxana Kempf, Johannes M. Foerster, Jonathan Haiss, Wulf Blankenfeld, G. Matthias Ullmann, Elisa Bombarda (2022): Structural and biophysical analysis of the phytochelatin-synthase-like enzyme from *Nostoc* sp. shows that its protease activity is sensitive to the redox state of the substrate. *ACS Chem. Biol.*, 17, 4, 883–897
DOI: 10.1021/acschembio.1c00941

All calculations on the reaction mechanism, the encounter complex simulations, the electrostatic interaction energies, and the protonation probabilities presented in the manuscript were performed by me and analyzed and interpreted by me, Matthias Ullmann, and Elisa Bombarda. The protein preparation protocol was established by Christian Feiler with support of Wulf Blankenfeld. Also the crystal structures were prepared, solved and refined by Christian Feiler with support of Wulf Blankenfeld. The comparison and analysis of the crystal structures were performed by me with support of Matthias Ullmann and Elisa Bombarda. The hydrogen bond network analysis was performed by me. The NMR activity measurements for NsPCS in its native fold and refolded were performed by me and Jonathan Haiss, analyzed by me, and interpreted by me, Matthias Ullmann and Elisa Bombarda. The binding studies with isothermal titration calorimetry were performed by me, Jonathan Haiss, and Elisa Bombarda. The binding studies with thermophoresis were performed by Jonathan Haiss and Elisa Bombarda, and analyzed by Jonathan Haiss, me, and Elisa Bombarda. The final analysis of all binding studies was performed by me and interpreted by me, Jonathan Haiss, and Elisa Bombarda. Kinetic measurements were performed by Johannes Foerster and Oxana Kempf, and analyzed by Johannes Foerster, Oxana Kempf and Elisa Bombarda. Final interpretation of the kinetic data was performed by me, Matthias Ullmann and Elisa Bombarda. This manuscript was written by me, Matthias Ullmann and Elisa Bombarda with support of Wulf Blankenfeld.

Manuscript D

Florian J. Gisdon, Elisa Bombarda, G. Matthias Ullmann (2022): Serine and cysteine peptidases – so similar, yet different. How the active-site electrostatics facilitates different reaction mechanisms. *J. Phys. Chem. B*, 126, 22, 4035–4048
DOI: 10.1021/acs.jpcb.2c01484

All calculations presented in the manuscript were performed by me, and discussed and interpreted by me, Elisa Bombarda, and Matthias Ullmann. This manuscript was written by me with support of Elisa Bombarda and Matthias Ullmann.

* These authors contributed equally to the manuscript

Manuscript A

Computational Biochemistry – Enzyme Mechanisms Explored

Martin Culka, Florian J. Gisdon, G. Matthias Ullmann, *Adv. Protein Chem. Struct. Biol.* 2017, 109, 77-112
DOI: 10.1016/bs.apcsb.2017.04.004

CHAPTER THREE

Computational Biochemistry—Enzyme Mechanisms Explored

Martin Culka[2], Florian J. Gisdon[2], G. Matthias Ullmann[1]
Computational Biochemistry, University of Bayreuth, Bayreuth, Germany
[1]Corresponding author: e-mail address: Matthias.Ullmann@uni-bayreuth.de

Contents

Abstract

Understanding enzyme mechanisms is a major task to achieve in order to comprehend how living cells work. Recent advances in biomolecular research provide huge amount of data on enzyme kinetics and structure. The analysis of diverse experimental results and their combination into an overall picture is, however, often challenging. Microscopic details of the enzymatic processes are often anticipated based on several hints from macroscopic experimental data. Computational biochemistry aims at creation of a computational model of an enzyme in order to explain microscopic details of the catalytic process and reproduce or predict macroscopic experimental findings. Results of such computations are in part complementary to experimental data and provide an explanation of a biochemical process at the microscopic level. In order to evaluate the mechanism of an enzyme, a structural model is constructed which can be analyzed by several theoretical approaches. Several simulation methods can and should be combined to get a reliable picture of the process of interest. Furthermore, abstract models of

[2] These authors contributed equally.

Advances in Protein Chemistry and Structural Biology, Volume 109
ISSN 1876-1623
http://dx.doi.org/10.1016/bs.apcsb.2017.04.004

biological systems can be constructed combining computational and experimental data. In this review, we discuss structural computational models of enzymatic systems. We first discuss various models to simulate enzyme catalysis. Furthermore, we review various approaches how to characterize the enzyme mechanism both qualitatively and quantitatively using different modeling approaches.

1. INTRODUCTION

"Certainly no subject or field is making more progress on so many fronts at the present moment than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms" (Feynman, 1964). These words were said by the well-known physicist Richard Feynman in his famous *Lectures on Physics* now more than 50 years ago. Until today, this sentence has not lost its validity and is the basis for much of the biomolecular research, maybe today even more than 50 years ago. Today, we are beginning to understand how the interplay of atoms and molecules lead to the complex processes that we find in living systems. Recent advances in genomics and systems biology help to gain more and more insights into the molecular organizations of the cell, which are nevertheless still too complex to allow a full overview of all the biochemical processes at atomic detail. New techniques in structural biology such as the free-electron laser allow to analyze structurally the kinetics of molecular processes at an atomic level (Nango et al., 2016; Pande et al., 2016). On the other hand, modern electron microscopy allows us to gain insights into large molecular assemblies that were not accessible to structural investigations a few years ago (Bartesaghi et al., 2015; Kühlbrandt, 2014). These new techniques complement traditional techniques of structural biology such as X-ray crystallography and NMR. In the past, a structural characterization of a protein was considered the final goal of an investigation. Nowadays to gain a deep understanding of an enzymatic mechanism, a structure is just the beginning. The structural information needs to be completed by other experimental information for instance from spectroscopy, kinetics, or electrochemistry. Often, all these different aspects are difficult to merge. Thus, it is important to approach the enzymatic mechanism also from a theoretical side. With the help of modern methods from computational chemistry, it is possible to gain insights into enzymatic mechanisms and to complete the picture.

In order to understand experimental data, we build models of reality and use these models in our theoretical considerations. Models are not only used for theoretical calculations, but also the interpretation of experimental data relies on models of the studied system. A model is a generalized hypothetical description used for analyzing or explaining a system. It is a simplified representation of a real system intended to enhance our ability to understand, predict, and control the behavior of the real system. When a model is made, there are always approximations required. Therefore a model is always an idealized representation of the real system. A model reproduces only certain aspects of the real system, only those that are relevant for the properties under study. Other aspects of the same real system may not be described equally well by the model, since it was originally made for another purpose. A model should be able to explain experimental data and make predictions about the outcome of new experiments. In order to be able to make predictions, a model has to be complicated enough to represent all important aspects of a real system at an appropriate level of description. However, the more complicated and complex a model is, the more difficult it becomes to interpret the results of the model. It is therefore required that the model is as complicated as necessary and not more complicated in order to give insights into the behavior of the real system. Most importantly, a model should promote our understanding of nature.

A model of a real system is generally constructed in several steps (see Fig. 1). The first and probably most important step is the construction of the conceptual model. In this step, the real system is translated into an idealized model system. Having in mind which properties of the system are interesting, the features of the system that are important to reproduce these properties are selected and it is decided how to use them to describe the desired aspects of the system. This first step requires a detailed inspection of the system that should be modeled in order to decide which details are required for the desired representation. When only a qualitative understanding is requested, the conceptual model is often enough to picture the system and thus it represents the final goal of the investigation. Instead, when a quantitative or semiquantitative understanding of the system is desired, the conceptual model needs to be translated into a mathematical model. A concrete physical theory is required to forge the mathematical model designed in the conceptual model. If this mathematical model can be solved analytically, the goal is reached and the behavior of the model system can be compared with the behavior of the real system. However, the mathematical models are often too complex to be solved analytically. In this case, the mathematical model needs to be translated into a discrete mathematical

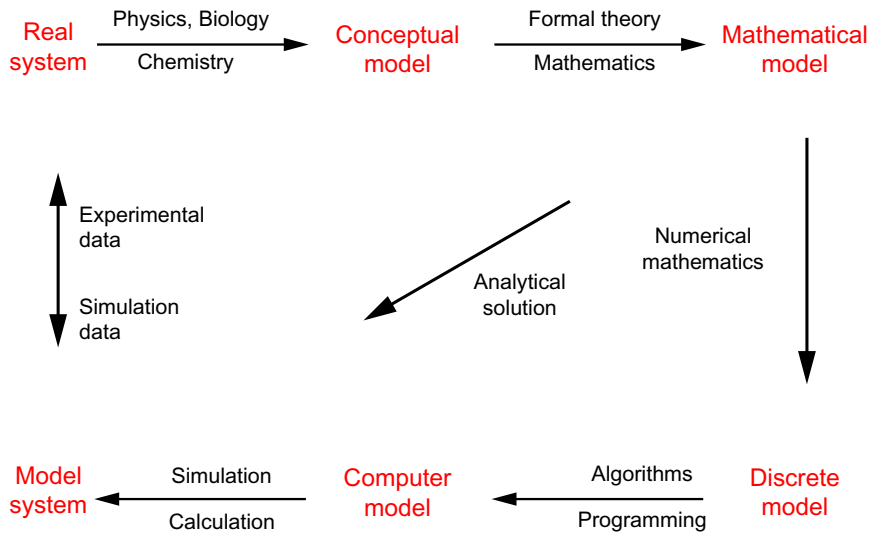


Fig. 1 Building theoretical models of real systems. This scheme depicts the stages of building theoretical models that can help to analyze real systems. These theoretical models are abstractions of the real systems that help to understand the behavior of the system. Mathematical modeling allows a quantitative comparison between the real system and the model.

model using methods from numerical mathematics. This discrete model can then be implemented in computer programs. The computer model can then be used to perform computer simulations and calculations using sets of parameters. The resulting simulation data can be compared with experimental data and so the quality of the model can be judged. The quantities that have been calculated from the model can lead to a better understanding of the real system. Moreover, the analysis of the effect of well-defined changes of the initial parameter can provide deeper insights into the behavior of the real system.

One of the most basic model that is widely used in the description of chemical systems is the Born–Oppenheimer approximation according to which one can separate the motion of nuclei from the motion of the electrons. This approximation allows us to describe molecular structures by defining the nuclear coordinates and to speak about different electronic states of a molecule. One fundamental concept that is used in molecular biophysics is the molecular energy landscape which describes the energy of a molecular system with N atoms in dependence of its $3N$ coordinates. This energy landscape contains minima which represent stable states of the systems, (first-order) saddle points which represent transition states of chemical reactions, and minimum energy paths which represent the trajectory along which chemical reactions occur. A large part of theoretical biochemistry is concerned with exploring this energy landscape and to extract different kind

of data from it. In this article, we review some of the most important methods to explore this energy landscapes and to explain how they can be used to explore enzymatic mechanisms.

2. STRUCTURAL MODELS

The most computational enzymology techniques construct a structural model of the enzyme in question, usually using its experiment-derived 3D structure as a starting point. A structural model provides a way to calculate energy of the given structure, by using various approximations of the real picture to stay computationally feasible. Based on energy differences between different states within structural computer models of enzymes, thermodynamics, and kinetics of enzyme catalysis can be addressed. With the use of structural models (Fig. 2), researchers aim to track the events within the

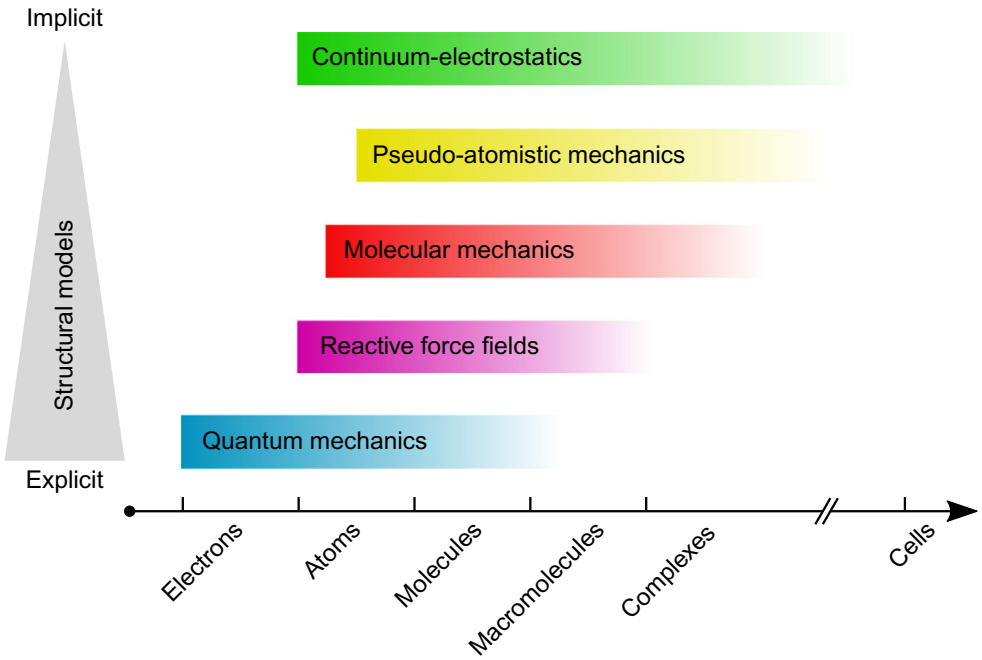


Fig. 2 Structural models for the simulation of catalytic events within enzymes. The *vertical axis* represents an ascending gradient from explicit to implicit models. The system description lose the detailed (explicit) character in favor of averaged (implicit) description of physical properties. The *horizontal axis* shows the information that can be obtained from the models, and together illustrates the size of the biological system, which is computationally feasible to treat. The *model bars* indicate possible applications. The *fading areas* show the potential of future developments enabled by improvements of algorithmic and computational power. All models can be combined into the so-called hybrid or multiscale models, which treat different parts of the biological system with the appropriate method, to enhance the effectiveness of the simulation and the quality of results.

enzyme that lead to catalysis. Models used to study enzyme mechanisms can be atomistic, i.e., they explicitly describe the behavior of electrons and nuclei, or pseudo-atomistic, i.e., they treat groups of atoms as one entity. In contrast to these particle-based models, continuum models describe a system in terms of continuous properties assigned to the space. Typical continuum models in enzymology are implicit solvent models. Their aim is to describe the average properties of a solvent environment, instead of discrete contributors of its parts. In this section, we briefly discuss various types of structural models and their combinations to simulate enzymatic activities.

2.1 Continuum-Electrostatics Models

Since most of the effects in biochemistry are dominated by electrostatics, a pure electrostatic model can be used to satisfactory describe many features of biomolecular systems while being computationally efficient. The most commonly used continuum-electrostatics model in biochemistry relies on the Poisson–Boltzmann equation. For a recent review the reader is referred to Ullmann and Bombarda (2014). The basic idea of this continuum-electrostatics model is to describe the protein as a low-dielectric region, which is embedded in (aqueous) solvent, described as a high-dielectric region. The charge distribution of the protein is described by a fixed charge distribution in the low-dielectric region, which is given by the molecular structure of the protein. The charge distribution of the protein is modeled by (fractional) point charges that are placed at the center of the atoms. The dissolved ions are represented by a Boltzmann-distributed charge density. The boundary of the low-dielectric region is defined by the solvent accessible surface of the protein.

The Poisson–Boltzmann equation is usually solved numerically (Honig & Nicholls, 1995; Warwicker & Watson, 1982). The solution of the Poisson–Boltzmann equation can be expressed as a potential $\phi(\mathbf{r})$, that is, composed of two contributions:

$$\phi(\mathbf{r}) = \sum_{i=1}^{M} \frac{q_i}{4\pi\varepsilon_p|\mathbf{r}-\mathbf{r}_i'|} + \phi_{\mathrm{rf}}(\mathbf{r}) \tag{1}$$

First, the Coulomb potential at the position $\mathbf{r}$ caused by M point charges q_i at positions $\mathbf{r}_i'$ in a medium with a permittivity ε_p, and second, the reaction field potential $\phi_{\mathrm{rf}}(\mathbf{r})$, arising from the M point charges q_i and the dielectric

boundary between the protein and the solvent, as well as from the distribution of ions in the solution. Electrostatic energy can be obtained by integrating the potential distribution over the space.

Continuum-electrostatics models such as the above described Poisson–Boltzmann model have various applications. The most straightforward is calculation of solvation energies and visualization of electrostatic potentials (Baker, Sept, Joseph, Holst, & McCammon, 2001). More advanced applications allow to calculate energies of different protonation and oxidation microstates of the protein, as well as ligand binding energies and many other applications (Ullmann et al., 2008). Furthermore, the Poisson–Boltzmann model or other popular continuum models such as COSMO (Klamt & Schüürmann, 1993) are often combined with more detailed models to simulate solvent effects in a computationally affordable way (Chen, Noodleman, Case, & Bashford, 1994; Li, Nelson, Peng, Bashford, & Noodleman, 1998; Liu et al., 2004).

2.2 Quantum-Mechanical Models

The most detailed structural models that treat both electrons and nuclei explicitly are based on quantum mechanics (QM). In QM models, the energy of the structure is derived from solving an approximate Schrödinger equation. One way is to employ ab initio molecular orbital wave function approaches such as the single-determinant Hartree–Fock (HF) method (Cramer, 2004). The major drawback of the HF method is the mean-field treatment of electron repulsion referred to as the electron correlation problem. Møller–Plesset theory (Møller & Plesset, 1934) addresses the electron correlation effects by means of Rayleigh–Schrödinger perturbation theory. The most common second-order version is called MP2 (Head-Gordon, Pople, & Frisch, 1988). In coupled cluster (CC) methods (Čížek, 1966), multielectron wave-functions are constructed using the exponential cluster operator to account for the electron correlation. The latter two more accurate approaches are, however, still very costly for larger systems. For complex biological systems, fast semiempirical (SE) approximations of the Hartree–Fock theory have become very popular in the past and are still used and developed up to now (Stewart, 2013; Řezáč & Hobza, 2012; Yilmazer & Korth, 2015). The SE methods achieve considerable calculation speed-up by parameterizing various parts of the HF theory in order to reproduce experimental or high-level ab initio QM results.

Density functional theory (DFT) is an alternative ab initio approach that has become very popular for biological systems due to its favorable price/performance ratio (van Mourik, Bühl, & Gaigeot, 2014). Instead of searching for a multidimensional wave-function that describes the position of every electron, the whole problem is solved from the point of view of total electron distribution (density) in the space. The electron correlation is taken into account in DFT methods, albeit on approximate level. In addition, hybrid DFT–HF methods (Becke, 1993; Zhao & Truhlar, 2008) combine the best of the both worlds and have turned out to be very successful (Bryantsev, Diallo, Van Duin, & Goddard, 2009). Also DFT has a parameterized alternative called density-functional tight-binding (DFTB). DFTB is designed to reproduce DFT results rather than fit empirical data like SE methods (Elstner, 2006). QM methods play an inevitable role in computational studies of enzyme-catalyzed reactions, since some amount of quantum-mechanical treatment or QM-based parameterization is always needed in the active site.

Even with the enormous growth of computational power in last decades, biological systems such as enzymes can hardly be treated in the full extend by QM methods. Successful attempts have been made to simulate whole proteins purely by QM methods (Cole & Hine, 2016; Todorović, Bowler, Gillan, & Miyazaki, 2013). Nevertheless, in order to make the calculations feasible, researches often reduce the QM model of the enzyme to the active site residues only, by constructing a so-called cluster model. Continuum solvent is used to mimic the protein as a low-dielectric environment beyond the shell of the active site residues. Cluster models of enzyme active sites have been used to elucidate the mechanisms of various enzymes, including difficult cases such as radical enzymes (Feliks & Ullmann, 2012) or metalloenzymes (Li & Ryde, 2014; Manta, Raushel, & Himo, 2014). Although a rather simple approach, due to the absence of various sources of errors that more complex methods may introduce, the cluster models retain their role in computational enzymology (Georgieva & Himo, 2010). Considering the influence of the protein environment, the continuum solvent might be replaced by the protein residues simulated on an empirical level, see QM/MM approaches later.

2.3 Empirical Molecular-Mechanical Models

Due to the enormous computational cost of the QM approaches, alternative empirical methods that use principles of classical physics have been

developed for treatment of complex systems. These methods are generally referred to as molecular mechanics (MM). In classical MM models, atoms are represented as spheres connected by springs. Empirical parameters have been derived to reproduce the expected behavior of biomolecules. These include spring constants for bond lengths, angles, and torsion angles as well as nonbonded interaction parameters such as van der Waals radii and partial atomic charges. The whole set of MM parameters is called force field (FF). The total potential energy of a system is calculated as a sum of all bonded and nonbonded contributions:

$$V(\vec{R}) = \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{torsions}} K_\varphi(1+\cos(n\varphi-\delta)) + \sum_{\substack{\text{improper}\\ \text{torsions}}} K_\omega(\omega-\omega_0)^2 + \sum_{\substack{\text{nonbonded}\\ \text{pairs}}} \left\{ \underbrace{\varepsilon_{ij}^{\min}\left[\left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}^{\min}}{r_{ij}}\right)^{6}\right]}_{\substack{\text{Lennard—Jones}\\ \text{potential}}} + \underbrace{\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r_{ij}}}_{\substack{\text{Coulomb}\\ \text{law}}} \right\} \quad (2)$$

where b is the bond length, θ is the bond angle, ϕ the torsion angle, ω the improper torsion angle. The 0 indices represent the equilibrium values of given parameters and K the respective spring constants. Van der Waals interactions are represented by the Lennard–Jones potential where $R_{ij}^{\min}$ is the distance energy minimum position, r_{ij} is the distance between two atoms and $\varepsilon_{ij}^{\min}$ is the interaction constant. Electrostatic interactions are described by the Coulomb law where q_i and q_j are the partial charges of the interacting atoms, r_{ij} is their distance and $\varepsilon_0\varepsilon$ is the environment permittivity.

Several empirical MM force fields for biomolecules have been developed over the years, e.g., CHARMM (MacKerell et al., 1998), AMBER (Cornell et al., 1996), GROMOS (Oostenbrink, Villa, Mark, & Van Gunsteren, 2004), or OPLS (Jorgensen, Maxwell, & Tirado-Rives, 1996). MM models proved to be very useful especially for studying protein dynamics by molecular dynamics (MD) simulations. In context of enzyme catalysis, pure MM models can be used to study conformational changes or substrate binding dynamics (Costa, Batista, Bisch, & Perahia, 2015;

Gilson & Zhou, 2007). When one wants to model details of a catalytic mechanism in the enzyme active site, however, the conventional MM models are insufficient. They are, however, often used to model the non-reactive parts of the enzyme in hybrid QM/MM approaches, as described below.

Nevertheless, attempts have been made to allow for chemical reactivity within the empirical force fields without employing QM principles. These approaches are referred to as reactive force fields. One of the most widely used reactive force field, that has been also utilized for biological systems, is called ReaxFF (Senftle et al., 2016; van Duin, Dasgupta, Lorant, & Goddard III, 2001). In ReaxFF, the empirical energy equation (Eq. 2) is modified to follow a more complex bond order formalism rather than the balls-on-springs formalism. Although reactive force fields have been originally developed for material chemistry applications (Liang et al., 2013), first attempts have been made to model peptides, small proteins (Golkaram, Shin, & van Duin, 2014; Monti et al., 2013), and DNA (Verlackt et al., 2015) using ReaxFF.

Empirical force fields are popular because of their speed and their often realistic description of molecules, which is a consequence of extensive parameterization to reproduce experimentally derived structures and behaviors. It has been shown that MM methods are able to reproduce experimentally derived molecular structures, except for nonstandard regions (Kulik, Luehr, Ufimtsev, & Martínez, 2012). In such regions, ab initio QM methods consistently perform better. Despite this benefit, as already mentioned above, full ab initio studies of proteins are restricted to rather small systems (Kulik et al., 2012). Instead, hybrid approaches (referred to as QM/MM) have been developed to combine accurate ab initio models and fast empirical models.

2.4 Hybrid QM/MM Models

In QM/MM, the active site, where the reaction takes place, is treated on QM level, while the rest of the enzyme is modeled by an empirical MM force field. Since its development in 1976 (Warshel & Levitt, 1976), QM/MM has become the most popular approach to study enzymatic reactions (Senn & Thiel, 2007, 2009).

The key feature that QM/MM methods have to deal with is the interaction between QM and MM regions. For the description of a QM/MM system, two energy schemes evolved, a subtractive scheme and an additive

scheme. The energy in the subtractive QM/MM scheme (Eq. 3) is obtained by an MM calculation of the entire system (*S*) with the inner part (*I*) cut out and replaced by a QM calculation.

$$E^{sub}_{\mathrm{QM/MM}} = E_{\mathrm{MM}}(S) + E_{\mathrm{QM}}(I) - E_{\mathrm{MM}}(I) \tag{3}$$

The QM energy of the quantum mechanically treated inner part ($E_{\mathrm{QM}}(I)$) is added to the MM energy of the entire system ($E_{\mathrm{MM}}(S)$), but the MM energy of the inner part ($E_{\mathrm{MM}}(I)$) has to be subtracted to avoid double counting. The most widely used approach that employs the subtractive scheme is ONIOM (Maseras & Morokuma, 1995; Svensson et al., 1996), which is capable of combining *n* layers of any implemented QM or MM approach. The potential drawback of the subtractive scheme is that a proper MM description of the active site region is required and this is often difficult to achieve for substrate or enzyme cofactors. In the additive QM/MM scheme (Field, 2007), the MM energy is only calculated on the outer system (*O*) instead of *S* (Eq. 4). The QM energy is added and additionally, a coupling term ($E_{\mathrm{QM/MM}}(I, O)$) is introduced, which treats the interaction of the QM and the MM part.

$$E^{add}_{\mathrm{QM/MM}} = E_{\mathrm{MM}}(O) + E_{\mathrm{QM}}(I) + E_{\mathrm{QM/MM}}(I, O) \tag{4}$$

The $E_{\mathrm{QM/MM}}(I, O)$ term itself is composed of van der Waals, electrostatic, and bonded contributions. Van der Waals interactions are fully described on the MM level. The common description with the Lennard–Jones potential is also applied to QM atoms. Therefore it is necessary to have suitable parameters for all QM atoms. Since the main impact of van der Waals interactions occurs in a short range, atoms near the boundary should be well parametrized. Thus, changes of QM atom properties during catalysis have no meaningful influence on $E_{\mathrm{QM/MM}}(I, O)$, as long as they appear not close to the boundary. In proteins, important catalytic contributions are mediated by electrostatic interactions. All charges and partial charges cause electrostatic forces. Within one model, the interactions of charges are well defined, but the coupling between two models needs to be adapted.

The electrostatic term within $E_{\mathrm{QM/MM}}(I, O)$ describes the coupling between the QM charge density and the MM charge model. One efficient but simple method to treat electrostatic interaction is mechanical embedding (Bakowies & Thiel, 1991), where the MM model is directly applied to the QM region. Since the charge density of the QM region is then represented

within the MM model, the MM region can react to it. A major drawback is, however, that the QM calculation is performed in the absence of the electrostatic MM environment, thus the atoms in the QM part cannot react to their full environment. But especially this electrostatic environment is crucial for enzymatic catalysis (Zhang, 2013). The electrostatic embedding approach (Bakowies & Thiel, 1991) treats this important interaction by providing the electrostatic environment for the QM calculation. The electrostatic environment appears as a one-electron term in the QM Hamiltonian. Thus the charge distribution in the QM region is polarized according to the MM charges. Despite the higher accuracy in the calculation, charge leakage effects can occur at the boundary of the QM region, where the QM charge density is polarized in immediate proximity by MM charges. In both embedding schemes the MM charges are rigid and do not react to the QM charge density. In a polarized embedding scheme the polarization happens in both, the QM region direction and the MM region direction. For the polarization of the MM region, a polarizable force field has to be applied on the MM level (Thompson & Schenter, 1995). This approach of polarized embedding is in general the most accurate, though computational demanding one.

In the majority of QM/MM simulations of enzymes, one has to deal with covalent bonds on QM/MM boundary (e.g., bond between an amino acid side chain and the protein backbone) in addition to the nonbonded interactions. The boundary QM atom valency needs to be saturated to allow for proper electronic structure calculation. The simplest approach to treat the QM/MM boundary is the link atom approach (Field, Bash, & Karplus, 1990; Singh & Kollman, 1986), where the QM atom is capped by an auxiliary atom (usually hydrogen) that is constraint in the direction of the QM/MM bond. This generates a problem of QM density overpolarization near the boundary that has to be treated. Alternative approaches are based on frozen hybrid orbitals (Amara, Field, Alhambra, & Gao, 2000; Théry, Rinaldi, Rivail, Maigret, & Ferenczy, 1994). A frontier atom is chosen in the QM/MM boundary and a set of suitably oriented localized orbitals is placed on it. This treatment allows to converge on more proper electronic structure in the boundary region. On the other hand frozen orbital approaches are more technically demanding and require calibration for the specific bond and QM method.

A conceptually different QM/MM approach that has been used to study enzyme catalysis is the empirical valence bond (EVB) (Kamerlin & Warshel, 2010, 2011; Warshel, 2003). EVB uses the valence-bond (VB) approach for quantum description of the enzyme active site. The QM methods discussed

so far are based on molecular orbital theory that combines the atomic orbitals into a molecular orbital wave-function. The VB methods instead describe the system as a linear combination of all possible states where electrons occupy localized orbitals. In EVB, the stationary points along the reaction path are described by an MM force field, while the transitions between them are treated by an SE valence-bond QM approach (Shurki, Derat, Barrozo, & Kamerlin, 2015). Thus for every reaction step simulated by EVB, a set of empirical parameters has to be derived. This is usually done on a model reaction in solution or in gas phase, where the parameters are fitted to reproduce the experimental data or ab initio QM results (Åqvist & Warshel, 1993). Once well calibrated, extensive conformational sampling along the reaction pathway can be achieved in reasonable time with EVB in order to get a proper free energy landscape (see Section 3.3.2). On the other hand, the major disadvantage of EVB in comparison to common QM and QM/MM methods is the need of specific calibration for every reaction step. In fact, prior knowledge of the reaction mechanism is required to perform an EVB simulation, while unknown mechanism alternatives can be discovered within conventional ab initio QM and QM/MM models.

2.5 Pseudo-Atomistic Models

Although the computational and algorithmic power is increasing, modeling of larger protein systems remains difficult. Bridging between molecular behavior and biological system function requires different levels of abstraction to manage the huge amount of data in a reasonable time. Pseudo-atomistic models provide one level of abstraction by introducing pseudo atoms, which comprise groups of atoms, several amino acids or even whole molecules. These pseudo atoms are represented as one entity with all atoms within such an entity considered to be frozen. They are modeled to simulate the essential or averaged behavior of such entities and their interactions. Pure pseudo-atomistic models contain merely homogeneous or heterogeneous pseudo atom species. Besides these pure models, multiscale models are developed, which combine atomistic resolution of some molecular parts with pseudo-atomistic simplification. Because of the simplification, to reduce multiple individual atoms to single entities, comparable to coarser grains, pseudo-atomistic models are referred to as coarse-grained (CG) models. The advantage of a CG model is, that the degrees of freedom are decreased. Therefore larger systems can be studied and longer time scales can be reached than using classical atomistic models. In the first application of such a

pseudo-atomistic model on proteins (Levitt & Warshel, 1975), it is pointed out that this simplification brings additional benefit in the interpretation of results. With the reduction in degrees of freedom, the energy landscape for the system is smoothened. That is, less important movements or details are averaged, and essential features can be focused. Since CG models allow the treatment of bigger biological systems, and can simplify the analysis, their development was increasing in the last years, as some recent reviews show (Ingólfsson et al., 2014; Kar & Feig, 2014; Kmiecik et al., 2016; Meier et al., 2013; Noid, 2013; Riniker, Allison, & van Gunsteren, 2012; Saunders & Voth, 2013).

The concept of a coarse-grained model is a reduction of degrees of freedom. This reduction is achieved by replacing several individual atoms by pseudo atoms. For an optimal simulation it is important that the coarse-grained system keeps the overall character of the all-atom system. Therefore it is necessary to provide rules for the coarse-grained particles to behave. Analogue to the all-atom MM force fields described above, the behavior is evaluated by an empirical potential energy function. Therefore a protein is represented as an elastic network of coarse-grained beads connected to each other by elastic springs. In a simple case, each bead represents one amino acid. Such a representation is useful to investigate domain motions in large systems, where longer time scales have to be achieved (Rücker, Wieninger, Ullmann, & Sticht, 2012). A harmonic bond potential is applied to each pair of beads, to allow them to move with respect to their surrounding. The potential energy function (Eq. 5) is a power series expansion near a minimum structure $\mathbf{r}_0$, represented as a $3N$ dimensional Cartesian coordinate vector.

$$V(\mathbf{r}) = V(0) + \nabla V(\mathbf{r}) + \frac{1}{2}\mathbf{r}^T\mathbf{H}\mathbf{r} \tag{5}$$

All movements of the system can be evaluated relative to that minimum. The constant term $V(0)$ describes the energy at the minimum position and can be set to zero. The first derivative of the potential $V(\mathbf{r})$ is the gradient, which is zero at a minimum. The elastic network potential simplifies to the second-order term, which is a sum of pairwise potentials, with the second derivative matrix $\mathbf{H}$ providing the force constants.

In recent years more and more complex models have been developed, which better and better represent the nature of biological systems (Tozzini, 2005). A further development in the treatment of big biological systems is the combination with more accurate methods, such as MM or

QM/MM. Such so-called multiscale or multiresolution models are suitable for the description of catalytic effects within a biological relevant surrounding. In a recent study (Sokkar, Boulanger, Thiel, & Sanchez-Garcia, 2015) chorismate mutase and p-hydroxybenzoate hydroxylase were investigated with a QM/MM/CG approach, by modeling the catalytically relevant part of the enzyme in QM, the remaining amino acids in MM and the surrounding solvent in CG. In such multiscale models it is important to define efficient data exchange between the different potentials. A proper exchange between the different resolutions becomes even more important in the adaptive resolution multiscale models (Heyden & Truhlar, 2008; Shen & Hu, 2014; Zavadlav, Melo, Marrink, & Praprotnik, 2015). Here, specific regions are defined, similarly to the QM/MM/CG approach, but with connecting buffer zones, which enable the entering particles to change their resolution. Such approaches incorporate all benefits of the multiscale models and the flexibility to adapt to major changes within the modeled biological system. A detailed list of CG models and available programs with an extensive description of the current state of the art can be found in a recent review (Kmiecik et al., 2016).

3. CALCULATING ENZYMATIC MECHANISMS

As discussed in the introduction, in computational enzymology one is interested in explaining macroscopic thermodynamic and kinetic data derived from experiment by microscopic models. In Section 2, we reviewed types of structural model environments to study the enzyme catalysis. In this section, we review how to derive thermodynamic and kinetic parameters within these models.

3.1 Thermodynamic Properties of Biological Systems

Each simulation of biological systems requires the analysis of its thermodynamic equilibrium states. At equilibrium, no driving forces act on the system. Thus, these states are thermodynamically stable. The relative stability of an initial state and a final state, for instance, allows to predict, if a reaction is endergonic or exergonic.

The direction in which a system changes can be evaluated by changes in free energy. Dependent on the simulation, the Gibbs free energy ΔG (Eq. 6) at constant temperature T and pressure P, or the Helmholtz free energy ΔA (Eq. 7) at constant temperature and volume V can be calculated. ΔA is comprised of the internal energy ΔU of the system, and temperature scaled

entropy ΔS, a measure for thermally accessible configurations. For ΔG the system energy is the enthalpy ΔH, which is comprised of internal energy ΔU of a system and the work for its volume expansion PV at adjusted pressure. Since experiments are mainly performed at constant pressure, we will hereafter refer to free energy as the Gibbs free energy ΔG.

$$\Delta G = \Delta H - T\Delta S = \Delta U + PV - T\Delta S \tag{6}$$

$$\Delta A = \Delta U - T\Delta S \tag{7}$$

The computational prediction of thermodynamic properties is based on the analysis of ensembles. An ensemble is a large number of virtual copies of a system with identical macroscopic properties. To obtain reasonable ensembles, sampling methods such as MD approaches or Monte Carlo (MC) approaches are applied (Paquet & Viktor, 2015). While MD approaches are usually based on Newtonian mechanics, MC methods are based on repeated random sampling using certain thermodynamics principles. MD simulations are used to investigate the time evolution of biological systems. Applying Newton's second law of motion (Eq. 8) the movements of atoms in a system are described with the classical equation of motion.

$$F_i = -\nabla_i E = m_i a_i \tag{8}$$

The force F_i acting on atom i with mass m_i determines its acceleration a_i, while the force is defined by the negative gradient of the potential energy function $\nabla_i E$ with respect to the coordinates of atom i. By integrating the equations of motion for all atoms over small time steps, it is possible to obtain their time dependent positions, just providing the initial positions and the initial velocities. The initial positions are directly given by the coordinates of the atoms, and the initial velocities are usually distributed randomly with a certain probability distribution. Only the acceleration is needed, which can be obtained by the gradient of the potential energy function and the atomic mass (Eq. 8). The velocities of the atoms define the temperature, which is an important thermodynamic property in MD simulations. An equilibration simulation is performed until the system reaches an equilibrium state, which is a global minimum on the energy surface. During this process it is important that the system has enough time and energy (i.e., temperature) to escape local minima. The system can in some cases anyway remain kinetically trapped in a local minimum. Annealing simulations start at higher temperature to overcome such energy barriers and gradually

decrease it to the desired temperature, allowing the system to reach an equilibrium state.

In contrast to MD simulations, statistical approaches, such as Monte Carlo sampling, are independent of force evaluation. The sampling is based on random movements of the system from a set of possible movements, while new configurations are accepted in the case of lower energy, $\Delta E < 0$. Configurations with higher energy are accepted according to the Metropolis criterion, which ensures importance sampling of the microstates by a Boltzmann factor $\exp\left({}^{-\Delta E}/_{k_B T}\right)$. Monte Carlo simulations create ensembles by energetically evaluating random movements of a system with the possibility to permit less favorable movements. In contrast to MD, Monte Carlo moves are not bound to small motions that can happen during the time integration step. Thus in Monte Carlo, it is often easier to overcome energy barriers, which prevents the system to be kinetically trapped. On the other hand in standard Metropolis Monte Carlo, dynamic properties of the system are not accessible, since the time dimension is not considered in Monte Carlo. If only sampling is considered, the same level of convergence is achieved faster by Monte Carlo simulations than with MD, provided efficient Monte Carlo moves are chosen (Jorgensen & Tirado-Rives, 1996). Monte Carlo sampling can also predict thermodynamic properties such as preferred protonation states or redox states under specific environmental conditions such as pH, solution redox potential, or membrane potential (Bombarda, Becker, & Ullmann, 2006; Calimet & Ullmann, 2004; Ullmann, 2000; Ullmann & Ullmann, 2012). These are equally important properties to define a state, which can be used for further kinetic analysis of biological systems.

3.2 Kinetic Properties of Biochemical Systems

Once thermodynamics of the stable conformations and protonation or oxidation states of the enzyme are identified, the interest of a computational enzymologist turns to kinetic properties. Enzyme kinetics aims at tracking the rates of chemical or physical processes in the enzyme. In computational enzymology, discrete reaction steps are first addressed within a structural model before the rate of overall reactions can be related to experimental measurements.

Reaction rate is characterized by a temperature-dependent rate constant, which can be viewed as a probability factor for overcoming a free energy

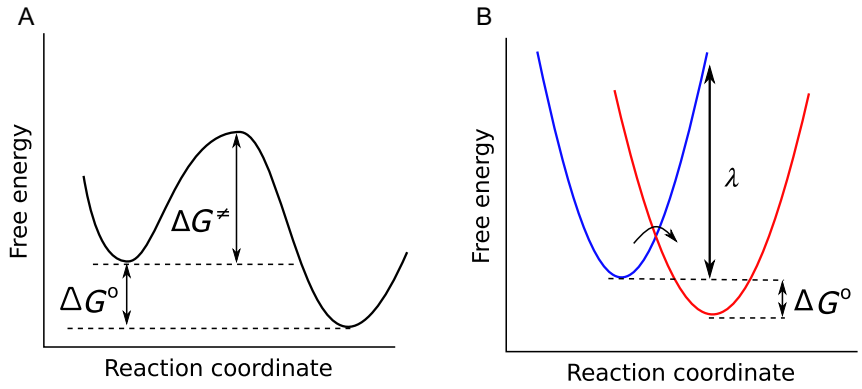


Fig. 3 (A) Reaction coordinate for interconversion between two stable states is thermodynamically characterized by reaction free energy ΔG° and kinetically by the free activation energy $\Delta G^{\neq}$. (B) Marcus model of two parabolas representing the two stable electronic configurations of the system. Reaction free energy ΔG° and the reorganization energy λ used for calculation of the free activation energy $\Delta G^{\neq}$ of the electron transfer are highlighted.

barrier between two stable states of the system (Fig. 3A). The reaction barrier can be related to the reaction rate by the Arrhenius law:

$$k(T) = A \quad exp\left(-\beta \Delta G^{\neq}\right) \tag{9}$$

where $\Delta G^{\neq}$ is the free activation energy, β is $^{1}/_{k_B T}$ (k_B is the Boltzmann constant, T is the absolute temperature), and A is a preexponential factor. Eq. (9) is further adopted to theoretically describe rates of different processes involved in the enzyme function.

The first category are long-range electron transfer reactions, which can be regarded as nonadiabatic processes. A Marcus model of two harmonic potentials representing an initial and a final electronic configuration (Fig. 3B) can be used to describe the reaction coordinate. The free activation energy can be calculated from reaction free energy ΔG° and reorganization energy λ, which is the energy needed to perform the structural changes within the system:

$$\Delta G^{\neq} = \frac{(\Delta G^{\circ} + \lambda)^2}{4\lambda} \tag{10}$$

The reaction rate constant (9) can then be adopted for electron transfers using Marcus theory for the activation energy and Fermi's golden rule (Marcus & Sutin, 1985) for the preexponential factor:

$$k_{ET}(T) = \frac{2\pi}{\hbar} H_{DA}^2 \frac{1}{\sqrt{4\pi\beta^{-1}\lambda}} \exp\left(-\beta \frac{(\Delta G^{\circ} + \lambda)^2}{4\lambda}\right) \tag{11}$$

where $\hbar$ is the reduced Planck constant and H_{DA}^2 is the electronic coupling between the reactant state and the product state, which, to a good approximation, decreases exponentially with the distance between donor and acceptor (Gamow, 1928; Gray & Winkler, 2005). Marcus theory for instance in the form of the Moser–Dutton ruler (Moser, Keske, Warncke, Farid, & Dutton, 1992; Page, Moser, Chen, & Dutton, 1999) can be combined with a microstate model used for calculating microscopic redox potentials (Ullmann, 2000), and an electrostatic model for calculating reorganization energies (Sharp, 1998) in order to calculate charge transfer energies in complex systems. This approach was successfully applied within continuum-electrostatics models to describe electron transfer reactions in complex systems such as photosynthetic center (Becker, Ullmann, & Ullmann, 2007; Bombarda & Ullmann, 2011) which was recently reviewed (Ullmann, Mueller, & Bombarda, 2016).

The second category, which will be discussed more extensively here, is chemical reactions occurring in the enzyme active site that can be described by transition state theory (TST). TST assumes that the free activation energy $\Delta G^{\neq}$ can be derived from quasi-equilibrium between the reactant and transition states. Thus, intermediate structures along the reaction coordinate (Fig. 3A) including the transition state structure are explicitly analyzed. The reaction rate constant (Eq. 9) can be expressed by Eyring–Polanyi equation:

$$k(T) = \zeta \beta^{-1} h^{-1} exp\left(-\beta \Delta G^{\neq}\right) \tag{12}$$

where h is the Planck constant and ζ in the transmission coefficient which corresponds to the probability of being reactive once the transition state is reached. Even though the transmission coefficient has to be considered in general (i.e., for quantum tunneling effects), the lowering of $\Delta G^{\neq}$ is a dominant effect of enzymes in catalysis (Gao et al., 2006) and majority of computational enzymology studies focus on it. To get the rate-limiting energy barrier of a chemical process, one has to investigate the reaction path between substrate and product state. On the basis of the models described in the previous section, potential energy of any conformation in a molecular structure (here enzyme–substrate complex) can be calculated. When searching for the reaction path, the $3N$-dimensional energy function (with N the number of atoms of the system) is portrayed as a potential energy surface (PES). A PES can be viewed as a landscape with valleys and mountain passes, which correspond to stable states and easiest transition pathways between them. A reaction coordinate of an elementary step follows the path

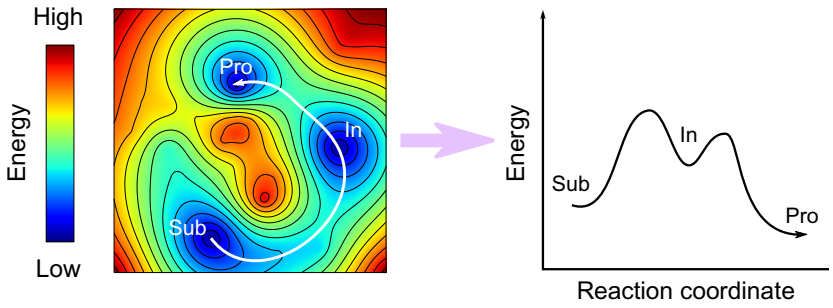


Fig. 4 Reaction path can be imagined as a trail through an energy landscape where basins represent the stable states and the passes represent the transition states. The path from substrate (*Sub*) to product (*Pro*) usually involves several intermediate (*In*) states. The energy change plotted against the position on the path is referred to as reaction profile.

from a minimum along the shallowest ascent toward the transition state and from there along the deepest descent toward the product state. Most of the chemical reactions are composed of several elementary steps forming together the minimum energy path (MEP) from reactant to product (Fig. 4).

Many methods have been developed to identify reaction paths and to distinguish among mechanism variants based on corresponding energy profiles. In general, one has to first identify the mechanistic options and then get corresponding accurate energy profiles. At the beginning of a computational kinetic study, optimization techniques are usually used to find minima and saddle points on the PES, which together define MEPs. The obtained energy profile corresponds to the enthalpic part of the free energy (Eq. 6). If the enthalpic differences among mechanism variants are big, one can already make qualitative conclusions about the preferred mechanism. The entropic part of Eq. (6) has to be considered to obtain proper reaction free energy profile at finite temperature in general, although the effect on the reaction barrier height is not always that profound (Kazemi, Himo, & Åqvist, 2016). The entropy-corrected PES is referred to as free energy surface at certain temperature. To get a quantitative free energy surface, one needs to perform extensive sampling along the reaction path, which can be extremely costly or even prohibitive in complex structural models. Therefore, the sampling is usually performed only when the most probable mechanism is identified on PES.

In the next part, we first discuss methods to identify a reaction path on PES. We then turn to the sampling methods that aim to get full free energy reaction profiles. In the end, we return to the preexponential transmission coefficient of Eq. (12) and discuss the role of nuclear quantum effects.

3.3 Mechanic and Energetic Properties of Molecular Reaction Paths

Computational investigation of an enzyme mechanism begins with a structural mechanistic model of the enzyme. As described earlier, stable experimentally determined structures corresponding to minima on the PES are commonly used as starting states. Approaches for searching reaction paths can be categorized as single ended or double ended. Double-ended approaches require the knowledge of two states, usually the initial state and the final state, to search for mechanistic possibilities in between. In most cases, however, one first explores the PES with a single-ended method and then loads the first path estimate into a double-ended method for further refinement.

3.3.1 Methods to Explore and Investigate Mechanistic Possibilities

3.3.1.1 Single-Ended Reaction Path Search Methods

The simplest single-ended approach is the adiabatic mapping of the PES. Based on chemical intuition and experimental data, a set of movements that lead toward a desired final state is applied. A function characterizing a desired movement is called collective variable (CV). For instance, in a cysteine protease, the catalytic cysteine has to be deprotonated by a neighboring histidine before it can attack the peptide bond of the substrate. Therefore, a first collective variable to be investigated is the distance between the thiol hydrogen of the cysteine and the closest nitrogen of the histidine. During adiabatic mapping along such a CV, the distance is shrunk in discrete steps. The enzyme structure is optimized at each step, while the CV is biased at given position by a harmonic potential. The mapping finishes by reaching the desired product state and the energies of the discrete scanning steps represent a rough potential energy reaction profile. The energy maxima and minima along this reaction profile should be further optimized without bias. Minima can be optimized by standard gradient-based energy minimization routines (Leach, 2001). Maxima can be refined to first-order saddle points using the eigenvector following method (Cerjan, 1981) that requires calculation of the Hessian matrix of second derivatives. The largest negative eigenvector of the Hessian points in direction of the reaction path and thus can be followed uphill to reach the saddle point. Since the saddle point is hard to reach in one step, the Hessian has to be updated. Although various approximate techniques exist (Munro & Wales, 1999; Wales & Walsh, 1998), the usage of purely single-ended eigenvector following for big systems with many degrees of freedom may be both unreliable and computationally unfeasible.

Alternatively, the single-ended growing string method (Zimmerman & Bowman, 2016) can be used to locate the nearest saddle point and next stable intermediate state. This method, like double-ended methods discussed further below, generates a chain of structures to sample the reaction coordinate. Unlike its double-ended sister, which uses both initial and final state structures for the chain generation, the single-ended growing string method uses a set of collective variables to indicate the initial direction from the initial state. Once, a saddle point estimate is located, the eigenvector following method is used to refine it. Interestingly, the single-ended growing string method can suggest the CVs automatically and thus explore various reaction possibilities from a given initial state.

In contrast to the above described PES mapping approaches, the reaction mechanism can be also explored at finite temperature without knowledge of the final state structure. In principle, sufficiently long MD simulations of the initial state at physiological temperature should lead to a reaction. Current computer power is, however, not sufficient to perform such a long simulation, especially for models with quantum essence that are required in most cases. Fortunately, MD simulations can be driven along predefined CVs in a similar way as in adiabatic PES mapping. One popular technique is called umbrella sampling (US) (Torrie & Valleau, 1977). A harmonic (umbrella) bias potential is applied in order to restrain the MD simulation around a certain value of the CV. The value of the CV can be changed in discrete steps, while a sufficiently long biased MD simulation is performed for each value till the desired final state is reached. Another elegant method how to sample reaction possibilities at finite temperature in accessible time is called metadynamics (Laio & Parrinello, 2002). In this method, Gaussian bias potential "hills" are added to the PES in certain time intervals along the values of a CV that has been already visited in the MD trajectory. In this fashion, the MD is discouraged to revisit already sampled CV values, and thus the simulation proceeds toward the desired final state. The bias introduced by metadynamics brings bigger flexibility compared to the US bias on one hand, but potentially slower convergence on the other.

Although the single-ended finite temperature methods provide reasonable conformational freedom and entropic aspects in one simulation, their use with higher QM models can be computationally unfeasible. Furthermore, while the choice of collective variables in stiff adiabatic PES mapping is usually trivial, in more flexible bias MD methods it can be rather elusive. Therefore, double-ended methods reviewed in the next paragraphs are often

preferred. The entropic aspects can also be accessed a posteriori once the PES path is established as we shall see below.

3.3.1.2 Double-Ended Reaction Path Search Methods

Once the final state structure is discovered by a single-ended method or when it is known from an experimental structure or guessed by chemical intuition, so-called double-ended methods can be employed to construct a path between them. Alternatively, an estimated path derived, e.g., from adiabatic PES mapping described above can be loaded in and refined. The major double-ended method category that will be described here comprises of chain-of-states (COS) methods. A COS method discretizes the reaction path between the initial and final state into a set of intermediate structures and optimizes them in a connected manner. Once this procedure achieves convergence, the resulting path should represent the minimal energy pathway (MEP) between the initial and final state.

One popular COS approach is nudged elastic band (NEB) (Jonsson, Mills, & Jacobsen, 1998). NEB approximates the reaction path by a set of structures that are connected by springs into a chain. The parallel component of the spring force keeps the images distributed along the whole path, while the perpendicular component helps to push images to the MEP valley. Because NEB does not directly seek for the saddle point structures, a climbing image variant of NEB (CI-NEB) (Henkelman, Uberuaga, & Jónsson, 2000) has been developed. The converged NEB path is usually loaded into CI-NEB. The parallel component of the NEB spring potential acting on the highest image of the path is inversed to drive it uphill, while the perpendicular component is kept unchanged in order to keep the image on the path. Once the path is converged again, the highest image should represent the saddle point structure.

The string method (SM) (E, Ren, & Vanden-Eijnden, 2002) uses a concept on equal image distribution to cover the path instead of introducing spring constants between the images. The images are equally redistributed along a spline fit of the reaction path every optimization cycle. The path is converged once the redistribution does not significantly change the image position. An SM variant called growing string method (Peters, Heyden, Bell, & Chakraborty, 2004) construct the chain of states gradually from the initial state toward the final state and thus reduces the bias potentially introduced by initial path guess (e.g., linear interpolation). Another SM variant optimizes the COS at finite temperature, allowing for a better conformational relaxation of the MEP (Vanden-Eijnden & Venturoli, 2009;

E, Ren, & Vanden-Eijnden, 2005). Note that the single-ended string method described above is a variant of the growing string method with even less initial bias.

The above described COS methods do not a priori search for the transition state structures. Although CI-NEB makes a step in this direction, it is not actually guaranteed that it will reach a first-order saddle point. One can, of course, employ the above described eigenvector following approaches to refine the maxima of a COS path to saddle points. However, these methods are both computationally costly and also unreliable once the input structure is not close to the actual saddle point. An elegant solution is the conjugate peak refinement (CPR) (Fischer & Karplus, 1992; Gisdon, Culka, & Ullmann, 2016) method that gradually constructs chain of states between the initial and final states while it aims at locating the first-order saddle points. In contrast to NEB or SM, the number of states along the path is not fixed and thus the sampling in the saddle point region can be increased to facilitate its proper location. The CPR method is based on the fact that in the vicinity of a saddle point, there is one direction, which points to an energetic maximum, while all others lead to a minimum. The CPR algorithm picks the highest energy structure along the discretized path and performs a line maximization along the corresponding tangential path vector. This corrected maximum is then minimized in conjugate space similar to the conjugated gradient minimization method. By staying conjugate to the original path, falling to the neighboring minimum is prevented. The optimized structure is added into the chain of states. Gradually the saddle point is approached by providing more sampling in the transition region and occasionally the conjugate optimization procedure can converge to locate the first-order saddle point. A successful run ends, when all maxima along the reaction path between the initial and final state are identified and optimized to first-order saddle points.

3.3.1.3 Path-Based Reaction Path Search Methods

Once a reaction path is found, it is sometimes necessary to correct it for conformational flexibility, since a biological system has many possible degrees of freedom. Path-based methods include the conformational flexibility to evaluate alternative reaction paths or a set of possible reactive coordinates. These methods usually require an initial path, which does not necessarily have to be properly refined. Any path generated by double-ended or single-ended methods can be used. The concept of metadynamics, introduced as a single-ended search method, can also be applied for

path-based approaches. Instead of using a trivial CV, such as bond distances or dihedral angles, a function that describes the whole reaction path is used (Branduardi, Gervasio, & Parrinello, 2007; Bernardi, Melo, & Schulten, 2015). The application of the path CV allows to dynamically sample the conformational space based on the input path estimate. This provides information on the conformational flexibility of the input intermediates and transition states, which is often crucial for reaction paths obtained from static approaches, such as adiabatic mapping, NEB, or CPR. If the distance from the initial path is taken as a second CV to be biased, even alternative paths can be identified and energetically evaluated.

A conceptually different approach is followed in transition path sampling (TPS) (Bolhuis, Dellago, & Chandler, 1998; Dellago, Bolhuis, Csajka, & Chandler, 1998). The aim of TPS is to connect two stable states by a collection of all likely transition pathways, which represent the transition path ensemble. As mentioned above, an input trajectory does not necessarily have to be properly refined, thus it may have a low weight in the transition path ensemble. Therefore, for a TPS simulation the initial trajectory has to be evaluated and equilibrated toward a representative transition path trajectory. Unlike US ensembles or metadynamics ensembles, a TPS simulation creates unbiased trajectories, since it does not enhance the sampling of rare events by bias potentials (Swenson & Bolhuis, 2014). TPS, however, is also dependent on a collective variable, yet it is not used to drive the simulation, but rather to discriminate the stable states of the system, and to monitor the progress along the trajectory. This makes the CV a crucial quantity in a TPS simulation. The enhancement of sampling rare events in TPS is achieved by importance sampling of trajectory space. All dynamic movements are preformed by a Monte Carlo approach, while a new trajectory is created from an existing one in the ensemble. For that, specific procedures are available to drive movements in trajectory space (Bolhuis, Chandler, Dellago, & Geissler, 2002; Rowley & Woo, 2009). TPS has been successfully applied to simulate biological systems (Dellago & Bolhuis, 2007) and is also recently used to study enzyme catalysis, such as the hydride transfer in a dihydrofolate reductase (Wang, Antoniou, Schwartz, & Schramm, 2016).

3.3.2 Methods to Obtain Free Energy

So far, we have been concentrating on the qualitative aspects of the reaction mechanism. Now we turn to the methods to get the free energy profile estimates. As noted earlier (Eq. 6), the free energy is composed of the potential energy (enthalpic) part that is calculated directly by the model, and the

entropic part, for which extensive sampling at finite temperature along the reaction path is needed. In case of US and metadynamics the sampling is already included in the path search procedure, so entropic influence is explicitly included and thus free energy profile can be directly estimated. For US, weighted histogram analysis method (WHAM) (Kumar, Rosenberg, Bouzida, Swendsen, & Kollman, 1992) can be used to remove the umbrella bias and to integrate the simulation windows into a free energy profile. In case of metadynamics, the sum of added Gaussian hills plotted against the collective variable directly represents the free energy profile. TPS collected pathways are dynamic trajectories, thus kinetic information, such as rate constants, can be extracted (Dellago, 2007). But since only trajectories are considered that connect certain regions in configuration space, the configurations are not distributed according to the equilibrium distribution of the system. To determine free energy profiles, one has to obtain equilibrium-distributed configurations. One possibility is to apply biasing procedures, such as US variants, to rarely visited states, and divide the CV into overlapping windows (Dellago, 2007).

In case of static methods such as adiabatic PES mapping or chain-of-states methods, the entropic aspect has to be added by sampling method (MD or Monte Carlo) to account for entropic part of the free energy profile. One option is to use these static paths as an input to above described path-based approaches. A possible obstacle can be that achieving extensive sampling when high-level QM methods are used to treat the active site may be computationally prohibitive. In the same time, usage of, e.g., SE QM models can introduce additional bias into the results. A way out of this dilemma offers the free energy perturbation (FEP) methods. For instance, one can find the reaction path by a chain-of-states method within QM/MM model (Kästner, Senn, Thiel, Otte, & Thiel, 2006). Subsequently, an MD simulation is performed for every state of the PES reaction path with QM region kept frozen. Perturbation energy is calculated as energy for moving one step forward in the PES QM reaction path while staying in the same MM conformational ensemble:

$$\Delta E_{pert} = E_{\mathrm{QM/MM}}(\mathbf{r}^{i+1}_{\mathrm{QM}}, \mathbf{r}^{i}_{\mathrm{MM}}) - E_{\mathrm{QM/MM}}(\mathbf{r}^{i}_{\mathrm{QM}}, \mathbf{r}^{i}_{\mathrm{MM}}) \tag{13}$$

where i and $i + 1$ are the indices of adjacent path steps and $\mathbf{r}$ are the coordinate vectors. Free energy for the $i \rightarrow i+1$ step is then calculated by averaging over the MM ensemble at step i using Zwanzig equation (Zwanzig, 1954):

$$\Delta G^{i \to i+1} = -\frac{1}{\beta} \ln \left\langle \exp\left(-\beta \Delta E_{pert}\right) \right\rangle_i \tag{14}$$

The reaction free energy profile is calculated by adding QM energy corrected for zero-temperature vibrations to the perturbation-derived free energy. Another example where FEP concept is exploited are the FEP/US studies in EVB model (Warshel, 1991).

3.3.3 Methods to Simulate Nuclear Quantum Effects

The majority of methods commonly used to locate and sample the enzyme reaction rely on atomistic models that treat the nuclei as classical objects. Even in the common QM methods, just the electrons are treated as quantum objects. This approximation is necessary to make the calculations on bigger systems computationally feasible and in many cases also justified, since the heavy nuclei hardly show any quantum behavior. The quantum nature becomes in practice significant in case of proton transfer processes, which are often part of enzyme-catalyzed reaction coordinates. If a proton transfer is the rate-limiting step, the nuclear quantum-mechanical effects (NQM) play important role in the overall reaction rate. NQM effect are most pronounced when comparing protium and deuterium variants and calculating kinetic isotope effects.

Although small number of protons can be treated as quantum particles using nuclear-electronic orbital (Webb, Iordanov, & Hammes-Schiffer, 2002) methods, most of the approaches that deal with NQM in enzymes can be classified as correction methods that are applied, e.g., on US path ensembles. One direction is the ensemble-averaged variational transition state theory with multidimensional tunneling (EA-VTST/MT) (Garcia-Viloca, Alhambra, Truhlar, & Gao, 2001; Truhlar et al., 2002) that corrects the free energy barrier $\Delta G^{\neq}$ for quantum-mechanical vibrations. EA-VTST/MT in addition also calculates the transmission coefficient ζ (see Eq. 12) that corrects mainly for quantum-mechanical tunneling through the free energy barrier. Another family of approaches to NQM is based on Feynman path integrals. Quantum nature of the nuclei is approximated by transforming the classical spheres into rings of quasiparticles connected by springs. Path integral MD simulation techniques include quantized classical path (QCP) (Hwang & Warshel, 1993), centroid molecular dynamics (CMD) (Cao & Voth, 1994a, 1994b), or ring-polymer molecular dynamics (RPMD) (Braams & Manolopoulos, 2006; Craig & Manolopoulos, 2004).

3.4 Path Search Strategy

Taken together, enormous amount of methods to locate reaction path and calculate its energy profile has been developed. It is not easy to objectively find the one and only correct strategy, although different groups certainly have their preferred approaches. The choice of the method is also influenced by the primary question that the researcher is asking. If the task is to find the correct catalytic mechanism, many potential variants of the reaction mechanism need to be tested in a reasonable time. If the individual steps of the mechanism are already known, the task might be to get proper rate constant to relate the model to experimental parameters. In many cases, big energy barriers on PES can rule the unfeasible mechanism variants out and find the most promising set of reaction steps in fraction of time in more accurate QM models compared to direct usage of sampling approaches in approximate models. The PES path can be further optimized and corrected by much more demanding sampling methods, or a reaction-specific SE potential (e.g., EVB) can be constructed based on the previous PES investigation in high-level QM models.

The kinetic and thermodynamic parameters determined by the aforedescribed methods should be combined in order to get a more reliable picture. The different methods should not be viewed as competitive approaches but rather as complementary and one should search for synergy among different methods with limitations of the models in mind. The limitations and synergy should also be considered when comparing computational results with experimental data.

4. GOING BEYOND THE EXPLORATION OF THE REACTION PATHS

In order to understand the biological systems, it is not enough to understand the mechanism of an enzyme. It is required to analyze the enzyme in its physiological context and how the rate of catalysis is influenced by parameters such as pH or metabolite concentration. Today's systems biology is using kinetic models with stretched exponential or noninteger stoichiometry in order to describe complex metabolic networks. Even if such models are of certain practical use to solve some research problems, they are not satisfactory from a theoretical point of view, since mass and energy conservation is not guaranteed. Other models are taking kinetic parameters from databases. However, such parameters were usually determined under

specific circumstances and cannot account for all possible effects. Considering the constantly increasing number of complete genomes and partially reconstructed metabolisms, it comes more and more important to get a more realistic view of the metabolic reaction in its context. The challenge for computational biochemistry today and in the future is to derive enzymatic parameters from structure models by using methods that we reviewed in this article. However, it will be required to go beyond such information. The kinetic parameters can be combined in master equation approaches (Becker et al., 2007; Bombarda & Ullmann, 2011) or kinetic Monte Carlo simulations (Till, Becker, Essigke, & Ullmann, 2008) in order to simulate complete catalytic cycles that are influenced by environmental parameters such as pH or membrane potential. To carry the approach further, it might be possible to model the whole cellular context in reaction–diffusion equations, which may allow in the future to model complex biochemical reactions. Combining structural biology and systems biology may thus be a promising direction, especially considering the pace in which both fields make progress in recent years. With the help of computer models that rely on a solid experimental basis, we may more and more understand how the jigglings and wigglings of atoms leads to the complex phenomenon we call *life*.

REFERENCES

Amara, P., Field, M. J., Alhambra, C., & Gao, J. (2000). The generalized hybrid orbital method for combined quantum mechanical/molecular mechanical calculations: Formulation and tests of the analytical derivatives. *Theoretical Chemistry Accounts*, *104*, 336–343.

Åqvist, J., & Warshel, A. (1993). Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches. *Chemical Reviews*, *93*, 2523–2544.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, *98*, 10037–10041.

Bakowies, D., & Thiel, W. (1991). MNDO study of large carbon clusters. *Journal of the American Chemical Society*, *113*, 3704–3714.

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L. S., & Subramaniam, S. (2015). 2.2 Å resolution cryo-EM structure of Î²-galactosidase in complex with a cell-permeant inhibitor. *Science*, *348*, 1147–1151.

Becke, A. (1993). Density functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, *98*, 5648–5652.

Becker, T., Ullmann, R. T., & Ullmann, G. M. (2007). Simulation of the electron transfer between the tetraheme subunit and the special pair of the photosynthetic reaction center using a microstate description. *The Journal of Physical Chemistry. B*, *111*, 2957–2968.

Bernardi, R. C., Melo, M. C. R., & Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta*, *1850*, 872–877.

Bolhuis, P. G., Chandler, D., Dellago, C., & Geissler, P. L. (2002). Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry*, *53*, 291–318.

Bolhuis, P. G., Dellago, C., & Chandler, D. (1998). Sampling ensembles of deterministic transition pathways. *Faraday Discussions*, *110*, 421–436.

Bombarda, E., Becker, T., & Ullmann, G. M. (2006). The influence of the membrane potential on the protonation of bacteriorhodopsin: Insights from electrostatic calculations into the regulation of proton pumping. *Journal of the American Chemical Society*, *128*, 12129–12139.

Bombarda, E., & Ullmann, G. M. (2011). Continuum electrostatic investigations of charge transfer processes in biological molecules using a microstate description. *Faraday Discussions*, *148*, 173–193.

Braams, B. J., & Manolopoulos, D. E. (2006). On the short-time limit of ring polymer molecular dynamics. *The Journal of Chemical Physics*, *125*, 124105.

Branduardi, D., Gervasio, F. L., & Parrinello, M. (2007). From A to B in free energy space. *The Journal of Chemical Physics, 126*, 054103. http://dx.doi.org/10.1063/1.2432340.

Bryantsev, V. S., Diallo, M. S., Van Duin, A. C. T., & Goddard, W. A. (2009). Evaluation of B3LYP, X3LYP, and M06-class density functionals for predicting the binding energies of neutral, protonated, and deprotonated water clusters. *Journal of Chemical Theory and Computation*, *5*, 1016–1026.

Calimet, N., & Ullmann, G. M. (2004). The influence of a transmembrane pH gradient on protonation probabilities of bacteriorhodopsin: The structural basis of the back-pressure effect. *Journal of Molecular Biology*, *339*, 571–589.

Cao, J., & Voth, G. A. (1994a). The formulation of quantum statistical mechanics based on the Feynman path centroid density. I. Equilibrium properties. *The Journal of Chemical Physics*, *100*, 5093–5105.

Cao, J., & Voth, G. A. (1994b). The formulation of quantum statistical mechanics based on the Feynman path centroid density. II. Dynamical properties. *The Journal of Chemical Physics*, *100*, 5106–5117.

Cerjan, C. J. (1981). On finding transition states. *The Journal of Chemical Physics*, *75*, 2800.

Chen, J. L., Noodleman, L., Case, D., & Bashford, D. (1994). Incorporating solvation effects into density functional electronic structure calculations. *The Journal of Physical Chemistry*, *98*, 11059–11068.

Čížek, J. (1966). On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods. *The Journal of Chemical Physics*, *45*, 4256–4266.

Cole, D. J., & Hine, N. D. M. (2016). Applications of large-scale density functional theory in biology. *Journal of Physics. Condensed Matter*, *28*, 393001.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., … Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, *118*, 2309–2309.

Costa, M. G. S., Batista, P. R., Bisch, P. M., & Perahia, D. (2015). Exploring free energy landscapes of large conformational changes: Molecular dynamics with excited normal modes. *Journal of Chemical Theory and Computation*, *11*, 2755–2767.

Craig, I. R., & Manolopoulos, D. E. (2004). Quantum statistics and classical mechanics: Real time correlation functions from ring polymer molecular dynamics. *The Journal of Chemical Physics*, *121*, 3368–3373.

Cramer, C. J. (2004). *Essentials of computational chemistry: Theories and models* (2nd ed.). Chichester: Wiley.

Dellago, C. (2007). Transition path sampling and the calculation of free energies. In C. Chipot & A. Pohorille (Eds.), *Free energy calculations* (pp. 249–276). Berlin, Heidelberg: Springer.

Dellago, C., & Bolhuis, P. G. (2007). Transition path sampling simulations of biological systems. In M. Reiher (Ed.), *Atomistic approaches in modern biology* (Vol. 268, pp. 291–317). Berlin, Heidelberg: Springer.

Dellago, C., Bolhuis, P. G., Csajka, F. S., & Chandler, D. (1998). Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics*, *108*, 1964.

E, W., Ren, W., & Vanden-Eijnden, E. (2002). String method for the study of rare events. *Physical Review B*, *66*, 052301.

E, W., Ren, W., & Vanden-Eijnden, E. (2005). Finite temperature string method for the study of rare events. *The Journal of Physical Chemistry B*, *109*, 6688–6693.

Elstner, M. (2006). The SCC-DFTB method and its application to biological systems. *Theoretical Chemistry Accounts*, *116*, 316–325.

Feliks, M., & Ullmann, G. M. (2012). Glycerol dehydratation by the B12-independent enzyme may not involve the migration of a hydroxyl group: A computational study. *The Journal of Physical Chemistry. B*, *116*, 7076–7087.

Feynman, R. (1964). *The Feynman lectures on physics*. New York, USA: Basic Books. Retrieved from http://www.feynmanlectures.caltech.edu/.

Field, M. J. (2007). *A practical introduction to the simulation of molecular systems* (2nd ed.). Cambridge, UK: Cambridge University Press.

Field, M. J., Bash, P. A., & Karplus, M. (1990). A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry*, *11*, 700–733.

Fischer, S., & Karplus, M. (1992). Conjugate peak refinement: An algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chemical Physics Letters*, *194*, 252–261.

Gamow, G. (1928). Zur quantentheorie des atomkernes. *Zeitschrift fr Physiotherapie*, *51*, 204–212.

Gao, J., Ma, S., Major, D. T., Nam, K., Pu, J., & Truhlar, D. G. (2006). Mechanisms and free energies of enzymatic reactions. *Chemical Reviews*, *106*, 3188–3209.

Garcia-Viloca, M., Alhambra, C., Truhlar, D. G., & Gao, J. (2001). Inclusion of quantum-mechanical vibrational energy in reactive potentials of mean force. *The Journal of Chemical Physics*, *114*, 9953–9958.

Georgieva, P., & Himo, F. (2010). Quantum chemical modeling of enzymatic reactions: The case of histone lysine methyltransferase. *Journal of Computational Chemistry*, *31*, 1707–1714.

Gilson, M. K., & Zhou, H.-X. (2007). Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, *36*, 21–42.

Gisdon, F. J., Culka, M., & Ullmann, G. M. (2016). PyCPR—A python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures. *Journal of Molecular Modeling*, *22*, 242.

Golkaram, M., Shin, Y. K., & van Duin, A. C. T. (2014). Reactive molecular dynamics study of the pH-dependent dynamic structure of α-helix. *The Journal of Physical Chemistry. B*, *118*, 13498–13504.

Gray, H. B., & Winkler, J. R. (2005). Long-range electron transfer. *Proceedings of the National Academy of Sciences*, *102*, 3534–3539.

Head-Gordon, M., Pople, J. A., & Frisch, M. J. (1988). MP2 energy evaluation by direct methods. *Chemical Physics Letters*, *153*, 503–506.

Henkelman, G., Uberuaga, B. P., & Jónsson, H. (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, *113*, 9901.

Heyden, A., & Truhlar, D. G. (2008). Conservative algorithm for an adaptive change of resolution in mixed atomistic/coarse-grained multiscale simulations. *Journal of Chemical Theory and Computation*, *4*, 217–221.

Honig, B., & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, *268*, 1144–1149.

Hwang, J. K., & Warshel, A. (1993). A quantized classical path approach for calculations of quantum mechanical rate constants. *The Journal of Physical Chemistry*, *97*, 10053–10058.

Ingólfsson, H. I., Lopez, C. A., Uusitalo, J. J., de Jong, D. H., Gopal, S. M., Periole, X., & Marrink, S. J. (2014). The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*, 225–248.

Jonsson, H., Mills, G., & Jacobsen, K. W. (1998). Nudged elastic band method for finding minimum energy paths of transitions. In B. J. Berne, G. Ciccotti, & D. F. Coker (Eds.), *Classical and quantum dynamics in condensed phase simulations* (pp. 385–404). Singapore: World Scientific.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, *118*, 11225–11236.

Jorgensen, W. L., & Tirado-Rives, J. (1996). Monte Carlo vs molecular dynamics for conformational sampling. *The Journal of Physical Chemistry*, *100*, 14508–14513.

Kamerlin, S. C. L., & Warshel, A. (2010). The EVB as a quantitative tool for formulating simulations and analyzing biological and chemical reactions. *Faraday Discussions*, *145*, 71–106.

Kamerlin, S. C. L., & Warshel, A. (2011). The empirical valence bond model: Theory and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *1*, 30–45.

Kar, P., & Feig, M. (2014). Recent advances in transferable coarse-grained modeling of proteins. *Advances in Protein Chemistry and Structural Biology*, *96*, 143–180.

Kästner, J., Senn, H. M., Thiel, S., Otte, N., & Thiel, W. (2006). QM/MM free-energy perturbation compared to thermodynamic integration and umbrella sampling: Application to an enzymatic reaction. *Journal of Chemical Theory and Computation*, *2*, 452–461.

Kazemi, M., Himo, F., & Åqvist, J. (2016). Enzyme catalysis by entropy without Circe effect. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 2406–2411.

Klamt, A., & Schüürmann, G. (1993). COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2* (799–805).

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical Reviews*, *116*, 7898–7936.

Kühlbrandt, W. (2014). Cryo-EM enters a new era. *eLife*, *3*, e03678.

Kulik, H. J., Luehr, N., Ufimtsev, I. S., & Martínez, T. J. (2012). Ab initio quantum chemistry for protein structures. *The Journal of Physical Chemistry. B*, *116*, 12501–12509.

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

Laio, A., & Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 12562–12566.

Leach, A. R. (2001). *Molecular modelling: Principles and applications* (2nd ed.). New Jersey: Prentice Hall.

Levitt, M., & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, *253*, 694–698.

Li, J., Nelson, M. R., Peng, C. Y., Bashford, D., & Noodleman, L. (1998). Incorporating protein environments in density functional theory: A self-consistent reaction field calculation of redox potentials of [2Fe2S] clusters in ferredoxin and phthalate dioxygenase reductase. *The Journal of Physical Chemistry. B*, *102*, 6311–6324.

Li, J., & Ryde, U. (2014). Comparison of the active-site design of molybdenum oxo-transfer enzymes by quantum mechanical calculations. *Inorganic Chemistry*, *53*, 11913–11924.

Liang, T., Shin, Y. K., Cheng, Y.-T., Yilmaz, D. E., Vishnu, K. G., Verners, O., … van Duin, A. C. (2013). Reactive potentials for advanced atomistic simulations. *Annual Review of Materials Research*, *43*, 109–129.

Liu, T., Han, W.-G., Himo, F., Ullmann, G. M., Bashford, D., Toutchkine, A., … Noodleman, L. (2004). Density functional vertical self-consistent reaction field theory for solvatochromism studies of solvent-sensitive dyes. *The Journal of Physical Chemistry. B*, *108*, 11157–11169.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., … Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry. B*, *102*, 3586–3616.

Manta, B., Raushel, F. M., & Himo, F. (2014). Reaction mechanism of zinc-dependent cytosine deaminase from *Escherichia coli*: A quantum-chemical study. *The Journal of Physical Chemistry. B*, *118*, 5644–5652.

Marcus, R., & Sutin, N. (1985). Electron transfers in chemistry and biology. *Biochimica et Biophysica Acta, Reviews on Bioenergetics*, *811*, 265–322.

Maseras, F., & Morokuma, K. (1995). IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *Journal of Computational Chemistry*, *16*, 1170–1179.

Meier, K., Choutko, A., Dolenc, J., Eichenberger, A. P., Riniker, S., & Van Gunsteren, W. F. (2013). Multi-resolution simulation of biomolecular systems: A review of methodological issues. *Angewandte Chemie, International Edition*, *52*, 2820–2834.

Møller, C., & Plesset, M. S. (1934). Note on an approximation treatment for many-electron systems. *Physics Review*, *46*, 618–622.

Monti, S., Corozzi, A., Fristrup, P., Joshi, K. L., Shin, Y. K., Oelschlaeger, P., … Bourne, P. E. (2013). Exploring the conformational and reactive dynamics of biomolecules in solution using an extended version of the glycine reactive force field. *Physical Chemistry Chemical Physics*, *15*, 15062.

Moser, C. C., Keske, J. M., Warncke, K., Farid, R. S., & Dutton, P. L. (1992). Nature of biological electron transfer. *Nature*, *355*, 796–802.

Munro, L. J., & Wales, D. J. (1999). Defect migration in crystalline silicon. *Physical Review B*, *59*, 3969–3980.

Nango, E., Royant, A., Kubo, M., Nakane, T., Wickstrand, C., Kimura, T., … Iwata, S. (2016). A three-dimensional movie of structural changes in bacteriorhodopsin. *Science*, *354*, 1552–1557.

Noid, W. G. (2013). Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, *139*, 090901.

Oostenbrink, C., Villa, A., Mark, A. E., & Van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, *25*, 1656–1676.

Page, C. C., Moser, C. C., Chen, X., & Dutton, P. L. (1999). Natural engineering principles of electron tunneling in biological oxidation-reduction. *Nature*, *402*, 47–52.

Pande, K., Hutchison, C. D. M., Groenhof, G., Aquila, A., Robinson, J. S., Tenboer, J., … Schmidt, M. (2016). Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science*, *352*, 725–729.

Paquet, E., & Viktor, H. L. (2015). Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: A computational review. *BioMed Research International*, *2015*, 183918.

Peters, B., Heyden, A., Bell, A. T., & Chakraborty, A. (2004). A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *The Journal of Chemical Physics*, *120*, 7877–7886.

Riniker, S., Allison, J. R., & van Gunsteren, W. F. (2012). On developing coarse-grained models for biomolecular simulation: A review. *Physical Chemistry Chemical Physics*, *14*, 12423.

Rowley, C. N., & Woo, T. K. (2009). New shooting algorithms for transition path sampling: Centering moves and varied-perturbation sizes for improved sampling. *The Journal of Chemical Physics*, *131*, 234102.

Rücker, P., Wieninger, S. A., Ullmann, G. M., & Sticht, H. (2012). pH-dependent molecular dynamics of vesicular stomatitis virus glycoprotein G. *Proteins*, *80*, 2601–2613.

Saunders, M. G., & Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annual Review of Biophysics*, *42*, 73–93.

Senftle, T. P., Hong, S., Islam, M. M., Kylasa, S. B., Zheng, Y., Shin, Y. K., … van Duin, A. C. T. (2016). The ReaxFF reactive force-field: Development, applications and future directions. *Computational Materials*, *2*, 15011.

Senn, H. M., & Thiel, W. (2007). QM/MM methods for biological systems. In M. Reiher (Ed.), *Atomistic approaches in modern biology* (Vol. 268, pp. 173–290). Berlin, Heidelberg: Springer.

Senn, H. M., & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angewandte Chemie, International Edition*, *48*, 1198–1229.

Sharp, K. E. (1998). Calculation of electron transfer reorganization energies using the Finite Difference Poisson-Boltzmann model. *Biophysical Journal*, *73*, 1241–1250.

Shen, L., & Hu, H. (2014). Resolution-adapted all-atomic and coarse-grained model for biomolecular simulations. *Journal of Chemical Theory and Computation*, *10*, 2528–2536.

Shurki, A., Derat, E., Barrozo, A., & Kamerlin, S. C. L. (2015). How valence bond theory can help you understand your (bio) chemical reaction. *Chemical Society Reviews*, *44*, 1037–1052.

Singh, U. C., & Kollman, P. A. (1986). A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH3Cl + Cl exchange reaction and gas phase protonation of polyethers. *Journal of Computational Chemistry*, 7, 718–730.

Sokkar, P., Boulanger, E., Thiel, W., & Sanchez-Garcia, E. (2015). Hybrid quantum mechanics/molecular mechanics/coarse grained modeling: A triple-resolution approach for biomolecular systems. *Journal of Chemical Theory and Computation*, *11*, 1809–1818.

Stewart, J. J. P. (2013). Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling*, *19*, 1–32.

Svensson, M., Humbel, S., Froese, R. D. J., Matsubara, T., Sieber, S., & Morokuma, K. (1996). ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels–Alder reactions and $Pt(P(t\text{-}Bu)_3)_2$ + H_2 oxidative addition. *The Journal of Physical Chemistry*, *100*, 19357–19363.

Swenson, D. W. H., & Bolhuis, P. G. (2014). A replica exchange transition interface sampling method with multiple interface sets for investigating networks of rare events. *The Journal of Chemical Physics*, *141*, 044101.

Théry, V., Rinaldi, D., Rivail, J.-L., Maigret, B., & Ferenczy, G. G. (1994). Quantum mechanical computations on very large molecular systems: The local self-consistent field method. *Journal of Computational Chemistry*, *15*, 269–282.

Thompson, M. A., & Schenter, G. K. (1995). Excited ates of the bacteriochlorophyll b dimer of Rhodopseudomonas viridis: A QM/MM study of the photosynthetic reaction center that includes MM polarization. *The Journal of Physical Chemistry*, *99*, 6374–6386.

Till, M. S., Becker, T., Essigke, T., & Ullmann, G. M. (2008). Simulating the proton transfer in Gramicidin A by a sequential dynamical Monte Carlo method. *The Journal of Physical Chemistry. B*, *112*, 13401–13410.

Todorović, M., Bowler, D. R., Gillan, M. J., & Miyazaki, T. (2013). Density-functional theory study of gramicidin A ion channel geometry and electronic properties. *Journal of The Royal Society Interface*, *10*, 20130547.

Torrie, G., & Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Chemistry*, *23*, 187–199.

Tozzini, V. (2005). Coarse-grained models for proteins. *Current Opinion in Structural Biology*, *15*, 144–150.

Truhlar, D. G., Gao, J., Alhambra, C., Garcia-Viloca, M., Corchado, J., Sánchez, M. L., & Villà, J. (2002). The incorporation of quantum effects in enzyme kinetics modeling. *Accounts of Chemical Research*, *35*, 341–349.

Ullmann, G. M. (2000). The coupling of protonation and reduction in proteins with multiple redox centers: Theory, computational method, and application to cytochrome c_3. *The Journal of Physical Chemistry. B*, *104*, 6293–6301.

Ullmann, G. M., & Bombarda, E. (2014). Continuum electrostatic analysis of proteins. In G. Naray-Szabo (Ed.), *Protein modelling* (pp. 135–163). Cham, Switzerland: Springer International Publishing.

Ullmann, G. M., Kloppmann, E., Essigke, T., Krammer, E.-M., Klingen, A. R., Becker, T., & Bombarda, E. (2008). Investigating the mechanisms of photosynthetic proteins using continuum electrostatics. *Photosynthesis Research*, *97*, 33–53.

Ullmann, G. M., Mueller, L., & Bombarda, E. (2016). Theoretical analysis of electron transfer in proteins: From simple proteins to complex machineries. In W. Cramer (Ed.), *Cytochrome complexes: Evolution, structures, energy transduction, and signaling* (Vol. 37, pp. 99–127). Dordrecht, Netherlands: Springer International Publishing.

Ullmann, R. T., & Ullmann, G. M. (2012). GMCT: A Monte Carlo simulation package for macromolecular receptors. *Journal of Computational Chemistry*, *33*, 887–900.

Vanden-Eijnden, E., & Venturoli, M. (2009). Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *The Journal of Chemical Physics*, *130*, 194103.

van Duin, A. C. T., Dasgupta, S., Lorant, F., & Goddard, W. A., III. (2001). ReaxFF: A reactive force field for hydrocarbons. *The Journal of Physical Chemistry. B*, *105*, 9396–9409.

van Mourik, T., Bühl, M., & Gaigeot, M.-P. (2014). Density functional theory across chemistry, physics and biology. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *372*, 20120488.

Verlackt, C. C. W., Neyts, E. C., Jacob, T., Fantauzzi, D., Golkaram, M., Shin, Y.-K., … Bogaerts, A. (2015). Atomic-scale insight into the interactions between hydroxyl radicals and DNA in solution using the ReaxFF reactive force field. *New Journal of Physics*, *17*, 103005.

Řezáč, J., & Hobza, P. (2012). Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *Journal of Chemical Theory and Computation*, *8*, 141–151.

Wales, D. J., & Walsh, T. R. (1998). Theoretical study of the water pentamer. *American Institute of Physics*, *105*, 6957–6971.

Wang, Z., Antoniou, D., Schwartz, S. D., & Schramm, V. L. (2016). Hydride transfer in DHFR by transition path sampling, kinetic isotope effects, and heavy enzyme studies. *Biochemistry*, *55*, 157–166.

Warshel, A. (1991). *Computer modeling of chemical reactions in enzymes and solutions*. New York: Wiley & Sons.

Warshel, A. (2003). Computer simulations of enzyme catalysis: Methods, progress, and insights. *Annual Review of Biophysics and Biomolecular Structure*, *32*, 425–443.

Warshel, A., & Levitt, M. (1976). Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, *103*, 227–249.

Warwicker, J., & Watson, H. C. (1982). Calculation of the electrostatic potential in the active site cleft due the α-Helix dipols. *Journal of Molecular Biology*, *186*, 671–679.

Webb, S. P., Iordanov, T., & Hammes-Schiffer, S. (2002). Multiconfigurational nuclear-electronic orbital approach: Incorporation of nuclear quantum effects in electronic structure calculations. *The Journal of Chemical Physics*, *117*, 4106–4118.

Yilmazer, N. D., & Korth, M. (2015). Enhanced semiempirical QM methods for biomolecular interactions. *Computational and Structural Biotechnology Journal*, *13*, 169–175.

Zavadlav, J., Melo, M. N., Marrink, S. J., & Praprotnik, M. (2015). Adaptive resolution simulation of polarizable supramolecular coarse-grained water models. *The Journal of Chemical Physics*, *142*, 244118.

Zhang, Y. (2013). Electrostatic interaction of the electrostatic-embedding and mechanical-embedding schemes for QM/MM calculations. *Communications in Computational Chemistry*, *1*, 109–117.

Zhao, Y., & Truhlar, D. G. (2008). The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other function. *Theoretical Chemistry Accounts*, *120*, 215–241.

Zimmerman, M., & Bowman, G. (2016). Chapter nine—How to run FAST simulations. *Methods in Enzymology*, *578*, 213–225.

Zwanzig, R. W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *The Journal of Chemical Physics*, *22*, 1420–1426.

Manuscript B

PyCPR – A Python-Based Implementation of the Conjugate Peak Refinement (CPR) Algorithm for Finding Transition State Structures

Florian J. Gisdon, Martin Culka, G. Matthias Ullmann, *J. Mol. Model.* 2016, 22, 242
DOI: 10.1007/s00894-016-3116-8

J Mol Model (2016) 22: 242
DOI 10.1007/s00894-016-3116-8

ORIGINAL PAPER

PyCPR – a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures

Florian J. Gisdon[1] · Martin Culka[1] · G. Matthias Ullmann[1]

Received: 27 July 2016 / Accepted: 2 September 2016 / Published online: 20 September 2016

Abstract Conjugate peak refinement (CPR) is a powerful and robust method to search transition states on a molecular potential energy surface. Nevertheless, the method was to the best of our knowledge so far only implemented in CHARMM. In this paper, we present PyCPR, a new Python-based implementation of the CPR algorithm within the pDynamo framework. We provide a detailed description of the theory underlying our implementation and discuss the different parts of the implementation. The method is applied to two different problems. First, we illustrate the method by analyzing the gauche to anti-periplanar transition of butane using a semiempirical QM method. Second, we reanalyze the mechanism of a glycyl-radical enzyme, namely of 4-hydroxyphenylacetate decarboxylase (HPD) using QM/MM calculations. In the end, we suggest a strategy how to use our implementation of the CPR algorithm. The integration of PyCPR into the framework pDynamo allows the combination of CPR with the large variety of methods implemented in pDynamo. PyCPR can be used in combination with quantum mechanical and molecular mechanical methods (and hybrid methods) implemented directly in pDynamo, but also in combination with external programs such as ORCA using pDynamo as interface. PyCPR is distributed as free, open source software and can be downloaded from http://www.bisb.uni-bayreuth.de/index.php?page=downloads.

Florian J. Gisdon and Martin Culka contributed equally.

✉ G. Matthias Ullmann
ullmann@uni-bayreuth.de

[1] Computational Biochemistry, University of Bayreuth, Universitätsstr. 30, NW I, 95447 Bayreuth, Germany

Keywords Potential energy surface · Saddle point · Transition state search · Minimum energy path · Reaction mechanism · pDynamo

Abbreviation list

CI-NEB	Climbing image - nudged elastic band
COS	Chain-of-states
CPR	Conjugate peak refinement
DFT	Density functional theory
HPD	4-hydroxyphenylacetate decarboxylase
MEP	Minimum energy pathway
NEB	Nudged elastic band
PES	Potential energy surface
QM/MM	Quantum mechanics / molecular mechanics
RMSD	Root mean square deviation
RMS	Root mean square
ZTS	Zero temperature string

Introduction

One of the most fruitful concepts in theoretical chemistry is the potential energy surface (PES) of a molecular system, which can be seen as a landscape with the valleys and mountain passes describing the states and reactions of a molecule. In a mathematical sense, the PES is a $3N$-dimensional function describing the energy of a molecular system in terms of the coordinates of its N atoms. Each point on the surface of this landscape can be identified with one particular structure that the molecular system adopts. Therefore the term point is often used as synonym for structure when energy landscapes are discussed. Many problems in thermodynamics and kinetics can be tackled by exploring this landscape.

Namely, stable conformations are identified as minima, and transition states are first order saddle points. For an elementary reaction, a reaction coordinate is the path from a minimum identified as the reactant state along the shallowest ascent towards the transition state and from there along the deepest descent towards the product state. Many chemical reactions do, however, not proceed in one elementary reaction, and thus several intermediates and transition states are involved. While the reactant state and the product state are usually low energy minima on the energy landscape, intermediate states are usually high energy minima. The mechanism of a chemical reaction can be identified as the sequence of intermediates and transition states that a molecular system passes through when going from the reactant state to the product state following a minimum energy path (MEP).

In order to investigate the mechanism of a chemical reaction theoretically, it is required to identify the various minima and saddle points along the reaction path on the PES. There are many techniques to explore the PES. One prominent method is molecular dynamics, which can be used in combination with enhanced sampling techniques to explore the options that a molecular system has [5, 44]. However, these methods are computational costly, especially if chemical reactions in proteins are explored for instance with quantum mechanical/molecular mechanical (QM/MM) techniques. For such applications, numerical optimization methods [33] are important alternatives. While it is relatively straightforward to find a nearby local minimum of any given structure, locating a transition state is much more challenging. The strategy to find a transition state structure is to search for a path connecting stable minima and get saddle point estimates along this path. In principle, it is possible to sample the PES by "walking" through it based on energy, gradient, and Hessian matrix calculations for instance by an eigenvector following method [7]. For more complex systems, however, these methods are inefficient and computational demanding. One approach used in more complex systems is to represent a reaction path as an estimated reaction coordinate of discrete points, a chain-of-states (COS) between the reactant and product state, which is optimized as a whole. A simple minimization of each state in the COS would cause that all structures optimize to a nearby local minimum structure or even the reactant or the product structure. In order to avoid this kind of behavior, various COS optimization approaches have been developed. Early methods used a line integral representation, where the whole path is minimized at once based on a non-linear objective function. [10] Later the self-penalty walk was developed by introducing repulsion terms between the path points. [8] The paths obtained by line integral methods often suffer from overestimation of the transition states, which is caused by non-zero force components perpendicular to the path. The nudged elastic band (NEB) method [24] tries to avoid this problem by subtracting the perpendicular components of the force and applying an elastic spring force along the tangential direction of the path. The problem of this method is, however, that kinetic barriers are usually underestimated, since the real transition states are often missed in the path search. The climbing image extension of the NEB (CI-NEB) [19] can be applied in order to move the highest point of the NEB path closer to the saddle point. In CI-NEB, the parallel force component is inverted in order to move the chosen point uphill along the path while keeping the perpendicular component unchanged. Another approach to search reaction paths is the zero temperature string (ZTS) method [9], which is not using spring forces between the points along the path, but instead represents the COS by a spline. After every optimization step of the COS, the points are equally redistributed along the spline to ensure even sampling. A method similar to ZTS is the growing string method [34], in which a COS between two minima is gradually constructed. As the string grows, the optimization proceeds analog to the ZTS method. In addition, several hybrid NEB-ZTS methods were developed as well [1, 16].

An alternative strategy is followed in the conjugate peak refinement (CPR), which was first introduced by Fischer & Karplus [14]. Also the CPR method can be classified as a COS method. In contrast to many other COS methods, the number of the states in the chain is not fixed and thus the path between substrate and product is constructed and modified gradually. The CPR algorithm interpolates between the path points in order to find high energy structures along the path. The highest point on such an interpolated path is optimized by a single maximization along the tangential direction of the path. The maximization is followed by a series of conjugate minimizations. By an iterative procedure, all peaks along the path are optimized and some will approximate first order saddle points.

Since its introduction in the year 1992, CPR has been successfully used in numerous studies to investigate reaction paths. [23, 27, 29, 32, 40, 46] To our knowledge, however, the CPR algorithm is so far only implemented in the program CHARMM. [6] We decided to implement the CPR algorithm within the pDynamo [12] framework to extend the set of reaction path optimization methods of this software. This python-based framework provides a good environment for molecular modeling using pure quantum mechanical (QM) potentials, various empirical molecular mechanical (MM) force fields, and hybrid QM/MM potentials. Many minimization methods, as well as reaction path optimization methods like NEB and string methods are already implemented in the pDynamo framework. Within pDynamo, our CPR implementation can be easily combined with the functionality of this framework. Since pDynamo

is open-source and its modules are mainly written in the modern language Python, the CPR algorithm can be easily modified and extended.

In this paper, we describe PyCPR, our new implementation of the CPR algorithm for the pDynamo framework. First, we summarize the theoretical background of the CPR algorithm as the basis of our implementation. Further, we provide the details of our implementation including a schematic overview of the program flow. Using two examples, we illustrate the application of the CPR algorithm and compare our results with the outcome of NEB calculations. In the first example, we look at the conformation change of butane using a semiempirical QM method. This example is straightforward, but nevertheless provides some idea about the applicability and strength of CPR. In the second example, we reanalyze the mechanism of a glycyl-radical enzyme, namely of 4-hydroxyphenylacetate decarboxylase (HPD) [11, 28, 41]. In the end, we give some practical guidance for the usage of PyCPR.

Theory

As a chain-of-states method, CPR tries to build a path between two minima on the potential energy landscape (PES). It focuses on finding true first order saddle points, which can be interpreted as chemical transition states [14]. The characteristic of a first order saddle point is that its gradient is zero and its Hessian matrix $\mathbf{H}$ has exactly one negative eigenvalue. The theoretical basis of the CPR algorithm is a method for finding first order saddle points of multidimensional functions, which was developed by Sinclair & Fletcher [42]. The method is an adaptation of the conjugate gradient method [15] and relies on the assumption that the function is quadratic, which is approximately the case in the vicinity of a stationary point. Sinclair & Fletcher take advantage of the fact that the Hessian at a saddle point has exactly one negative eigenvalue. The basic idea is the following: the Hessian can be diagonalized by $\mathbf{E}^{-1}\mathbf{HE} = \mathbf{D}$ yielding a matrix $\mathbf{D}$ with the eigenvalues of $\mathbf{H}$ as its diagonal elements and zero as its off-diagonal elements. The matrix $\mathbf{E}$ is the matrix of the eigenvectors of $\mathbf{H}$ and $\mathbf{E}^{-1}$ is its inverse. Since the Hessian $\mathbf{H}$ is always real and symmetric, its eigenvectors form an orthogonal basis and thus $\mathbf{E}^{-1} = \mathbf{E}^{T}$ is valid. In general, one can find other basis sets $\mathbf{S}$ that transform the matrix $\mathbf{H}$ into another diagonal matrix $\mathbf{B}$: $\mathbf{S}^{T}\mathbf{HS} = \mathbf{B}$. Each of these basis sets are called conjugate basis sets and the basis vectors have the property $\mathbf{s}_i^{T}\mathbf{H}\mathbf{s}_j = 0$ for $i \neq j$. Sylvester's law of inertia [45] states that diagonalization always results in the same number of positive, negative and zero entries in the diagonal matrix independent of the conjugate basis. Consequently, each set of conjugate basis vectors of the Hessian at the saddle point contains exactly one direction with a negative diagonal entry and its associated vector must be equivalent to the eigenvector associated with the negative eigenvalue. The idea of this saddle point search method is that near the saddle point one can follow the vector associated with the negative entry in the diagonal matrix to locate the saddle point. Thus, starting from a point $\mathbf{x}_0$ which is close to a saddle point, Sinclair & Fletcher maximize the function along a vector $\mathbf{s}_0$ that should point towards the saddle point. This direction is thought to be associated with the negative eigenvalue. After the maximization along $\mathbf{s}_0$, all remaining directions ($\mathbf{s}_1$ to $\mathbf{s}_i$) are associated with positive values in the diagonal matrix and thus the energy function has to be minimized in order to find a true first order saddle point. Starting from an arbitrary initial direction $\mathbf{s}_0$, Sinclair & Fletcher applied the formulae of Beale [3] to construct the remaining basis vectors $\mathbf{s}_i$ that are conjugate to each other with respect to $\mathbf{H}$. If the initial vector $\mathbf{s}_0$ is given, the first conjugate direction is calculated by Eq. 1.

$$\mathbf{s}_1 = -\mathbf{g}_1 + \frac{\mathbf{g}_1^{T}(\mathbf{g}_1 - \mathbf{g}_0)}{\mathbf{s}_0^{T}(\mathbf{g}_1 - \mathbf{g}_0)}\mathbf{s}_0 \tag{1}$$

All subsequent directions are obtained by Eq. 2

$$\mathbf{s}_i = -\mathbf{g}_i + \frac{\mathbf{g}_i^{T}(\mathbf{g}_1 - \mathbf{g}_0)}{\mathbf{s}_0^{T}(\mathbf{g}_1 - \mathbf{g}_0)}\mathbf{s}_0 + \beta_i^{FR}\mathbf{s}_{i-1},\ i = 2, \ldots, 3N-1 \tag{2}$$

$$\beta_i^{FR} = \frac{\mathbf{g}_i^{T}\mathbf{g}_i}{\mathbf{g}_{i-1}^{T}\mathbf{g}_{i-1}} \tag{3}$$

For the first conjugate direction $\mathbf{s}_1$, it is necessary to know the gradient ($\mathbf{g}_0$) at an initial point ($\mathbf{x}_0$) and the gradient ($\mathbf{g}_1$) after maximization along the direction $\mathbf{s}_0$. For all other directions $\mathbf{s}_i$, the gradient of the current point $\mathbf{g}_i$ and previous point $\mathbf{g}_{i-1}$ as well as the previous direction $\mathbf{s}_{i-1}$ are used in addition. The term β_i^{FR} refers to the so-called beta-type originally proposed by Fletcher & Reeves [15] which was used by Fischer & Karplus [14] in the original implementation of CPR. A beta-type is a scaling factor adjusting the influence of the previous direction. Different beta-types exist that may perform better in the case of more complex functions. [25, 37] Therefore in PyCPR, we adapted Eq. 2 with two different well-established beta-types that are commonly used in conjugate gradient optimizations [17]. The additionally implemented types are those of Polak & Ribière [35] and of Polyak [36] (Eq. 4) and Hestenes & Stiefel [21] (Eq. 5).

$$\beta_i^{PRP} = \frac{\mathbf{g}_i^{T}(\mathbf{g}_i - \mathbf{g}_{i-1})}{\mathbf{g}_{i-1}^{T}\mathbf{g}_{i-1}} \tag{4}$$

$$\beta_i^{HS} = \frac{\mathbf{g}_i^{T}(\mathbf{g}_i - \mathbf{g}_{i-1})}{\mathbf{s}_{i-1}^{T}(\mathbf{g}_i - \mathbf{g}_{i-1})} \tag{5}$$

With Eqs. 1–5, a set of conjugate basis vectors with respect to a matrix can be defined starting with an arbitrary initial direction $\mathbf{s}_0$. Since the Hessian at a saddle point has exactly one negative eigenvalue, starting at a point $\mathbf{x}_0$ close to that saddle point, one direction $\mathbf{s}_0$ exists, where the function has to be maximized, and minimization has to be performed in all directions $\mathbf{s}_i$ conjugate to $\mathbf{s}_0$ (with respect to $\mathbf{H}$) in order to find the saddle point.

The method of Sinclair & Fletcher is guaranteed to find saddle points, if the function is purely quadratic. In practice, a multidimensional function can be approximated by a quadratic expansion in proximity of a saddle point. However, it may happen that the approximation fails, which causes unstable behavior. A measure for the quality of the approximation is the parameter τ_i, which is the normalized squared scalar product of the respective conjugate gradient vector $\mathbf{g}_i$ with the initial search direction $\mathbf{s}_0$ [42].

$$\tau_i = \frac{(\mathbf{g}_i^T \mathbf{s}_0)^2}{(\mathbf{g}_i^T \mathbf{g}_i)(\mathbf{s}_0^T \mathbf{s}_0)} \tag{6}$$

The quadratic approximation is good if all τ_i have values close to zero. If this criterion is not justified, it may happen that the optimization diverges too far from the saddle point and may lead to a nearby minimum. Thus τ_i can be utilized to define a stop criterion for conjugate optimization.

The CPR algorithm relies on the theory described above which resembles to a certain extent the basic ideas of the eigenvector following algorithm. However, CPR is robust and may find transition states, even if the initial guess is

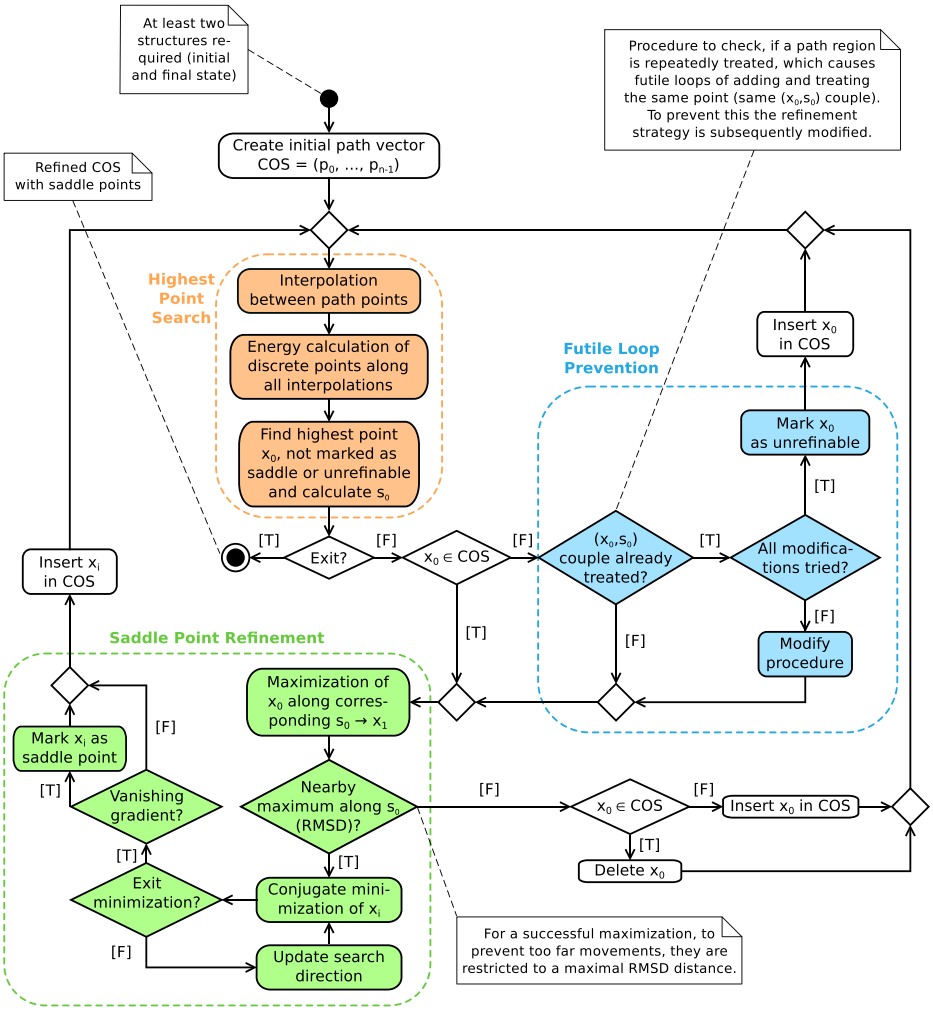


Fig. 1 General overview of the PyCPR implementation. The algorithm can be subdivided into three major parts. The *Highest Point Search* determines the next point ($\mathbf{x}_0$) for the *Saddle Point Refinement* and communicates with the exit decision of the algorithm. Before refinement of that chosen point $\mathbf{x}_0$, the *Futile Loop Prevention* is performed to prevent repeated refinement of the same couple ($\mathbf{x}_0$, $\mathbf{s}_0$) without success. To escape these futile loops, the *Saddle Point Refinement* procedure is modified

not optimal. CPR gains this better performance by gradually approaching closer to the transition state in several CPR iterations as will be explained below.

Algorithm of CPR and implementation

In the section Theory, we describe how a saddle point can be determined, if a starting point $\mathbf{x}_0$ and an initial search direction $\mathbf{s}_0$ are provided. The described theory relies on the assumption that $\mathbf{x}_0$ is located near a saddle point, because there the energy landscape can be well approximated by a quadratic function. The CPR algorithm, originally proposed by Fischer & Karplus [14], provides a strategy how the promising point $\mathbf{x}_0$ with its initial search direction $\mathbf{s}_0$ is chosen and refined to find eventually a saddle point using the method of Sinclair & Fletcher [42]. As a starting point of the transition state search, the initial and the final path point have to be given and, if required, also an initial guess of the reaction path can be provided. The CPR algorithm constructs a chain-of-states and adds more and more states to the chain, especially close to the saddle point, and thus approaches closer and closer to a real first order saddle point. Our implementation of the algorithm is summarized in Fig. 1.

The CPR algorithm requires at least two structures to create an initial COS which represents the reaction path (see Fig. 1). Usually these structures are the reactant state and the product state of the reaction and if required additional discrete intermediates. These structures are saved in a list holding several states that approximate the reaction path. The algorithm can be subdivided into three major parts. First, an initial path is generated by interpolating linearly between the path points and the point with the highest energy along this path is determined. We call this step *Highest Point Search*. Second, the point is refined to approach closer to a real saddle point. We call this step *Saddle Point Refinement*. During this step, it is not necessary that a saddle point is found, but in the course of the CPR run, saddle points will be gradually approached until the algorithm is able to locate transition states. After the *Saddle Point Refinement*, the path is modified and a new highest path point is searched. During this iterative process, futile loops can occur in which the same points are refined without

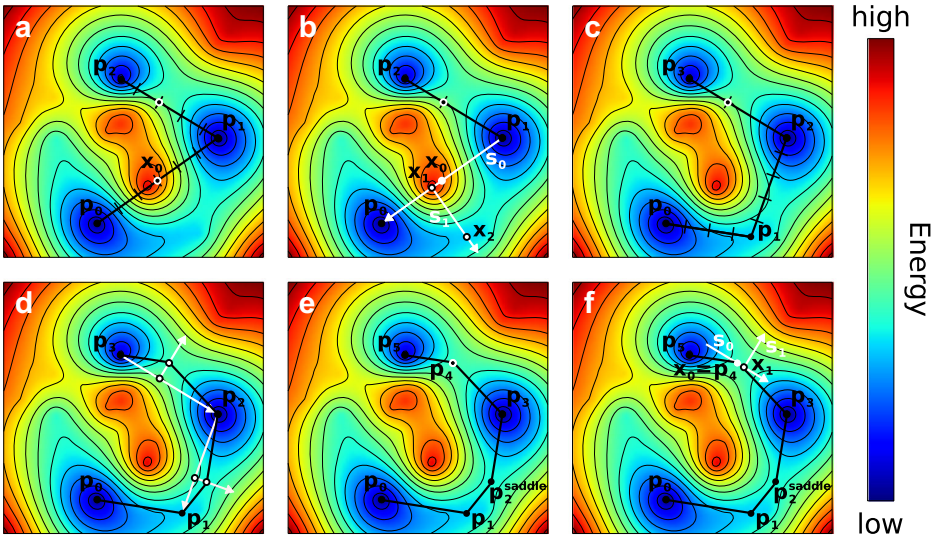


Fig. 2 Schematic illustration of the CPR algorithm on an artificial energy landscape starting from the initial path connecting the minima (p_0, p_1, p_2). **a** The whole initial path is approximated by piecewise linear interpolations (segments) between the path points. For each segment, the linear interpolation is discretized and the point with the highest energy is determined. In the end, the point with the highest energy of the whole reaction path (x_0) is determined. **b** The initial search direction s_0 corresponding to the highest point x_0 is determined. The point x_0 is maximized along s_0 yielding x_1. For each subsequently calculated search direction s_i, a line minimization is performed. In 2D-space, just one subsequent search direction s_1 exists along which x_1 is minimized yielding x_2. **c** The optimized point x_2 is inserted into the path as p_1. After extending the path, only the modified segments are updated. **d** The algorithm proceeds with finding the highest point along the path and optimizes it respectively. While the point between p_2 and p_3 was optimized to a normal path point, which lies along the path but is not yet a saddle point, the point between p_1 and p_2 was optimized to a saddle point. **e** The first part of the path from point p_0 to point p_3 is refined. It contains one saddle point connecting two minima and has no higher points along the linear interpolations connecting the path points. Thus this part will not be treated by the algorithm anymore. The second part of the path from point p_3 to p_5 is not completely refined. The path point p_4 is the highest energy point of the whole path, which will be optimized next. **f** Since p_4 is an existing path point, the initial search direction s_0 is calculated according to Eq. 8. The optimization of this point in the conjugate direction results in the second saddle point of the path. The reaction path is fully refined and the algorithm terminates

success again and again. These futile loops are prevented in the third part of our implementation, which we call *Futile Loop Prevention*. A schematic illustration of a transition path search using CPR on a fictive energy landscape is provided in Fig. 2 and explained in detail in the caption of this figure. After CPR has finished successfully, the path contains the transition states of different path regions. The algorithm is not performing optimizations to connect the transition states with stable intermediates on the PES. Thus, additional calculations should follow in order to generate the full MEP. In the following subsections, the three main parts of the algorithm are described in detail.

Highest point search This procedure determines the starting point $\mathbf{x}_0$ for the subsequent optimization and an initial search direction $\mathbf{s}_0$. The path from the reactant to the product state is approximated by piecewise linear interpolations between the path points. Each piece between existing path points is called segment. These segments are a rough estimate of the true reaction path between path points. The number of interpolation steps is calculated dependent on the segment length. For each segment, the structure with the highest energy is identified. In the first iteration of the CPR algorithm, all segments are linearly interpolated, while in the subsequent cycles only modified segments are updated. The point $\mathbf{x}_0$ is determined as the point with the highest energy of the whole reaction path that is not already a saddle point or another type of stationary point or that is not marked as unrefinable (for explanation, see Section *Futile Loop Prevention*). If no such point is found, the reaction path is considered to be refined and the algorithm stops.

In addition to the search of $\mathbf{x}_0$, the initial search direction $\mathbf{s}_0$ is determined, which is a vector in tangential direction of the reaction path at the point $\mathbf{x}_0$. If $\mathbf{x}_0$ is located between two path points $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, $\mathbf{s}_0$ is the vector connecting these two points (Eq. 7).

$$\mathbf{s}_0 = \mathbf{p}_{i+1} - \mathbf{p}_i \tag{7}$$

If $\mathbf{x}_0$ is an already existing path point $\mathbf{p}_i = \mathbf{x}_0$, $\mathbf{s}_0$ is calculated by Eq. 8 [14].

$$\mathbf{s}_0 = \frac{\mathbf{p}_{i+1} - \mathbf{p}_i}{|\mathbf{p}_{i+1} - \mathbf{p}_i|} - \frac{\mathbf{p}_{i-1} - \mathbf{p}_i}{|\mathbf{p}_{i-1} - \mathbf{p}_i|} \tag{8}$$

The point $\mathbf{x}_0$ and the initial search direction $\mathbf{s}_0$ are passed to the *Saddle Point Refinement*.

Saddle point refinement In this procedure, the point $\mathbf{x}_0$ is optimized towards a saddle point by the theory described above. At the beginning of each *Saddle Point Refinement*, we employ a line search algorithm [30] to search more thoroughly for a nearby maximum along the search direction $\mathbf{s}_0$. The direction $\mathbf{s}_0$ is considered to approximate the direction of the eigenvector corresponding to the negative eigenvalue of the Hessian. Whether a maximum is nearby or not is determined on the basis of the root mean square deviation (RMSD) between the structure corresponding to $\mathbf{x}_0$ and the potential maximum $\mathbf{x}_1$. If no proper maximum was found, the optimization ends. Depending on where $\mathbf{x}_0$ is located on the reaction path, the algorithm will proceed differently. If $\mathbf{x}_0$ was found within a segment, it will be inserted into the path as a normal path point. If $\mathbf{x}_0$ was an existing path point $\mathbf{p}_i$, it will get deleted from the reaction path. If a nearby maximum $\mathbf{x}_1$ was found, the optimization continues with a conjugate minimization along the direction $\mathbf{s}_1$, using again a line search algorithm. The optimization continues along the other conjugate directions $\mathbf{s}_i$ and stops, if one of the following three criteria is fulfilled:

(1) A saddle point has been found: This criterion is fulfilled, if the gradient of the energy function is vanishing in all $3N$ directions where N is the total number of atoms of the system and the second derivative of the energy function is negative in one direction and non-negative in all others. In practice this criterion is too stringent for large molecules. To reduce the computational cost, it is reasonable to assume, that the optimization reached a point sufficiently close to a saddle point, if the gradient is close to zero in M successive dimensions. Noé et al. [32] suggested to set M to $\sqrt{N}$. In our implementation, the default value of M is $\sqrt{N}$ but can be replaced by any user-defined value. If a saddle point is found, it is inserted into the reaction path and marked as a saddle point.

(2) The new search direction $\mathbf{s}_i$ is not anymore conjugate to the initial search direction: This criterion is fulfilled, if the quadratic approximation does not hold anymore. A measure for the quality of the quadratic approximation is the parameter τ_i, which was described in the theory section. Ideally τ_i is zero. If τ_i reaches a predefined threshold value t, we stop the optimization and insert the point $\mathbf{x}_i$ into the path but do not mark it as a saddle point. In our implementation, we have also the possibility to continue the optimization. If the continuation of the optimization leads to a point with vanishing gradients in M successive dimensions, the point will be inserted into the path and marked as stationary point but not as saddle point.

(3) A predefined maximum number of optimization steps are performed: This criterion is only reached if criterion (1) or (2) did not lead to the stopping of the optimization. The default value of the maximum number of optimization steps is $3N - 1$, but it can be set by the user. However it should be noted, that this number

should be greater than the number M of criterion (1). Since a low gradient region is usually not found from the beginning of the optimization, the maximal number of steps should be at least in the order of $2M$.

After the *Saddle Point Refinement* is finished, the algorithm continues with the *Highest Point Search*. Only the segments that have changed because of the modification of the path are updated, i. e. only in those changed segments a new linear interpolation is performed and a new highest segment point is determined. This procedure repeats until no refinable highest point was found in the *Highest Point Search*. Since points are inserted into and deleted from the COS by the algorithm, futile loops can occur in which the same point is inserted and deleted repeatedly. This behavior is prevented because of the implementation of the *Futile Loop Prevention*.

Futile loop prevention If the *Highest Point Search* returns a point $\mathbf{x}_0$ (together with a direction $\mathbf{s}_0$), which is not already part of the reaction path, it will be checked if the same ($\mathbf{x}_0$, $\mathbf{s}_0$) couple was already treated in a previous *Saddle Point Refinement*. This check is done by calculating the RMSD between the corresponding vectors of the new ($\mathbf{x}_0$, $\mathbf{s}_0$) couple and of each of the previously treated ($\mathbf{x}_0$, $\mathbf{s}_0$) couples. A futile loop is predicted, if the RMSD values for $\mathbf{x}_0$ and $\mathbf{s}_0$ are both below a certain threshold which can be adjusted. In order to prevent such a futile loop, we have implemented several modifications in the optimization procedure, which can be optionally applied. First we increase the number of discretization steps along the considered segment and choose the highest point among these points as $\mathbf{x}_0$. If this treatment was not successful, the next step is to increase the threshold t for τ_i (Eq.: 6) to $2t$. This option is only executed, if the τ_i parameter is considered as an exit criterion for the *Saddle Point Refinement*. As a final modification of the optimization strategy, the search directions $\mathbf{s}_i$ ($i \neq 0$) are determined using orthogonal instead of conjugate directions (Eqs. 9 and 10) [13].

$$\mathbf{s}_1 = -\mathbf{g}_1 + \frac{\mathbf{g}_1^T \mathbf{s}_0}{\mathbf{s}_0^T \mathbf{s}_0}\mathbf{s}_0 \tag{9}$$

$$\mathbf{s}_i = -\mathbf{g}_i + \frac{\mathbf{g}_i^T \mathbf{s}_0}{\mathbf{s}_0^T \mathbf{s}_0}\mathbf{s}_0 + \beta_i^{FR}\mathbf{s}_{i-1},\ i = 2, ..., 3N-1 \tag{10}$$

If all modifications do not prevent the futile loop, the point is included in the reaction path and marked as unrefinable. An unrefineable point will be ignored in the *Highest Point Search* in order to concentrate on different regions of the path. In PyCPR, we provide an option that the points which were marked as unrefinable are optimized once before quitting the CPR algorithm. If the optimization was not successful, these points will be kept in the path still marked as unrefinable.

The result of a CPR run is a COS, connecting the structures that were given as an initial and final state. The COS contains, besides the product state and the reactant state, transition states as well as other states that lie approximately on the MEP.

Variation of PyCPR from CPR algorithm by Fischer & Karplus Although our PyCPR implementation is based on the original implementation of the CPR algorithm by Fischer & Karplus [14], it features several differences. As already noted in the section Theory, we introduced alternative scaling factors – beta-types. In PyCPR, the beta-type of Polak & Ribière [35] and of Polyak [36] (Eq. 4) was chosen as the default one, because it was reported [37] as more efficient than the originally used one of Fletcher & Reeves [15]. We use the original expression for conjugacy measure τ_i as proposed by Sinclair & Fletcher [42], while in the Fischer & Karplus CPR an empirical version that varies with system size was used [14]. In contrast to the original CPR, we also introduced an option to not use τ_i as an exit criterion for the conjugate optimization and thus allow generation of more intermediates in an early stage of the refinement. In the futile loop detection process, we introduced storing of every ($\mathbf{x}_0$, $\mathbf{s}_0$) couple already treated in a list. By comparing the currently treated couple with this list, we are able to detect a futile loop of any length. This strategy turned out to be extremely effective.

Implementation in pDynamo We have implemented the above described CPR algorithm within the open source framework pDynamo [12]. Within this framework, simulations of molecular systems can be performed using hybrid quantum mechanical (QM) and molecular mechanical (MM) potentials. In pDynamo, computational expensive operations are implemented in the programming language C, while the higher order algorithms are organized in Python modules and packages. The accessibility of these modules makes the code highly adaptable and convenient to extend.

PyCPR is structured in three Python modules, which can be directly added to a working pDynamo version. We tested the compatibility of our implementation for version 1.8.0 and 1.9.0 of pDynamo. The first module *ConjugatePeakRefinement* contains the main CPR algorithm. The second module *CPRSaddlePointRefinement* is performing one complete conjugate optimization as described in the *Saddle Point Refinement* paragraph. The last module is inherited from pDynamo's *MoreThuenteLineSearch*. This module contains a line search algorithm with the additional feature of a line maximization *MoreThuenteLineSearchWithMax*.

Since pDynamo is a highly flexible framework, also alternatives for the optimizers could be used if they comply the necessary data requirements.

We provide the full source code of PyCPR free of charge. It can be downloaded from our web page http://www.bisb.uni-bayreuth.de/index.php?page=downloads.

Computational details

Conformational analysis of butane For analyzing the conformational change of butane from the anti-periplanar conformation to the gauche conformation, both conformations were optimized to a root-mean-square (RMS) gradient of $2.5 \cdot 10^{-4}$ kcal/(mol Å) with the conjugate gradient algorithm using pDynamo [12]. For all calculations, the RM1 [39] method was used in combination with ADIIS/DIIS SCF conversion [22]. During CPR, the threshold for the successive low gradient structures was $2.5 \cdot 10^{-4}$ kcal/(mol Å). The option to quit the *Saddle Point Refinement*, if the threshold of τ is reached, was not set. For the prevention of futile loops, the option to increase the number of discrete interpolation steps was disabled. The beta-type of Polak & Ribière [35] and of Polyak [36] (Eq. 4) was used to construct the conjugate directions. The resulting path consisted of ten points. For comparison, we performed a search with nudged elastic band (NEB) and climbing image (CI) NEB [1] as implemented in pDynamo. Since NEB is restricted to the initially adjusted number of path points, the PyCPR generated path defined the sampling accuracy. Thus, ten points were used to construct the COS using NEB. These structures were generated by the growing string [34] procedure implemented in pDynamo using an RMS gradient tolerance of 0.35 kcal/(mol Å). The NEB optimization was performed with an RMS gradient tolerance of $2.5 \cdot 10^{-4}$ kcal/(mol Å), a spring force constant of 120 kcal/(mol Å^2) and no spline redistribution. The same parameters were chosen for CI-NEB. The highest point of the prior NEB was used as the climbing image.

Refinement of the reaction mechanism of HPD Optimized stable intermediate structures from the previous work on HPD in our group [11] were used as a basis for this study. The transition states that were obtained by an NEB method [16] implemented in pDynamo were used here only for structural comparison. The biggest QM region (designated M4 in Ref. [11]) was used for all the QM/MM studies here. Same QM/MM setup as in the previous work was employed, namely the CHARMM27 [26] force field combined with UB3LYP [4]:6-31G* [18] QM method. We used the pDynamo [12] framework in combination with ORCA [31] for all the QM/MM calculations. PyCPR was used to find all the transition states. The threshold for the low gradient was set to 0.02 kcal/(mol Å). The beta-type of Polak & Ribière [35] and of Polyak [36] (Eq. 4) was chosen. For the vibration frequencies analysis of the transition states, the whole MM region was fixed to reduce the computational load.

Results and discussion

Conformational analysis of butane The conformational change of butane from the anti-periplanar conformation to the gauche conformation is a simple example for an MEP search that passes one saddle point. The structures of both states were minimized using the RM1 method [39]. The anti-periplanar conformation has a dihedral angle of the carbon chain of −180.0°, while this dihedral angle adopts a value of −64.7° in the gauche conformation. Our value for the gauche angle is matching well with values found in the literature ranging from −60° to −66°, which were obtained from high-level *ab initio* calculations [2, 20, 38, 43]. We calculated the transition between these two states using three different methods: CPR, NEB, and CI-NEB (Fig. 3). For the transition state search with PyCPR, no intermediates were provided, so the algorithm found the complete transition from just the initial and final structures resulting in a path of ten points. For consistency also for NEB and CI-NEB, we used ten points and optimized the gradients to the same tolerance. The transition state structure obtained by CPR has a torsional angle of −119.7° and a barrier of 2.00 kcal/mol. The value of the torsional angle is close to the ideal value of 120.0° and agrees with other studies [2, 38]. Frequency analysis confirms a true saddle point showing exactly one negative frequency. Pure NEB without the climbing image modification misses the true transition state. Note that pure

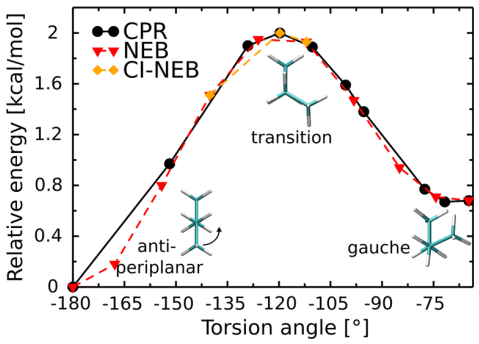


Fig. 3 Reaction energy profiles of the conformational change of butane from anti-periplanar to gauche. Full black line: CPR started from initial and final state; dashed red line: NEB started from ten points; dashed orange line: CI-NEB of point 4 of the NEB-derived path. The calculations were done on the semiempirical RM1 level. The angles of the shown states are: anti-periplanar −180.0°, transition −119.7° and gauche −64.7°

NEB is able to locate the correct transition state for some different numbers of initial path points, yet one can hardly derive this number rationally. Thus a successful transition state search is dependent on the initially set number of path points. This example nicely shows the different behavior of NEB and CPR. Since CPR is a method identifying transition states, the resulting path has a higher point density around the transition state. Further the flexibility to add and delete path points allows CPR to adapt the path to the respective reaction. NEB distributes points along the whole reaction path and thus represents the path as a whole very well. However, especially the sampling around the transition state is not optimal. For bigger systems, NEB is known to miss saddle points, which was the reason why the CI-NEB modification was developed [19]. CI-NEB was able to find the same transition as CPR at −119.7°. But we experienced problems finding a proper transition state with the CI-NEB method for more complex reactions. Identifying proper transition states gets more difficult with a higher number of degrees of freedom of a system, unless the starting point for CI-NEB is close to the saddle point. A problematic case, where the NEB procedure and also the CI-NEB procedure failed to locate the transition state, is shown in the following section.

Mechanism of 4-hydroxyphenylacetate decarboxylase

To test our CPR implementation on a more complex example, we have returned to the mechanism of 4-hydroxyphenylacetate decarboxylase (HPD) which was recently studied in our group [11]. 4-hydroxyphenylacetate decarboxylase is a glycyl radical enzyme that cleaves 4-hydroxyphenylacetate into p-cresol and CO_2 [28, 41]. Glycyl radical enzymes transfer an electron and a proton from a cysteine (Cys503 in HPD) to a glycine radical (Gly873 in HPD) in the activation phase, which results in a neutral cysteinyl radical species. This neutral radical state is the initial state of the analysis of the reaction cycle. The reaction mechanism of HPD starts with transfer of a single electron from the substrate to the Cys503. This electron transfer occurs in a concerted manner with a proton transfer from phenolic hydroxyl group of the substrate to a nearby glutamate (Glu637). The resulting deprotonated radical intermediate reprotonates causing a cleavage of the intermediate into CO_2 and p-hydroxybenzene radical. The thiolate of Cys503 receives a proton from a nearby glutamic acid (Glu505). In the final step, a hydrogen atom (i. e. a proton and an electron) are transferred from Cys503 to the methylene carbon of the p-hydroxybenzene radical producing p-cresol and regenerating the neutral cysteinyl radical.

In this paper, we employ PyCPR to reanalyze transition states between the stable intermediates found in the previous study and compared them with the NEB-derived transition states presented previously [11]. We use the same computational setup as in Ref. [11]. For all but one reaction step, we found essentially the same transition states both from structural and energetic point of view (Fig. 4). However, for the last reaction step, where a hydrogen atom is transferred from the Cys503 to the methylene carbon (C7) of the substrate, a barrier (TS4) of 7.2 kcal/mol relative to the last intermediate (In3) or 16.3 kcal/mol relative to the initial (Sub) state

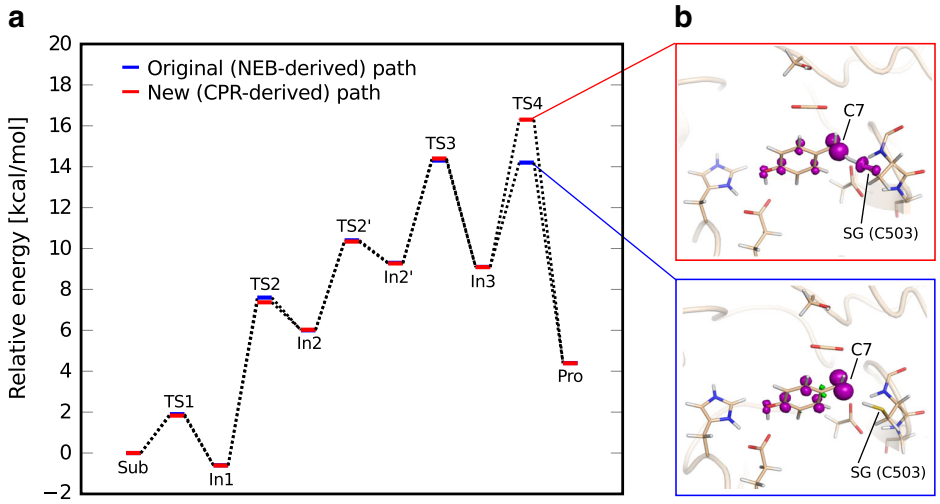


Fig. 4 Reaction energy profiles of the enzymatic cleavage of 4-hydroxyphenylacetate by the enzyme 4-hydroxyphenylacetate decarboxylase. The two reaction energy profiles derived by NEB (*blue*) and CPR (*red*) are based on identical intermediates. Data for the NEB-derived energy profile were taken from previous work [11]. A frequency analysis showed, that the NEB failed to find a real transition state and thus underestimates the barrier. **a**: NEB- and CPR-derived energy profiles. **b**: Map of the spin density distribution (purple positive, green negative) in the active site for the energetically different TS4 states shown at isovalue of 0.01 au

was found. This transition state has a significantly higher energy than the original TS4 barrier found with the same QM setup by NEB – 5.1 or 14.2 kcal/mol relative to In3 or Sub, respectively (Fig. 4a). The two TS4 structures differ in several ways. First, the hydrogen being transferred is 1.5 Å away from both Cys503 SG and the methylene carbon (C7) of the substrate for the CPR-derived TS4, while in case of original NEB-derived TS4 it is 1.4 Å and 2.1 Å away from SG and C7, respectively. More insight to the bonding situation can be derived from the Mayer bond order analysis. For the CPR-derived TS4, the hydrogen that is transferred in this step has a partial bond to both, C7 of the substrate (bond order 0.32) and SG of the Cys503 (bond order 0.62), while the original NEB-derived TS4 has only a bond to C7 (bond order 0.85). The next difference lies in the distribution of the unpaired spin. Although the total amount of the spin $\langle S^2 \rangle$ is comparable in both systems (0.76, and 0.77 for the CPR-derived and NEB-derived TS4, respectively), it is distributed differently. The CPR-derived TS4 has the unpaired spin density partitioned between the substrate (mostly on methylene C7 carbon, Mulliken population 0.52) and SG of Cys503 (Mulliken population 0.24). In contrast, the unpaired spin density is located exclusively on the substrate (again mostly on C7, Mulliken population 0.69) in the original NEB-derived TS4. The differences are shown in Fig. 4b and the Mulliken populations are summarized to Table 1. Finally, we performed vibrational frequency analysis on the both TS4 structures. The CPR-derived TS4 has one imaginary frequency and all other frequencies positive, a sign of a first order saddle point. In contrast, the original NEB-derived TS4 does not show this feature. To make the picture complete, we tried to refine the original NEB-derived TS4 using CI-NEB approach. For these calculations, we obtained a distorted structure with unrealistically high energy of more than 350 kcal/mol above the initial state indicating that the CI-NEB failed to stay on the original path. All these hints indicate that the NEB approach may have actually missed the real transition state TS4 and therefore underestimated the overall reaction barrier, while the CPR was able to find a transition state structure.

Table 1 Mulliken spin populations on selected atoms of the transition state of the last step (TS4) of substrate-bound 4-Hydroxyphenylacetate Decarboxylase

	NEB-derived TS4	CPR-derived TS4
C1 (subst.)	−0.15	−0.08
C2 (subst.)	0.20	0.14
C3 (subst.)	−0.10	−0.05
C4 (subst.)	0.21	0.15
C5 (subst.)	−0.10	−0.07
C6 (subst.)	0.21	0.14
C7 (subst.)	0.69	0.52
O4 (subst.)	0.06	0.06
SG (Cys503)	0.03	0.24

Strategy of path search using PyCPR PyCPR can be used to find a reaction path using only the reactant and product structure as an input. However, one has to keep in mind that there may be more reaction paths connecting the reactant and product state (see Fig. 5). Since the algorithm uses linear interpolations between the path points, the CPR approach may not find the path with the lowest overall activation barrier for going from the reactant state to the product state. Instead, the path that is near the initial linear interpolation is likely to be found as illustrated in Fig. 5. In this example, we use the same energy landscape as in Fig. 2 but provide just the reactant state and the product state. This starting scenario results in a different path with higher barrier (see Fig. 5). Once this passage is approached during the path refinement, CPR will hardly escape from there. But this problem is not specific for CPR. Especially for more

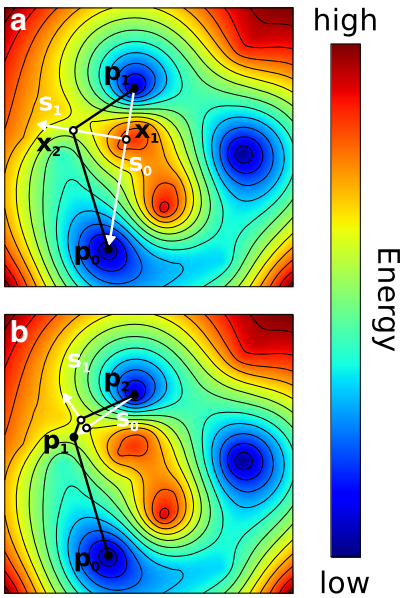


Fig. 5 Schematic illustration for finding the closest MEP to the initial path, which is represented by the initial COS with the linear interpolations connecting the points. Using the same artificial energy landscape as in Fig. 2 but providing just the reactant and product state without the intermediate, CPR fails to find the lowest MEP. Instead a less favorable path with a higher transition state is found. **a** The point with the highest energy found along the linear interpolation s_0 between p_1 and p_0 is optimized along the conjugate direction s_1, the intermediate state (minimum on the right-hand side of the energy landscape) is not found. **b** Further refinement of the path takes place only in the region around the path point p_1 until the saddle point is found. The refinement ends without approaching the lowest MEP

complex systems, we are not aware of an algorithm to find the correct MEP from scratch. The results always have to be analyzed critically. A crucial point for finding a promising reaction path is a good initial guess of the path. Starting from the pure linear estimation of the reaction pathway, the most concerted mechanism is usually found by CPR. A reasonable initial guess of the reaction path can consist of structures from an adiabatic PES scan guided by chemical intuition or it can be a preoptimized path from some other COS method. Within pDynamo, the growing string method can for instance be used for rapid gradual construction of the first path guess. Both, a path from adiabatic PES scan intermediates or a growing string path can be further optimized by NEB methods implemented in pDynamo prior to passing it to CPR. The initial guess can be obtained with computational cheap methods, since it is just a guidance for CPR. But even when a well-optimized NEB path is used as an initial guess, it can still have low sampling in the transition state regions where CPR might find higher structures and refine these to saddle points.

Another problem that is caused by the linear interpolation between two states is that the highest point on the initial (interpolated) path might have an unrealistically distorted structure with a high energy. Optimization may fail on such a distorted structure if quantum chemical methods are used, since the self-consistent field iteration during the electronic structure calculation may not converge, as also discussed in the CHARMM manual [13]. Also these problems can be prevented by adding less distorted geometries as intermediate structures. The intermediates do not need to be very realistic intermediates of the path, since they will be treated during the refinement of the MEP [13].

If a prerefined path is not available, it may help to run PyCPR with the option not to quit the *Saddle Point Refinement* procedure when τ (Eq. 6) reaches the tolerance t. This setting leads to a less restricted search that can divert the path further from the initial linear interpolation by adding new points to the path between the starting points. Once new structures are found, it is advisable to restart the CPR but now exiting the *Saddle Point Refinement* procedure if τ reaches the tolerance criterion, which leads to better path sampling in the transition state regions. As also discussed in the CHARMM manual [13], using an orthogonal construction of the search directions instead of a conjugate construction in the first few cycles of the CPR run may help to converge to better initial guesses for the path.

Once a first complete CPR path is found, potential stable intermediates should be identified and further minimized. Then the path can be split into discrete mechanistic steps connecting the minima, which can be refined separately to optimize different path regions in parallel. In general, it is easier to get transition states between two intermediates representing discrete mechanistic steps than trying to search the whole mechanism between two stable states *de novo*. In order to refine the reaction path, one can also benefit from a combination of CPR with different methods like NEB, which usually relaxes the chain-of-states nicely to the MEP, but tends to miss the transition states. CPR instead is transition state oriented. Thus the transition states found by CPR can be used as input for NEB. In the NEB runs, the transition states and the minima need to be kept fixed. This identified reaction path can then be a useful input for sampling methods such as umbrella sampling.

Conclusions

CPR [14] is a powerful method for finding transition states starting from a preliminary path, that was guessed initially by chemical intuition. The method is ideal for finding transition states even if the initial path points are far from the transition state. The strength of the method is that the theory developed by Sinclair & Fletcher [42] to find first order saddle points of multidimensional functions is applied in an approximate way to approach closer and closer to real first order saddle points. We implemented PyCPR as a part of the pDynamo framework [12], a versatile program designed for the simulation of molecular systems using quantum mechanical, molecular mechanical and hybrid QM/MM potential energy functions. As such, PyCPR can be used in a pure MM setup with various force fields, or in a pure QM setup using either pDynamo's own QM implementation or its interface to the program ORCA [31], which is more efficient for usage of higher QM methods like DFT. The biggest merit of pDynamo lies, however, in the hybrid QM/MM approach. Thus, PyCPR can be applied for exploring enzymatic mechanisms in combination with a wide variety of theoretical methods. Initial guesses of the mechanism guided by chemical intuition or by other theoretical methods are a useful starting point of the mechanistic explorations. While the outcome of CPR is transition state oriented, NEB can relax the path as a whole to the MEP using the transition states identified by CPR as fixed points.

CPR is a conceptual different path search method compared to existing approaches within the pDynamo framework, since it focuses on saddle point refinement. PyCPR supplements the available approaches in pDynamo and in return benefits from the python-based environment. Structured in three Python modules, which contain the three major parts of the algorithm, the code can be accessed and modified easily. We hope that with this CPR implementation in pDynamo, we provide a valuable reaction path search tool to the computational chemistry community which can serve as a powerful alternative and complement to other established methods.

Acknowledgments This work was supported by the DFG grant UL 174/9.

References

1. Aleksandrov A, Field MJ (2012) A hybrid elastic band string algorithm for studies of enzymatic reactions. Phys Chem Chem Phys 14:12,544–12,553
2. Allinger NL, Fermann JT, Allen WD, Schaefer IIIHF (1997) The torsional conformations of butane: Definitive energetics from ab initio methods. J Chem Phys 106:5143–5150
3. Beale EML (1972) A derivation of conjugate gradients. In: Lootsma FA (ed) Numerical methods for nonlinear optimization. Academic Press, London, pp 39–43
4. Becke A (1993) Density functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652
5. Bernardi RC, Melo MCR, Schulten K (2015) Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta 1850:872–877
6. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
7. Cerjan CJ (1981) On finding transition states. J Chem Phys 75:2800
8. Czerminski R, Elber R (1990) Self-avoiding walk between two fixed points as a tool to calculate reaction paths in large molecular systems. Int J Quantum Chem 38:167–185
9. E W, Ren W, Vanden-Eijnden E (2002) String method for the study of rare events. Phys Rev B 66:052,301
10. Elber R, Karplus M (1987) Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. Science 235:318–321
11. Feliks M, Martins BM, Ullmann GM (2013) Catalytic mechanism of the glycyl radical enzyme 4-hydroxyphenylacetate decarboxylase from continuum electrostatic and QC/MM calculations. J Am Chem Soc 135:14,574–14,585
12. Field MJ (2008) The pdynamo program for molecular simulations using hybrid quantum chemical and molecular mechanical potentials. J Chem Theory Comput 4:1151–1161
13. Fischer S (2003) TReK: A program for trajectory refinement and kinematics. In: CHARMM c40b1 documentation. https://www.charmm.org/charmm/documentation/by-version/c40b1/params/doc/trek/ [Accessed: 2016-06-02]
14. Fischer S, Karplus M (1992) Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. Chem Phys Lett 194:252–261
15. Fletcher R, Reeves CM (1964) Function minimization by conjugate gradients. Comput J 7:149–154
16. Galván IF, Field MJ (2008) Improving the efficiency of the NEB reaction path finding algorithm. J Comput Chem 29:139–143
17. Gilbert JC, Nocedal J (1992) Global convergence properties of conjugate gradient methods for optimization. SIAM J Optim 2:21–42
18. Hehre WJ, Ditchfield R, Pople J (1972) Self-consistent molecular orbital methods. XII. Further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. J Chem Phys 56:2257–2261
19. Henkelman G, Uberuaga BP, Jónsson H (2000) A climbing image nudged elastic band method for finding saddle points and minimum energy paths. J Chem Phys 113:9901–9904
20. Herrebout WA, van der Veken BJ, Wang A, Durig JR (1995) Enthalpy difference between conformers of n-butane and the potential function governing conformational interchange. J Phys Chem 99:578–585
21. Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. J Res Natl Bur Stand 49:409–436
22. Hu X, Yang W (2010) Accelerating self-consistent field convergence with the augmented Roothaan-Hall energy function. J Chem Phys 132:1–7
23. Imhof P, Fischer S, Smith JC (2009) Catalytic mechanism of DNA backbone cleavage by the restriction enzyme EcoRV: A quantum mechanical/molecular mechanical analysis. Biochemistry 48:9061–9075
24. Jonsson H, Mills G, Jacobsen KW (1998) Nudged elastic band method for finding minimum energy paths of transition. In: BJ Berne GC, Coker DF (eds) Classical and quantum dynamics in condensed phase simulations, world scientific, chap 16, pp 385–404
25. Leach AR (2003) Molecular modelling – principles and applications, 2nd edn. Pearson Education Ltd
26. MacKerell AD, Bashford D, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
27. Madhumalar A, Smith DJ, Verma C (2008) Stability of the core domain of p53: Insights from computer simulations. BMC Bioinformatics 9 Suppl 1:S17
28. Martins BM, Blaser M, Feliks M, Ullmann GM, Buckel W, Selmer T (2011) Structural basis for a kolbe-type decarboxylation catalyzed by a glycyl radical enzyme. J Am Chem Soc 133:14,666–14,674
29. Mixcoha E, Garcia-Viloca M, Lluch JM, González-Lafont A (2012) Theoretical analysis of the catalytic mechanism of helicobacter pylori glutamate racemase. J Phys Chem B 116:12,406–12,414
30. Moré JJ, Thuente DJ (1994) Line search algorithms with guaranteed sufficient decrease. ACM T Math Softw 20:286–307
31. Neese F (2012) The ORCA program system. Wiley Interdiscip Rev Comput Mol Sci 2:73–78
32. Noé F, Ille F, Smith JC, Fischer S (2005) Automated computation of low-energy pathways for complex rearrangements in proteins: Application to the conformational switch of Ras p21. Proteins 59:534–544
33. Peng T, Larkin JD, Brooks BR (2012) Reaction path optimization and sampling methods and their applications for rare events. In: Pahlavani MR (ed) Some applications of quantum mechanics, intech, pp 27–66
34. Peters B, Heyden A, Bell AT, Chakraborty A (2004) A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. J Chem Phys 120:7877–7886
35. Polak E, Ribiere G (1969) Note sur la convergence de methods de directions conjugues. ESAIM: Math Model Num 3:35–43
36. Polyak BT (1969) The conjugate gradient method in extremal problems. USSR Comp Math Math+ 9:807–821
37. Powell MJD (1986) Convergence properties of algorithms for nonlinear optimization. SIAM Rev 28(4):487–500
38. Raghavachari K (1984) Rotational potential surface for alkanes: Basis set and electron correlation effects on the conformations of nbutane. J Chem Phys 81:1383–1388
39. Rocha BG, Freire RO, Simas AM, Stewart JJP (2006) RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. J Comput Chem 27:1101–1111
40. Rodríguez A, Oliva C, González M (2010) A comparative QM/MM study of the reaction mechanism of the Hepatitis C virus NS3/NS4A protease with the three main natural substrates NS5A/5B, NS4B/5A and NS4A/4B. Phys Chem Chem Phys 12:8001–8015

41. Selvaraj B, Buckel W, Golding BT, Ullmann GM, Martins BM (2016) Structure and function of 4-hydroxyphenylacetate decarboxylase and its cognate activating enzyme. J Mol Microb Biotech 26:76–91
42. Sinclair JE, Fletcher R (1974) A new method of saddle-point location for the calculation of defect migration energies. J Phys C Solid State 7:864–870
43. Smith GD, Jaffe RL (1996) Quantum chemistry study of conformational energies and rotational energy barriers in n-alkanes. J Phys Chem 100:18,718–18,724
44. Spiwok V, Šučur Z, Hošek P (2014) Enhanced sampling techniques in biomolecular simulations. Biotechnol Adv 33:1130–1140
45. Sylvester JJ (1852) A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. Philos Mag 4:138–142
46. Zheng M, Xu D (2013) Catalytic mechanism of hyaluronate lyase from streptococcus pneumonia [corrected]: Quantum mechanical/molecular mechanical and density functional theory studies. J Phys Chem B 117:10,161–10,172

Manuscript C

Structural and Biophysical Analysis of the Phytochelatin-Synthase-Like Enzyme From *Nostoc* sp. Shows that its Protease Activity is Sensitive to the Redox State of the Substrate

Florian J. Gisdon, Christian G. Feiler, Oxana Kempf, Johannes M. Foerster, Jonathan Haiss, Wulf Blankenfeld, G. Matthias Ullmann, Elisa Bombarda, *ACS Chem. Biol.* 2022, 17, 4, 883–897
DOI: 10.1021/acschembio.1c00941

pubs.acs.org/acschemicalbiology

Articles

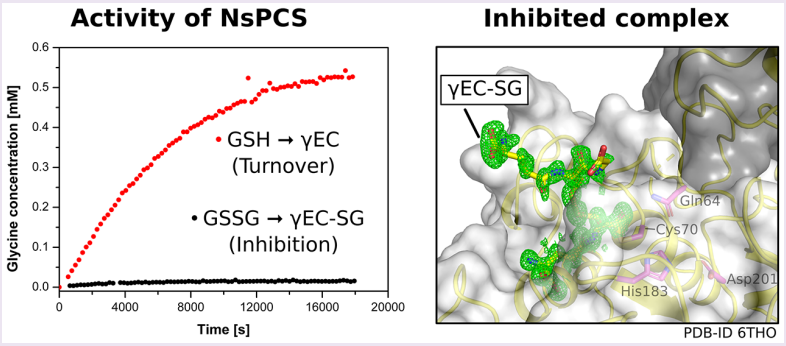
Structural and Biophysical Analysis of the Phytochelatin-Synthase-Like Enzyme from *Nostoc* sp. Shows That Its Protease Activity is Sensitive to the Redox State of the Substrate

Florian J. Gisdon, Christian G. Feiler, Oxana Kempf, Johannes M. Foerster, Jonathan Haiss, Wulf Blankenfeldt, G. Matthias Ullmann, and Elisa Bombarda*

Cite This: *ACS Chem. Biol.* 2022, 17, 883–897

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Phytochelatins (PCs) are nonribosomal thiol-rich oligopeptides synthetized from glutathione (GSH) in a γ-glutamylcysteinyl transpeptidation reaction catalyzed by PC synthases (PCSs). Ubiquitous in plant and present in some invertebrates, PCSs are involved in metal detoxification and homeostasis. The PCS-like enzyme from the cyanobacterium *Nostoc* sp. (NsPCS) is considered to be an evolutionary precursor enzyme of genuine PCSs because it shows sufficient sequence similarity for homology to the catalytic domain of the eukaryotic PCSs and shares the peptidase activity consisting in the deglycination of GSH. In this work, we investigate the catalytic mechanism of NsPCS by combining structural, spectroscopic, thermodynamic, and theoretical techniques. We report several crystal structures of NsPCS capturing different states of the catalyzed chemical reaction: (i) the structure of the wild-type enzyme (wt-NsPCS); (ii) the high-resolution structure of the γ-glutamyl-cysteine acyl-enzyme intermediate (acyl-NsPCS); and (iii) the structure of an inactive variant of NsPCS, with the catalytic cysteine mutated into serine (C70S-NsPCS). We characterize NsPCS as a relatively slow enzyme whose activity is sensitive to the redox state of the substrate. Namely, NsPCS is active with reduced glutathione (GSH), but is inhibited by oxidized glutathione (GSSG) because the cleavage product is not released from the enzyme. Our biophysical analysis led us to suggest that the biological function of NsPCS is being a part of a redox sensing system. In addition, we propose a mechanism how PCS-like enzymes may have evolved toward genuine PCS enzymes.

INTRODUCTION

Phytochelatins (PCs) are cysteine-rich nonribosomal peptides involved in metal homeostasis and detoxification with the typical structure $(\gamma\text{-GluCys})_n\text{Gly}$ (n between 2 and 4).[1] PCs are synthesized by the enzyme PC synthase (PCS) by linking glutathione under the release of glycine. At the first glance, this enzyme shows high resemblance to cysteine proteases. However, the catalytic repertoire of PCS goes beyond the hydrolytic cleavage of peptides because it can also work as a transpeptidase under physiological conditions and thus forms peptide bonds without the use of ATP. In eukaryotes, PCS appears to be ubiquitous in the plant kingdom.[2,3] It is also present in many invertebrates such as protozoa and nematodes,[4−6] but it has not been found in vertebrates. This peculiar distribution makes PCS an interesting drug target against parasitic representatives of these animal groups.[7] In prokaryotes, a number of cyanobacteria and proteobacteria[8] possess genes that encode for proteins that show approximately 30% similarity to the PCS consensus Pfam domain 05023[9] but no significant similarity to any other group of proteins. Thus, these bacterial proteins have been assigned as

Received: November 30, 2021
Accepted: March 14, 2022
Published: April 4, 2022

being PCS-like. Similar to PCS, PCS-like proteins use GSH as the substrate. However, while they are able to cleave off the glycine residue to form γ-Glu-Cys (γEC), the formation of PC in PCS-like proteins was detected only in low amounts and thus the transpeptidase activity of these proteins is still debated.[10−13]

Several features allow us to assign PCS and its prokaryotic homologue to clan CA of cysteine peptidases, whose archetype is papain according to the peptidase database MEROPS (https://www.ebi.ac.uk/merops/, release 12.4).[14,15] The striking feature of clan CA is the presence of a catalytic triad with a catalytic cysteine assisted by a histidine and an asparagine (as in papain) or an aspartic acid (as in PCS), which seems to be a permissible substitution from a catalytic viewpoint.[16] Moreover, these key residues in clan CA peptidases are not only the same but they also follow the same order in the sequence, as would be expected for divergent evolution from a common ancestor rather than convergent evolution from unrelated progenitors.[17,18] The catalytic triad of the PC synthase of *Nostoc* sp. (NsPCS) is Cys70-His183-Asp201, as confirmed by crystal structures[19] and is in line with the catalytic triad Cys25-His162-Asn175 in papain.[20,21] The catalytic cysteine (Cys70 in NsPCS) is conserved with no exception within all eukaryotic and prokaryotic PCS enzymes, supporting its essential role in catalysis. The substitution of only one of the residues of the catalytic triad abolishes the enzymatic activity.[22] In particular, the mutation of Cys70 to Ser in NsPCS was shown to inactivate the enzyme.[23]

The eukaryotic PCS consists of two domains: the N-terminal conserved catalytic domain and the more variable C-terminal domain.[24] Prokaryotic PCSs are homologous to the N-terminal domain but miss the C-terminal domain. A sequence alignment shows that NsPCS possesses 34.5% identity and 53.0% similarity with the N-terminal domain of PCS from *Arabidopsis thaliana* (AtPCS1). Notably, the truncated N-terminal domain of AtPCS1 is sufficient for catalysis, that is, it cleaves GSH into γ-Glu-Cys and Gly, and it is even able to synthetize PCs from GSH in media containing Cd^{2+}.[25] It is therefore not surprising that the ability to cleave GSH is shared by the prokaryotic PCS.

Although extensively studied, the spectrum of functions of PCS is still a matter of debate.[13,17,26] The sole role in heavy metal detoxification seems not sufficient to explain the ubiquity of this protein in the plant kingdom as well as in yeast and nematodes. Amounting evidence supports the hypothesis that PCS serves functions besides cadmium and arsenic detoxification, for example, through roles in essential metal homeostasis[3] and in innate immunity.[27] The role of the PCS-like proteins is even less clear because they are seemingly not able to produce larger amounts of PC. In order to better identify the tasks of PCS-like proteins, their detailed mechanistic characterization is a priority.

In this work, we focus on the alr0975 protein from *Nostoc* sp. strain PCC 7120 (syn. *Anabaena* sp. strain PCC 7120),[28] which is a PCS-like protein[10] that we name NsPCS. We report several crystal structures of the NsPCS: (i) the structure of wild-type enzyme (wt-NsPCS); (ii) the high-resolution structure of the γ-glutamyl-cysteine acyl-enzyme intermediate (acyl-NsPCS); and (iii) the structure of an inactive variant of NsPCS, with the catalytic cysteine mutated into serine (C70S-NsPCS). Furthermore, we analyze the affinity of the enzyme for the substrate and the catalytic activity experimentally and support this work by theoretical calculations. This investigation provides the first quantitative analysis of the enzymatic mechanism of a PCS-like protein, with some implications for possible roles of such enzymes in prokaryotes.

RESULTS

Our goal is to dissect the mechanism of the peptidase reaction catalyzed by NsPCS and to identify its biological implications. To reach a deep understanding of the enzymatic mechanism, it is necessary to relate the thermodynamic and kinetic data to the structural features of the enzyme. In order to merge all these different aspects, it is important to approach the enzymatic mechanism also from a theoretical point of view. With electrostatic and quantum chemical calculations, it is possible to gain insights into enzymatic mechanisms.[29] To this purpose, the knowledge of the 3D structure of the enzyme capturing different states of the chemical reaction is a prerequisite.

Crystal Structure of NsPCS with and without the Substrate. We cloned the gene of alr0975 from *Nostoc* sp. strain PCC 7120 excluding the previously predicted signal sequence responsible for the periplasmic secretion of the protein,[10] see the Supporting Information. Additionally, we introduced the mutation Cys70 to Ser (C70S-NsPCS) in order to analyze the reasons of its inactivity[23] at the atomic level and to investigate binding independent from catalysis. The structure of both variants, wt-NsPCS and C70S-NsPCS have been determined by X-ray crystallography. In addition, we crystallized an acyl-form of NsPCS (acyl-NsPCS), which has the γEC moiety covalently bound. At variance to previous work,[19] in which the acyl-enzyme was crystallized at an extremely acidic pH (pH 2.6−3.4), we obtained these crystals at a significantly higher pH (pH 5.5). Moreover, the higher resolution of our acyl-enzyme structure allows us to obtain more details about the bound ligand. We could not only confirm the oxidation at the sulfur atom of the cysteine of the covalently bound γEC moiety that was previously reported[19] but we could also resolve a second glutathione bound via a disulfide bond to the acylated γEC moiety. The most relevant features of the reported crystal structures are represented in Figure 1.

The structure of the native protein (PDB ID: 6TH5) has a resolution of 1.99 Å with R_{work} = 19.9% and R_{free} = 23.8% and is very similar to the one previously resolved (PDB ID: 2BTW).[19] The structure of the acyl-form (PDB ID: 6THO) has a higher resolution, 1.09 Å (R_{work} = 11.8%; R_{free} = 14.8%) than the one previously resolved (PDB ID: 2BU3).[19] The structure of the serine mutant C70S-NsPCS (PDB ID: 6TJL) was refined to a resolution of 1.87 Å (R_{work} = 20.5%, R_{free} = 26.8%). Data collection statistics are summarized in Table 1.

All the three NsPCS structures are homodimers, displaying root-mean-square deviations (RMSDs) lower than 0.25 Å between the two chains [RMSD(C70S-NsPCS) = 0.19 Å, RMSD(wt-NsPCS) = 0.19 Å, RMSD(acyl-NsPCS) = 0.23 Å].

The structure of the mutated protein C70S-NsPCS is nearly identical to that of the native enzyme, indicating that rather than structural changes, the substitution of the sulfur atom of the cysteine by the oxygen of a serine is the key point for the loss of the activity.

Unexpectedly, we observed electron density for one molecule of glutathione (GSH) in the active site of C70S-NsPCS, even if GSH was not added during crystallization. The ligand therefore seems to have been co-purified after

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

ACS Chemical Biology pubs.acs.org/acschemicalbiology Articles

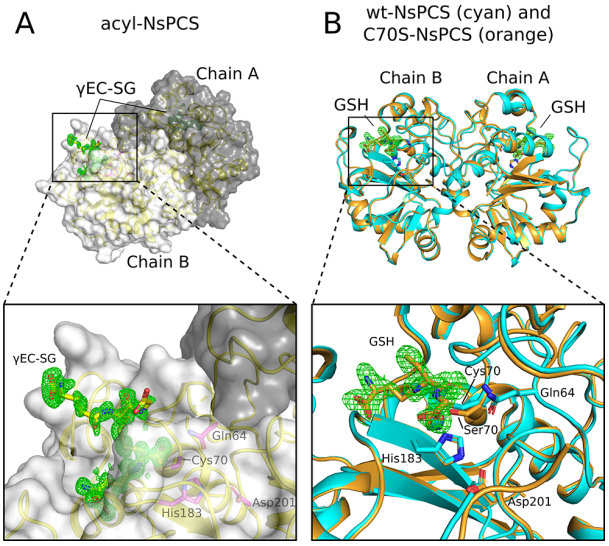


Figure 1. (A) Binding pocket of acyl-NsPCS (PDB ID: 6THO). Active site residues are shown as magenta sticks. The electron density of the mono-deglycinated oxidized form of glutathione (γEC-SG) is shown. (B) Superimposition of wt-NsPCS (PDB ID: 6TH5) and C70S-NsPCS (PDB ID: 6TJL). The ligand GSH is present in the mutant structure C70S-NsPCS and is represented with its electron density. For all structures, the $F_o - F_c$ electron density before incorporation of the substrate (omit map) is represented as a mesh with a sigma-level of 3.

overexpression in *E. coli* and remained tightly bound during purification and crystallization of the protein.

Investigation of Binding of GSH and GSSG to the Active Site. In order to measure the affinity of an enzyme for its substrate, the catalytic activity has to be separated from binding. Therefore, in our case, C70S-NsPCS represents the variant of choice to focus on the binding process. The fact that the uncleaved substrate is stably captured by C70S-NsPCS confirms that this mutation abolishes the activity of the enzyme without precluding binding. On the other hand, the presence of GSH in the binding pocket of the recombinant protein hampers binding studies, and removal of the ligand is an unavoidable prerequisite for such experiments.

In order to eliminate the substrate from the binding pocket, the protein was unfolded, and then dialyzed to remove the substrate from the solution and finally refolded. To check if the native fold is reached, a spectroscopic analysis of the non-refolded (i.e., before unfolding) and refolded proteins was performed.

We recorded the CD spectrum of the C70S-NsPCS during the unfolding/refolding process. As shown in Figure 2, the evolution of the CD signature indicates the loss of the secondary structure during unfolding and its restoration after refolding. An additional confirmation of the occurrence of the unfolding and refolding process was provided by the analysis of the fluorescence signal. Due to the presence of eight tyrosines, an excitation wavelength of 295 nm was chosen in order to selectively excite the tryptophan residues. The emission spectrum of the non-refolded C70S-NsPCS presents a maximum at 336 nm, which indicates a clear hypsochromic shift when compared to the emission of a tryptophan residue in aqueous solution whose maximum is at 350 nm (Figure 2B). Emission at a lower wavelength is typical of nonsolvent

Table 1. Data Collection and Refinement Statistics

	wt-NsPCS	acyl-NsPCS	C70S-NsPCS
PDB ID	6TH5	6THO	6TJL
	Data Collection		
Space group	$P2_1$	$P2_12_12_1$	$P2_1$
a, *b*, *c* (Å)	61.36, 46.67, 68.58	49.28, 57.96, 139.47	61.39, 47.83, 67.9
α, *β*, *γ* (deg)	90.00, 92.26, 90.00	90.00, 90.00, 90.00	90.00, 91.80, 90.00
resolution (Å)	34.26−1.99 (2.07−1.99)[a]	29.21−1.09 (1.10−1.09)[a]	46.25−1.87 (1.92−1.87)[a]
R_{merge} (%)	11.5 (66.5)[a]	5.6 (42.7)[a]	8.4 (84.1)[a]
R_{pim} (%)	9.8 (57.0)[a]	3.4 (26.2)[a]	5.3 (53.3)[a]
$I/\sigma(I)$	6.6 (1.3)[a]	12.1 (2.7)[a]	10.2 (1.34)[a]
completeness (%)	97.4 (98.5)[a]	98.6 (99.3)[a]	99.3 (99.5)[a]
redundancy	2.2 (1.9)[a]	3.6 (3.4)[a]	3.3 (3.3)[a]
$CC_{1/2}$	0.987 (0.845)[a]	0.998 (0.815)[a]	0.997 (0.621)[a]
	Refinement		
number of reflections	26210 (2924)[a]	164587 (5490)[a]	32749 (2683)[a]
R_{work}/R_{free}	0.199 (0.282)[a] /0.238 (0.313)[a]	0.118 (0.200)[a] /0.148 (0.205)[a]	0.205 (0.311)[a] /0.268 (0.323)[a]
	Number of Atoms		
protein	6790	8451	7027
ligand/ion	5	124	74
water	318	768	118
	B-Factors (Å^2)		
protein	26.8	14.3	48.5
ligand/ion	27.8	28.4	43.9
water	28.8	32.0	33.1
	RMSD		
bond length (Å)	0.006	0.007	0.016
bond angles (deg)	0.554	1.042	1.413

[a]Values in parentheses are for the highest-resolution shell.

exposed tryptophan residues in line with the buried location of both tryptophan residues in NsPCS. During the unfolding process we observe the shift of the emission band toward higher wavelengths until reaching the profile of the typical spectrum of fully water-exposed tryptophan residues. This hyperchromic shift reports the loss of structural features leading to the disruption of the hydrophobic core and the consequent exposure of the tryptophan residues to the aqueous solution (Figure 2B). The shift of the emission band was reversed during the refolding process until the emission spectrum regained the original profile, indicating the restoration of the original structure.

The spectroscopic analysis of the refolded protein made us confident that the protein refolds to the native conformation. Nevertheless, we desired to test if the refolded protein is also functional. Because the C70S mutant is inactive, we unfolded and refolded the wild-type enzyme following the same protocol and checked the quality of the procedure with CD and fluorescence spectroscopy, obtaining the confirmation of the restoration of the original structure as for C70S-NsPCS. We performed a real-time ^{1}H NMR experiment to monitor the time course of the production of free glycine at pH 8. The activity of the refolded wt-NsPCS was compared to the activity of the non-refolded wt-NsPCS from the same batch.

The activity assay shows that the refolded protein is active. However, although the refolded protein is able to cleave the

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

ACS Chemical Biology pubs.acs.org/acschemicalbiology Articles

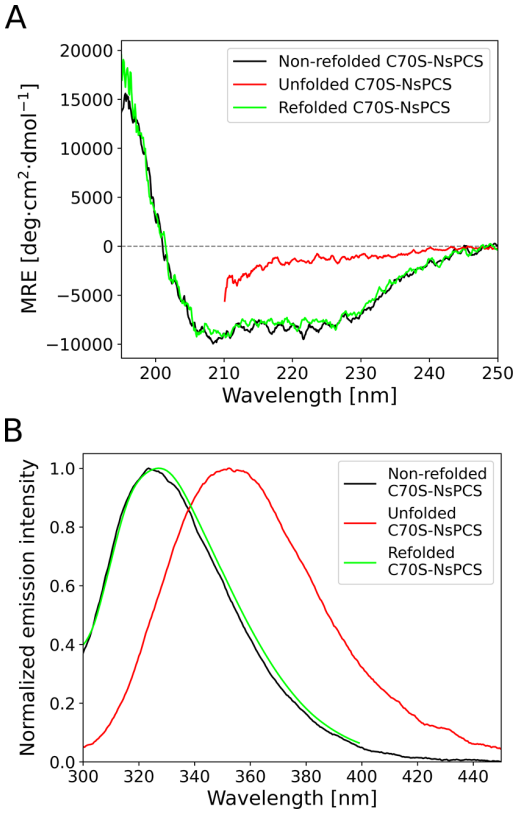


Figure 2. (A) CD spectra of C70S-NsPCS: non-refolded (black), unfolded in 3 M GdmCl (red), and refolded by elimination of GdmCl (green). The spectrum in 3 M GdmCl was cut at λ = 210 nm because the high salt concentration causes a high noise at shorter wavelengths. (B) Normalized fluorescence spectra ($\lambda_{Ex.}$ = 295 nm) of non-refolded (black), unfolded (red), and refolded (green) C70S-NsPCS. Temperature was set at 25 °C.

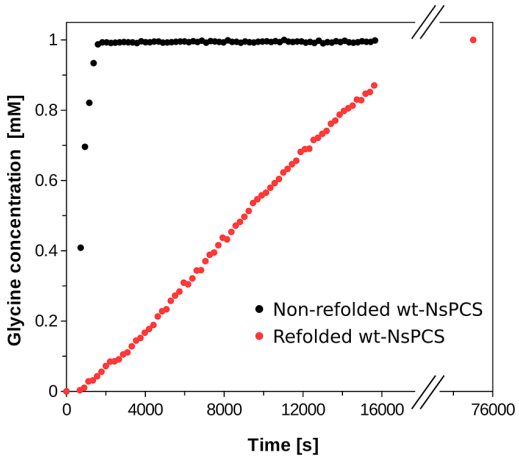


Figure 3. Time-dependent production of free glycine resulting from GSH cleavage mediated by NsPCS. The concentration of free glycine was measured by ^{1}H NMR and quantified by comparison with sodium trimethylsilylpropanesulfonate (DSS) as the internal NMR standard. Because the used GSH concentration has been 1 mM, the final concentration of detected glycine was normalized to 1 mM.

entire amount of the substrate, it appears to work much more slowly than the non-refolded protein. The time dependence of free glycine release (Figure 3) indicates that in the time during which the non-refolded enzyme has processed all the substrate (after about 1600 s), the refolded enzyme processed only about 10% of it. However, both the refolded and the non-refolded enzymes are able to process the entire amount of the substrate.

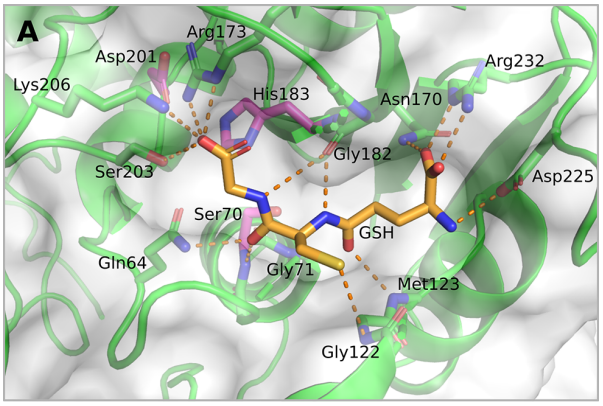
To characterize the binding pocket, we analyzed the hydrogen bond network (Figure 4). The experimental binding study was performed with isothermal titration calorimetry (ITC). First of all, we measured the binding of GSH to the purified recombinant non-refolded C70S-NsPCS, that is, the protein that retained the substrate in the binding site. The thermogram shows only the effect of dilution, confirming that the mutated enzyme is saturated with GSH. Because GSH is often oxidized to GSSG, we decided to determine the affinity of the C70S-NsPCS to both molecules, GSH and GSSG. The binding experiments have been performed under an argon atmosphere to prevent the oxidation of the thiol groups. Because the protein is a homodimer, we initially fitted the experimental data to a two-site model. No significant interaction between the two binding sites has been detected in line with the fact that the two binding pockets are far apart from each other. Therefore, the data have been further analyzed with a one-site model (Table 2).

The affinities of C70S-NsPCS for GSH (5×10^5 M^{-1}) and GSSG (1×10^5 M^{-1}) are similar. The fit of the experimental data indicates that only 10 to 20% of the protein is able to bind the substrate (see parameter p in Table 2). Notably, also the activity assay by NMR showed that a similar small amount of refolded protein retained activity. This similarity between the percentages of active and binding protein was replicated with different batches of protein indicating the significance of this finding.

In order to consolidate the result of ITC we measured the affinity of the refolded C70S-NsPCS for GSH with microscale thermophoresis. First of all, we tested the binding of GSH to the purified recombinant non-refolded C70S-NsPCS which is expected to retain the substrate in the binding pocket. As expected, no binding could be detected because the binding pocket is occupied. Afterward, we measured the binding of GSH to the refolded C70S-NsPCS and obtained a binding constant of $K_a = (8 \pm 4) \times 10^5$ M^{-1}. Noticeably, the binding constant measured by microscale thermophoresis is fully in line with the binding constant measured by ITC (Table 2).

The affinity of the enzyme for the reduced (GSH) and oxidized (GSSG) forms of glutathione is similar within 1 order of magnitude, suggesting that the oxidation of glutathione does not substantially affect binding. This observation prompted us to test if GSSG can be a substrate for the enzyme. We performed a real-time ^{1}H NMR experiment to monitor and compare the time course of the production of free glycine at pH 8 using GSSG and GSH as substrates, respectively. Only a very small amount of free glycine was detected when GSSG was added in the reaction tube containing the enzyme (Figure 5), indicating that GSSG is not an optimal substrate for NsPCS. Additionally, the quantification of the free glycine which is present when the reaction is terminated (plateau of the time trace) reveals that the concentration of the free glycine is similar to the concentration of active sites of the enzyme present in the reaction tube. This result indicates that the free glycine was produced in a stoichiometric amount and leads us to suggest that each monomer of the enzyme interacts with one molecule of GSSG and is able to cleave one of its two

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897



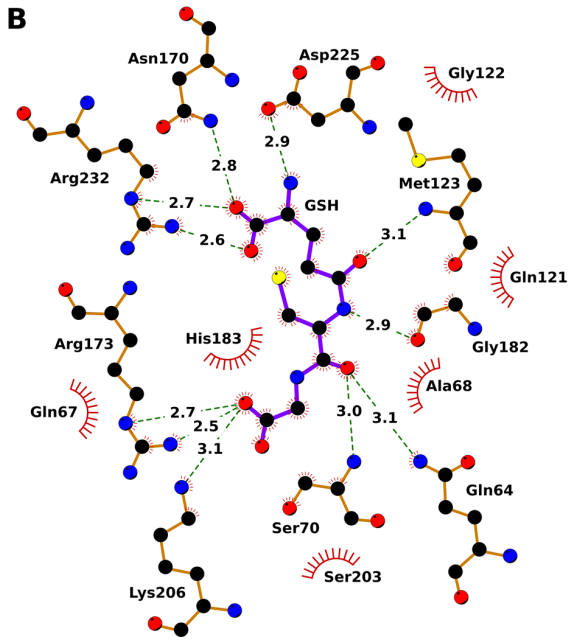


Figure 4. Hydrogen bond network of GSH in the binding pocket of NsPCS. (A) Possible hydrogen bond interactions of GSH with NsPCS in the structure C70S-NsPCS (PDB ID: 6TJL), subunit A. Interaction possibilities were obtained with the in-house program Hbond. (B) Binding site representation of GSH in the binding pocket of NsPCS, subunit A calculated with LigPlot+ (https://www.ebi.ac.uk/thornton-srv/software/LigPlus/, Version1.4).[30]

Table 2. Thermodynamic Quantities Obtained from ITC[a]

	C70S-NsPCS + GSH	C70S-NsPCS + GSSG
ΔH [kcal mol^{-1}]	−12 ± 1	−11.6 ± 0.4
K_a [M^{-1}]	(5 ± 3) × 10^5	(1.0 ± 0.2) × 10^5
ΔS [kcal mol^{-1} K^{-1}]	−0.014	−0.016
ΔG [kcal mol^{-1}]	−7.9	−6.9
R^2	0.935	0.990
p	0.20 ± 0.01	0.118 ± 0.003

[a]The correction parameter p corresponds to the fraction of binding protein (see Material and Methods). ΔH and K_a are experimentally determined. ΔS and ΔG are calculated from $\Delta G = -RT \ln K_a$ and $\Delta G = \Delta H - T\Delta S$.

glycine residues, without the subsequent release of the γEC-SG moiety. This hypothesis is supported by the crystal structure of

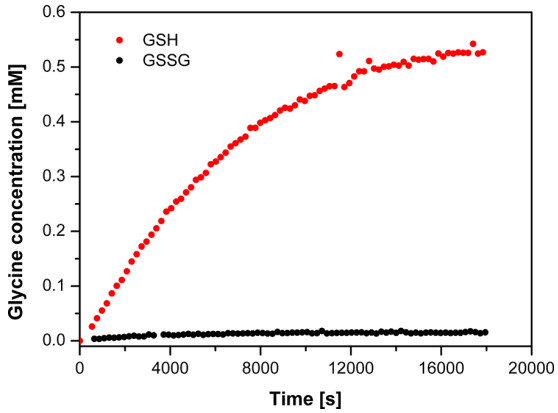


Figure 5. Time-dependence of the production of free glycine resulting from the substrate cleavage, GSH (red dots), and its oxidized form GSSG (black dots). The concentration of free glycine was measured under the anoxic condition by ^{1}H NMR and quantified by comparison with DSS as an internal NMR standard.

acyl-NsPCS which reveals that γEC-SG and not γEC is bound in the acyl-enzyme.

Electrostatic Analysis of the Substrate Binding. Our experimental observations indicate that NsPCS catalyzes the cleavage of GSH but is trapped in the acylated state when GSSG is used as the substrate. In order to understand the different reactivities of the enzyme, we performed electrostatic calculations. These calculations are particularly suited to this purpose because, due to the charges present on both GSH and GSSG, the binding of both molecules to the enzyme has an exquisite electrostatic character. For our calculations, we used the crystal structures C70S-NsPCS and acyl-NsPCS to create models of the complexes with GSH, GSSG, and their cleaved forms γEC and γEC-SG, respectively. All the molecules have been modeled as noncovalently bound in the binding pocket of the enzyme (see Material and Methods for a detailed description of the procedure).

We calculated the electrostatic interaction energies between the enzyme and the analyzed ligands using the Poisson−Boltzmann equation (see Table 3). The so-calculated energies contribute mainly to the enthalpic part of the binding energy.

The electrostatic interactions stabilize the binding of GSSG and γEC-SG by more than 20 kcal mol^{-1}, whereas only −16.5 and −14.1 kcal mol^{-1} are the electrostatic energies for the binding of the reduced forms GSH and γEC, respectively. The complex NsPCS + γEC was modeled in two ways and slightly different electrostatic interaction energies have been obtained:

Table 3. Electrostatic Contribution to the Interaction Energy between NsPCS and Different Variants of Its Substrate

crystal structure	structural model	electrostatic contribution (kcal mol^{-1})
C70S-NsPCS (GSH bound)	NsPCS + GSH	−16.5
	NsPCS + γEC	−14.1
	NsPCS + GSSG	−22.4
	C70S-NsPCS + GSH	−16.3
acyl-NsPCS (γEC bound)	NsPCS + γEC	−18.2
	NsPCS + γEC-SG	−22.2

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

−14.1 kcal mol^{-1} for the complex based on the crystal structure of C70S-NsPCS and −18.2 kcal mol^{-1} for the complex based on the crystal structure of the acyl-NsPCS (Table 3). The difference between the interaction energy calculated for the complex NsPCS + GSH (−16.5 kcal mol^{-1}) and the one calculated for the complex NsPCS + γEC (−14.1 kcal mol^{-1}) can be considered as the interaction of the glycine moiety of GSH with PCS showing that γEC when hydrolytically cleaved from C70 interacts less strongly with the enzyme than GSH.

In addition, we have simulated the encounter of GSH with NsPCS using a Monte Carlo technique.[31] We considered three forms of NsPCS differing in the occupancy of the binding pocket: (i) NsPCS bearing empty binding pockets, (ii) NsPCS with bound GSH in the binding pockets, and (iii) NsPCS with one acylated binding pocket and GSH bound in the other (see Figure 6). In NsPCS with empty binding pockets, we could observe the formation of the encounter complex in the area around the active site (Figure 6A). In contrast, NsPCS with GSH bound in both binding pockets shows no specific encounter complex formation (Figure 6B), indicating that the binding of GSH prevents another GSH molecule from approaching the active site region. Finally, Figure 6C shows the results of the calculation in which γEC is covalently bound in one pocket of the homodimer and GSH is noncovalently bound in the other pocket. In line with the finding for NsPCS with GSH bound to both binding pockets, at the pocket with bound GSH, no encounter complex formation is observed. Instead the pocket with bound γEC, representing the acyl-enzyme state, shows an encounter complex formation.

Michaelis−Menten Analysis of the Catalytic Reaction. In order to measure the catalytic activity, we used the wild-type enzyme. We focused on the peptidase activity that is the only activity that is doubtlessly recognized to be performed by NsPCS. The reaction consists in the cleavage of glycine from the substrate (GSH) concomitant to the formation of the acyl-enzyme intermediate, followed by the hydrolysis of the acyl-adduct.

We designed a new test to quantify the peptidase activity of NsPCS based on real-time ^{1}H NMR under anoxic conditions. Due to the fast oxidation of the substrate GSH, the catalytic activity of NsPCS at pH 8 was highly sensitive to oxygen exposure. Therefore, we elaborated a protocol to eliminate the presence of oxygen in the reaction mixture (see Materials and Methods for details).

The deglycination of GSH was monitored and a Michaelis−Menten analysis was performed. For comparison, the reference spectra of the substrate GSH, its oxidized form GSSG, and the expected products γEC and free glycine have been measured. During the reaction, the newly appearing resonances can be attributed to free glycine and γEC and the disappearing resonance corresponds to GSH. No resonance indicating oxidation of GSH was detectable during the 4 h of the experiment, indicating that the protocol to maintain the anoxic condition was successful. The time-dependence of the accumulation of free glycine is used to estimate the initial velocities of the reaction. A plot of the initial velocity versus the substrate concentration displays the typical Michaelis−Menten behavior (Figure 7), with a Michaelis−Menten constant K_M of 0.2 mM and a turnover number k_{cat} of 3.5 s^{-1}.

QM/MM Analysis of the Reaction Mechanism. NsPCS belongs to the superfamily of papain-like cysteine peptidases. Hence, it is expected to cleave its natural substrate GSH according to a mechanism similar to the one of papain. To test this assumption, we performed quantum mechanics/molecular mechanics (QM/MM) calculations on both enzymes, NsPCS and papain. In the case of papain, we chose the substrate Phe-Ser-Ile with an acetylated N-terminus and N-methylated C-terminus based on previous mechanistic studies and experimental binding data.[32]

The peptidase reaction performed by both enzymes can be divided into two steps: the acylation reaction and the hydrolytic cleavage. In NsPCS, the acylation reaction consists in the cleavage of the substrate resulting in glycine and γEC. The latter remains covalently linked to the enzyme as so-called acyl-adduct. In papain, the investigated acylation reaction

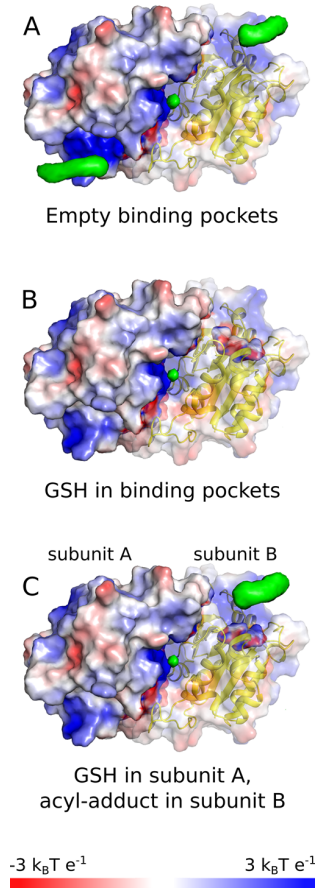


Figure 6. Comparative representation of encounter complex densities (shown in green) of GSH around NsPCS. NsPCS is shown in the surface representation, colored by electrostatic potential on the solvent accessible surface. One subunit of each homodimer is opaque, and the other is transparent, with a translucent cartoon representation of the protein. The electrostatic potential was calculated with 0.1 M ionic strength. (A) NsPCS with empty binding pockets. Encounter complexes form at both binding pockets. (B) GSH bound to both binding pockets. No encounter complexes form near the binding pockets indicating that a second GSH molecule is not able to approach the active site when a GSH is already bound in the binding pocket. (C) GSH bound to subunit A and γEC bound covalently to subunit B. The simulation shows that after deglycination, a second GSH molecule can approach the active pocket in subunit B. All the simulations were performed with 100,000 runs and 1,000,000 Monte Carlo steps each. The isovalue is set to 5500 and describes the minimal number of encounters which occurred as a visible green surface.

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

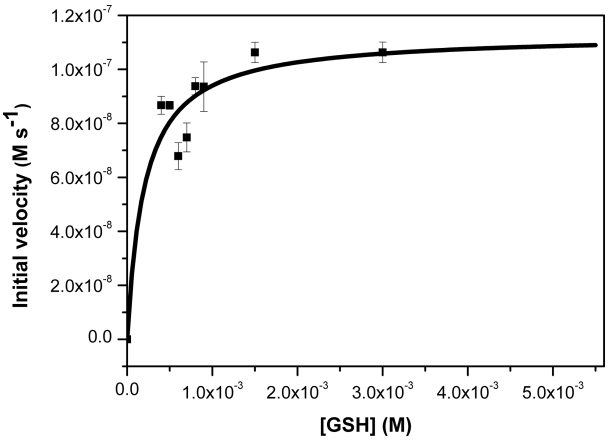


Figure 7. Plot of the initial velocities of the peptidase reaction catalyzed by NsPCS as a function of the concentration of the substrate GSH. For each substrate concentration, the accumulation of free glycine is monitored and the initial velocity of the reaction is determined and plotted against the concentration of the substrate. The solid line represents the fit of the experimental data according to the Michaelis−Menten equation.

consists in the cleavage of the substrate between serine and isoleucine. Similarly, the latter remains covalently linked to the enzyme. In both enzymes, a subsequent hydrolytic reaction leads to the cleavage of the acyl-adduct. The calculations for the acylation reaction are performed on the crystal structure of C70S-NsPCS, which has the full substrate GSH bound to its active site and therefore represents the ideal starting point to study the cleavage reaction. In order to simulate the enzymatic activity, the serine was converted computationally into the native cysteine. The calculations for the deacylation have been performed on the crystal structure of the acyl-enzyme. We compared the obtained energy profiles with the energy profile of the corresponding reaction steps in papain (Figure 8).

The acylation reaction can be divided into two steps: (i) the activation of the nucleophile by proton transfer from cysteine to the catalytic histidine resulting in the ion-pair state (Figure 8, reaction coordinates 1−3) and (ii) the nucleophilic attack and cleavage of the substrate via a tetrahedral state (Figure 8, reaction coordinates 3−7). The comparison of the reaction profiles of the two enzymes shows that the energy barrier for the activation of the nucleophile in NsPCS (6.1 kcal mol^{-1}) is higher than in papain (0.9 kcal mol^{-1}). Additionally, the ion-pair state is less stable in NsPCS than in papain, as the less pronounced minimum in the energy profile of NsPCS indicates (reaction coordinate 3). In papain, the catalysis proceeds via the formation of a stable tetrahedral intermediate state (reaction coordinate 5). The values of the classical potential energy suggest that also the reaction path for NsPCS displays a tetrahedral intermediate, which reacts further to form the acyl-enzyme state; nevertheless, the occurrence of a stable tetrahedral intermediate for NsPCS is not observed after zero-point-energy correction. The zero point energy of a quantum mechanical system accounts for the ground-state fluctuations according to the Heisenberg uncertainty principle; therefore, after the zero-point-energy correction, the energy of

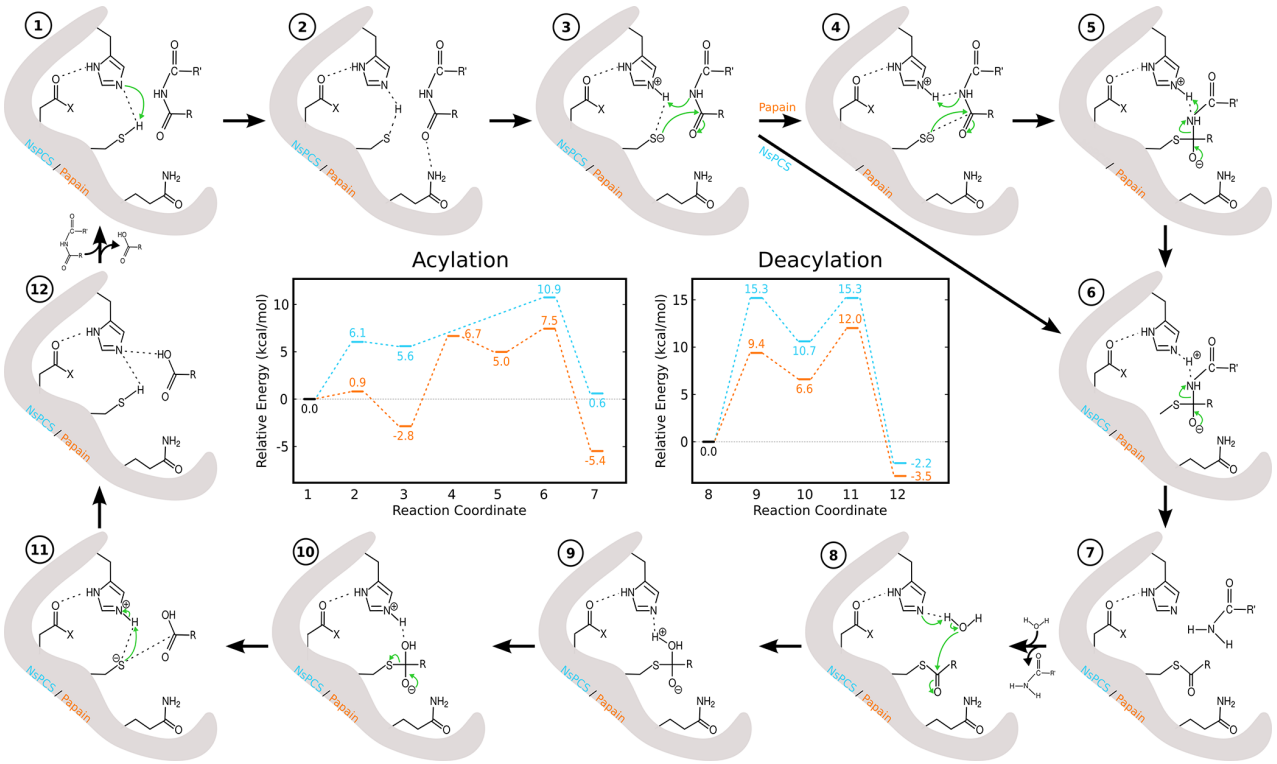


Figure 8. Comparison of the enzymatic mechanism of NsPCS (cyan) and papain (orange). Comparable states along the reaction coordinate are shown in the energy diagrams and structurally represented with the corresponding numbers. States 4 and 5 have been identified only for papain and not for NsPCS. The energies were calculated with a QM/MM approach and zero-point energy corrected. The reaction coordinate corresponds to the number of the states.

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

the system is higher than its energy in the classical limit, which is the mathematical minimum of the potential well. Because different states may have different zero point energies, energetically close states in the reaction path may appear indistinguishable, when zero-point-energy correction is applied. According to the Eyring equation, the rate constants at 303.15 K for the rate-limiting step in the acylation reaction of NsPCS and papain are 9.2×10^4 and 4.6×10^7 s^{-1}, respectively, suggesting a faster process in papain. The calculated rate constants represent the actual cleavage reaction without binding of the substrate GSH and release of the cleaved-off glycine. However, substrate binding and product release contribute to the measured rate constants. Thus, only a qualitative comparison with the experimentally obtained rate constants is meaningful, also because we consider only the enthalpic and not the entropic contribution to the energy barrier.

During the hydrolytic cleavage of the acyl-enzyme, a water molecule in the active site attacks and cleaves the thioester bond of the acyl-enzyme, leading to the release of γEC in NsPCS and Phe-Ser in papain. The hydrolytic cleavage can be divided into two steps: (i) the formation of a tetrahedral state (Figure 8, reaction coordinates 8−10) and (ii) the cleavage of the covalent enzyme substrate complex and re-protonation of the catalytic nucleophile (Figure 8, reaction coordinates 10−12). The energy profiles of NsPCS and papain look similar. The two identified transition states and the intermediate of the reaction catalyzed by NsPCS are roughly 3−6 kcal mol^{-1} higher than the corresponding state in the reaction catalyzed by papain. For the hydrolytic cleavage, the rates of NsPCS and papain calculated according to the Eyring equation at 303.15 K are 6.3×10^1 and 1.3×10^4 s^{-1}, respectively.

DISCUSSION

The function of an enzyme is interconnected to its catalytic behavior, which has its rationale in the conformational and chemical features of the enzyme at the atomic level. Therefore, in our attempt to understand the mechanism of NsPCS and to identify its biological functions, we started by determining the crystal structure of three relevant forms of the enzyme: the ligand-free enzyme, the acyl-enzyme, and a mutant enzyme with the substrate bound noncovalently.

New Structural Features in NsPCS and the Acyl-Enzyme. The structure of the native protein (NsPCS) resolved in this study is similar to the one in the literature[19] (RMSD of 0.28 Å). However, a careful optimization of the hydrogen-bond network led to the re-orientation of several sidechains in comparison to the previous work, in particular for amino acids close to the active site, such as Asn170 or Gln67. The structure of the acyl-enzyme is very well superimposable to the structure of the native enzyme. We could resolve all loop regions, including the loop comprising the residues 83−106, which was partly undefined in the previous structure[19] (PDB ID: 2BU3). Because of its vicinity to the active site of the other monomer, this loop was previously called "protruding loop" and its incompleteness was attributed to the flexibility required by the catalytic reaction. However, the lack of unresolved regions in our structure together with the absence of a significant structural difference between the acylated and the ligand-free forms of the enzyme indicates that the catalytic reaction does not require much flexibility in the active site and the active site of NsPCS is rather preshaped to accommodate the substrate. Additionally, because we do not detect significant structural differences between the protruding loop regions in both monomers, we have no reason to invoke mechanistic differences between the two monomers.

Based on our binding study, we exclude a cooperative behavior between the two monomers at variance to suggestions of Vivares et al.[19] This conclusion is supported by our structural study. In fact, the two binding pockets are far apart from each other and the apparent absence of mechanistic differences between the two monomers pleads for their independence.

We succeeded to crystallize the acyl-enzyme at pH 5.5, which is significantly higher than the pH at which the acyl-enzyme was previously trapped.[19] Therefore, we think that the acidic pH and the consequent protonation of the catalytic histidine is not the reason for the stabilization of the acyl-enzyme, as previously suggested. In fact, the electrostatic calculations indicate that His183 titrates at extremely low pH (see the Supporting Information). Moreover, in our high-resolution structure, we could detect a second molecule of GSH bound via a disulfide bond to the acylated γEC moiety. Therefore, we suggest that the reason of the stabilization of the acyl-enzyme resides in the larger size of the ligand (γEC-SG instead of γEC) and thus in the larger number of interactions. The hydrolytic cleavage reaction occurs through the formation of a tetrahedral complex due to the arrival of a water molecule attacking the thioester bond. In contrast to the structure of Vivares et al.,[19] where a water molecule was found only in one monomer, we see this water molecule ideally placed in both monomers.

The substrate is retained in the binding pocket of the inactive mutant C70S-NsPCS.

Unexpectedly in the crystal structure of the catalytically inactive C70S-NsPCS, the active site is occupied by a molecule of GSH. Considering that GSH was not added during crystallization, we assume that the protein has been exposed to GSH during expression in *E. coli*. Because the mutation impairs activity, the substrate is not processed. Surprisingly, it is also not released during purification. This finding suggests a tight complex between the enzyme and the substrate, in contrast to what was previously insinuated on the basis of the Michaelis−Menten analysis of the catalytic mechanism of PCS from *Silene cucubalus*.[33] However, in order to quantify the affinity of an enzyme for its substrate, it is required to separate binding from catalysis. For this purpose, the C70-NsPCS was taken as the variant of choice to focus on binding. However, the ideality of this variant is partly downscaled by the necessity to remove the substrate from the binding pocket through the invasive procedure of unfolding/refolding. Although successful in regaining the original conformation, the unfolding/refolding procedure showed one drawback: when applied to the wild-type, the enzymatic activity slowed down significantly. The hampered catalysis indicates that the features responsible for the enzymatic activity were not properly restored during the refolding process, at least not in all molecules. However, because no difference in the spectroscopic properties of the refolded and non-refolded proteins emerged, these unrestored features are such to be unable to affect the spectroscopic properties. We can explain the reduced velocity of cleavage in two ways: (i) the entire amount of the refolded enzyme is able to process the substrate at a lower speed or (ii) the kinetics of the catalytic reaction is not modified, but only a small amount of the refolded protein is active. We consider the second hypothesis more likely because the percentage of refolded

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

enzyme able to bind the substrate (as shown by the binding study) correlates well with the amount of refolded enzyme for which activity was restored. We conclude that the refolding protocol is able to generate a small amount of enzyme which is fully functional and suitable for further analysis.

Oxidation of GSH Does Not Substantially Affect Binding but It Drastically Affects Activity. The affinity of GSH to C70S-NsPCS is 5×10^5 M^{-1} and no cooperative behavior was observed between the two monomers. This finding is not surprising because the two binding pockets are far apart from each other. The binding constant is higher than believed in earlier studies.[33] A relatively high binding constant is important because NsPCS is found in the periplasm where the concentration of GSH is not as high as in the cytoplasm.[34,35] However, the binding constant is not as high as expected, considering that the substrate GSH remains in the binding pocket during purification.

The enthalpic contribution to the binding interaction is very high (−12 kcal mol^{-1} for GSH binding, see Table 2), pleading for a strong interaction. Accordingly, the entropic component of the binding energy is small, which can be explained with a compensation between favorable and unfavorable contributions. On the one hand, the binding pocket is largely solvent exposed which minimizes the favorable entropic contribution due to water release. On the other hand, the binding pocket with the ligand bound in the C70S-NsPCS structure is superimposable to the empty binding pocket in the wild-type structure, indicating that the binding pocket is prepared to accommodate the substrate; hence, the unfavorable entropic contribution due to the loss of conformational degrees of freedom upon binding is also expected to be marginal. However, thermodynamic reasons are not sufficient to justify the persistence of GSH in the binding pocket during purification steps and kinetic arguments have to be evoked. In fact, if the substrate is not released, it means that the dissociation is sufficiently slow not to occur during the time of the purification. Normally, a very low dissociation rate is associated with a high affinity constant. This statement, however, is not exclusive. In fact, reminding that the association equilibrium constant can be expressed as the ratio of the association and dissociation rates as follows, $K = k_{on}/k_{off}$, the dissociation process can be slow also with a moderate equilibrium constant, provided the association process is likewise slow. According to the affinity constant of NsPCS for its substrate, the overall dissociation rate, k_{off}, will be about 6 orders of magnitude slower than the overall association rate, k_{on}. Therefore, if k_{on} is sufficiently slow, k_{off} can potentially compete with the time of the purification process.

In our case, the purification process lasts about 3 days, that is, 2.6×10^5 s. Therefore, strictly speaking, the dissociation rate has to be lower than about 4×10^{-6} s^{-1} to ensure the persistence of the substrate in the binding pocket during these 3 days. To allow this happening and according to an affinity constant of 5×10^5 M^{-1}, the association process needs to be in the order of a second.

Such a slow association can be explained with the necessity of releasing the solvation water prior to binding. In general, water molecules solvating the binding pocket act as a significant obstacle to ligand binding.[36] In our system, this effect is enhanced by the presence of numerous charges in both the substrate and the binding pocket, which render the surface of contact highly polar and therefore strengthen the interaction between the molecular surface and the water of solvation. However, in order to accommodate GSH in the binding pocket in view of acylation, both the substrate and the binding pocket have to be fully desolvated. The encounter of solvated molecules is a fast process; in contrast, the release of the water of solvation is a slow process,[36] which can be particularly slow when the water molecules are retained in a polar environment, as in NsPCS. The combination of all such processes may result in a very low overall association rate, k_{on}. Additionally, once the molecular partners are desolvated, they will interact strongly, due to their charged profile no more shielded by the solvation shell, reducing the probability of dissociation even more. The strong stabilization of the GSH in the binding pocket is confirmed by the large enthalpy of association measured by ITC (Table 2) and by the interaction energy obtained by the electrostatic calculation (Table 3). These considerations about the kinetic behavior of NsPCS will be tested in further studies.

The affinity of the enzyme for the reduced and oxidized forms of glutathione is similar within 1 order of magnitude. Moreover, the binding of both substrates to the enzyme has a similar enthalpic character, in line with the extended network of hydrogen bonds that is found between protein residues and the substrate (see Figure 4). Our results indicate that the oxidation of GSH does not substantially affect binding, although it drastically affects activity because GSSG acts as suicide inhibitor. We can conclude that γEC-SG stabilizes the acyl-enzyme and inhibits the catalytic turnover. This hypothesis is confirmed by our success to trap the acyl-enzyme with γEC-SG, without applying the extreme acidic conditions that had previously been assumed as necessary for the stabilization of the acyl-enzyme.[19]

Mechanistic Aspects of the Catalytic Reaction. To our knowledge, the only quantitative analysis of the catalytic reaction of a PCS protein up to date is the study of the catalytic mechanism of PCS from *S. cucubalus*[33] in which the dependence of the rate of PC formation on glutathione concentration yielded a K_M value of 6.7 mM and k_{cat} = 0.2 s^{-1} in the presence of 0.1 mM Cd^{2+}. The study of Grill et al.[33] was performed on a eukaryotic PCS and the Michaelis−Menten analysis concerns the production of PC, that is, it is applied to a reaction that includes both peptidase and transpeptidase reactions. Instead, our study is performed on a prokaryotic PCS and concerns the deglycination of GSH, that is, in our case, the reaction consists only in the peptidase activity. Moreover, the effect of oxygen on catalysis was not considered in the study of Grill et al.,[33] and we show here that oxygen plays an important role for the activity of NsPCS.

Despite the catalytic process monitored in this earlier study is different from ours, the Michaelis−Menten parameters are similar. In fact, our analysis of the catalytic reaction of NsPCS leads to a Michaelis−Menten constant K_M of 0.2 mM and a turnover number k_{cat} of 3.5 s^{-1}, which are 1 order of magnitude lower and 1 order of magnitude higher than the respective values reported in the study of the catalytic mechanism of PCS from *S. cucubalus*.[33] This similarity can be rationalized by our finding that the rate-limiting step is most likely the deacylation reaction, which is common to both catalytic processes.

We performed QM/MM calculations to dissect the enzymatic mechanism and to identify the transition states. The comparison between the energy barriers obtained for the acylation and the deacylation leads us to conclude that the deacylation is the slowest chemical step. In the literature, the experimental rate measured for the cleavage of peptidic

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

substrates by papain[37] is about 2×10^3 s^{-1} which is within a factor of six in agreement with our calculated rate 1.3×10^4 s^{-1}. Also in the case of NsPCS, the catalytic rate k_{cat} of 3.5 s^{-1}, resulting from our Michaelis−Menten analysis of the peptidase reaction, is in good agreement with our calculated rate for the deacylation of 6.3 s^{-1}. Because the measured rate reflects the rate-limiting step, the good agreement between the experimental and the calculated rate suggests that deacylation is the rate-limiting step of the peptidase reaction catalyzed by NsPCS.

Our findings indicate that the binding of the substrate to the active site is a high-energy-barrier process, both in association and in dissociation. In fact, the binding of the substrate implies the energy-costly disruption of an extended and well-ordered network of water molecules solvating the binding site. On the other side, the binding of the substrate is stabilized by strong electrostatic interactions as confirmed by the high enthalpy, which implies a high energy barrier also for the complex dissociation. The kinetic consequences of such an energetic profile consist in a slow dissociation rate and a comparatively slow association rate, consistent with the moderate binding constant that was measured.

NsPCS Activity is Sensitive to the Redox State of the Substrate, Suggesting a Potential Function of NsPCS in Cyanobacteria. Our electrostatic calculations contribute to rationalize the catalytic reaction. In fact, on the one hand, we could show that the deglycination of GSH leads to a product (γEC), which binds less strongly than the substrate and therefore can more easily leave the catalytic site to be replaced by a new unprocessed substrate. The cleavage disrupts a covalent bond in the substrate, leading to the destabilization of the complex. Moreover, the cleavage induces an amino group in the cleaved-off glycine which tends to be positively charged, leading to the repulsion of the glycine from the positively charged binding pocket. On the other hand, the tighter complex of NsPCS with GSH compared to the one with the cleaved form γEC appears to be stable enough to allow efficient cleavage also at low concentrations of GSH. A tight complex is in line with the location of NsPCS in the periplasm, where the concentration of GSH is not as high as in the cytoplasm.[34,35]

Our binding study shows that GSH binds the enzyme also when the catalytic cysteine is mutated into serine, as indicated by the presence of the uncleaved substrate GSH in the binding pocket of C70S-NsPCS, revealed by crystallography. The outcome of the electrostatic calculations further explains this finding. In fact, because no deglycination occurs in the inactive serine mutant, the complex will not be destabilized by the loss of glycine and GSH remains trapped.

Our docking simulations show two interesting features. Particularly inspiring is the simulation in which one active site is occupied by GSH, that is, the intact substrate, and the other with γEC covalently bound to Cys70, that is, the state in which the glycine moiety is cleaved off. At the active site with bound GSH, a second GSH molecule is prevented from approaching the active site and thus oxidation of GSH to GSSG cannot occur. At the active site with γEC bound, a second GSH molecule is able to approach the pocket and thus the disulfide bond formation between the second GSH molecule and the bound γEC to form γEC-SG may occur. Accordingly, in the crystal structure of acyl-NsPCS, γEC-SG is found and not γEC. These observations explain also why in the binding site of the crystal structure of C70S-NsPCS, where the substrate cannot be cleaved and the acyl-enzyme is not formed, the reduced GSH and not the oxidized form GSSG is found. Furthermore, the different reactivity of NsPCS for the reduced and the oxidized forms of glutathione indicates that NsPCS is sensitive to the redox state of the substrate leading to the conclusion that the activity of NsPCS depends on the redox potential of the solution. Namely, under reducing conditions, glutathione is present in its reduced form GSH. GSH can bind to NsPCS and is cleaved into γEC and glycine. Thus, under reducing conditions, free glycine is continuously produced. If the conditions are getting more oxidizing, the oxidized form of glutathione GSSG builds up. GSSG binds to NsPCS and one of its two glycines is cleaved off. The so-generated γEC-SG remains covalently bound to the enzyme and inhibits the enzyme as we have shown by kinetic NMR experiments (Figure 5). Alternatively, GSH can bind to the acyl-enzyme via a disulfide bond forming γEC-SG. The result would be the same: NsPCS is inhibited by γEC-SG and no further production of free glycine occurs. Therefore, because NsPCS is not active anymore under oxidizing conditions, GSSG accumulates in the periplasm. In other words, if the redox conditions are sufficiently reducing to maintain glutathione in the reduced form (GSH), the cleavage reaction occurs and glycine is produced by the activity of PCS; instead, in oxidizing conditions, when GSSG is formed, the enzyme is blocked. This scenario suggests that NsPCS may be involved in redox sensing, opening a new hypothesis in the search of the physiological role of the PCS-like enzyme.

These considerations allow further speculation about the physiological role of NsPCS. The analysis of the sequence of NsPCS using SignalP (https://services.healthtech.dtu.dk/service.php?SignalP-5.0)[38] shows that the enzyme has an N-terminal Sec signal peptide, indicating that the enzyme is secreted into the periplasm. The genome of *Nostoc* sp. PCC 7120 is sequenced and annotated (NCBI Accession code: NC_003272[28]). The gene for NsPCS is located in an operon together with the genes for a potential ABC transporter and a hypothetical protein, which seems to be a membrane protein with homology to ABC transporters (see analysis in the Supporting Information). Thus, the following scenario seems possible: one ABC transporter exports GSH to the periplasm, and the other ABC transporter imports the glycine that results from the NsPCS activity to the cytoplasm. If the conditions in the periplasm get oxidizing, the flux of glycine into the cell will stop, providing a signal inside the cell, for example, to switch between oxic and anoxic metabolism. Alternatively, the hypothetical protein may function as a receptor sensing the glycine level directly. Another option would be that GSSG would be the signaling molecule, which might be sensed directly in the periplasm or after being imported into the cytoplasm. The idea that NsPCS is a part of a redox signaling chain could be tested experimentally and may solve the mystery of the function of PCS-like enzymes. This hypothesis may also explain why PCS is not found in insects or vertebrates which are strictly aerobic, but can be found in protozoa or nematodes, which are also able to survive under anoxic conditions.

Furthermore, the rationalization of the binding mode of the substrate with NsPCS allows also speculation about the evolution of genuine PCS in higher organisms because the approach of a second GSH molecule and the formation of a γEC-SG may sporadically lead to transpeptidation and thus to the formation of phytochelatin, which was found as a low-yield

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

product of NsPCS.[10,12,23] If this hypothesis is correct, the potential function of PCS-like enzymes as a redox-sensor may have led to a transpeptidase activity in the genuine PCS and the formation of longer PCs during evolution.

CONCLUSIONS

The enzyme PCS in plants is known to synthesize the metal chelating peptide PC and thus is involved in metal homeostasis and detoxification. The bacterial homologues of this enzyme, so-called PCS-like enzymes, can serve as a structural model for PCS. However, the function of these PCS-like enzymes in bacteria is not known. In this work, we combined structural, spectroscopical, thermodynamic, and theoretical techniques to investigate the mechanism of the PCS-like enzyme NsPCS from *Nostoc* sp. We delineated the energetic profile of the catalytic reaction whose rate-limiting step appears to be the deacylation. The high energy barriers are compatible with the low reaction rate. Several interesting features have been uncovered by our study. When the active-site cysteine is mutated into serine, the substrate GSH binds to the protein such that GSH was co-purified with the mutant enzyme and crystallized. Our analysis indicates that the substrate binds tightly and shows a slow exchange. Interestingly, in the crystal, no indication of oxidation of the substrate GSH bound to the mutant enzyme is found (i.e., no GSSG formation), even though the crystal was grown under aerobic conditions. In contrast, in the crystal structure of the acyl-enzyme (an intermediate of the catalytic cycle), γEC-SG is bound. We explained this behavior by the differences in the electrostatics between GSH and γEC as corroborated by Monte Carlo simulations. These findings prompted us to test whether GSH and GSSG can both serve as a substrate for NsPCS. Our results were unexpected. In fact, we found that GSH is cleaved into γEC and glycine under an enzymatic turnover. Instead when GSSG binds to the enzyme, one glycine is cleaved off resulting in γEC-SG covalently bound to the enzyme and the reaction stops. From this finding, we conclude that NsPCS is active under reducing conditions, but gets inhibited under oxidizing conditions. Our observation together with the analysis of the genetic context of the NsPCS gene and the fact that NsPCS has a signal peptide directing the enzyme to the periplasm lead us to suggest that NsPCS may be a part of a signaling system that senses the redox state of the periplasm.

From an evolutionary point of view, PCS-like enzymes may be considered as precursors of the genuine PCSs found in plants, that is, the enzymes that form PC by transpeptidation. Our finding that γEC-SG is bound stably to NsPCS and basically inhibits the enzyme is also interesting in this evolutionary context. In fact, one could imagine that the close proximity between the covalently bound γEC and a second glutathione molecule in γEC-SG may have sporadically led to transpeptidation, which eventually became a new function.

Taken together, our biophysical analysis allows us to characterize NsPCS as a relatively slow enzyme which may work as a part of a redox sensing system in cyanobacteria. Moreover, we propose a mechanism how PCS-like enzymes may have gained the function of transpeptidation.

MATERIALS AND METHODS

Cloning. Wild-type NsPCS gene (alr0975) lacking the natural signal sequence (ΔssNsPCS) was amplified using the primer pair NsPCS_for 5′-TTA TTA CAT ATG CAA ACT TTG ACA CTT TCA CC-3′ and NsPCS_rev 5′-TAA TAA CTC GAG CTA ATC TTG TGT TTT ACT TAC-3′. The purified PCR product was ligated with the plasmid p10$[39] using NdeI and XhoI insertion sides, generating a plasmid coding for an N-terminally His6-lysozyme-tagged fusion protein. The plasmid carrying the wild-type gene served as a template for standard QuikChange (Strategene) mutagenesis using the complementary primer set C70S_NsPSC_for 5′-GTT AAT CAA GCT TAC TCT GGT GTA GCT AGT ATA ATT ATG-3′ and C70S_NsPCS_rev 5′-CAT AAT TAT ACT AGC TAC ACC AGA GTA AGC TTG ATT AAC-3′ to generate a variant of NsPCS in which the catalytic Cys is mutated into a Ser. The correct integrity of the gene fragment and the introduction of the mutation were confirmed by DNA sequencing.

Protein Expression and Purification. Proteins were recombinantly expressed in Rosetta 2 (DE3) plysS cells harboring the respective plasmids. Ampicillin-supplemented (100 μg mL^{-1}) LB broth was inoculated with an overnight culture at an OD600 of 0.02. Cells were grown until the mid-log phase at 310 K before the temperature was lowered to 293 K and gene expression was induced with 0.5 mM IPTG overnight. Harvested cells were resuspended in 150 mM phosphate buffer pH 8.0, 300 mM NaCl (buffer A) supplemented with 100 μM phenylmethylsulfonylfluoride and disrupted with a microfluidizer (Microfluidics). Cleared lysate (100,000 g, 30 min) was passed over a HiTrap chelating column (GE Healthcare) charged with $NiSO_4$ using an ÄKTA prime system (GE Healthcare). The column was developed with a gradient to buffer B (buffer A containing 500 mM imidazole), and fractions were analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. NsPCS-containing fractions were pooled in a dialysis bag to which His-tagged human rhinovirus 3C peptidase (in house production) was added in a ratio of 1:40. Dialysis against 50 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 8.0, 150 mM NaCl was carried out overnight at 277 K. After the dialysis, the recovered protein was again passed over a Ni-charged HiTrap chelating column to remove uncleaved protein and peptidase. The flow-through was concentrated and applied to size-exclusion chromatography using a Superdex 75 column (GE Healthcare) in the same buffer. Fractions containing pure protein were concentrated to 25 mg mL^{-1} and either flash frozen in liquid nitrogen prior to storage at −80 °C or used immediately.

Protein Crystallization. Initial crystallization conditions were identified using commercial screens (Qiagen) and a Phoenix robot (Art Robbins Instruments). Initial hits were optimized using a hanging drop vapor diffusion setup with a 500 μL reservoir and drops consisting of 1 μL of protein (12 mg mL^{-1}) and 1 μL of reservoir solution. Large rod-shaped crystals were obtained using 17% PEG 8000, 0.1 M MES pH 5.5, and 0.2 M calcium acetate as mother liquor. The acyl-enzyme was obtained by co-crystallization of wild-type NsPCS protein (12 mg mL^{-1}) preincubated with glutathione at a final concentration of 5 mM using the same mother liquor composition. All crystals grew within 2 days to an approximate size of 150 × 150 × 600 μm. Prior to data collection, crystals were washed briefly with cryoprotectant containing mother liquor supplemented with 20% glycerol before plunging into liquid nitrogen.

Data Collection. Diffraction data have been collected on BL14.1 of the BESSY II electron storage ring (Berlin-Adlershof, Germany) on a Rayonix MX-225 CCD and Pilatus 6M detector.[40] C70S-NsPCS mutant data were collected in nonoverlapping 1° (Rayonix MX-225 CCD) and the acyl-enzyme in nonoverlapping 0.1° (Pilatus 6M) oscillation images, respectively.

Data were integrated with XDS,[41] scaled with AIMLESS from CCP4 suite.[42] A value of 0.5 of CC1/2[43] in the highest shell was chosen as the cutoff criterion in respect to completeness of data. Data collection statistics are summarized in Table 1.

Structure Determination and Refinement. The crystal structure of C70S-NsPCS was determined by molecular replacement with phenix.phaser[44] searching with a monomer of NsPCS (PDB ID: 2BTW)[19] as a model. Initial phases for the acyl-enzyme structure were obtained by molecular replacement searching with a monomer of the C70S mutant. Iterative refinement was carried out with phenix.re-

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

fine[44] and manual adjustments by hand were done in Coot.[45] Coordinates and structure factor amplitudes have been deposited in the Protein Data Bank[46] with access codes indicated in Table 1.

Protein Unfolding and Refolding. NsPCS variants were unfolded at room temperature in buffer containing 6 M guanidinium chloride (GdmCl), 50 mM Na_3PO_4, and 10% glycerol at pH 8. As the final protein concentration, 100 $\mu g\ mL^{-1}$ was obtained. For refolding, the obtained solution was dialyzed at 277 K overnight against a buffer containing 3 M GdmCl and then for at least 4 h against buffers containing 1 M GdmCl and no GdmCl, respectively.

ITC Measurement. ITC measurements were performed using a Nano-ITC instrument (TA instruments). Prior to measurements, the buffer for sample preparation (50 mM Na_3PO_4 at pH 8, and 300 mM NaCl) was degassed in an ultrasonic bath for 5 min and kept under an argon atmosphere. The concentrations of refolded proteins were adjusted to 1 mM (C70S-NsPCS) and 0.1 mM (wt-NsPCS) with respective concentrations for ligand solutions of 1 mM GSSG and 0.5 mM GSH prepared in buffer. Measurements were performed at 303.15 K without stirring to prevent protein aggregation. The time interval between two successive ligand injections was extended to ensure equilibration. Each 500 s 2.01 μL injections were made for C70S-NsPCS and injections with varied volumes were made for wt-NsPCS (injections 1−2: 0.17 μL, injections 3−8: 0.52 μL, injections 9−10: 0.75 μL, injections 11−13: 0.98 μL, and injections 15−16: 2.01 μL). The syringe and the instrument cell were purged with argon to prevent any trace of oxygen. The data were processed using NanoAnalyze (TA instruments) and corrected by subtracting the enthalpy of dilution of the ligand in buffer. The measured differential heat per mole, dH, was analyzed according to the one-site-differential binding model[47] and fitted to the equation

$$dH = \frac{\Delta H \Delta M_T \cdot \frac{K_a \cdot [X]}{1 + K_a \cdot [X]} \cdot \left(1 - \frac{K_a \cdot [X]}{1 + K_a \cdot [X]}\right) + \Delta h_{dil} \cdot [X]}{[X] + M_T \cdot \left(\frac{K_a \cdot [X]}{1 + K_a \cdot [X]} - \left(\frac{K_a \cdot [X]}{1 + K_a \cdot [X]}\right)^2\right)}$$

where ΔH is the enthalpy change upon binding, K_a is the association equilibrium constant, and Δh_{dil} is the enthalpy of dilution of the binding species.

The free ligand concentration [X] is calculated as follows

$$[X] = M_T \cdot \frac{\frac{X_T}{M_T} - \frac{1}{M_T \cdot K_a} - 1 + \left(\left(1 + \frac{1}{M_T \cdot K_a} - \frac{X_T}{M_T}\right)^2 + \frac{4 \cdot \frac{X_T}{M_T}}{M_T \cdot K_a}\right)^{1/2}}{2}$$

where M_T and X_T are the total receptor concentration and the total ligand concentration, respectively, given by

$$M_T = p \cdot M_{t_0} \cdot (e^{\Sigma dV_{injected}/V_0})$$

$$X_T = X_{t_0} \cdot (1 - e^{\Sigma dV_{injected}/V_0})$$

where M_{t_0} and X_{t_0} are the respective concentrations before injections, $dV_{injected}$ is the injection volume at a time, and V_0 is the initial volume within the measured cell. The model allows the correction of the total receptor concentration M_T to estimate the concentration of the receptor that is truly able to bind the ligand (parameter p).

Thermophoresis. The microscale thermophoresis experiments have been performed with the Monolith NT.LabelFree (Nanotemper Technologies). The assay buffer was 50 mM HEPES pH 8 with 100 mM NaCl.

The samples were left to incubate for 5 min being loaded into the standard treated capillaries (Nanotemper Technologies). Data collection was carried out at 25 °C. The solution inside the capillary is locally heated with a focused IR-laser, which is coupled into the path of exciting light using a hot mirror, at 20, 30, and 40% power.

The sample is excited at 270 nm. The resulting fluorescence from the aromatic residues of the protein is detected with a photodiode at 370 nm. The IR-laser is switched on after 5 s from the beginning of the detection and let on for 30 s. The fluorescence intensity detected at the beginning of the heating procedure is called F_{cold} and the fluorescence detected during the heating procedure is called F_{hot}. The normalized fluorescence $F_{norm} = F_{hot}/F_{cold}$ was plotted against the ligand concentration. Data analysis was carried out using MO.Affinity Analysis v2.3 (Nanotemper Technologies).

NMR Spectroscopy. NMR measurements were carried out on a Bruker AVANCE III 600 MHz spectrometer with a sample volume of 600 μL. Spectra were recorded at 600.2 MHz 1H frequency and a calibrated temperature of 303.15 K. Data processing was performed with NMRPipe.[48] Buffer conditions were 50 mM phosphate buffer pH 8.0. The NMR buffer was degassed in an ultrasonic bath under vacuum for 5 min. All solutions were prepared and stored under an argon atmosphere, and all experimental steps were performed under an argon atmosphere to prevent the oxidation of glutathione. To compare the activity of NsPCS with different substrates, the concentration of GSH and GSSG was 600 μM. For the Michaelis−Menten kinetics, different samples of GSH were prepared with concentrations of 500 μM, 600 μM, 700 μM, 800 μM, 900 μM, 1 mM, 1.5 mM, and 3 mM of GSH. To compare the activity of wt-NsPCS and refolded wt-NsPCS, the concentration of GSH was 1 mM, and 300 mM NaCl was added to the buffer. All samples contained 0.1 mM sodium trimethylsilylpropanesulfonate (DSS) and 10% D_2O. In all cases, the reaction was started by adding 30 nM enzyme to the substrate solution. Each sample was rapidly collected and poured in a NMR tube purged with argon using a syringe purged with argon as well. An NMR spectrum was taken roughly every 110 s. The progress of the observed reaction was monitored using the increasing glycine signal. The DSS signal was used as reference for the glycine concentration. 1D spectra were normalized by the DSS concentration, number of scans, and length of the 90° proton pulse. To perform the Michaelis−Menten analysis, the time traces describing the accumulation of free glycine have been fitted with one exponential term for each substrate concentration. The initial increase of the exponential curve could be approximated with a straight line whose slope gives the initial velocity of the reaction in the presence of a given concentration of substrate. The initial velocities were then plotted versus the corresponding substrate concentrations.

Computational Preparation of Protein Structures. Protein structures for all simulations were prepared with the program CHARMM[49] using the CHARMM27[50] force field. Present disulfide bonds were set. The protein was surrounded by a 6 Å explicit water layer, whereas available water molecules from the crystal structure were included. All hydrogen atoms were added with the HBUILD routine in CHARMM. An optimization of all water molecules was performed to adapt to the protein, followed by an optimization of all hydrogen atoms. Protonation probabilities of prepared structures were calculated using MEAD[51,52] and GMCT.[53] Thereby, a Metropolis Monte Carlo titration algorithm[54,55] is applied on a Poisson−Boltzmann continuum electrostatic model. Protonation probabilities are dependent on pH and were calculated in steps of 0.25 in the pH range 0−14. 200 equilibration scans and 100,000 production scans were performed at 300 K, with 0.1 M ionic strength and permittivity 4 for protein and 80 for solvent. The protonation states of titratable groups were set according to this calculation (see Table S2 in the Supporting Information). For investigation of NsPCS, the crystal structure of C70S-NsPCS and acyl-NsPCS were used. For the simulations, the active site serine of C70S-NsPCS was mutated to cysteine with PyMOL (The PyMOL Molecular Graphics System, Version 2.4.0a0 Schrödinger, LLC). Simulations for papain were performed on a crystal structure with synthetic inhibitor E-64-c (2.1 Å, PDB ID: 1PE6).[56] Analogously to other studies, the substrate Phe-Ser-Ile with acetylated N-terminus and N-methylated C-terminus was used.[32,57] The substrate in the crystal structure 1PE6 is attached covalently to papain. For modeling of the noncleaved substrate, a structure with a substrate analogue (2.8 Å, PDB ID: 1PAD)[20] was superimposed. The modeled substrate was optimized with CHARMM.

Ligand Binding Calculations. Docking of the ligand GSH to NsPCS variants was performed with MCMap,[31] which applies Monte Carlo sampling of ligand movements within the electrostatic field of a

https://doi.org/10.1021/acschembio.1c00941
ACS Chem. Biol. 2022, 17, 883−897

receptor. For each simulation, 100,000 runs were performed with 1,000,000 Monte Carlo steps at 300 K. The initial center of mass separation of the protein and the ligand was set to 130 Å, the maximum separation was set to 180 Å. The Monte Carlo run was reset after 50 rejected steps in a row. The maximal displacement was set to 3 Å, the maximal rotation was set to 5 rad. Electrostatic potential maps with 2 Å grid spacing for the outer potential grid and with 1 Å for the inner potential grid were used. APBS[58] was used for electrostatic potential map calculations, with permittivity 4 for protein and 80 for solvent at 300 K. An ion concentration of 0.1 M was adjusted, with a single positive or a single negative charge and radius 2.0 Å. The solvent radius was set to 1.4 Å. For calculating the density of GSH around PCS, the center of mass of GSH was recorded every 100th Monte Carlo step. The interaction energies between PCS and the various ligands were calculated with the program solinprot from the MEAD package. The dielectric constants of the protein and the water were set at 4 and 80, respectively. The ionic strength was set to 0.1 M.

Reaction Path Search Simulation. The reaction path search was performed with a QM/MM model within pDynamo[59] and the CHARMM27[50] force field in combination with ORCA[60,61] on the B3LYP:6-31+G** level.[62−65]

The protein structures, which were prepared with CHARMM, were treated with MM, whereas a small relevant region for the catalysis was treated with QM. The QM/MM boundary was treated with a link-atom scheme and electrostatic embedding, as implemented in pDynamo. All atoms in the QM region and within an 8 Å MM layer around the QM region were set flexible. Beyond that, a linearly increasing force constant from 0 to 12 kcal mol^{-1} $Å^{-1}$ was applied for further 8 Å on the atoms. All other atoms were restraint with 12 kcal mol^{-1} $Å^{-1}$. The sidechains included in the QM region were truncated between Cα and Cβ. Exceptions are mentioned explicitly. For NsPCS, the QM region consisted of the catalytic triad residues Cys70, His183 (protonated at position ε), and Asp201. Furthermore, Gln64 (truncated between Cβ and Cγ) for the oxyanion hole and Arg173 (truncated between Cγ and Cδ) were included. The complete residues Gly (only present for the acylation reaction), and Cys of the substrate GSH were included into the QM region, with addition of the atoms Cδ, Oε, and Cγ with both hydrogen atoms of residue γGlu. For the acylation reaction, an interacting water molecule near the carboxyl group of the substrate was included. For the deacylation reaction two interacting water molecules were included. All remaining parts belong to the MM region.

For papain, the QM region contained the catalytic triad resides Cys25, His159 (protonated on position ε), and Asn175. Furthermore, Gln19 (truncated between Cβ and Cγ) for the oxyanion hole was included. The complete residue Ser of the substrate together with backbone atoms O and C of residue Phe (only present for the acylation reaction), and atoms N, H, Cα, and Hα of residue Ile were included into the QM region. For the deacylation reaction, two interacting water molecules were included.

Reaction path search was performed with PyCPR.[29,66] Reaction path exploration was performed by adiabatic surface scans with an root-mean-square gradient criterion of 0.002 kcal mol^{-1} $Å^{-1}$, and structures of stable states were optimized by a conjugate gradient minimizer with the root-mean-square gradient threshold set to 0.002 kcal mol^{-1} $Å^{-1}$. As collective variables for the adiabatic scans, the proton transfer from the cysteine or serine nucleophile to the catalytic histidine and the nucleophilic attack of the cysteine or serine nucleophile on the substrate were used. Initial path estimates between the stable states were obtained by a growing string method,[67] as implemented in pDynamo. Transition paths were obtained by PyCPR. States were characterized by vibrational frequency calculations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acschembio.1c00941.

Sequence alignments, sequence analysis of the NsPCS operon, protonation probability of the catalytic histidine, list of models used for the simulations, and protonation states of all the titratable residues (PDF)

Accession Codes

PDB IDs: 6TH5—wild-type NsPCS; 6THO—wild-type NsPCS with covalently bound γEC-SG; 6TJL—mutant C70S-NsPCS with noncovalently bound GSH.

■ AUTHOR INFORMATION

Corresponding Author

Elisa Bombarda − *Department of Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany;* orcid.org/0000-0002-1385-3710; Email: elisa.bombarda@uni-bayreuth.de

Authors

Florian J. Gisdon − *Department of Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany; Computational Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany*

Christian G. Feiler − *Department Structure and Function of Proteins, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany*

Oxana Kempf − *Department of Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany*

Johannes M. Foerster − *Computational Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany;* orcid.org/0000-0003-0442-9413

Jonathan Haiss − *Department of Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany*

Wulf Blankenfeldt − *Department Structure and Function of Proteins, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany; Institute for Biochemistry, Biotechnology and Bioinformatics, Technische Universität Braunschweig, 38106 Braunschweig, Germany;* orcid.org/0000-0001-9886-9668

G. Matthias Ullmann − *Computational Biochemistry, University of Bayreuth, 95440 Bayreuth, Germany;* orcid.org/0000-0002-6350-798X

Complete contact information is available at: https://pubs.acs.org/10.1021/acschembio.1c00941

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank K. Schweimer for the highly competent support during NMR experiments and S. Shanmugaratnam for the great and friendly all-around technical support. We are very thankful to S. Clemens for fruitful discussions at the beginning of this project. Diffraction data have been collected on BL14.1 at the BESSY II electron storage ring operated by the Helmholtz-Zentrum Berlin.[68] The present work was supported by the DFG grant BO 3578/1 and DFG Priority Program 1927 through grant UL 174/9.

■ REFERENCES

(1) Cobbett, C.; Goldsbrough, P. Phytochelatins and Metallothioneins: Roles in Heavy Metal Detoxification and Homeostasis. *Annu. Rev. Plant Biol.* **2002**, *53*, 159−182.

(2) Gekeler, W.; Grill, E.; Winnacker, E.-L.; Zenk, M. H. Survey of the Plant Kingdom for the Ability to Bind Heavy Metals through Phytochelatins. *Z. Naturforsch., C: J. Biosci.* **1989**, *44*, 361−369.

(3) Filiz, E.; Saracoglu, I. A.; Ozyigit, I. I.; Yalcin, B. Comparative Analyses of Phytochelatin Synthase (PCS) Genes in Higher Plants. *Biotechnol. Biotechnol. Equip.* **2019**, *33*, 178−194.
(4) Clemens, S.; Kim, E. J.; Neumann, D.; Schroeder, J. I. Tolerance to Toxic Metals by a Gene Family of Phytochelatin Synthases from Plants and Yeast. *EMBO J.* **1999**, *18*, 3325−3333.
(5) Cazalé, A. C.; Clemens, S. Arabidopsis Thaliana Expresses a Second Functional Phytochelatin Synthase. *FEBS Lett.* **2001**, *507*, 215−219.
(6) Vatamaniuk, O. K.; Bucher, E. A.; Ward, J. T.; Rea, P. A. Worms take the "phyto" out of 'phytochelatins'. *Trends Biotechnol.* **2002**, *20*, 61−64.
(7) Ray, D.; Williams, D. L. Characterization of the Phytochelatin Synthase of Schistosoma Mansoni. *PLoS Neglected Trop. Dis.* **2011**, *5*, No. e1168.
(8) Clemens, S. Evolution and Function of Phytochelatin Synthases. *J. Plant Physiol.* **2006**, *163*, 319−332.
(9) Marchler-Bauer, A.; Anderson, J. B.; DeWeese-Scott, C.; Fedorova, N. D.; Geer, L. Y.; He, S.; Hurwitz, D. I.; Jackson, J. D.; Jacobs, A. R.; Lanczycki, C. J.; et al. CDD: A Curated Entrez Database of Conserved Domain Alignments. *Nucleic Acids Res.* **2003**, *31*, 383−387.
(10) Harada, E.; von Roepenack-Lahaye, E.; Clemens, S. A cyanobacterial protein with similarity to phytochelatin synthases catalyzes the conversion of glutathione to ?-glutamylcysteine and lacks phytochelatin synthase activity. *Phytochemistry* **2004**, *65*, 3179−3185.
(11) Tsuji, N.; Nishikori, S.; Iwabe, O.; Shiraki, K.; Miyasaka, H.; Takagi, M.; Hirata, K.; Miyamoto, K. Characterization of Phytochelatin Synthase-like Protein Encoded by Alr0975 from a Prokaryote, Nostoc Sp. PCC 7120. *Biochem. Biophys. Res. Commun.* **2004**, *315*, 751−755.
(12) Rea, P. A. Phytochelatin synthase, papain's cousin, in stereo. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 507−508.
(13) Bellini, E.; Varotto, C.; Borsò, M.; Rugnini, L.; Bruno, L.; Sanità di Toppi, L. Eukaryotic and Prokaryotic Phytochelatin Synthases Differ Less in Functional Terms than Previously Thought: A Comparative Analysis of Marchantia Polymorpha and Geitlerinema Sp. PCC 7407. *Plants* **2020**, *9*, 914−12.
(14) Rawlings, N. D.; Waller, M.; Barrett, A. J.; Bateman, A. MEROPS: The Database of Proteolytic Enzymes, Their Substrates and Inhibitors. *Nucleic Acids Res.* **2014**, *42*, D503−D509.
(15) Rawlings, N. D.; Barrett, A. J.; Thomas, P. D.; Huang, X.; Bateman, A.; Finn, R. D. The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database. *Nucleic Acids Res.* **2018**, *46*, D624−D632.
(16) Barrett, A. J.; Rawlings, N. D. Evolutionary Lines of Cysteine Peptidases. *Biol. Chem.* **2001**, *382*, 727−733.
(17) Rea, P. A. Phytochelatin Synthase: Of a Protease a Peptide Polymerase Made. *Physiol. Plant.* **2012**, *145*, 154−164.
(18) Rea, P. A. Phytochelatin Synthase. *eLS* **2020**, *3*, 1−15.
(19) Vivares, D.; Arnoux, P.; Pignol, D. A Papain-like Enzyme at Work: Native and Acyl-Enzyme Intermediate Structures in Phytochelatin Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18848−18853.
(20) Drenth, J.; Kalk, K. H.; Swen, H. M. Binding of Chloromethyl Ketone Substrate Analogs to Crystalline Papain. *Biochemistry* **1976**, *15*, 3731−3738.
(21) Vernet, T.; Tessier, D. C.; Chatellier, J.; Plouffe, C.; Lee, T. S.; Thomas, D. Y.; Storer, A. C.; Ménard, R. Structural and Functional Roles of Asparagine 175 in the Cysteine Protease Papain. *J. Biol. Chem.* **1995**, *270*, 16645−16652.
(22) Romanyuk, N. D.; Rigden, D. J.; Vatamaniuk, O. K.; Lang, A.; Cahoon, R. E.; Jez, J. M.; Rea, P. A. Mutagenic Definition of a Papain-like Catalytic Triad, Sufficiency of the N-Terminal Domain for Single-Site Core Catalytic Enzyme Acylation, and C-Terminal Domain for Augmentative Metal Activation of a Eukaryotic Phytochelatin Synthase. *Plant Physiol.* **2006**, *141*, 858−869.
(23) Tsuji, N.; Nishikori, S.; Iwabe, O.; Matsumoto, S.; Shiraki, K.; Miyasaka, H.; Takagi, M.; Miyamoto, K.; Hirata, K. Comparative Analysis of the Two-Step Reaction Catalyzed by Prokaryotic and Eukaryotic Phytochelatin Synthase by an Ion-Pair Liquid Chromatography Assay. *Planta* **2005**, *222*, 181−191.
(24) Li, M.; Barbaro, E.; Bellini, E.; Saba, A.; Sanità di Toppi, L.; Varotto, C. Ancestral Function of the Phytochelatin Synthase C-Terminal Domain in Inhibition of Heavy Metal-Mediated Enzyme Overactivation. *J. Exp. Bot.* **2020**, *71*, 6655−6669.
(25) Ruotolo, R.; Peracchi, A.; Bolchi, A.; Infusini, G.; Amoresano, A.; Ottonello, S. Domain Organization of Phytochelatin Synthase. Functional Properties of Truncated Enzyme Species Identified by Limited Proteolysis. *J. Biol. Chem.* **2004**, *279*, 14686−14693.
(26) Clemens, S.; Peršoh, D. Multi-Tasking Phytochelatin Synthases. *Plant Sci.* **2009**, *177*, 266−271.
(27) Clay, N. K.; Adio, A. M.; Denoux, C.; Jander, G.; Ausubel, F. M. Glucosinolate Metabolites Required for an Arabidopsis Innate Immune Response. *Science* **2009**, *323*, 95−101.
(28) Kaneko, T.; Nakamura, Y.; Wolk, C. P.; Kuritz, T.; Sasamoto, S.; Watanabe, A.; Iriguchi, M.; Ishikawa, A.; Kawashima, K.; Kimura, T.; et al. Complete Genomic Sequence of the Filamentous Nitrogen-Fixing Cyanobacterium Anabaena Sp. Strain PCC 7120. *DNA Res.* **2001**, *8*, 205−213.
(29) Culka, M.; Gisdon, F. J.; Ullmann, G. M. Computational Biochemistry-Enzyme Mechanisms Explored. *Advances in Protein Chemistry and Structural Biology*, 1st ed.; Elsevier Inc., 2017; Vol. *109*.
(30) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778−2786.
(31) Foerster, J. M.; Poehner, I.; Ullmann, G. M. MCMap-A Computational Tool for Mapping Energy Landscapes of Transient Protein-Protein Interactions. *ACS Omega* **2018**, *3*, 6465−6475.
(32) Harrison, M. J.; Burton, N. A.; Hillier, I. H. Catalytic Mechanism of the Enzyme Papain: Predictions with a Hybrid Quantum Mechanical/Molecular Mechanical Potential. *J. Am. Chem. Soc.* **1997**, *119*, 12285−12291.
(33) Grill, E.; Löffler, S.; Winnacker, E.-L.; Zenk, M. H. Phytochelatins, the heavy-metal-binding peptides of plants, are synthesized from glutathione by a specific γ-glutamylcysteine dipeptidyl transpeptidase (phytochelatin synthase). *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 6838−6842.
(34) Zechmann, B.; Tomašić, A.; Horvat, L.; Fulgosi, H. Subcellular Distribution of Glutathione and Cysteine in Cyanobacteria. *Protoplasma* **2010**, *246*, 65−72.
(35) Eser, M.; Masip, L.; Kadokura, H.; Georgiou, G.; Beckwith, J. Disulfide bond formation by exported glutaredoxin indicates glutathione's presence in the E. coli periplasm. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 1572−1577.
(36) Mondal, J.; Friesner, R. A.; Berne, B. J. Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. *J. Chem. Theory Comput.* **2014**, *10*, 5696−5705.
(37) Diaz-Mochon, J. J.; Planonth, S.; Bradley, M. From 10,000 to 1: Selective Synthesis and Enzymatic Evaluation of Fluorescence Resonance Energy Transfer Peptides as Specific Substrates for Chymopapain. *Anal. Biochem.* **2009**, *384*, 101−105.
(38) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37*, 420−423.
(39) Bock, T.; Luxenburger, E.; Hoffmann, J.; Schütza, V.; Feiler, C.; Müller, R.; Blankenfeldt, W. AibA/AibB Induces an Intramolecular Decarboxylation in Isovalerate Biosynthesis by Myxococcus Xanthus. *Angew. Chem., Int. Ed.* **2017**, *56*, 9986−9989.
(40) Mueller, U.; Darowski, N.; Fuchs, M. R.; Förster, R.; Hellmig, M.; Paithankar, K. S.; Pühringer, S.; Steffien, M.; Zocher, G.; Weiss, M. S. Facilities for Macromolecular Crystallography at the Helmholtz-Zentrum Berlin. *J. Synchrotron Radiat.* **2012**, *19*, 442−449.
(41) Kabsch, W. XDS. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 125−132.

(42) Potterton, E.; Briggs, P.; Turkenburg, M.; Dodson, E. A graphical user interface to theCCP4 program suite. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 1131−1137.
(43) Karplus, P. A.; Diederichs, K. Linking Crystallographic Model and Data Quality. *Science* **2012**, *336*, 1030−1033.
(44) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; et al. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 213−221.
(45) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 486−501.
(46) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(47) Herrera, I.; Winnik, M. A. Differential Binding Models for Isothermal Titration Calorimetry: Moving beyond the Wiseman Isotherm. *J. Phys. Chem. B* **2013**, *117*, 8659−8672.
(48) Delaglio, F.; Grzesiek, S.; Vuister, G.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* **1995**, *6*, 277−293.
(49) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(50) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(51) Bashford, D.; Gerwert, K. Electrostatic Calculations of the PKa Values of Ionizable Groups in Bacteriorhodopsin. *J. Mol. Biol.* **1992**, *224*, 473−486.
(52) Bashford, D. An object-oriented programming suite for electrostatic effects in biological molecules An experience report on the MEAD project. In *Lecture Notes in Computer Science*; Ishikawa, Y., Oldehoeft, R. R., John, V. W., Reynders, M. T., Eds.; SpringerBerlin Heidelberg, 1997; pp 233−240.
(53) Ullmann, R. T.; Ullmann, G. M. GMCT : A Monte Carlo simulation package for macromolecular receptors. *J. Comput. Chem.* **2012**, *33*, 887−900.
(54) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Protonation of Interacting Residues in a Protein by a Monte Carlo Method: Application to Lysozyme and the Photosynthetic Reaction Center of Rhodobacter Sphaeroides. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804−5808.
(55) Ullmann, G. M.; Knapp, E.-W. Electrostatic Models for Computing Protonation and Redox Equilibria in Proteins. *Eur. Biophys. J.* **1999**, *28*, 533−551.
(56) Yamamoto, D.; Matsumoto, K.; Ohishi, H.; Ishida, T.; Inoue, M.; Kitamura, K.; Mizuno, H. Refined x-ray structure of papain.E-64-c complex at 2.1-A resolution. *J. Biol. Chem.* **1991**, *266*, 14771−14777.
(57) Arad, D.; Langridge, R.; Kollman, P. A. A simulation of the sulfur attack in catalytic pathway of papain using molecular mechanics and semiempirical quantum mechanics. *J. Am. Chem. Soc.* **1990**, *112*, 491−502.
(58) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(59) Field, M. J. The PDynamo Program for Molecular Simulations Using Hybrid Quantum Chemical and Molecular Mechanical Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1151−1161.
(60) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(61) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1327.
(62) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724−728.
(63) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.
(64) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.
(65) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(66) Gisdon, F. J.; Culka, M.; Ullmann, G. M. PyCPR - a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures. *J. Mol. Model.* **2016**, *22*, 242.
(67) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A Growing String Method for Determining Transition States: Comparison to the Nudged Elastic Band and String Methods. *J. Chem. Phys.* **2004**, *120*, 7877−7886.
(68) Mueller, U.; Förster, R.; Hellmig, M.; Huschmann, F. U.; Kastner, A.; Malecki, P.; Pühringer, S.; Röwer, M.; Sparta, K.; Steffien, M.; et al. The Macromolecular Crystallography Beamlines at BESSY II of the Helmholtz-Zentrum Berlin: Current Status and Perspectives. *Eur. Phys. J. Plus* **2015**, *130*, 141−150.

Structural and biophysical analysis of the phytochelatin-synthase-like enzyme from *Nostoc sp*. shows that its protease activity is sensitive to the redox state of the substrate

Florian J. Gisdon [1, 2], Christian G. Feiler [3], Oxana Kempf [1], Johannes M. Foerster [2], Jonathan Haiss [1], Wulf Blankenfeldt [3, 4], G. Matthias Ullmann [2], Elisa Bombarda [1] *

[1] Department of Biochemistry and [2] Computational Biochemistry, University of Bayreuth, Universitätsstr. 30, 95440, Bayreuth, Germany.

[3] Department Structure and Function of Proteins, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany

[4] Institute for Biochemistry, Biotechnology and Bioinformatics, Technische Universität Braunschweig, Spielmannstr. 7, 38106 Braunschweig, Germany

***Corresponding Author**: elisa.bombarda@uni-bayreuth.de

Supporting Information

Table of contents

I. Sequence alignment of AtPCS1 (AAD16046.1) and NsPCS (WP_044520790.1)

Full alignment using Needle EMBOSS:6.6.0.0 (https://www.ebi.ac.uk/Tools/psa/emboss_needle/)[1]

```
#=======================================
# Aligned_sequences: 2
# 1: WP_044520790.1
# 2: AAD16046.1
# Matrix: EBLOSUM30
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
# Length: 506
# Identity:      79/506 (15.6%)
# Similarity:   128/506 (25.3%)
# Gaps:         288/506 (56.9%)
# Score: 524.0
#=======================================
WP_044520790.      1 MKLFIPVTVIGLCLSSSQVLAQTLTLSPNLIGFNSNEGEKLL---LTSRS     47
                       ::.:.::.::|. ||..|.|.|.||:.::   |...:
AAD16046.1         1 ------------MAMASLYRRSLP-SPPAIDFSSAEGKLIFNEALQKGT     36

WP_044520790.     48 REDFFPLSMQFVTQVNQAYCGVASIIMVLNSLGINAPETAQYSPYRVFTQ     97
                     .|.||.|...|.||...||||:||:..|||:|.|: |......|:|.|.:
AAD16046.1        37 MEGFFRLISYFQTQSEPAYCGLASLSVVLNALSID-PGRKWKGPWRWFDE     85

WP_044520790.     98 DNFFSNEKTKAVIAPEVVARQGMTLDELGRLIASYGVKVKVNHASDTNIE    147
                     ..:...|:.      |||...|:::.::..|....|:||::..:|.::|:
AAD16046.1        86 SMLDCCEPL------EVVKEKGISFGKVVCLAHCSGAKVEAFRTSQSTID    129

WP_044520790.    148 DFRKQVAENLKQDGNFVIVNYLRKEIGQERGGHISPLAAYNEQTDRFLIM    197
                     ||||.|::....:....|.:|.|....|...||.||:..||.:.|..||:
AAD16046.1       130 DFRKFVVKCTSSENCHMISTYHRSVFKQTGNGHFSPIGGYNAERDMALIL    179

WP_044520790.    198 DVSRYKYPPVWVKTTDLWKAMNTVDSVSQKTRGFVFVSKTQD--------    239
                     ||:|:||||.||:...||:||:::|..:.|.|||.::|:...
AAD16046.1       180 DVARFKYPPHWVPLKLLWEAMDSIDQSTGKRRGFMLISRPHREPGLLYTL    229

WP_044520790.    239 --------------------------------------------------    239

AAD16046.1       230 SCKDESWIEIAKYLKEDVPRLVSSQHVDSVEKIISVVFKSLPSNFNQFIR    279

WP_044520790.    239 --------------------------------------------------    239

AAD16046.1       280 WVAEIRITEDSNQNLSAEEKSRLKLKQLVLKEVHETELFKHINKFLSTVG    329

WP_044520790.    239 --------------------------------------------------    239

AAD16046.1       330 YEDSLTYAAAKACCQGAEILSGSPSKEFCCRETCVKCIKGPDDSEGTVVT    379

WP_044520790.    239 --------------------------------------------------    239

AAD16046.1       380 GVVVRDGNEQKVDLLVPSTQTECECGPEATYPAGNDVFTALLLALPPQTW    429

WP_044520790.    239 --------------------------------------------------    239

AAD16046.1       430 SGIKDQALMHEMKQLISMASLPTLLQEEVLHLRRQLQLLKRCQENKEEDD    479

WP_044520790.    239 ------    239

AAD16046.1       480 LAAPAY    485
```

Alignment of the N-Terminal Domain of AtPCS1 (AAD16046.1) and NsPCS (WP_044520790.1) without the signal peptide, truncated at the amino acid 221 to match with the sequence length of the bacterial protein.

```
########################################
#
# Aligned_sequences: 2
# 1: WP_044520790.1
# 2: AAD16046.1
# Matrix: EBLOSUM30
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
# Length: 229
# Identity:      79/229 (34.5%)
# Similarity:   123/229 (53.7%)
# Gaps:          19/229 ( 8.3%)
# Score: 521.0
#
#
########################################

WP_044520790.      1 --------QTLTLSPNLIGFNSNEGEKLL---LTSRSREDFFPLSMQFVT     39
                       ::|. ||..|.|.|.||:.::   |...:.|.||.|...|.|
AAD16046.1         1 MAMASLYRRSLP-SPPAIDFSSAEGKLIFNEALQKGTMEGFFRLISYFQT     49

WP_044520790.     40 QVNQAYCGVASIIMVLNSLGINAPETAQYSPYRVFTQDNFFSNEKTKAVI     89
                     |...||||:||:..|||:|.|: |......|:|.|.:..:...|:.
AAD16046.1        50 QSEPAYCGLASLSVVLNALSID-PGRKWKGPWRWFDESMLDCCEPL----     94

WP_044520790.     90 APEVVARQGMTLDELGRLIASYGVKVKVNHASDTNIEDFRKQVAENLKQD    139
                       |||...|:::.::..|....|:||::..:|.::|:||||.|::....:
AAD16046.1        95 --EVVKEKGISFGKVVCLAHCSGAKVEAFRTSQSTIDDFRKFVVKCTSSE    142

WP_044520790.    140 GNFVIVNYLRKEIGQERGGHISPLAAYNEQTDRFLIMDVSRYKYPPVWVK    189
                     ....|.:|.|....|...||.||:..||.:.|..||:||:|:||||.||:
AAD16046.1       143 NCHMISTYHRSVFKQTGNGHFSPIGGYNAERDMALILDVARFKYPPHWVP    192

WP_044520790.    190 TTDLWKAMNTVDSVSQKTRGFVFVSKTQD    218
                     ...||:||:::|..:.|.|||.::|:...
AAD16046.1       193 LKLLWEAMDSIDQSTGKRRGFMLISRPHR    221


#---------------------------------------
#---------------------------------------
```

Using the SignalP 5.0 server [2], we confirmed the prediction of a signal sequence of 24 amino acids (cleavage site VLA-QT) see below.

Because of the absence of the signal sequence, the number of amino acids differs from the one of the original gene product. In this sequence alignment (without the signal sequence) the catalytic cysteine is numbered with 46. However, along our investigation we kept the counting of the original gene product, i.e. the catalytic cysteine is still numbered with 70.

II. Sequence analysis of the NsPCS operon

NsPCS-Operon

The genome of Nostoc sp. PCC 7120 is completely sequenced and annotated (NCBI Accession Code: NC_003272). The genes for the following proteins are located on this operon:

Protein ID	**Annotation**	**Length**	**SignalP**	**TMHMM**
WP_010995144.1	ABC transporter ATP-binding protein	316 AA	no signal	no TM
WP_044520788.1	hypothetical protein	562 AA	no signal	12 TM
WP_010995146.1	ATP-binding cassette domain-containing protein	251 AA	no signal	no TM
WP_010995147.1	ABC transporter permease	210 AA	no signal	5 TM
WP_011319704.1	ABC transporter substrate-binding protein	302 AA	Sec/SPII	0 TM
WP_044520790.1	glutathione gamma-glutamylcysteinyltransferase (NsPCS)	239 AA	Sec/SPI	0 TM

Table S1: Information on the NsPCS Operon. The 3rd column gives the total length of the gene product in number of amino acids (AA) including signal peptides. The 4th column shows the results from a SignalP analysis. The 5th column shows the number of transmembrane helices (TM) found with TMHMM [3,4].

The protein annotated as **[hypothetical protein] (WP_044520788.1)** is a transmembrane protein as the analysis using the TMHMM server suggests (see below).

An analysis using SignalP indicates that the **[ABC transporter substrate-binding protein] (WP_011319704.1)** is a lipoprotein protein (i.e. with lipid covalently linked to the cysteine at position 21; see below for the data). The protein is located in the periplasm (see Juncker et al.[5] for a more detailed discussion of the signal peptide).

From BLAST searches, we find the following interesting connections:

- **[ABC transporter substrate-binding protein] (WP_011319704.1)** shows a very high similarity to a protein annotated as [glycine/betaine ABC transporter substrate-binding protein] from Nostoc sp. ATCC 29411 (WP_015137004) suggesting that a binding of glycine to this substrate-binding protein is likely
- The protein annotated as **[hypothetical protein] (WP_044520788.1)** shows similarities to proteins annotated as [ABC transporter permease] (for instance WP_026734415.1 -- E-value 8×10^{-169}, HAZ49862.1 – E-value 10^{-142}) suggesting that this protein may also function as a transporter.

From these findings we hypothesize that the NsPCS-operon may contain two transport systems: an import system that might be responsible for the import of glycine involving the periplasmatic ABC transporter substrate-binding protein(WP_011319704.1) and as well as the other ABC-transporter related proteins (WP_010995144.1, WP_010995146.1, WP_010995147.1) and a second transporter system involving the protein annotated as hypothetical protein (WP_044520788.1) which is of unknown function, but we suggest is responsible for the export of glutathione.

SignalP analysis of the ABC transporter substrate-binding protein (WP_011319704.1)

SignalP (https://services.healthtech.dtu.dk/service.php?SignalP-5.0)

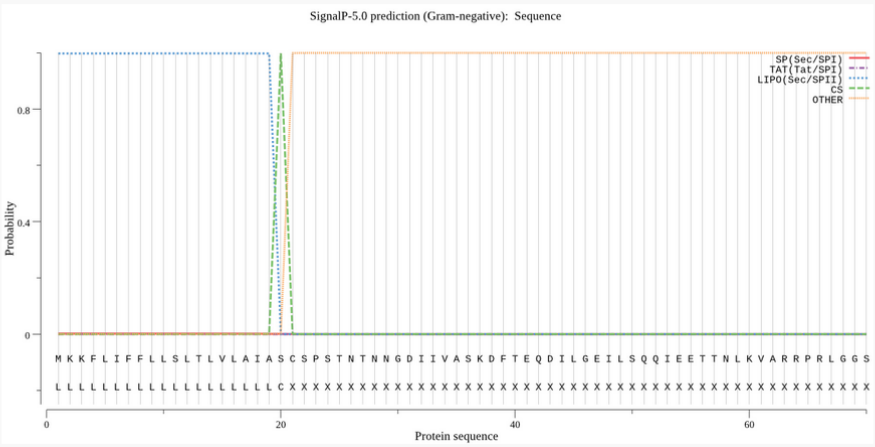
This result indicates that this protein is a lipoprotein (i.e. with diacylglyceride linked to the cysteine at position 21) that is located in the periplasm [5].

Sequence
Prediction: Lipoprotein signal peptide (Sec/SPII)

Cleavage site between pos. 20 and 21: IAS-CS.
Probability: 0.9982

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.002	0.0001	0.9976	0.0002

Download: PNG / EPS / Tabular



SignalP-analysis of the Protein Sequence of NsPCS

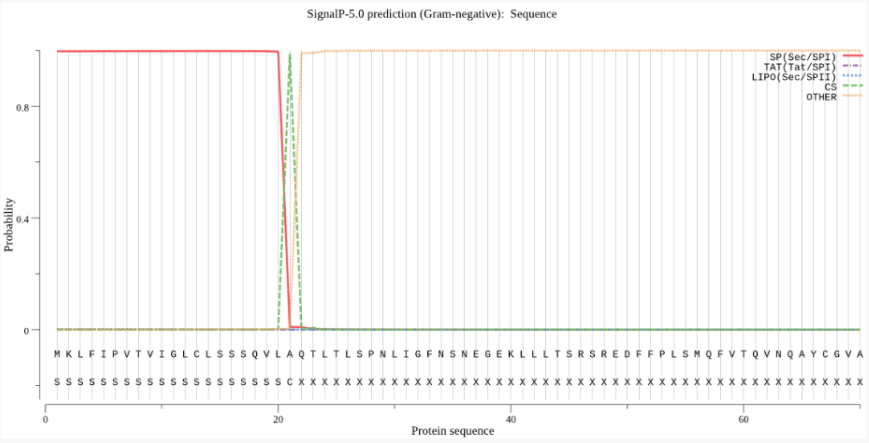
This result indicates that NsPCS is exported to the periplasm.

Sequence
Prediction: Signal peptide (Sec/SPI)

Cleavage site between pos. 21 and 22: VLA-QT.
Probability: 0.9864

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.9966	0.0004	0.0022	0.0008

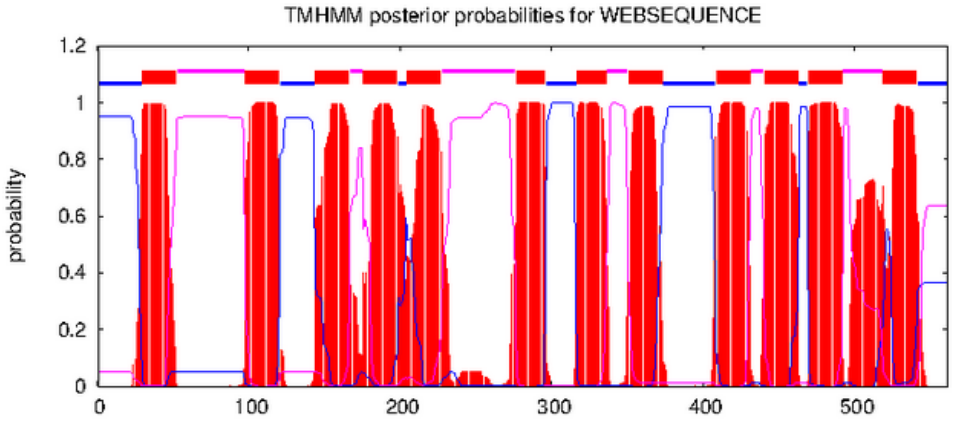
Download: PNG / EPS / Tabular



TMHMM results of the hypothetical protein (WP_044520788.1)

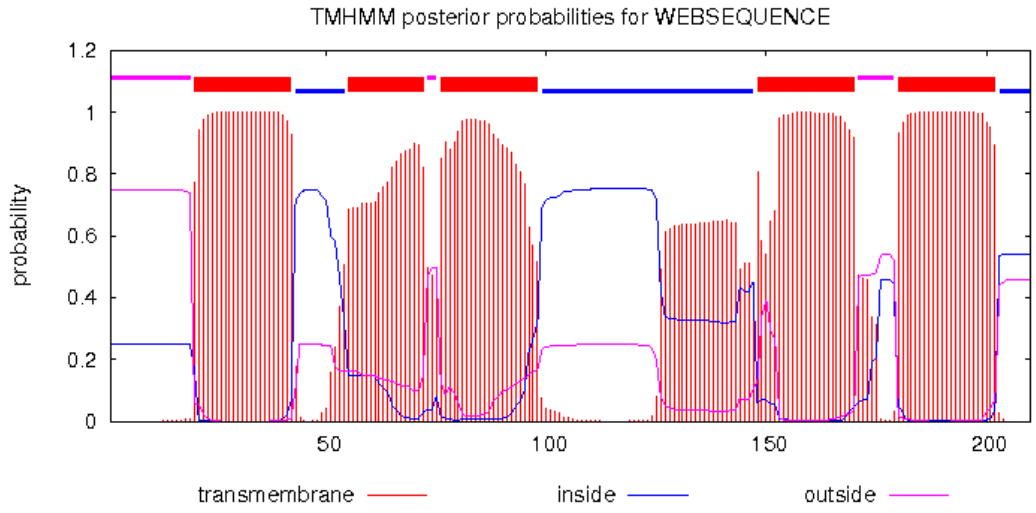
TMHMM (https://services.healthtech.dtu.dk/service.php?TMHMM-2.0)

```
# WEBSEQUENCE Length: 562
# WEBSEQUENCE Number of predicted TMHs:  12
# WEBSEQUENCE Exp number of AAs in TMHs: 277.37851
# WEBSEQUENCE Exp number, first 60 AAs:  21.21563
# WEBSEQUENCE Total prob of N-in:        0.95095
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        inside       1    29
WEBSEQUENCE     TMHMM2.0        TMhelix     30    52
WEBSEQUENCE     TMHMM2.0        outside     53    97
WEBSEQUENCE     TMHMM2.0        TMhelix     98   120
WEBSEQUENCE     TMHMM2.0        inside     121   143
WEBSEQUENCE     TMHMM2.0        TMhelix    144   166
WEBSEQUENCE     TMHMM2.0        outside    167   175
WEBSEQUENCE     TMHMM2.0        TMhelix    176   198
WEBSEQUENCE     TMHMM2.0        inside     199   204
WEBSEQUENCE     TMHMM2.0        TMhelix    205   227
WEBSEQUENCE     TMHMM2.0        outside    228   276
WEBSEQUENCE     TMHMM2.0        TMhelix    277   296
WEBSEQUENCE     TMHMM2.0        inside     297   316
WEBSEQUENCE     TMHMM2.0        TMhelix    317   336
WEBSEQUENCE     TMHMM2.0        outside    337   350
WEBSEQUENCE     TMHMM2.0        TMhelix    351   373
WEBSEQUENCE     TMHMM2.0        inside     374   408
WEBSEQUENCE     TMHMM2.0        TMhelix    409   431
WEBSEQUENCE     TMHMM2.0        outside    432   440
WEBSEQUENCE     TMHMM2.0        TMhelix    441   463
WEBSEQUENCE     TMHMM2.0        inside     464   469
WEBSEQUENCE     TMHMM2.0        TMhelix    470   492
WEBSEQUENCE     TMHMM2.0        outside    493   518
WEBSEQUENCE     TMHMM2.0        TMhelix    519   541
WEBSEQUENCE     TMHMM2.0        inside     542   562
```



TMHMM results of the ABC transporter permease (WP_010995147.1)

```
# WEBSEQUENCE Length: 210
# WEBSEQUENCE Number of predicted TMHs:  5
# WEBSEQUENCE Exp number of AAs in TMHs: 118.87821
# WEBSEQUENCE Exp number, first 60 AAs:  28.29846
# WEBSEQUENCE Total prob of N-in:        0.25108
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        outside      1    19
WEBSEQUENCE     TMHMM2.0        TMhelix     20    42
WEBSEQUENCE     TMHMM2.0        inside      43    54
WEBSEQUENCE     TMHMM2.0        TMhelix     55    72
WEBSEQUENCE     TMHMM2.0        outside     73    75
WEBSEQUENCE     TMHMM2.0        TMhelix     76    98
WEBSEQUENCE     TMHMM2.0        inside      99   147
WEBSEQUENCE     TMHMM2.0        TMhelix    148   170
WEBSEQUENCE     TMHMM2.0        outside    171   179
WEBSEQUENCE     TMHMM2.0        TMhelix    180   202
WEBSEQUENCE     TMHMM2.0        inside     203   210
```



III. Protonation probability of the catalytic histidine (His183)

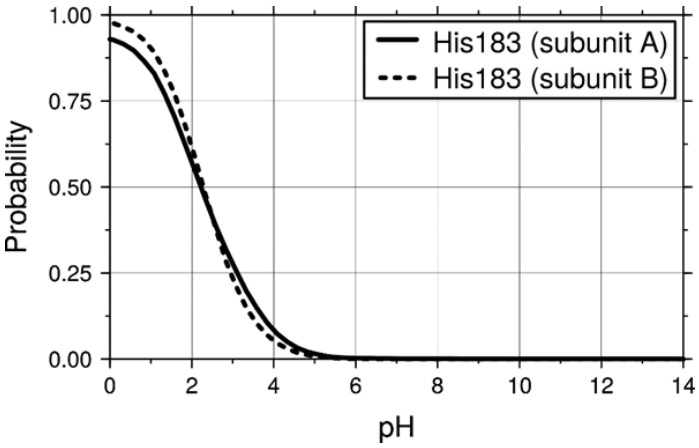


Figure S1: Calculated protonation probability for the protonated state of the catalytic histidine His183 for both subunits of the NsPCS in the acylated form.

We calculated the protonation probability of the catalytic histidine (His183) in the acylated form of the enzyme (Figure S1). The structures were prepared with CHARMM [6], using the CHARMM27 [7] force field. Protonation probabilities were calculated using MEAD [8,9] and GMCT [10]. A Poisson-Boltzmann continuum electrostatic model has been used applying a Metropolis Monte Carlo titration algorithm [11,12]. Protonation probabilities were calculated in steps of 0.25 in the pH range 0 to 14. 200 equilibration scans and 100000 production scans were performed at 300 K, with 0.1 M ionic strength and relative permittivity 4 for the protein and 80 for the solvent.

The probability for His183 changes from fully protonated to single protonated at around pH 2.2 in both subunits of the homodimer NsPCS in its acylated form. The pH at which the structure has been crystallized is pH 5.5. Our protonation probability calculations show that the His183 is single protonated, which shows that His 183 can act as proton acceptor at this pH. This finding supports our hypothesis that the arrest of the enzyme in the acylated form is not caused by the low pH but rather by the oxidized form of the substrate, γEC-SG.

IV. Protonation states of all the titratable residues

We have calculated the protonation state of all the titratable residues for the protein models described in Table S2. The reference pH was pH 8 for NsPCS and pH 7 for papain.

Table S2: List of models used for the simulations

Model	**Description**
NsPCS+GSH	Crystal structure of C70S-NsPCS with noncovalently bound substrate GSH in the binding pocket (PDB ID: **6TJL**). For the simulations the active site serine was mutated to cysteine to model the active wildtype enzyme.
acyl-NsPCS	Crystal structure of acyl-NsPCS with covalently bound γEC-SG (PDB ID: **6THO**). For the simulations only the γEC portion was used.
Papain + Phe-Ser-Ile	Crystal structure with the synthetic inhibitor E-64-c (PDB ID: **1PE6**). Based on the covalently attached inhibitor and the noncovalently attached substrate analogue of another crystal structure of papain (PDB ID: **1PAD**), we modeled the substrate Phe-Ser-Ile.
acyl-papain	Crystal structure with the synthetic inhibitor E-64-c (PDB ID: **1PE6**). Based on the covalently attached inhibitor, we modeled the covalently attached Phe-Ser.

Abbreviations used in the following tables,

1 = protonated

0 = deprotonated

HSE = neutral histidine with Nε protonated and Nδ deprotonated

HSP = protonated histidine with both Nε and Nδ protonated

Model: NsPCS+GSH			
Subunit A		**Subunit B**	
Residue	**Protonation state**	**Residue**	**Protonation state**
Glu40	0	Glu40	0
Glu42	0	Glu42	0
Lys43	1	Lys43	1
Arg49	1	Arg49	1
Arg51	1	Arg51	1
Glu52	0	Glu52	0
Asp53	1	Asp53	1
Tyr69	1	Tyr69	1
Cys70	1	Cys70	1
Glu88	0	Glu88	0
Tyr92	1	Tyr92	1
Tyr95	1	Tyr95	1
Arg96	1	Arg96	1
Asp101	0	Asp101	0
Glu107	0	Glu107	0
Lys108	1	Lys108	1
Lys110	1	Lys110	1
Glu116	0	Glu116	0
Arg120	1	Arg120	1
Asp126	0	Asp126	0
Glu127	0	Glu127	0
Arg130	1	Arg130	1
Tyr135	1	Tyr135	1
Lys138	1	Lys138	1
Lys140	1	Lys140	1
His143	HSP	His143	HSP
Asp146	0	Asp146	0
Glu150	0	Glu150	0
Asp151	0	Asp151	0
Arg153	1	Arg153	1
Lys154	1	Lys154	1
Glu158	0	Glu158	0
Lys161	1	Lys161	1
Asp163	0	Asp163	0
Tyr171	1	Tyr171	1
Arg173	1	Arg173	1
Lys174	1	Lys174	1
Glu175	0	Glu175	0
Glu179	0	Glu179	0
Arg180	1	Arg180	1
His183	HSE	His183	HSE
Tyr190	1	Tyr190	1
Glu192	0	Glu192	0
Asp195	0	Asp195	0
Arg196	1	Arg196	1
Asp201	0	Asp201	0
Arg204	1	Arg204	1
Tyr205	1	Tyr205	1
Lys206	1	Lys206	1
Tyr207	1	Tyr207	1
Lys213	1	Lys213	1
Asp216	0	Asp216	0
Lys219	1	Lys219	1
Asp225	0	Asp225	0
Lys230	1	Lys230	1
Asp232	1	Asp232	1
Gly-C-terminus (GSH)	0	Gly-C-terminus (GSH)	0

Model: acyl-NsPCS			
Subunit A		Subunit B	
Residue	**Protonation state**	**Residue**	**Protonation state**
His23	HSE	Glu40	0
Glu40	0	Glu42	0
Glu42	0	Lys43	1
Lys43	1	Arg49	1
Arg49	1	Arg51	1
Arg51	1	Glu52	0
Glu52	0	Asp53	1
Asp53	1	Tyr69	1
Tyr69	1	Glu88	0
Glu88	0	Tyr92	1
Tyr92	1	Tyr95	1
Tyr95	1	Arg96	1
Arg96	1	Asp101	0
Asp101	0	Glu107	0
Glu107	0	Lys108	1
Lys108	1	Lys110	1
Lys110	1	Glu116	0
Glu116	0	Arg120	1
Arg120	1	Asp126	0
Asp126	0	Glu127	0
Glu127	0	Arg130	1
Arg130	1	Tyr135	1
Tyr135	1	Lys138	1
Lys138	1	Lys140	1
Lys140	1	His143	HSP
His143	HSE	Asp146	0
Asp146	0	Glu150	0
Glu150	0	Asp151	0
Asp151	0	Arg153	1
Arg153	1	Lys154	1
Lys154	1	Glu158	0
Glu158	0	Lys161	1
Lys161	1	Asp163	0
Asp163	0	Tyr171	1
Tyr171	1	Arg173	1
Arg173	1	Lys174	1
Lys174	1	Glu175	0
Glu175	0	Glu179	0
Glu179	0	Arg180	1
Arg180	1	His183	HSE
His183	HSE	Tyr190	1
Tyr190	1	Glu192	0
Glu192	0	Asp195	0
Asp195	0	Arg196	1
Arg196	1	Asp201	0
Asp201	0	Arg204	1
Arg204	1	Tyr205	1
Tyr205	1	Lys206	1
Lys206	1	Tyr207	1
Tyr207	1	Lys213	1
Lys213	1	Asp216	0
Asp216	0	Lys219	1
Lys219	1	Asp225	0
Asp225	0	Lys230	1
Lys230	1	Asp232	1
Asp232	1	Lys239	1
		Asp242	0

Model: papain + Phe-Ser-Ile

Residue	Protonation state
Glu3	0
Tyr4	1
Asp6	0
Arg8	1
Lys10	1
Lys17	1
Cys25	1
Glu35	0
Lys39	1
Arg41	1
Tyr48	1
Glu50	0
Glu52	0
Asp55	0
Asp57	0
Arg58	1
Arg59	1
Tyr61	1
Tyr67	1
Tyr78	1
His81	HSP
Tyr82	1
Arg83	1
Tyr86	1
Tyr88	1
Glu89	0
Arg93	1
Tyr94	1
Arg96	1
Arg98	1
Glu99	0
Lys100	1
Tyr103	1
Lys106	1
Asp108	0
Arg111	1
Tyr116	1
Tyr123	1
Lys139	1
Asp140	0
Tyr144	1
Arg145	1
Lys156	1
Asp158	0
His159	HSE
Tyr166	1
Tyr170	1
Lys174	1
Glu183	0
Tyr186	1
Arg188	1
Lys190	1
Arg191	1
Tyr197	1
Tyr203	1
Tyr208	1
Lys211	1

Model: acyl-papain

Residue	Protonation state
Glu3	0
Tyr4	1
Asp6	0
Arg8	1
Lys10	1
Lys17	1
Glu35	0
Lys39	1
Arg41	1
Tyr48	1
Glu50	0
Glu52	0
Asp55	0
Asp57	0
Arg58	1
Arg59	1
Tyr61	1
Tyr67	1
Tyr78	1
His81	HSP
Tyr82	1
Arg83	1
Tyr86	1
Tyr88	1
Glu89	0
Arg93	1
Tyr94	1
Arg96	1
Arg98	1
Glu99	0
Lys100	1
Tyr103	1
Lys106	1
Asp108	0
Arg111	1
Tyr116	1
Tyr123	1
Lys139	1
Asp140	0
Tyr144	1
Arg145	1
Lys156	1
Asp158	0
His159	HSE
Tyr166	1
Tyr170	1
Lys174	1
Glu183	0
Tyr186	1
Arg188	1
Lys190	1
Arg191	1
Tyr197	1
Tyr203	1
Tyr208	1
Lys211	1

V. References

(1) Madeira, F.; Park, Y. M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A. R. N.; Potter, S. C.; Finn, R. D. et al. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47* (W1), W636–W641. https://doi.org/10.1093/nar/gkz268.

(2) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37*, 420–423. https://doi.org/10.1038/s41587-019-0036-z.

(3) Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E. L. L. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **2001**, *305* (3), 567–580. https://doi.org/10.1006/jmbi.2000.4315.

(4) Sonnhammer, E. L.; von Heijne, G.; Krogh, A. A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences. In *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*; J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. S., Ed.; AAAI Press: Menlo Park, CA, USA, 1998; pp 175–182.

(5) Juncker, A. S.; Willenbrock, H.; von Heijne, G.; Brunak, S.; Nielsen, H.; Krogh, A. Prediction of Lipoprotein Signal Peptides in Gram-Negative Bacteria. *Protein Sci.* **2003**, *12* (8), 1652–1662. https://doi.org/10.1110/ps.0303703.

(6) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**. https://doi.org/10.1002/jcc.540040211.

(7) MacKerell A. D., J.; Bashford, D.; Bellott, M.; Dunbrack R. L., J.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al. All-Atom Empirical Potential for Molecular and Dynamics Studies of Protein. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616. https://doi.org/https://doi.org/10.1021/jp973084f.

(8) Bashford, D.; Gerwert, K. Electrostatic Calculations of the PKa Values of Ionizable Groups in Bacteriorhodopsin. *J. Mol. Biol.* **1992**, *224* (2), 473–486. https://doi.org/10.1016/0022-2836(92)91009-E.

(9) Bashford, D. An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules. An Experience Report on the MEAD Project. In *Lecture Notes in Computer Science*; Yutaka Ishikawa, Rodney R. Oldehoeft, John V. W. Reynders, M. T., Ed.;

Springer: Berlin Heidelberg, 1997; pp 233–244.

(10) Ullmann, R. T.; Ullmann, G. M. GMCT: A Monte Carlo Simulation Package for Macromolecular Receptors. *J. Comput. Chem.* **2012**, *33* (8), 887–900. https://doi.org/10.1002/jcc.22919.

(11) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Protonation of Interacting Residues in a Protein by a Monte Carlo Method: Application to Lysozyme and the Photosynthetic Reaction Center of Rhodobacter Sphaeroides. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88* (13), 5804–5808. https://doi.org/10.1073/pnas.88.13.5804.

(12) Ullmann, G. M.; Knapp, E. W. Electrostatic Models for Computing Protonation and Redox Equilibria in Proteins. *Eur. Biophys. J.* **1999**, *28* (7), 533–551.

Manuscript D

Serine and Cysteine Peptidases – So Similar, Yet Different. How the Active-Site Electrostatics Facilitates Different Reaction Mechanisms

Florian J. Gisdon, Elisa Bombarda, G. Matthias Ullmann, *J. Phys. Chem. B* 2022, 126, 22, 4035–4048
DOI: 10.1021/acs.jpcb.2c01484

THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Serine and Cysteine Peptidases: So Similar, Yet Different. How the Active-Site Electrostatics Facilitates Different Reaction Mechanisms

Published as part of The Journal of Physical Chemistry virtual special issue "Biomolecular Electrostatic Phenomena".

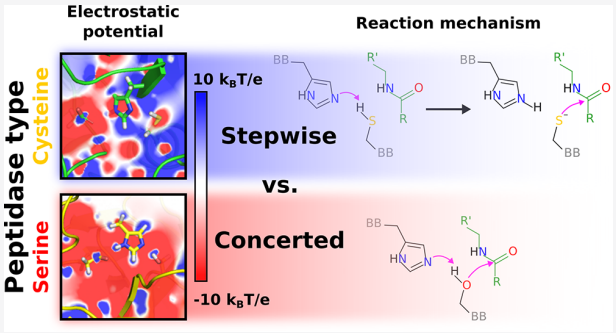
Florian J. Gisdon, Elisa Bombarda, and G. Matthias Ullmann*

Cite This: *J. Phys. Chem. B* 2022, 126, 4035−4048

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The catalytic mechanisms of serine and cysteine peptidases are similar: the proton of the nucleophile (serine or cysteine) is transferred to the catalytic histidine, and the nucleophile attacks the substrate for cleavage. However, they differ in an important aspect: cysteine peptidases form a stable ion-pair intermediate in a stepwise mechanism, while serine peptidases follow a concerted mechanism. While it is known that a positive electrostatic potential at the active site of cysteine peptidases stabilizes the cysteine anion in the ion-pair state, the physical basis of the concerted mechanism of serine peptidases is poorly understood. In this work, we use continuum electrostatic analysis and quantum mechanical/molecular mechanical (QM/MM) simulations to demonstrate that a destabilization of an anionic serine by a negative electrostatic potential in combination with a compact active site geometry facilitates a concerted mechanism in serine peptidases. Moreover, we show that an anionic serine would destabilize the protein significantly compared to an anionic cysteine in cysteine peptidases, which underlines the necessity of a concerted mechanism for serine peptidases. On the basis of our calculations on an inactive serine mutant of a natural cysteine peptidase, we show that the energy barrier for the catalytic mechanism can be substantially decreased by introducing a negative electrostatic potential and by reducing the relevant distances indicating that these parameters are essential for the activity of serine peptidases. Our work demonstrates that the concerted mechanism of serine peptidases represents an evolutionary innovative way to perform catalysis without the energetically expensive need to stabilize the anionic serine. In contrast in cysteine peptidases, the anionic cysteine is energetically easily accessible and it is a very efficient nucleophile, making these peptidases mechanistically simple. However, a cysteine is highly oxygen sensitive, which is problematic in an aerobic environment. On the basis of the analysis in this work, we suggest that serine peptidases represent an oxygen-insensitive alternative to cysteine peptidases.



INTRODUCTION

Cysteine and serine hydrolases are specific enzymes that serve the same basic purpose: the hydrolytic cleavage of polar bonds. The functions of these hydrolases in the cell span from peptide degradation over specific regulation in cellular pathways and the defense against pathogens up to the degradation of substance used as energy source.[1] These different tasks require a broad range of substrate specificities for the different enzymes. However, aside from substrate recognition, the basic features of substrate cleavage appear to be the same for serine and cysteine hydrolases. Both enzyme families contain a catalytic triad or dyad and function through a catalytic mechanism, which appears very similar at first sight (Figure 1). In what follows, we mainly concentrate on peptidases, although most of the statements are valid for cysteine and serine hydrolases in general.

Cysteine and serine pepdidases differ from each other in their nucleophile, which affects the catalytic process. The cleavage in cysteine peptidases proceeds in a stepwise manner; namely, an ion-pair intermediate forms, in which the catalytic histidine has accepted the proton of the thiol group and the thiolate anion acts as activated nucleophile.[2] In serine peptidases instead, the cleavage is a concerted process with nucleophilic attack of the substrate and proton transfer to the catalytic histidine occurring simultaneously.[3] The very similar amino acids cysteine and

Received: March 2, 2022
Revised: April 24, 2022
Published: May 24, 2022

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

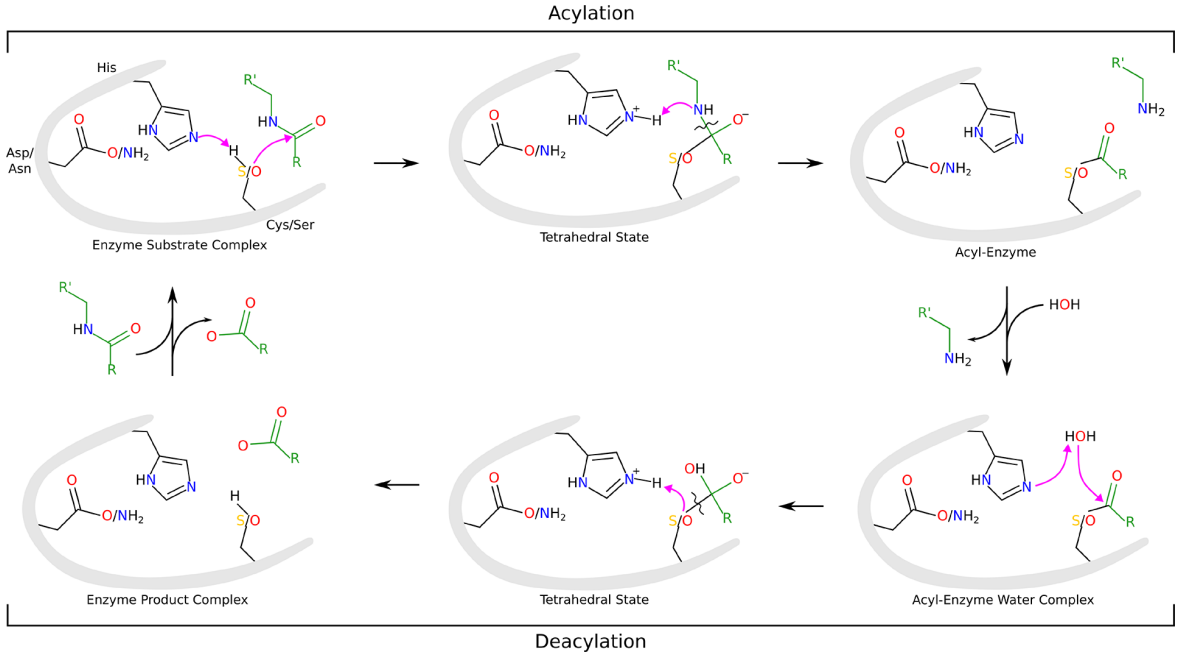


Figure 1. General catalytic mechanism of cysteine and serine peptidases. During acylation the substrate (green) is attacked by the respective nucleophile, cysteine or serine, and its proton is accepted by the catalytic histidine resulting in a tetrahedral state. Subsequently, the substrate is cleaved and the C-terminal part of it leaves the active site while the N-terminal part forms an ester bond with the enzyme (acyl−enzyme). After release of the C-terminal part of the substrate, a water molecule initiates the hydrolytic cleavage in the deacylation reaction, which again results in a tetrahedral state. The acyl−enzyme ester bond is cleaved (enzyme product complex), and also the N-terminal part of the substrate is released.

serine differ only slightly in geometry because of the different atomic radii of sulfur and oxygen. However, their side-chain pK_a values differ substantially, namely, around 9.5 for the thiol group[4,5] and 15.9 for the hydroxyl group.[6] Consequently, deprotonation of the catalytic cysteine is easier and makes the thiol group an effective nucleophile at physiological pH.[7,8] The formation of a reactive thiolate anion is stabilized by hydrogen bonds and a positive electrostatic environment around the catalytic cysteine.[2,7,9] In serine peptidases, it was found that a negative electrostatic potential from the catalytic aspartate contributes to the transition state stabilization.[10] It was further suggested that the ionized aspartate would provide a stabilizing effect to the concerted mechanism.[10] An interesting finding was reported by a recent study,[11] which compares soluble serine peptidases with membrane peptidases, which only possess a catalytic dyad consisting of serine and histidine. This study points out that serine peptidases with a catalytic dyad lack the electrostatic stabilization of the transition state due to the aspartate, which can be correlated with the observed slower reaction rate. Thus, it seems that the electrostatic contribution from the catalytic aspartate is not essential for the activity of serine peptidases. Nevertheless, it is not clear which conditions for the catalytic serine at the active site are important for efficient serine peptidase catalysis.

The effect of different active site environments on cysteine and serine peptidase catalysis can be studied by changing the electrostatics of the binding pocket. A complementary approach is to maintain the electrostatics of the binding pocket, which differs in cysteine peptidases and serine peptidases, and exchange the nucleophile. It was found that the cysteine peptidase papain shows no activity toward natural substrates after mutating its nucleophile to serine.[12] Similarly, the cysteine mutant of the serine peptidases subtilisin loses activity toward natural substrates,[13] and thiol−trypsin, the cysteine mutant of the serine peptidase trypsin, shows reduced substrate turnover by 6 orders of magnitude.[14] Nevertheless in both cases, a perceptible activity has been observed with activated ester substrates. Recently, an adapted cysteine−lipase mutant was presented with increased activity compared to the wild-type protein with serine.[15] In this case, however, the characterization was done using a bulky activated ester substrate, which is more easily cleaved compared to natural substrates and has difficulties binding to the wild-type enzyme. Thus, the increased activity of this cysteine−lipase mutant arose most likely from improved enzyme−substrate interactions and not from changes of the active site environment.

A computational study compared the mechanism of thiol−trypsin with that of the cysteine peptidase papain using various quantum chemical calculations on cluster models.[7] The results of this study suggest that thiol−trypsin does not possess a positive electrostatic potential at the active site, which could stabilize the negatively charged thiolate intermediate as in papain. Therefore, in thiol−trypsin, the formation of an ion-pair intermediate is difficult, and thus the nucleophilic attack occurs most likely by a protonated thiol. But the thiol group is a less efficient nucleophile compared to the hydroxyl group of a serine in line with the lower efficiency of thiol−trypsin compared to its wild type with serine. This reduced activity can be attributed to the low nucleophilic character of the thiol group compared to the hydroxyl group,[14] which is explained by the lower electronegativity of sulfur (electronegativity of 2.4) compared to oxygen (electronegativity of 3.5). These results indicate that it is not only the pK_a of the catalytic side chain that determines the efficiency of the reaction.

In this paper, we investigate the geometry and the electrostatic properties of the active site of cysteine and serine peptidases and

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

correlate our findings with a QM/MM analysis of the mechanism. We analyze the catalytic mechanisms of a cysteine peptidase (phytochelatin synthase from *Nostoc* sp., NsPCS), its inactive serine mutant C70S-NsPCS, and the serine peptidase trypsin. Starting from the inactive serine mutant, where the concerted step of the acylation reaction shows the highest energy barrier, we computationally introduced additional mutations in order to lower the transition state energy of this rate-determining step. This analysis provides a better understanding of the active site of trypsin-like serine peptidases. It is suggested that the main difference in the catalytic cycle between cysteine and serine peptidases is a stepwise versus a concerted mechanism. We have found that this difference is a consequence of the electrostatic nature of the active sites, which is necessary to support the different character of the respective nucleophiles. We conclude that the concerted mechanism of trypsin-like serine peptidases is an evolutionary requirement for their enzymatic activity.

COMPUTATIONAL METHODS AND SETTINGS

Preparation of Protein Structures. For all studies in this work, protein structures were prepared using the program CHARMM[16] with the CHARMM27[17] force field parameters. If present, disulfide bonds were built using CHARMM. Missing hydrogen atoms were added with CHARMM using HBUILD. Protonation states of titratable groups were adjusted according to protonation probabilities, which were obtained by continuum-electrostatic calculations described below. The catalytic histidine was always set to the active form, i.e., in the neutral form to accept the proton from the nucleophile. A water layer of 6 Å was added around the protein, and all water molecules available from the crystal structure were kept. Positions of all water molecules were optimized. *In silico* mutations were introduced using the mutagenesis wizard of PyMOL.[18] For the mutated structures, hydrogen atoms and water molecules together with the mutated residues were optimized with CHARMM, while all other atoms were kept fixed.

For studies on the serine peptidase trypsin, the X-ray crystal structure with the PDB code 1MCT was used as a starting structure.[19] Similar to other studies,[3,20] the dipeptide Arg-Ile with acetylated N-terminus and N-methylated C-terminus was used as substrate. This substrate was modeled in the active site on the basis of the inhibiting peptide MCTI-A present in the structure 1MCT. Six disulfide bonds were taken into account connecting Cys22−Cys152, Cys40−Cys56, Cys124−Cys225, Cys131−Cys198, Cys163−Cys177, and Cys188−Cys212.

For studies on the cysteine peptidase papain, PDB code 1PE6 was used as a starting structure.[21] As in other studies,[22,23] we use the substrate Phe-Ser-Ile with acetylated N-terminus and N-methylated C-terminus. The coordinates of the substrate were modeled on the basis of the substrate analog from the structure with PDB code 1PAD.[24] The modeled substrate was optimized with CHARMM. Three disulfide bonds were taken into account connecting Cys56−Cys95, Cys22−Cys63, and Cys200−Cys153.

For studies on the cysteine peptidase NsPCS, the crystal structure of the serine mutant of NsPCS (C70S-NsPCS,[25] PDB code 6TJL) was used as a starting structure. The substrate glutathione (GSH) is bound noncovalently in the binding pocket. For computational studies on wild-type NsPCS, the active site serine in 6TJL was mutated to cysteine with PyMOL.

Continuum-Electrostatic Calculations. The contiuum electrostatic calculations are based on the Poisson−Boltzmann continuum electrostatic model.[26] In all calculations, the dielectric permittivity was set to 4 for the protein and to 80 for the solvent using a probe sphere radius of 1.4 Å. The ionic strength was set to 0.1 M, and the ionic radius was set to 2.0 Å. The temperature was set to 300 K. Electrostatic potential maps were calculated with the program APBS.[27]

Protonation probabilities of titratable residues were calculated using MEAD[28,29] and GMCT.[30] These calculations are based on a Poisson−Boltzmann continuum electrostatic model with Monte Carlo titration[31] using a Metropolis Monte Carlo algorithm.[32] Protonation probabilities were calculated in the pH range from 0 to 14 in steps of 0.25 pH units. Calculations for every pH step included 200 equilibration scans and 100 000 production scans at 300 K.

Protonation energies were calculated based on the Tanford−Roxby pK_a value (pK_{TR})[33] using the protonation probabilities obtained from the Monte Carlo calculations.[34] pK_{TR} values were calculated using multiflex from the MEAD package[35] to perform the Poisson−Boltzmann calculations and using cmct[36] to perform Monte Carlo titration. Titrations were performed in pH range 0−14 with a step size of 0.01 pH units. For every pH step, 100 equilibration scans and 500 000 production scans were performed at 300 K. The states of catalytic histidines were set to their reactive neutral state so that only the pK_{TR} value of that histidine site accepting the proton during catalysis was calculated.

Setup of QM/MM Models. The protein structures, which were prepared with CHARMM, were divided into a quantum-mechanical (QM) and a surrounding molecular-mechanical (MM) region. QM/MM calculations were performed with pDynamo[37] in combination with ORCA.[38] A link-atom scheme was used for the QM/MM boundary, and the influence of the MM part on the QM part was treated with electrostatic embedding as implemented in pDynamo. The QM region was surrounded by flexible MM atoms, while a harmonic restraint was applied on atoms beyond 8 Å from any QM atom. Restraints of MM atoms within 8−16 Å around any QM atom were treated with linearly increasing force constants from 0 kcal·mol^{-1}·Å^{-1} to 12 kcal·mol^{-1}·Å^{-1}. Restraint force constants for MM atoms further away were set to 12 kcal·mol^{-1}·Å^{-1}. For QM calculations, the B3LYP[39] functional with def2-TZVP[40,41] basis set and RIJCOSX approximation[42,43] was used, and for MM calculations, the CHARMM27[17] force field was used. All side chains included in the QM region were truncated between C_α and C_β unless mentioned differently. The QM regions of trypsin, papain, and NsPCS are illustrated in Figure S1 of the Supporting Information. The QM region of trypsin comprises the catalytic triad residues Ser192 (with atoms N, H, C_α and H_α), His55 (protonated on δ position), Asp99, and the backbone of Asp191 and Gly190, which form the oxyanion hole, without side chains (truncated between C_α and C_β), together with backbone atoms C and O of residue Gln189. For the substrate Arg-Ile, the backbone with N-terminal acetylation of the residue Arg together with the atoms N, H, C_α and H_α of the residue Ile was included into the QM region. The QM region of papain comprised the catalytic triad residues Cys25, His159 (protonated on ε position), and Asn175, as well as Gln19 (truncated between C_β and C_γ), which forms the oxyanion hole. For the substrate Phe-Ser-Ile, the residue Ser with backbone, together with atoms O and C of the residue Phe and atoms N, H, C_α and H_α of the residue Ile, belongs to the QM region. For the NsPCS homodimer, QM/MM calculations were performed on the active site of subunit B. The QM region of NsPCS and its

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

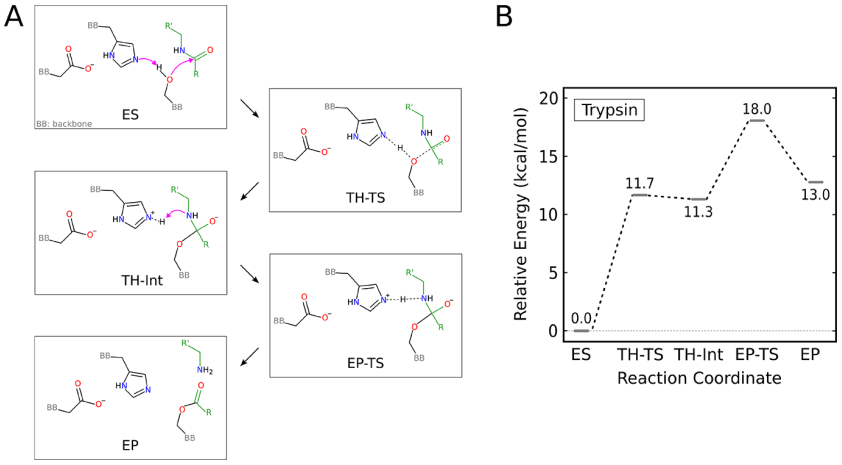


Figure 2. (A) Concerted catalytic mechanism of the acylation reaction in the serine peptidase trypsin. Nucleophilic attack of the substrate (green) in the enzyme substrate complex (ES) is facilitated by concerted proton transfer to the catalytic histidine in the tetrahedral transition state (TH-TS), which results in a metastable tetrahedral intermediate state (TH-Int). In the second transition state (EP-TS), the protonated histidine has reoriented toward the substrate and proton transfer to the substrate occurs. Upon protonation, the substrate gets cleaved, which results in the acylated enzyme product complex (EP). (B) QM/MM energy profile of the acylation reaction of trypsin.

mutants comprised the catalytic triad residues Cys70/Ser70, His183 (protonated on ε position), and Asp201, as well as Gln64 (truncated between C_β and C_γ), which forms the oxyanion hole. Due to its interactions with the catalytic aspartate, Arg173 (truncated between C_γ and C_δ) was included in the QM region. The complete residues Gly and Cys of the substrate GSH together with the atoms C_δ, O_ε, and C_γ (with both its hydrogens) of the residue γGlu belong to the QM region. Further, one water molecule, which interacts with the carboxyl group of the substrate, was included in the QM region. For all models, the parts of the system, which are not part of the QM region, form the MM region.

Reaction Path Search. Initial structures for searches of the reaction paths were obtained by optimizing the prepared protein structures, which were set up as a QM/MM model with pDynamo. Optimization was performed by a conjugate gradient minimizer with RMS gradient threshold set to 0.005 kcal·mol^{-1}·Å^{-1}. Mayer bond orders[44] were obtained directly from QM/MM optimizations. Adiabatic surface scans were performed to find a reaction path estimate. For this scan, one distance constraint was applied between the proton of the nucleophile and the nitrogen of the catalytic histidine that accepts the proton. A second distance constraint was applied between the catalytic nucleophile and the carbonyl carbon of the cleavable bond of the substrate. The RMS gradient for surface scans was set to 0.02 kcal·mol^{-1}·Å^{-1}. Found intermediate states were optimized by a conjugate gradient minimizer. Reaction path estimates, which connect stable states, were obtained from a growing string method,[45] as it is implemented in pDynamo. Stable intermediates and reaction path estimates were used as input for PyCPR,[46] an implementation of the conjugate peak refinement[47,48] (CPR) algorithm for pDynamo, to obtain transition states. Transition states were characterized by vibrational frequency analysis. For all stationary points of the reaction path, the commonly used zero point energy correction was applied.[49] The Supporting Information provides tables, which supplement the reaction path energy profiles with absolute energies, zero point energy corrections, and information about imaginary frequencies (Tables S3−S13).

RESULTS AND DISCUSSION

Computational Analysis of Mechanisms of Cysteine and Serine Peptidases. As explained above (see Figure 1) the mechanism of cysteine and serine peptidases consists of two parts, the acylation and the deacylation. Because we are interested in the properties of the active-site nucleophiles, i.e., cysteine and serine, we focus on the acylation reaction, i.e., the step in which the nucleophile is essential. The deacylation step consisting in the hydrolytic cleavage of the enzyme−substrate ester is instead initiated by a water molecule. For the comparative investigation, we analyze the mechanisms of papain and trypsin, whose catalytic triads are Cys-His-Asn and Ser-His-Asp, respectively. Moreover, we analyze NsPCS, a member of the papain superfamily, whose catalytic triad Cys-His-Asp includes an aspartate in analogy to trypsin (Table S1). The comparison of the active sites of cysteine peptidases and serine peptidases reveals different properties of the surroundings of the respective nucleophiles. We investigated the differences in geometry and electrostatics, and we analyzed our findings by comparing QM/MM reaction paths. We consider that the acylation reaction follows a stepwise mechanism if the reaction path includes an ion-pair intermediate, i.e., an additional energy minimum in which the proton is transferred from the nucleophile to the histidine. Otherwise we consider the reaction to be concerted. Reaction paths for trypsin, papain, and NsPCS obtained from QM/MM calculations are shown in supplementary movies.

Trypsin. Our calculations on the serine peptidase trypsin show that the nucleophilic attack of the substrate and the proton transfer from the active-site nucleophile to the catalytic histidine proceeds in a concerted manner with an energy barrier of 11.7 kcal·mol^{-1} (Figure 2). The resulting tetrahedral intermediate is metastable and only 0.4 kcal·mol^{-1} lower than the preceding transition state. In the next step, the acyl−enzyme is formed by proton transfer from the catalytic histidine to the substrate, which is subsequently cleaved with an overall activation barrier of 18.0 kcal·mol^{-1}. This activation barrier is in the range of values estimated from experimental data[50] and is in good agreement with values of other computational studies.[3,11,20] The

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

N-terminal part of the substrate remains bound covalently to the enzyme via an ester bond.

The occurrence of a stable tetrahedral intermediate, which corresponds to a minimum in the energy landscape, is still debated.[51] In our reaction path, we observed a metastable tetrahedral intermediate state. However, after zero point energy correction, the energy difference between the tetrahedral intermediate and the preceding transition state is just 0.4 kcal·mol^{-1} as illustrated in Figure 2B. Thus, a minimum can be found but it is not very pronounced. A recent QM/MM molecular dynamics study identifies also a metastable intermediate as local minimum on the energy surface with a subsequent barrier of about 1.2 kcal·mol^{-1}.[20] However, it should be kept in mind that in QM/MM simulations, nuclei are treated classically and thus zero point energy correction is not taken into account. According to the Heisenberg uncertainty principle, a quantum mechanical system fluctuates in the ground state, which leads to an increase of the energy compared to a classical system (zero point energy correction). Therefore, the minimum of a state found on a Born−Oppenheimer surface is lower than its real quantum mechanical energy and has to be corrected. As a consequence, reaction path states, which are close in energy, may change their relative energy when zero point energy correction is applied. In our work after zero point energy correction, the first transition state (TH-TS) becomes almost equal in energy compared to the metastable intermediate (TH-Int) (see Figure 3). We conclude that an additional tetrahedral intermediate is not always present, and in case a tetrahedral intermediate is found, its energy minimum might be very shallow after zero point energy correction.

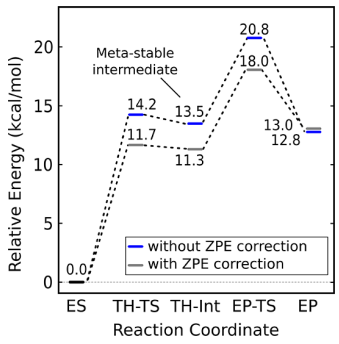


Figure 3. QM/MM energy profile of the acylation reaction of trypsin with (gray) and without (blue) zero point energy (ZPE) correction. The values with the ZPE correction are the same as in Figure 2B.

Papain. In contrast to the concerted mechanism of the serine peptidase trypsin, the cysteine peptidase papain shows a stepwise mechanism (Figure 4). In the first step, the proton transfer from cysteine to histidine leads to the formation of the ion-pair state. In the second step, the cysteine anion attacks the substrate and the histidine reorients to allow a proton transfer to the substrate.

In agreement with previous computational studies,[23,52] we observe a low energy barrier of 1.9 kcal·mol^{-1} for the formation of the ion pair (IP-TS and IP-Int in Figure 4B). In our calculations, the resulting ion-pair intermediate has an energy of 1.8 kcal·mol^{-1} after zero point energy correction. After the formation of this ion-pair intermediate, the cysteine anion attacks the substrate. For this step, some studies of peptidase mechanism mention that the attack of the substrate occurs via a tetrahedral intermediate,[53−55] while other studies do not confirm it.[52,56,57] Our calculations indicate a metastable tetrahedral intermediate, which remains energetically distinguishable after zero point energy correction (TH-Int in Figure 4B), even if this minimum is not very pronounced. Since this minimum is not very pronounced, we think that both possibilities can be found, i.e., a tetrahedral intermediate with subsequent proton transfer from the catalytic histidine to the substrate or the path in which both steps occur at once in a tetrahedral transition state. However, while the occurrence of a tetrahedral intermediate is debated, the occurrence of an ion-pair intermediate seems to be mandatory for cysteine peptidases.

NsPCS. The peptidases papain and trypsin differ not only in the nature of their nucleophile (cysteine vs serine) but also in the nature of the third residue in the catalytic triad. While trypsin possesses a negatively charged aspartate in this position, papain has a neutral asparagine. Mutation of asparagine in papain to aspartate introduces a negative charge, which stabilizes the positively charged state of histidine.[7,10,22] The formation of the ion-pair state would then be promoted, however with the consequence that the proton transfer from the histidine to the substrate becomes more difficult. Since this proton transfer step is essential,[7,22] this mutation would have negative effects for catalysis. To avoid such problems, we used NsPCS for further studies. NsPCS is another member of the papain superfamily, which contains an aspartate in the catalytic triad and is thus directly comparable to trypsin (Table S1). Analogously to papain, the mechanism of NsPCS proceeds stepwise: first the formation of an ion pair and second the nucleophilic attack resulting in the acyl−enzyme. The formation of the ion pair in NsPCS has an energy barrier, which is about 4.5 kcal·mol^{-1} higher than in papain (Figure 4B). Consequently, NsPCS needs slightly more energy to reach the ion-pair state, but the overall barrier of the reaction (9.7 kcal·mol^{-1}) is comparable to that of papain. In contrast to papain, the mechanism of NsPCS proceeds in one step from the ion-pair state to the acyl enzyme. As in trypsin, the path search indicates a metastable tetrahedral intermediate. However, this intermediate appears to be energetically indistinguishable after zero point energy correction. Because of the stepwise mechanism, both papain and NsPCS can be considered as typical representatives of peptidases that use cysteine as nucleophile. In the following discussion we concentrate on NsPCS and trypsin as model enzymes for cysteine peptidases and serine peptidases, respectively, because of the similarity of their catalytic triad.

For NsPCS, a cysteine to serine mutant exists (C70S-NsPCS), which is catalytically inactive.[25] In fact, our calculations showed that the energy barrier of the acylation reaction in C70S-NsPCS is 31.5 kcal·mol^{-1}, rendering this reaction unfeasible. Nevertheless, this reaction path can be analyzed in terms of a reaction mechanism. The energy barrier in C70S-NsPCS is about 22 kcal·mol^{-1} higher than the barrier in the wild type. This energy difference is large considering that the pK_a values of serine and cysteine in aqueous solution differ only by about 6−7 pK_a units, which corresponds to about 8−11 kcal·mol^{-1}. Thus, the large barrier cannot be attributed only to the pK_a value of the nucleophile. Nevertheless, C70S-NsPCS shows the typical concerted mechanism of serine peptidases in which the proton transfer from the catalytic serine to the catalytic histidine and the attack of the substrate occur simultaneously. We therefore conclude that the exchange of the nucleophile from a cysteine to a serine is already sufficient to convert the mechanism of the reaction from stepwise to concerted. However, in order to make

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

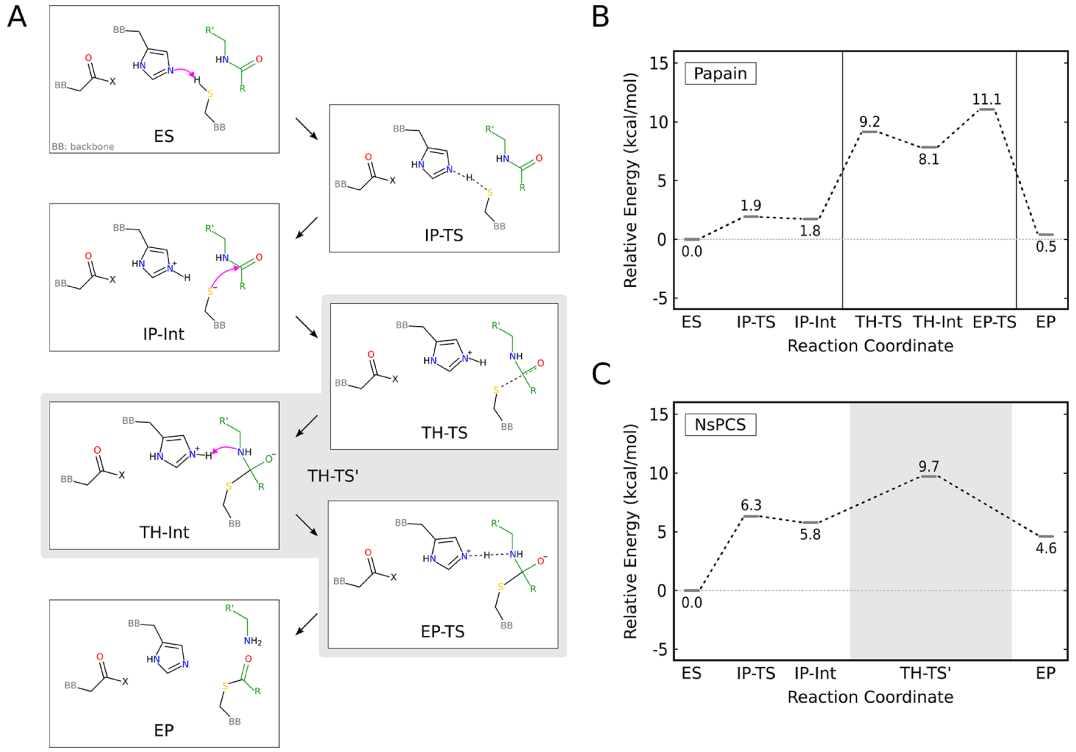


Figure 4. (A) Stepwise catalytic mechanism of the acylation reaction of cysteine peptidases. Since the third residue of the catalytic triad is Asn in papain and Asp in NsPCS, atom label X denotes NH_2 or O^-, respectively. The nucleophilic attack of the substrate (green) in the enzyme substrate complex (ES) occurs stepwise. The first step is the proton transfer from the nucleophile to the catalytic histidine via an ion-pair transition state (IP-TS). The resulting intermediate state IP-Int is the ion pair, which facilitates nucleophilic attack of the substrate. Afterward, the reaction proceeds differently in papain and NsPCS. Papain forms a tetrahedral intermediate TH-Int via a tetrahedral transition state TH-TS and cleaves the substrate via a third transition state EP-TS (states collected in black frame (B)), which results in the enzyme product complex (EP). In NsPCS, the steps, in which the protonated histidine has reoriented toward the substrate and the proton transfer to the substrate occurs, proceed within one step via a tetrahedral transition state TH-TS′ (gray-shaded area in (A) and (C)). The QM/MM energy profiles of papain (B) and NsPCS (C) reflect the states sketched in (A).

the reaction possible, the energy barrier needs to be lowered, which can be achieved by adapting the environment of the active site.

Comparison of the Active Site of NsPCS and Trypsin. The catalytic mechanism of cysteine and serine peptidases is based on a nucleophilic attack which is connected to a proton transfer from the nucleophile to the catalytic histidine. The major difference is that in the catalytic mechanism of cysteine peptidases the proton transfer occurs via an ion-pair intermediate in a stepwise mechanism, while in serine peptidases the proton transfer from the serine to histidine and the nucleophilic attack of the substrate occur in a concerted fashion. We try to find the reason for this mechanistic difference. Mutation studies of cysteine and serine peptidases in which the nucloephile is mutated into the other type report a substantial loss of activity.[12,14] Also in NsPCS, the mutation of the catalytic cysteine to serine causes an inactive enzyme.[25] Our computational analysis of a potential reaction path of the acylation reaction of C70S-NsPCS showed a concerted mechanism as it is characteristic for serine peptidases with an energy barrier of 31.5 kcal·mol^{-1}. This transition state energy corresponds to a reaction rate constant of 1.2×10^{-10} s^{-1} according to the Eyring−Polanyi equation. Such a low reaction rate constant represents the loss of enzyme activity for C70S-NsPCS, in agreement with experimental findings.[25] This high barrier suggests that while the mutation of the nucleophile is sufficient to induce the conversion between a concerted and a stepwise mechanism, it is the environment of each nucleophile that is affecting the catalytic efficiency of the enzymes.

Active Site Geometries. The active site of the serine peptidases trypsin has a compact conformation. From our QM/MM calculations, we find that the hydroxyl hydrogen of the serine is 0.4 Å closer to the hydrogen-bond accepting nitrogen of the catalytic histidine in trypsin (distance 1.8 Å), compared to that in the serine mutant C70S-NsPCS (distance 2.2 Å) (Figure 5A,C). In the cysteine peptidase NsPCS, the thiolyl hydrogen of the cysteine is 2.3 Å apart from the nitrogen of the catalytic histidine. In the serine mutant C70S-NsPCS, the corresponding distance of the hydroxyl hydrogen decreased to 2.2 Å. This finding is counterintuitive, since the introduced serine has shorter bond distances compared to cysteine (Figure 5B,C,F). Namely, according to the CHARMM parameter set,[17] the bond distances $C_\beta-O_\gamma$ and $O_\gamma-H_\gamma$ in serine are 1.42 Å and 0.96 Å, respectively, while the analogous distances in cysteine ($C_\beta-S_\gamma$ and $S_\gamma-H_\gamma$) are 1.81 Å and 1.32 Å, respectively. Thus, the surrounding enzyme must possess some ability to adapt to the introduced serine mutation by orientation of the active site histidine closer to the smaller serine nucleophile. Moreover, we observe also a reorientation of the substrate in C70S-NsPCS toward the nucleophile, which reduces the distance for the nucleophilic attack from 3.2 Å to 3.0 Å.

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

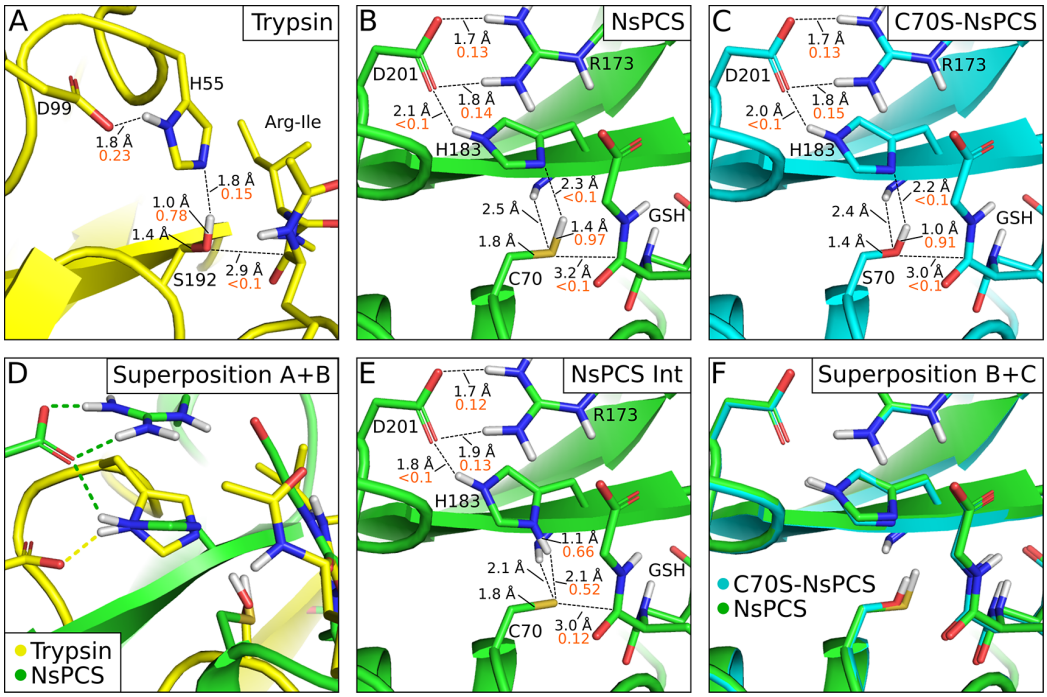


Figure 5. QM/MM optimized active site structures of (A) trypsin (yellow), (B) NsPCS (green), and (C) C70S-NsPCS (cyan), with relevant distances (black numbers) and Mayer bond orders (orange numbers). (D) Superposition of active sites of trypsin and of NsPCS to compare the catalytic triad structures of cysteine and serine peptidases. (E) QM/MM optimized active site structure of the ion-pair intermediate state of NsPCS. (F) Superposition of NsPCS and C70S-NsPCS active sites for catalytic triad structure comparison.

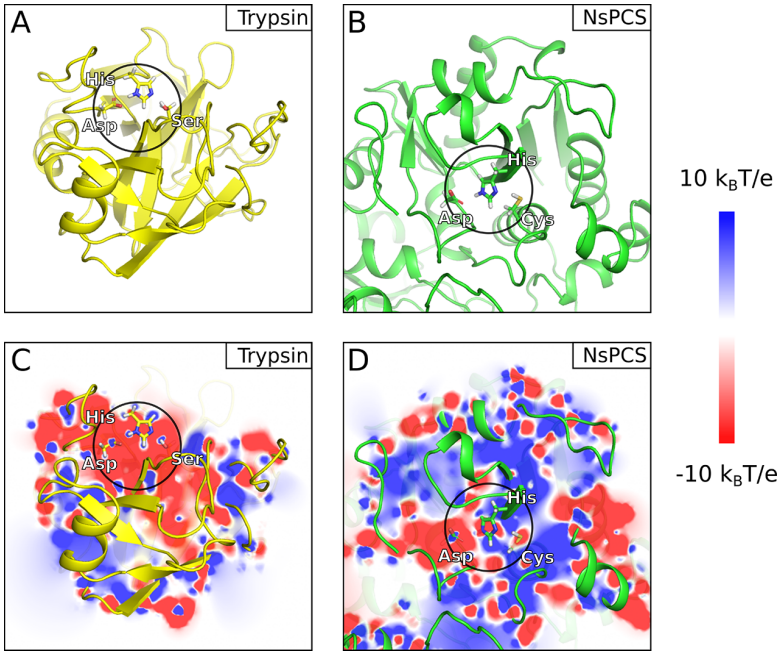


Figure 6. Representations of (A) the serine peptidase trypsin (yellow) and (B) the cysteine peptidase NsPCS (green) with the respective electrostatic potentials (C, D). The electrostatic potential is represented as a slice through the protein along the plane, which contains the catalytic triad residues. The active site region for comparison is encircled. The orientation of the proteins in (C) and (D) is the same as in (A) and (B), respectively.

In addition, we analyze Mayer bond orders, which are a measure of the degree of covalency for the interaction between two atoms (i.e., 0, no covalency; 1, single bond; 2, double bond; etc.). In trypsin, the interaction between one carboxylate oxygen of the catalytic aspartate and the δ-hydrogen of the catalytic histidine has a bond order of 0.23 in the hydrogen bond. This result indicates that the δ-hydrogen is drawn to the aspartate, making the ε-nitrogen available to accept the proton from the catalytic serine during catalysis. The interaction between the γ-hydrogen of the serine hydroxyl group and the ε-nitrogen of the

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

Table 1. Proton Transfer Energies ($\Delta G^{TR}_{pt,Cys/Ser\rightarrow His}$), Interaction Energies $W_{Cys/Ser,His}$ and $W_{sum,Cys/Ser\rightarrow His}$,[a] and Intrinsic p$K_a$ Values (pK_{int}) at pH 8. All Energies Are Given in pK Units[b]

protein (nucleophile)	trypsin (Ser)	NsPCS (Cys)	C70S-NsPCS (Ser)	NsPCS-mut1 (Ser)	NsPCS-mut4 (Ser)
$\Delta G^{TR}_{pt,Cys/Ser\rightarrow His}$	10.7	8.3	13.7	12.0	12.5
$W_{Cys/Ser,His}$	−11.0	−9.5	−9.7	−10.2	−11.3
$W_{sum,Cys/Ser\rightarrow His}$[a]	−2.9	−2.7	−3.0	−4.4	−3.5
pK_{int}					
Cys/Ser	27.2	20.1	26.2	27.0	30.4
His	2.7	−0.4	−0.1	0.4	3.1
Asp	2.0	11.6	11.4	10.2	9.7

[a] $W_{sum,i\rightarrow j} = \sum_{k=1;k\neq j,i}^{N} (\langle x_k(pH)\rangle - x_k^o)(W_{ik} - W_{jk})$, where $\langle x(pH)\rangle$ is the average protonation probability of a site at certain pH (see also eq 1). [b] Conversion factor for pK units to kcal·mol^{-1}: $RT \ln 10 \approx 1.37$, with gas constant R and T = 300 K.

histidine has a bond character of 0.15. Since in serine peptidases, the proton transfer from the catalytic serine to the catalytic histidine does not occur as one separate step of the catalysis, the catalytic histidine has to be able to accept the proton from the serine during the nucleophilic attack (concerted mechanism). Accordingly, our findings show that all atoms in the hydrogen bond network of the catalytic triad are strongly interacting in the reactant state (Figure 5A) so that the catalytic system in trypsin is well-prepared to transfer the proton from the serine to the histidine once the nucleophilic attack proceeds.

The situation is different in cysteine peptidases, where the proton transfer from the catalytic cysteine to the catalytic histidine occurs prior to the nucleophilic attack resulting in the formation of a stable ion-pair intermediate. The thiol orients toward the imidazole to reduce the distance for proton transfer. Afterward the thiolate orients toward the substrate for the nucleophilic attack (stepwise mechanism). Therefore, the catalytic system does not have to be prepared to accept the proton during the nucleophilic attack. This circumstance is reflected in a larger distance and a lower bond order value between the respective imidazole nitrogen and the proton of the nucleophile in NsPCS (Figure 5B) compared to those in trypsin (Figure 5A). Also in the inactive serine mutant C70S-NsPCS, the respective distance is large compared to that in the active trypsin and result in a low bond order (Figure 5C). Thus, shorter distances between the catalytic residues seem beneficial in order to have a lower transition state for an energetically accessible concerted mechanism. To obtain further insights into the ability of the catalytic residues to donate or accept protons, we analyzed the electrostatic potential and the protonation energetics at the active site.

Electrostatic Potentials at the Catalytically Active Site. In order to be able to populate the characteristic ion-pair state, the catalytic site environment in cysteine peptidases has to ensure stabilization of the ion pair. This stabilization is achieved by a positive electrostatic potential around the catalytic cysteine.[2,7,9] Accordingly, we observed a positive electrostatic potential in the surroundings of the nucleophile in NsPCS (Figure 6D). In the whole active site of NsPCS, the electrostatic potential shows positive and negative contributions. In contrast, the active site of trypsin is dominated by a negative electrostatic potential (Figure 6C). In both proteins, the aspartate contributes a negative potential, which enables the stabilization of a protonated histidine in the catalytic cycle. The positive electrostatic contribution in the active site of NsPCS stabilizes the cysteine anion in the ion-pair state as required for the stepwise mechanism. In contrast, the negative electrostatic potential in trypsin prevents the formation of a serine anion as required for the concerted mechanism. Furthermore, the negative potential in the active site makes the histidine a better proton acceptor during the catalytic cycle, enabling serine to act as a strong nucleophile.

Protonation Characteristics of the Residues of the Catalytic Triad. The proton transfer from cysteine or serine to the catalytic histidine is crucial for the mechanism in cysteine and serine peptidases, respectively. Since the nature of this proton transfer is different in the two peptidases, as discussed above, its energetics is important for understanding the mechanism of these enzymes. For this discussion, we need some theoretical considerations. The energy of the proton transfer from residue i to residue j can be estimated using the Tanford−Roxby approximation (eq 1), which assumes that all other titratable residues k remain in their equilibrium protonation.[34]

$$\Delta G^{TR}_{pt,i\rightarrow j}(pH) = -RT \ln 10(pK_{int,j} - pK_{int,i}) + W_{ij}(x_j^o - x_i^o) + W_{sum,i\rightarrow j} \quad (1)$$

The symbols in eq 1 have the following meaning: R is the gas constant, T is the absolute temperature, p$K_{int,i}$ is the intrinsic pK_a of residue i, x_i^o is the reference protonation state of site i, W_{ij} is the interaction energy of sites i and j, and $W_{sum,i\rightarrow j}$ (eq 2) is the pH-dependent interaction energy with the remaining titratable sites, i.e.,

$$W_{sum,i\rightarrow j} = \sum_{k=1;k\neq j,i}^{N} (\langle x_k(pH)\rangle - x_k^o)(W_{jk} - W_{ik}) \quad (2)$$

where $\langle x_k(pH)\rangle$ is the average protonation probability of site k at certain pH (see Supporting Information for derivation). Since we are interested in the energies for the transfer of the proton from the nucleophile to the catalytic histidine in trypsin and NsPCS, i is the respective nucleophile cysteine or serine (Cys/Ser) and j is the catalytic histidine (His). These energies ($\Delta G^{TR}_{pt,Cys/Ser\rightarrow His}$), together with the interaction energies $W_{Cys/Ser,His}$ and $W_{sum,Cys/Ser\rightarrow His}$, and p$K_{int}$ values of the relevant residues are listed in Table 1.

At pH 8 at which both trypsin and NsPCS are active, the calculated proton transfer energy ($\Delta G^{TR}_{pt,Cys/Ser\rightarrow His}$) for trypsin is higher by only 3.3 kcal·mol^{-1} compared to that in NsPCS. This difference in the proton transfer energies is surprisingly small,

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

considering that the pK_a values of free cysteine and free serine in solution differ by about 7 pK_a units, which would correspond to 9.6 kcal·mol^{-1}. Since also the pK_a values of cysteine and serine in NsPCS and trypsine respectively differ by about 7 pK_a units (Table 1), the reason for this relatively low protonation transfer energy in trypsin has to be found elsewhere. The intrinsic pK_a value of the active site histidine in trypsin compared to NsPCS is lower by about 3 pK_a units, which corresponds to a lower protonation energy of about 4.2 kcal·mol^{-1}, making the histidine a better proton acceptor in trypsin. In addition, the interaction between the catalytic residues is stronger in trypsin than in NsPCS (Table 1, $W_{Cys/Ser,His}$). The interaction of the catalytic residues with other titratable residues does not play a major role (see $W_{sum,Cys/Ser\rightarrow His}$ in Table 1). In conclusion, the major reason why this proton transfer energy for trypsin is so surprisingly low is that the histidine in trypsin is a better proton acceptor.

The pK_{int} of serine in the serine mutant C70S-NsPCS is about 6 pK_a units higher than that of cysteine in the wild type, reflecting the difference between the respective solution pK_a values. All other energetic parameters remain about the same. Thus, the mutation of cysteine into serine leads to a significantly higher proton transfer energy. This increased proton transfer energy is also reflected in the high energy barrier of 31.5 kcal·mol^{-1} for C70S-NsPCS obtained from our QM/MM calculations.

The catalytic mechanism of cysteine peptidases relies on the formation of an ion pair, which is stabilized within the active site. Serine peptidases, however, developed a mechanism without the formation of an ion pair, since the deprotonation of serine needs a significantly higher energy compared to cysteine. To maintain an ion-pair state in serine peptidases, the higher energy required to deprotonate the nucleophile would have to be compensated by an increase in protein stability.

Deprotonation and Protein Stability. The mechanism of serine peptidases proceeds concertedly without the formation of an ion pair. In order to have an ion pair in serine peptidases, it would be required to deprotonate the catalytic serine. However, the deprotonation of residues within a protein affects protein stability. To describe the pH dependence of protein stability, we use the following model. We consider a protein with only one titratable site. In the unfolded state, this titratable residue has a very high pK_a value in comparison to physiological pH and it is thus protonated when the protein is unfolded. In the folded state, the pK_a value decreases and the residue deprotonates. The folding energy ΔG_{fold} is given by eq 3,

$$\Delta G_{fold} = \Delta G_{conf} - RT \ln \frac{1 + e^{\ln 10(pK_a + \Delta pK_a - pH)}}{1 + e^{\ln 10(pK_a - pH)}} \quad (3)$$

where ΔG_{conf} is the energy difference between the folded and the unfolded state, if the folding would not be coupled to a protonation event, and ΔpK_a is the difference between the pK_a values in the folded and the unfolded state.

We apply this model in order to investigate which stabilization energy would be required to accommodate a deprotonated cysteine or serine in a protein. It is generally accepted that the pK_a value of an amino acid in the unfolded state is about the same as the pK_a value of the free amino acid. To facilitate the deprotonation of serine in the folded state, the high pK_a value of serine in the unfolded state would have to be drastically lowered by the protein upon folding. In order to keep the protein folded, the conformational energy ΔG_{conf} needs to compensate for the difference in pK_a between the folded and the unfolded state as can be seen in eq 3 (derivation in Supporting Information). To deprotonate a residue at pH 8 with an appreciable probability, the pK_a should not be higher than 10. For cysteine in the unfolded state, the pK_a is about 10. Thus no pK_a shift would be required upon folding in order to deprotonate in the folded state at pH 8, i.e., $\Delta pK_a = 0$ and ΔG_{fold} equals ΔG_{conf}. In contrast, for serine in the unfolded state, the pK_a is about 16. Therefore, in order to deprotonate serine with an appreciable probability in the folded state at pH 8, the pK_a needs to be shifted to about 10, i.e., $\Delta pK_a = -6$, which corresponds to 8.2 kcal·mol^{-1}. Consequently, to stabilize a deprotonated serine in the folded state, ΔG_{conf} would have to be more negative by that amount. Such a large stabilization energy is very difficult to obtain making an ion-pair intermediate unfavorable in serine peptidases.

A recent study[8] pointed out that a stepwise mechanism as it is found in cysteine peptidases is energetically more favorable compared to a concerted mechanism as it is found in serine peptidases. Nevertheless, serine peptidases follow a concerted mechanism, where the environment of the active site favors the uncharged state of the catalytic serine avoiding the explicit formation of an ion pair. Due to the concerted mechanism of serine peptidases, no conformational stabilization of a deprotonated serine is required. It is thus likely that serine peptidases evolved a concerted mechanism to avoid the evolutionary pressure to maintain a large conformational stabilization.

Computational Modification of C70S-NsPCS to Adapt to a Serine−Peptidase-like Active Site. We have observed, in agreement with other studies,[2,54] that cysteine peptidases perform a stepwise mechanism, which involves an ion-pair intermediate (Figure 4). This intermediate state is populated because the cysteine anion is stabilized by a positive electrostatic potential (Figure 6D). Instead in the concerted mechanism of serine peptidases, the uncharged state of the catalytic serine is favored. Accordingly, the nucleophilic attack in serine peptidases occurs in a concerted manner with the proton transfer to the catalytic histidine.

In the following paragraphs, we attempt to mimic a suitable environment around the catalytic triad for serine-based catalysis within the cysteine peptidase NsPCS to computationally restore peptidase activity for its serine mutant C70S-NsPCS. Adaptations are realized for the electrostatic environment of the active site and for the active site distances. The spatial arrangements of the active site residues are shown in Figures S4−S6 in the Supporting Information. An overview of the introduced mutations for the discussed NsPCS mutants is provided in Table 2.

Requirements for an Active Variant of the Cysteine to Serine Mutant C70S-NsPCS. In order to reactivate the serine mutant C70S-NsPCS, we have to reduce the energy barrier of its concerted reaction, where the proton transfer from the

Table 2. Overview of Presented NsPCS Mutant Structures with Introduced Mutations

name	mutation
C70S-NsPCS	C70S
NsPCS-mut1	C70S-S185D-D201S
NsPCS-mut2	C70S-I184P
NsPCS-mut3	C70S-S74P
NsPCS-mut4	C70S-S185D-D201S-I184P-S74P

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

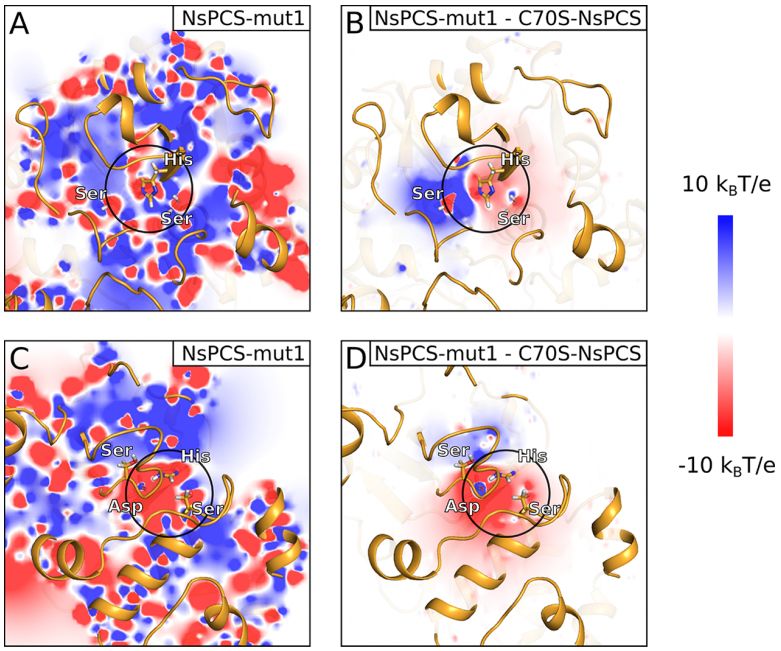


Figure 7. Analysis of the electrostatic potential of NsPCS-mut1 (orange) and C70S-NsPCS (cyan). The active site region is encircled. The electrostatic potential is represented as a slice through the protein. In the upper panels, the residues with the numbers 70, 183, and 201 were used to define the plane (plane I, which is the plane of the catalytic triad residues in NsPCS), and in the lower panels, the residues with the numbers 70, 183, and 185 were used to define the plane (plane II, which is the plane of the new catalytic triad in NsPCS-mut1). (A) Electrostatic potential in NsPCS-mut1 in plane I. (B) Difference of the electrostatic potentials of NsPCS-mut1 and C70S-NsPCS in plane I. (C) Electrostatic potential in NsPCS-mut1 in plane II. (D) Difference of the electrostatic potentials of NsPCS-mut1 and C70S-NsPCS in plane II.

nucleophile to the catalytic histidine is crucial. Therefore, our strategy is to reduce this proton transfer energy.

On the basis of our findings, a decrease of the proton transfer energy $\Delta G^{TR}_{pt,Ser\rightarrow His}$ might be obtained in several ways (or a combination of them): (1) decrease of pK_{int} of the catalytic serine, (2) increase of pK_{int} of the catalytic histidine, (3) stronger interaction between the catalytic serine and the catalytic histidine, or (4) stronger interaction between the catalytic serine and all other sites. The shift of the pK_{int} values can be achieved by modification of the electrostatic potentials at the catalytically active site as analyzed before. Stronger interactions can be obtained also with structural modifications.

Adaptation of the Electrostatic Environment of the Catalytic Triad. The positive electrostatic potential in the cysteine peptidase NsPCS favors ion-pair formation by stabilizing the thiolate. For serine-based catalysis, the predominant electrostatic potential at the active site has to be negative. To achieve a negative potential around the nucleophile in C70S-NsPCS, we change the position of the catalytic aspartate. Namely, the nearby Ser185 is mutated to aspartate, and as a counteracting mutation to maintain the same overall charge within that region, the former catalytic Asp201 is mutated to serine (NsPCS-mut1). By this modification, the electrostatic potential around the catalytic histidine and the nucleophile becomes more negative (Figure 7) as required for a serine-peptidase-like mechanism. Interestingly, the resulting spatial arrangement of the catalytic triad becomes comparable to that of trypsin. Our QM/MM calculations indicate that also the Mayer bond order of the hydrogen bond between the introduced Asp185 and the catalytic histidine is comparable to that in trypsin and higher compared to that in wild-type NsPCS. The mutations introduced in NsPCS-mut1 reduces $\Delta G^{TR}_{pt,Ser\rightarrow His}$ compared to C70S-NsPCS by 1.7 pK_a units (2.3 kcal·mol^{-1}) (Table 1). However, the calculated QM/MM energy for the reaction of NsPCS-mut1 decreases just by 0.5 kcal·mol^{-1} compared to C70S-NsPCS, which indicates that the adaptation of the electrostatic environment of the catalytic triad alone is not sufficient.

Adaptation of the Active Site Geometry. To facilitate the concerted mechanism, which is characteristic for serine peptidases, the serine has to come close to both the catalytic histidine and the substrate. In NsPCS, the amide hydrogen of the peptide bond of Ile184 is oriented toward the catalytic cysteine and contributes to the stabilization of the cysteine–histidine ion pair (Figure 5B). By mutating Ile184 to proline (NsPCS-mut2), this interaction is eliminated, which allows the serine to orient toward the substrate for nucleophilic attack. In addition the C_δ of the side chain of prolin pushes the serine closer to the substrate. The distances of the catalytic serine to the catalytic histidine and of the catalytic serine to the substrate are decreased. Namely, the distance between H_γ of the catalytic serine and N_ε of the catalytic histidine is decreased by 0.2 Å and the distance of O_γ of the catalytic serine and C of the substrate is decreased by 0.1 Å. The calculated QM/MM energy for the reaction in NsPCS-mut2 decreases by almost 4 kcal·mol^{-1} compared to that in C70S-NsPCS, indicating that this mutation is more effective than that in NsPCS-mut1 but still not sufficient.

To decrease the relevant distances in the active site of C70S-NsPCS even further, we modified the α-helix, which has the catalytic Cys70 → Ser on its N-terminal side. This modification was inspired by the serine peptidase subtilisin.[58,59] The mutation of the catalytic serine to cysteine in subtilisin changes the catalytic activity from a peptidase to a ligase. This change in activity was attributed to the fact that the larger cysteine

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

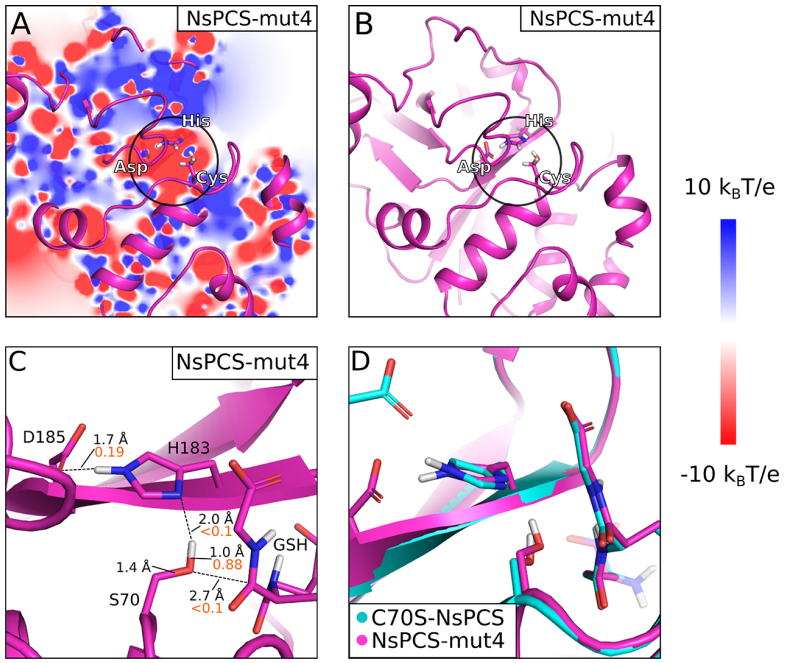


Figure 8. (A) Slice through the electrostatic potential of NsPCS-mut4 in which the protein has the same orientation as in (B). The active site region is encircled. The represented slice through the electrostatic potential map shows the plane, which contains the catalytic triad residues. (C) QM/MM optimized structure of NsPCS-mut4, with relevant distances (black numbers) and Mayer bond orders (orange numbers). (D) Geometrical comparison between the serine mutant C70S-NsPCS (cyan) and NsPCS-mut4.

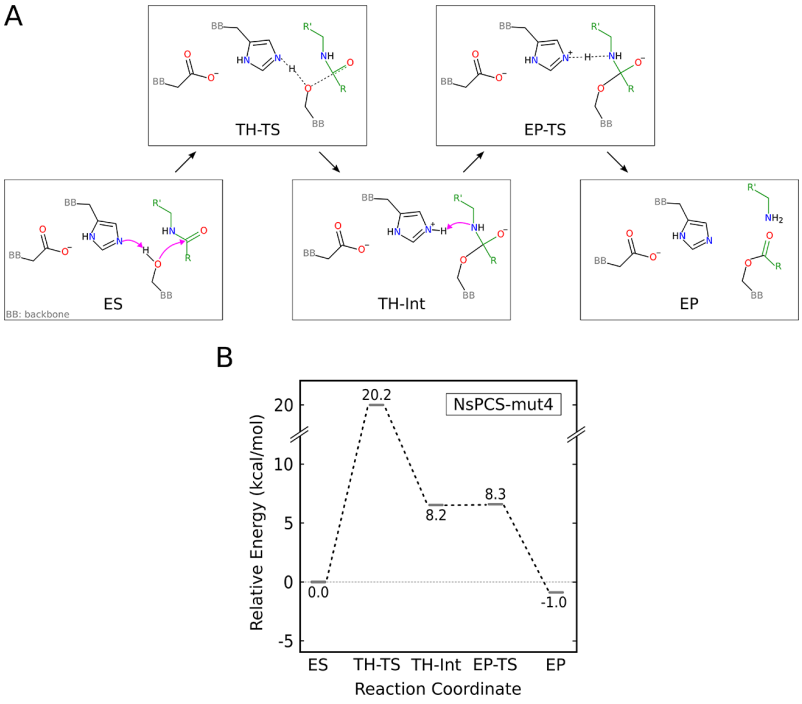


Figure 9. (A) Catalytic mechanism of the acylation reaction of the serine mutant NsPCS-mut4. The nucleophilic attack of the substrate (green) in the enzyme substrate complex (ES) occurs concerted with the proton transfer to the catalytic histidine in the tetrahedral transition state (TH-TS) leading to a tetrahedral intermediate state (TH-Int). In the second transition state (EP-TS), the substrate receives a proton from the catalytic histidine. This proton transfer initiates substrate cleavage, which results in the enzyme product complex (EP). (B) QM/MM energy profile of the acylation reaction of NsPCS-mut4.

nucleophile is too close to the substrate to allow proper substrate orientation as compared to the smaller serine nucleophile. Subtilisin possesses a proline within its active site helix. Mutation of this proline to alanine brought the helix turns closer together and increased the distance between the nucleophile at the end of the helix and the substrate. We need to achieve the opposite

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

effect, i.e., to decrease the distance between the catalytic serine and the substrate. Therefore, we replaced Ser74 in the helix of the active site of C70S-NsPCS by proline (NsPCS-mut3), which is at the same position in the helix as the proline in subtilisin. The distance between O_γ of the catalytic serine and C of the substrate is decreased by 0.1 Å. Additionally, also the distance between H_γ of the catalytic serine and N_ε of the catalytic histidine is decreased by 0.1 Å. However, the calculated QM/MM barrier for the reaction of NsPCS-mut3 increases by almost 1 kcal·mol^{-1} compared to C70S-NsPCS, showing that this single mutation is reducing the relevant distances in the active site but not the energy barrier.

Combining All Four Adaptations for the Active Site Environment. The introduction of all four mutations discussed above into the serine mutant C70S-NsPCS leads to NsPCS-mut4. This mutant has a negative electrostatic environment around the catalytic histidine and the catalytic serine, which is suitable for serine-based catalysis (Figure 8A). Further, NsPCS-mut4 has a compact active site geometry, which increases the interactions among the active site residues compared to those in C70S-NsPCS (Figure 5 and Figure 8C). In fact the distance between the catalytic serine and the substrate, and the proton transfer distance between the catalytic serine and the catalytic histidine decrease from 3.0 Å and 2.2 Å in C70S-NsPCS to 2.7 Å and 2.0 Å in NsPCS-mut4, respectively, as shown in the superposition of C70S-NsPCS with NsPCS-mut4 (Figure 8D). As a consequence of these modifications, the QM/MM barrier for the reaction of NsPCS-mut4 decreases considerably by about 11 kcal·mol^{-1}, i.e., from 31.5 kcal·mol^{-1} for C70S-NsPCS to 20.2 kcal·mol^{-1} for NsPCS-mut4 (Figure 9), which is close to the barrier of active enzymes (for instance we obtained 18.0 kcal·mol^{-1} for trypsin, Figure 2).

All introduced mutations were evaluated by reaction path searches using QM/MM. Each single effect, which we introduced in the mutants NsPCS-mut1, NsPCS-mut2, and NsPCS-mut3, did not significantly reduce the QM/MM energy for catalysis, and also combinations were not sufficient (see Supporting Information, NsPCS-mut5 to NsPCS-mut7). For some combinations, the calculated QM/MM reaction path energy is even higher compared to C70S-NsPCS. Instead, the combination of all presented adaptations leads to a transition state energy of 20.2 kcal·mol^{-1} for NsPCS-mut4 (Figure 9), which is a reasonable barrier for an enzymatic reaction. Thus, the contribution of all four mutations leads to an enzyme that could be active. Taken together, these results show that cysteine and serine peptidases need specifically adapted active site environments, which are essentially different in how they support catalysis. The difference between the two kinds of peptidases is larger than one might think from the many similarities between the two groups of enzymes and shows that a conversion between the two enzyme families is not easily possible.

CONCLUSIONS

In this work, we have analyzed the essential differences of the catalytic mechanisms of cysteine and serine peptidases. At the first sight, the nature of the peptide bond cleavage in cysteine and serine peptidases appears to be very similar. However, catalysis proceeds substantially differently. Cysteine peptidases show a stepwise mechanism with an ion-pair intermediate. In contrast, serine peptidases work through a concerted mechanism avoiding the formation of a deprotonated serine that would significantly destabilize the folded protein. Consequently, the concerted mechanism of serine peptidases leads to an evolutionary advantage, since it does not require a large conformational stabilization. In line with the literature,[10,11] we have shown that a negative electrostatic potential is crucial in the active site of trypsin-like serine peptidases. In fact, a negative electrostatic potential stabilizes the histidine cation in the tetrahedral intermediate. Additionally, we pointed out that also a negative potential around the catalytic serine is important to destabilize the serine anion and enhance the nucleophilicity of the serine toward the substrate in a concerted mechanism without the need of an ion pair.

Our findings can provide an insight also into the nature and the occurrence of cysteine and serine hydrolases in general. The nature of the nucleophile in these two types of enzymes are very different. They differ not only in their protonation properties but also in their redox properties. Namely, cysteine can easily be oxidized, while serine can not. Thus, it is advantageous to use serine as a nucleophile in an aerobic environment. That is probably why serine hydrolases are so widespread in nature. Our results show that the change of the nucleophile from a cysteine to a serine is only possible by adapting the active-site environment to enable a concerted mechanism, which avoids the formation of an ion pair. Namely, the active-site environment has to be changed from "stabilizing the cysteine anion" to "destabilizing the serine anion". The destabilization of a deprotonated serine appears to be counterintuitive, since the catalytic mechanism of peptidases requires a proton transfer from the nucleophile to the catalytic histidine. However, a negative electrostatic potential enhances the nucleophilicity of the protonated serine and allows the attack of the substrate with a simultaneous proton transfer from serine to the catalytic histidine. Moreover, avoiding the appearance of an anionic serine in the concerted mechanism eliminates the need to stabilize the protein harboring the deprotonated nucleophile. Finally, although cysteine and serine peptidases appear to be so similar, the nature of their nucleophiles requires qualitatively different mechanisms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c01484.

Derivations of the equation for the proton transfer energy (eq 1) and of the equation for the folding energy (eq 3); all QM/MM reaction paths supplied as diagrams and tables, which contain absolute energies, zero point energy corrections, and imaginary frequencies; Cartesian coordinates of all presented stationary points; additional figures for some mutants of C70S-NsPCS (PDF)

Movie of reaction path of trypsin (yellow) (AVI)

Movie of reaction path of papain (salmon) (AVI)

Movie of reaction paths of NsPCS (green) (AVI)

Structure information files in PDB format (ZIP)

AUTHOR INFORMATION

Corresponding Author

G. Matthias Ullmann − *Computational Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany;* orcid.org/0000-0002-6350-798X; Email: Matthias.Ullmann@uni-bayreuth.de

The Journal of Physical Chemistry B pubs.acs.org/JPCB Article

Authors

Florian J. Gisdon − *Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany; Computational Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany*

Elisa Bombarda − *Computational Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany;* orcid.org/0000-0002-1385-3710

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c01484

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SFB1357 Project C03.

■ REFERENCES

(1) Schaller, A. A cut above the rest: the regulatory function of plant proteases. *Planta* **2004**, *220*, 183−197.

(2) Mladenovic, M.; Fink, R. F.; Thiel, W.; Schirmeister, T.; Engels, B. On the Origin of the Stabilization of the Zwitterionic Resting State of Cysteine Proteases: A Theoretical Study. *J. Am. Chem. Soc.* **2008**, *130*, 8696−8705.

(3) Ishida, T.; Kato, S. Theoretical Perspectives on the Reaction Mechanism of Serine Proteases: The Reaction Free Energy Profiles of the Acylation Process. *J. Am. Chem. Soc.* **2003**, *125*, 12035−12048.

(4) Nozaki, Y.; Tanford, C. *Methods in Enzymology*; Elsevier, 1967; pp 715−734.

(5) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **2006**, *15*, 1214−1218.

(6) Ballinger, P.; Long, F. A. Acid Ionization Constants of Alcohols. II. Acidities of Some Substituted Methanols and Related Compounds1, 2. *J. Am. Chem. Soc.* **1960**, *82*, 795−798.

(7) Beveridge, A. J. A theoretical study of the active sites of papain and S195C rat trypsin: Implications for the low reactivity of mutant serine proteinases. *Protein Sci.* **1996**, *5*, 1355−1365.

(8) Cuesta, S. A.; Mora, J. R.; Zambrano, C. H.; Torres, F. J.; Rincón, L. Comparative study of the nucleophilic attack step in the proteases catalytic activity: A theoretical study. *Mol. Phys.* **2020**, *118* (14), DOI: 10.1080/00268976.2019.1705412.

(9) Dardenne, L. E.; Werneck, A. S.; de Oliveira Neto, M.; Bisch, P. M. Electrostatic properties in the catalytic site of papain: A possible regulatory mechanism for the reactivity of the ion pair. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 236−253.

(10) Warshel, A.; Naray-Szabo, G.; Sussman, F.; Hwang, J. K. How do serine proteases really work? *Biochemistry* **1989**, *28*, 3629−3637.

(11) Asadi, M.; Oanca, G.; Warshel, A. Effect of environmental factors on the catalytic activity of intramembrane serine protease. *J. Am. Chem. Soc.* **2022**, *144*, 1251−1257.

(12) Clark, P. I.; Lowe, G. Conversion of the Active-Site Cysteine Residue of Papain into a Dehydro-serine, a Serine and a Glycine Residue. *Eur. J. Biochem.* **1978**, *84*, 293−299.

(13) Neet, K. E.; Nanci, A.; Koshland, D. E. Properties of thiol-subtilisin. The consequences of converting the active serine residue to cysteine in a serine protease. *J. Biol. Chem.* **1968**, *243*, 6392−6401.

(14) Higaki, J. N.; Evnin, L. B.; Craik, C. S. Introduction of a cysteine protease active site into trypsin. *Biochemistry* **1989**, *28*, 9256−9263.

(15) Cen, Y.; Singh, W.; Arkin, M.; Moody, T. S.; Huang, M.; Zhou, J.; Wu, Q.; Reetz, M. T. Artificial cysteine-lipases with high activity and altered catalytic mechanism created by laboratory evolution. *Nat. Commun.* **2019**, *10*, 3198.

(16) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(17) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(18) *The PyMOL Molecular Graphics System*, version 2.4.0a0; Schrödinger.

(19) Huang, Q.; Liu, S.; Tang, Y. Refined 1.6 Å Resolution Crystal Structure of the Complex Formed between Porcine β-Trypsin and MCTI-A, a Trypsin Inhibitor of the Squash Family. *J. Mol. Biol.* **1993**, *229*, 1022−1036.

(20) Zhou, Y.; Zhang, Y. Serine protease acylation proceeds with a subtle re-orientation of the histidine ring at the tetrahedral intermediate. *Chem. Commun.* **2011**, *47*, 1577−1579.

(21) Yamamoto, D.; Matsumoto, K.; Ohishi, H.; Ishida, T.; Inoue, M.; Kitamura, K.; Mizuno, H. Refined X-ray Structure of Papain·E-64-c Complex at 2.1-Å Resolution. *J. Biol. Chem.* **1991**, *266*, 14771−14777.

(22) Arad, D.; Langridge, R.; Kollman, P. A. A simulation of the sulfur attack in catalytic pathway of papain using molecular mechanics and semiempirical quantum mechanics. *J. Am. Chem. Soc.* **1990**, *112*, 491−502.

(23) Harrison, M. J.; Burton, N. A.; Hillier, I. H. Catalytic Mechanism of the Enzyme Papain: Predictions with a Hybrid Quantum Mechanical/Molecular Mechanical Potential. *J. Am. Chem. Soc.* **1997**, *119*, 12285−12291.

(24) Drenth, J.; Kalk, K. H.; Swen, H. M. Binding of chloromethyl ketone substrate analogs to crystalline papain. *Biochemistry* **1976**, *15*, 3731−3738.

(25) Gisdon, F. J.; Feiler, C. G.; Kempf, O.; Foerster, J. M.; Haiss, J.; Blankenfeldt, W.; Ullmann, G. M.; Bombarda, E. Structural and biophysical analysis of the phytochelatin-synthase-like enzyme from *Nostoc sp.* shows that its protease activity is sensitive to the redox state of the substrate. *ACS Chem. Biol.* **2022**, *17*, 883−897.

(26) Ullmann, G. M.; Bombarda, E. *Protein Modelling*; Springer International Publishing, 2014; pp 135−163.

(27) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.

(28) Bashford, D.; Gerwert, K. Electrostatic calculations of the pKa values of ionizable groups in bacteriorhodopsin. *J. Mol. Biol.* **1992**, *224*, 473−486.

(29) Bashford, D. *Lecture Notes in Computer Science*; Springer: Berlin, 1997; pp 233−240.

(30) Ullmann, R. T.; Ullmann, G. M. GMCT: A Monte Carlo simulation package for macromolecular receptors. *J. Comput. Chem.* **2012**, *33*, 887−900.

(31) Ullmann, G. M.; Knapp, E.-W. Electrostatic models for computing protonation and redox equilibria in proteins. *Eur. Biophys. J.* **1999**, *28*, 533−551.

(32) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804−5808.

(33) Tanford, C.; Roxby, R. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **1972**, *11*, 2192−2198.

(34) Bombarda, E.; Ullmann, G. M. pH-Dependent pKaValues in Proteins—A Theoretical Analysis of Protonation Energies with Practical Consequences for Enzymatic Reactions. *J. Phys. Chem. B* **2010**, *114*, 1994−2003.

(35) Bashford, D.; Karplus, M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* **1990**, *29*, 10219−10225.

(36) Ullmann, G. M. *CMCT: A Monte Carlo Titration Program Dealing with Conformational Variability*, version 0.6; 2019.

(37) Field, M. J. The pDynamo Program for Molecular Simulations using Hybrid Quantum Chemical and Molecular Mechanical Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1151−1161.

(38) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8*, No. e1327.
(39) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(40) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829−5835.
(41) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057−1065.
(42) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree−Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree−Fock exchange. *Chem. Phys.* **2009**, *356*, 98−109.
(43) Kossmann, S.; Neese, F. Comparison of two efficient approximate Hartee−Fock approaches. *Chem. Phys. Lett.* **2009**, *481*, 240−243.
(44) Mayer, I. Charge, bond order and valence in the AB initio SCF theory. *Chem. Phys. Lett.* **1983**, *97*, 270−274.
(45) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877−7886.
(46) Gisdon, F. J.; Culka, M.; Ullmann, G. M. PyCPR − a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures. *J. Mol. Model.* **2016**, *22*, 242.
(47) Fischer, S.; Karplus, M. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.* **1992**, *194*, 252−261.
(48) Sinclair, J. E.; Fletcher, R. A new method of saddle-point location for the calculation of defect migration energies. *J. Phys. C Solid State* **1974**, *7*, 864−870.
(49) Henkelman, G.; Arnaldsson, A.; Jónsson, H. Theoretical calculations of CH4 and H2 associative desorption from Ni(111): Could subsurface hydrogen play an important role? *J. Chem. Phys.* **2006**, *124*, 044706.
(50) Fersht, A. *Structure and Mechanism in Protein Science*; World Scientific, 2017.
(51) Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* **2002**, *102*, 4501−4524.
(52) Wei, D.; Huang, X.; Liu, J.; Tang, M.; Zhan, C.-G. Reaction Pathway and Free Energy Profile for Papain-Catalyzed Hydrolysis of N-Acetyl-Phe-Gly 4-Nitroanilide. *Biochemistry* **2013**, *52*, 5145−5154.
(53) Shankar, R.; Kolandaivel, P. Reaction mechanism of O-acylhydroxamate with cysteine proteases. *J. Chem. Sci.* **2007**, *119*, 533−544.
(54) Ma, S.; Devi-Kesavan, L. S.; Gao, J. Molecular Dynamics Simulations of the Catalytic Pathway of a Cysteine Protease: A Combined QM/MM Study of Human Cathepsin K. *J. Am. Chem. Soc.* **2007**, *129*, 13633−13645.
(55) Verma, S.; Dixit, R.; Pandey, K. C. Cysteine Proteases: Modes of Activation and Future Prospects as Pharmacological Targets. *Front. Pharmacol.* **2016**, *7*, 107.
(56) Harrison, M. J.; Burton, N. A.; Hillier, I. H.; Gould, I. R. Mechanism and transition state structure for papain catalysed amide hydrolysis, using a hybrid QM/MM potential. *Chem. Commun.* **1996**, 2769.
(57) Han, W.-G.; Tajkhorshid, E.; Suhai, S. QM/MM Study of the Active Site of Free Papain and of the NMA-Papain Complex. *J. Biomol. Struct. Dyn.* **1999**, *16*, 1019−1032.
(58) Nakatsuka, T.; Sasaki, T.; Kaiser, E. T. Peptide segment synthesis catalyzed by the semisynthetic enzyme thiolsubtilisin. *J. Am. Chem. Soc.* **1987**, *109*, 3808−3810.
(59) Abrahmsen, L.; Tom, J.; Burnier, J.; Butcher, K. A.; Kossiakoff, A.; Wells, J. A. Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. *Biochemistry* **1991**, *30*, 4151−4159.

https://doi.org/10.1021/acs.jpcb.2c01484
J. Phys. Chem. B 2022, 126, 4035−4048

Serine and Cysteine Peptidases – So Similar, Yet Different. How the Active Site Electrostatics Facilitates Different Reaction Mechanisms.

Supporting Information

Florian J. Gisdon,[1,2] Elisa Bombarda,[2] G. Matthias Ullmann[2,*]

[1] Biochemistry, University of Bayreuth, Universitätsstr. 30, BGI, 95447 Bayreuth, Germany

[2] Computational Biochemistry, University of Bayreuth, Universitätsstr. 30, BGI, 95447 Bayreuth, Germany

[*] to whom correspondence should be addressed;

e-mail: Matthias.Ullmann@uni-bayreuth.de

S1 Supplementary Derivations

Protonation Properties of the Residues of the Catalytic Triad. For the description of proton transfer processes as they occur in peptidases, knowledge about the protonation energetics is required. The pK_{a} of a titratable site is influenced by the protein environment and by other titratable sites. The pK_{a} value of a titratable site i, if all other sites $j \neq i$ are in their reference state, is described by the intrinsic p$K_{\mathrm{a},i}$ (p$K_{\mathrm{int},i}$). However, depending on pH and the environment, titratable sites do not necessarily occupy their reference state. One way to characterize the individual protonation behavior of site i is the use of the Tanford-Roxby pK_{a} value (p$K_{\mathrm{TR},i}$), which represents an average microscopic pK_{a} of a titratable site.[1] The Tanford-Roxby p$K_{\mathrm{TR},i}$ of site i can be calculated as

$$\mathrm{p}K_{TR,i}(\mathrm{pH}) = \mathrm{p}K_{\mathrm{int},i} - \sum_{j \neq i} (\langle x_j(\mathrm{pH}) \rangle - x_j^o) W_{ij} \tag{1}$$

where $\langle x_j(\mathrm{pH}) \rangle$ is the average protonation probability of site j at certain pH, x_j^o is the reference protonation state of site j, and W_{ij} is the interaction energy of site i with site j. Thus, the pK_{TR} of one site depends on the pK_{int} and on the interactions with all other titratable residues. Such interactions are weighted by the proton probability of the involved residues and are therefore pH-dependent. The protonation energy can be approximated from

$$\Delta \mathrm{G}_{prot,i} = RT \ \ln 10(\mathrm{pH} - \mathrm{p}K_{TR,i}(\mathrm{pH})) \tag{2}$$

where R is the gas constant and T is the absolute temperature, assuming that only residue i changes its protonation from deprotonated ($x_i = 0$) to protonated ($x_i = 1$) and all the other residues remain at the equilibrium protonation $\langle x_j(\mathrm{pH}) \rangle$ at this pH. Analogously, the proton transfer energy from residue i to residue j can be approximated using the Tanford-Roxby approximation, assuming that only residue i changes its protonation from protonated ($x_i^{\mathrm{begin}} = 1$) to deprotonated ($x_i^{\mathrm{end}} = 0$) and residue j changes its protonation from deprotonated ($x_j^{\mathrm{begin}} = 0$)

to protonated ($x_j^{\mathrm{end}} = 1$) and all the other residues remain at the equilibrium protonation $\langle x_k(\mathrm{pH})\rangle$ at this pH.

$$
\begin{aligned}
\Delta G^{TR}_{pt,i\to j}(\mathrm{pH}) = \quad & G^{TR}_{(x_i^{\mathrm{end}},x_j^{\mathrm{end}})}(\mathrm{pH}) - G^{TR}_{(x_i^{\mathrm{begin}},x_j^{\mathrm{begin}})}(\mathrm{pH}) \\
= \quad & (x_i^{\mathrm{end}} - x_i^{\mathrm{o}})RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}) + (x_j^{\mathrm{end}} - x_j^{\mathrm{o}})RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},j}) \\
- \quad & (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}})RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}) - (x_j^{\mathrm{begin}} - x_j^{\mathrm{o}})RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},j}) \\
+ \quad & (x_i^{\mathrm{end}} - x_i^{\mathrm{o}})(x_j^{\mathrm{end}} - x_j^{\mathrm{o}})W_{ij} - (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}})(x_j^{\mathrm{begin}} - x_j^{\mathrm{o}})W_{ij} \\
+ \quad & \sum_{k=1;k\neq j,i}^{N} (x_i^{\mathrm{end}} - x_i^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}})W_{ik} + \sum_{k=1;k\neq j,i}^{N} (x_j^{\mathrm{end}} - x_j^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}})W_{jk} \\
- \quad & \sum_{k=1;k\neq j,i}^{N} (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}})W_{ik} - \sum_{k=1;k\neq j,i}^{N} (x_j^{\mathrm{begin}} - x_j^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}})W_{jk} \\
= \quad & ((x_i^{\mathrm{end}} - x_i^{\mathrm{o}}) - (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}}))RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}) \\
+ \quad & ((x_j^{\mathrm{end}} - x_j^{\mathrm{o}}) - (x_j^{\mathrm{begin}} - x_j^{\mathrm{o}}))RT\ \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},j}) \\
+ \quad & ((x_i^{\mathrm{end}} - x_i^{\mathrm{o}})(x_j^{\mathrm{end}} - x_j^{\mathrm{o}}) - (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}})(x_j^{\mathrm{begin}} - x_j^{\mathrm{o}}))W_{ij} \\
+ \quad & \sum_{k=1;k\neq j,i}^{N} ((x_i^{\mathrm{end}} - x_i^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}}) - (x_i^{\mathrm{begin}} - x_i^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}}))W_{ik} \\
+ \quad & \sum_{k=1;k\neq j,i}^{N} ((x_j^{\mathrm{end}} - x_j^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}}) - (x_j^{\mathrm{begin}} - x_j^{\mathrm{o}})(\langle x_k(\mathrm{pH})\rangle - x_k^{\mathrm{o}}))W_{jk} \qquad (3)
\end{aligned}
$$

With $x_i^{\mathrm{begin}} = 1$, $x_i^{\mathrm{end}} = 0$, $x_j^{\mathrm{begin}} = 0$, and $x_j^{\mathrm{end}} = 1$ we obtain eq. 4, again assuming that the only event is the transfer of a proton from residue i to residue j while all the other titratable

residues k remain at the equilibrium protonation at that pH.

$$\begin{aligned}
\Delta G^{TR}_{pt,i\to j}(\text{pH}) = & \quad ((0-x^{\text{o}}_i)-(1-x^{\text{o}}_i))RT\ \ln 10(\text{pH}-\text{p}K_{\text{int},i}) \\
+ & \quad ((1-x^{\text{o}}_j)-(0-x^{\text{o}}_j))RT\ \ln 10(\text{pH}-\text{p}K_{\text{int},j}) \\
+ & \quad (0-x^{\text{o}}_i)(1-x^{\text{o}}_j)-(1-x^{\text{o}}_i)(0-x^{\text{o}}_j))W_{ij} \\
+ & \quad \sum_{k=1;k\neq j,i}^{N}((0-x^{\text{o}}_i)(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k)-(1-x^{\text{o}}_i)(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k))W_{ik} \\
+ & \quad \sum_{k=1;k\neq j,i}^{N}((1-x^{\text{o}}_j)(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k)-(0-x^{\text{o}}_j)(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k))W_{jk} \\
= & \quad -RT\ \ln 10(\text{pH}-\text{p}K_{\text{int},i})+RT\ \ln 10(\text{pH}-\text{p}K_{\text{int},j}) \\
+ & \quad (x^{\text{o}}_j-x^{\text{o}}_i))W_{ij} \\
+ & \quad \sum_{k=1;k\neq j,i}^{N}-(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k))W_{ik}+\sum_{k=1;k\neq j,i}^{N}(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k))W_{jk} \\
= & - \quad RT\ \ln 10(\text{p}K_{\text{int},j}-\text{p}K_{\text{int},i}) \\
+ & \quad W_{ij}(x^{\text{o}}_j-x^{\text{o}}_i)+\sum_{k=1;k\neq j,i}^{N}(\langle x_k(\text{pH})\rangle - x^{\text{o}}_k)(W_{jk}-W_{ik}) \qquad (4)
\end{aligned}$$

Protein Stability and Deprotonation. We assume to have a protein that has only one titratable site, which is uncharged in the protonated state. We now want to analyze how the protein stability depends on pH. We can define four states of this protein, namely unfolded-deprotonated (ud), unfolded-protonated (up), folded-deprotonated (fd), and folded-protonated (fp). We define the energy of the protein in these four states in general as

$$G_i = \delta_{conf}\Delta G_{conf}+RT\ln 10(x-x^{\circ})(\text{p}K_{\text{a}}+\delta_{conf}\Delta\text{p}K_{\text{a}}-\text{pH}) \qquad (5)$$

where δ_{conf} is 1 if the protein is in the folded state and 0 if the protein is in the unfolded state; x is the protonation state and x° is the reference state ($x^{\circ}=1$ for this example), $\text{p}K_{\text{a}}$ is the $\text{p}K_{\text{a}}$ of the titratable group in the unfolded state (usually this value is more or less identical to the $\text{p}K_{\text{a}}$ value of the titratable group in solution), $\Delta\text{p}K_{\text{a}}$ is the shift of the $\text{p}K_{\text{a}}$ value from the unfolded to the folded state (due to desolvation and interactions with other charges of the protein), ΔG_{conf}

is the energy, which stabilizes the folded state in comparison to the unfolded state. Thus for the four states defined above, we obtain the following energies:

$$G_{up} = G_u(x=1) = 0 \tag{6}$$

$$G_{ud} = G_u(x=0) = -RT\ln 10\ (\mathrm{p}K_a - \mathrm{pH}) \tag{7}$$

$$G_{fp} = G_f(x=1) = \Delta G_{conf} \tag{8}$$

$$G_{fd} = G_f(x=0) = \Delta G_{conf} - RT\ln 10\ (\mathrm{p}K_\mathrm{a} + \Delta\mathrm{p}K_\mathrm{a} - \mathrm{pH}) \tag{9}$$

If we now want to calculate the free energy of stabilization of the folded state, we can define the partition function of the folded and the unfolded state and obtain as free energy change

$$\Delta G_{fold} = -RT\Big(\ln Z_f - \ln Z_u\Big) \tag{10}$$

$$= -RT\ln\frac{Z_f}{Z_u} \tag{11}$$

$$= -RT\ \ln\frac{e^{-(RT)^{-1}\Delta G_{conf}} + e^{-(RT)^{-1}(\Delta G_{conf} - RT\ \ln 10(\mathrm{p}K_\mathrm{a}+\Delta\mathrm{p}K_a-\mathrm{pH}))}}{1+e^{-(RT)^{-1}(-RT\ \ln 10\ (\mathrm{p}K_\mathrm{a}-\mathrm{pH}))}} \tag{12}$$

$$= \Delta G_{conf} - RT\ln\frac{1+e^{\ln 10(\mathrm{p}K_\mathrm{a}+\Delta\mathrm{p}K_\mathrm{a}-\mathrm{pH})}}{1+e^{\ln 10\ (\mathrm{p}K_\mathrm{a}-\mathrm{pH})}} \tag{13}$$

Suppose we have now an enzyme that needs to have its titratable site residue deprotonated in order to be active, it becomes already obvious from eq. 9, that the conformational stability needs to compensate for the pK_a shift. Namely, lets rewrite eq. 9

$$G_{fd} = \Delta G_{conf} - RT\ \ln 10\Delta\mathrm{p}K_\mathrm{a} - RT\ \ln 10(\mathrm{p}K_\mathrm{a} - \mathrm{pH}) \tag{14}$$

For a residue where the pK_a shifts from 16 in the unfolded state to 10 in the folded state, $\Delta\mathrm{p}K_\mathrm{a} = -6$, thus the second term in eq. 14 becomes positive and increases the energy of the folded and deprotonated (fd) state.

S2 Supplementary Tables

The catalytic triads of trypsin, papain and NsPCS differ in the nucleophile or in the third catalytic triad residue or in both, instead C70S-NsPCS and trypsin have identical catalytic triad residues (Table S1).

Table S1: Catalytic triad residues of papain, NsPCS and C70S-NsPCS, and trypsin.

Protein	Catalytic Triad Residues		
Papain	Cys25	His159	Asn175
NsPCS	Cys70	His183	Asp201
C70S-NsPCS	Ser70	His183	Asp201
Trypsin	Ser192	His55	Asp99

Table S2 provides an overview of all mutant variants of C70S-NsPCS, for which QM/MM reaction paths are calculated. The Tables S3-S13 supplement the presented reaction path energy profiles and show the relevant energies and imaginary frequencies for all combinations of the mutations for NsPCS. The listed energies correspond to the provided coordinates in the PDB format of all stationary points, which are named according to the names in the respective table. Additional reaction path energy profiles are shown in Figure S2.

Table S2: Overview of the NsPCS mutant structures with introduced mutations.

Name	Mutation
C70S-NsPCS	C70S
NsPCS-mut1	C70S-S185D-D201S
NsPCS-mut2	C70S-I184P
NsPCS-mut3	C70S-S74P
NsPCS-mut4	C70S-S185D-D201S-I184P-S74P
NsPCS-mut5	C70S-S185D-D201S-I184P
NsPCS-mut6	C70S-S185D-D201S-S74P
NsPCS-mut7	C70S-I184P-S74P

Table S3: QM/MM energies for the acylation reaction of the cysteine peptidase papain, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
papain-1	0.0	257.2	0.0	-
papain-2	3.7	255.4	1.9	-856.7
papain-3	0.5	258.5	1.8	-
papain-4	7.3	259.1	9.2	-61.1
papain-5	5.9	259.5	8.1	-
papain-6	11.6	256.7	11.1	-1139.1
papain-7	-1.2	258.9	0.5	-

Table S4: QM/MM energies for the acylation reaction of the cysteine peptidase NsPCS, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-1	0.0	358.6	0.0	-
nspcs-2	8.0	356.9	6.3	-882.2
nspcs-3	5.0	359.4	5.8	-12.7[a]
nspcs-4	10.7	357.5	9.7	-1003.5
nspcs-5	3.1	360.1	4.6	-

[a] The small imaginary frequency for this intermediate state was ascribed to numerical noise from integration.

Table S5: QM/MM energies for the acylation reaction of the serine mutant peptidase C70S-NsPCS, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
c70s-nspcs-1	0.0	362.1	0.0	-
c70s-nspcs-2	32.3	361.3	31.5	-220.7
c70s-nspcs-3	5.5	362.5	5.9	-

Table S6: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut1, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut1-1	0.0	362.2	0.0	-
nspcs-mut1-2	32.3	360.8	31.0	-164.0
nspcs-mut1-3	19.3	361.5	18.6	-
nspcs-mut1-4	20.2	359.9	17.9	-926.1
nspcs-mut1-5	7.4	362.6	7.8	-

Table S7: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut2, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut2-1	0.0	361.8	0.0	-
nspcs-mut2-2	27.5	360.8	26.6	-224.5
				-49.1[a]
nspcs-mut2-3	4.3	362.3	4.7	-

[a] The small imaginary frequency for this intermediate state was ascribed to numerical noise from integration.

Table S8: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut3, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut3-1	0.0	361.5	0.0	-
nspcs-mut3-2	30.1	360.6	32.1	-245.0
nspcs-mut3-3	4.1	361.9	4.5	-

Table S9: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut4, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut4-1	0.0	362.8	0.0	-
nspcs-mut4-2	21.9	361.1	20.2	-120.2
nspcs-mut4-3	8.3	362.7	8.2	-
nspcs-mut4-4	10.4	360.7	8.3	-1054.2
nspcs-mut4-5	-1.3	363.2	-1.0	-

Table S10: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut5, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut5-1	0.0	362.7	0.0	-
nspcs-mut5-2	27.1	361.1	25.6	-109.4
nspcs-mut5-3	13.0	361.6	12.0	-
nspcs-mut5-4	15.4	359.6	12.3	-1013.0
nspcs-mut5-5	4.7	362.7	4.8	-

Table S11: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut6, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut6-1	0.0	362.6	0.0	-32.3[a]
nspcs-mut6-2	30.6	361.5	29.4	-238.2
nspcs-mut6-3	0.9	363.2	1.4	-

[a] The small imaginary frequency for this intermediate state was ascribed to numerical noise from integration.

Table S12: QM/MM energies for the acylation reaction of the adapted serine mutant peptidase NsPCS-mut7, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
nspcs-mut7-1	0.0	362.2	0.0	-
nspcs-mut7-2	32.3	361.0	31.2	-237.3
nspcs-mut7-3	4.6	362.4	4.8	-

Table S13: QM/MM energies for the acylation reaction of the serine peptidase trypsin, with zero point energy (ZPE) contribution and imaginary frequencies.

Name	Relative Energy (kcal·mol^{-1})	Absolute ZPE (kcal·mol^{-1})	Relative Energy (ZPE corrected) (kcal·mol^{-1})	Imaginary Frequencies (cm^{-1})
trypsin-1	0.0	292.4	0.0	-
trypsin-2	14.2	289.9	11.7	-193.2
trypsin-3	13.7	290.0	11.3	-
trypsin-4	20.8	289.7	18.0	-1094.5
trypsin-5	12.8	292.7	13.1	-

S3 Supplementary Figures

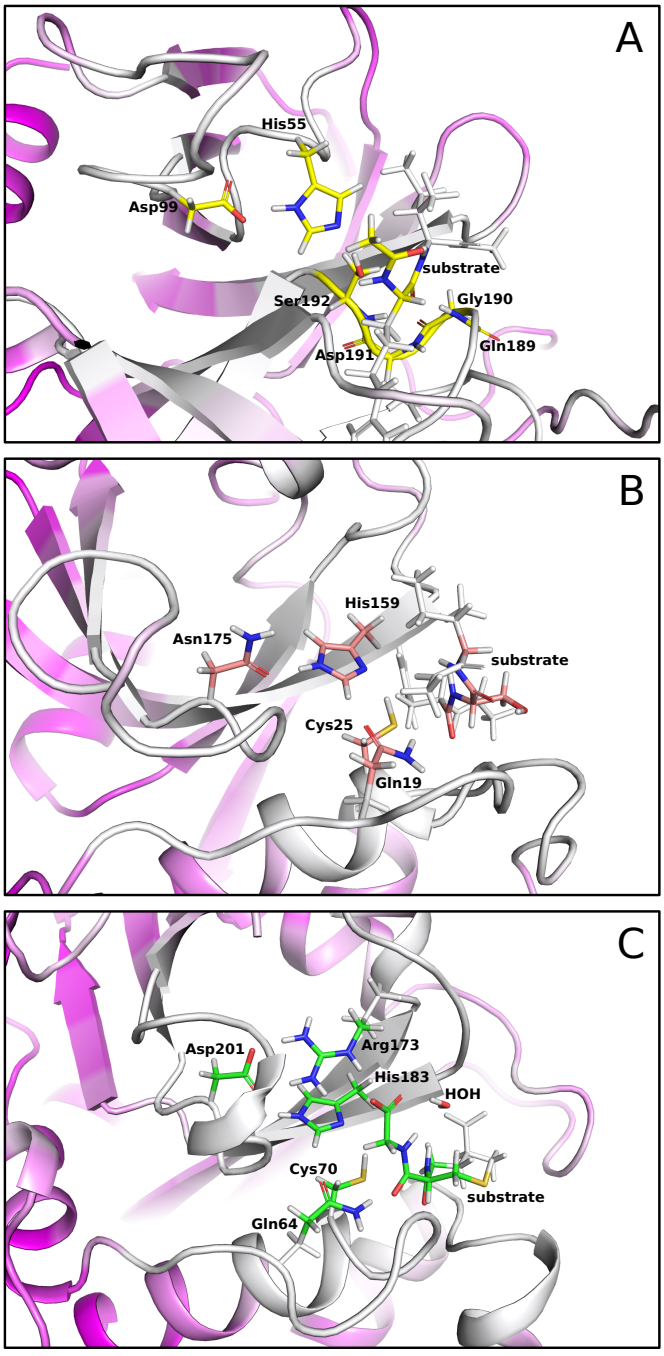


Figure S1: Representation of the QM regions of (A) trypsin, (B) papain, and (C) NsPCS. The QM treated atoms are shown as colored sticks. MM treated atoms are characterized with a gradient from gray to magenta, which indicates zero constraints to full constraints, respectively (see main text for details).

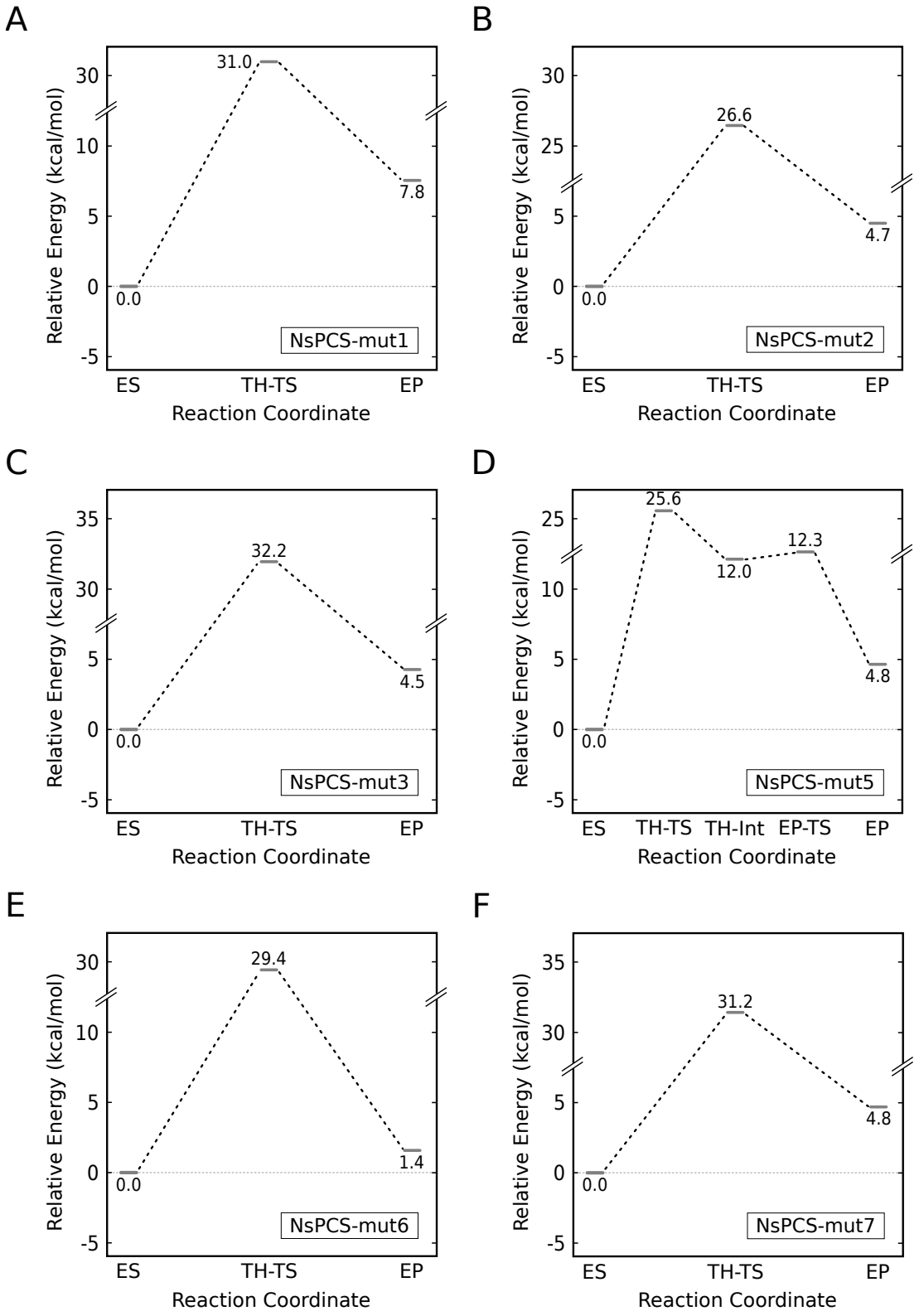


Figure S2: Calculated QM/MM energy diagrams of modeled NsPCS mutant structures.

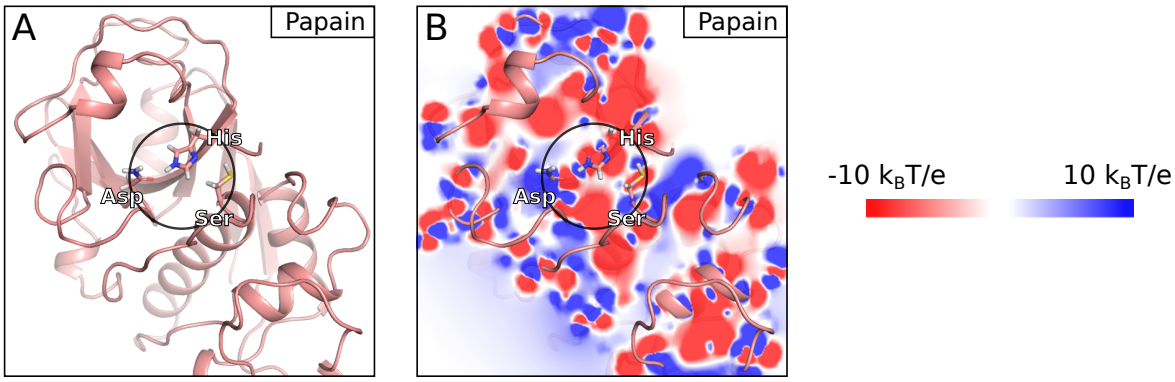


Figure S3: Representation of the cysteine peptidase papain without (A) and with (B) the electrostatic potential. The electrostatic potential is represented as a slice through the protein along the plane, which contains the catalytic triad residues. The active site region is encirled. The orientation of the protein in (A) is the same as in (B). The electrostatic potential map was calculated with APBS based on the protein structure, which was prepared with CHARMM.

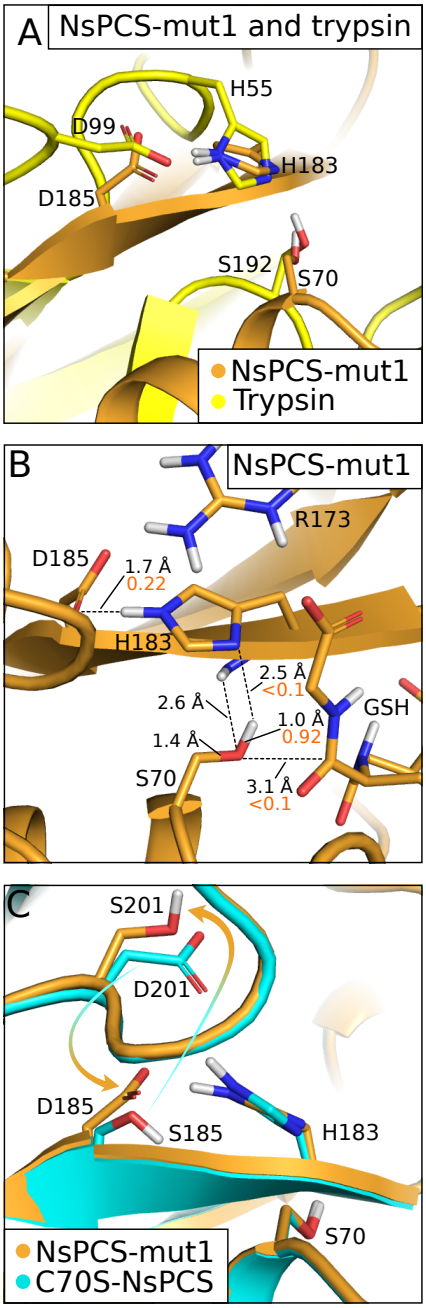


Figure S4: (A) Superposition of the QM/MM optimized structures of NsPCS-mut1 (orange) and trypsin (yellow), which show the geometrical similarities of the catalytic triad in trypsin with the adapted catalytic triad in NsPCS-mut1. (B) QM/MM optimized structure of NsPCS-mut1, showing relevant distances (black numbers) and Mayer bond orders (orange numbers). (C) Structural comparison of NsPCS-mut1 with C70S-NsPCS (cyan), showing the positional interchange mutations S185D and D201S.

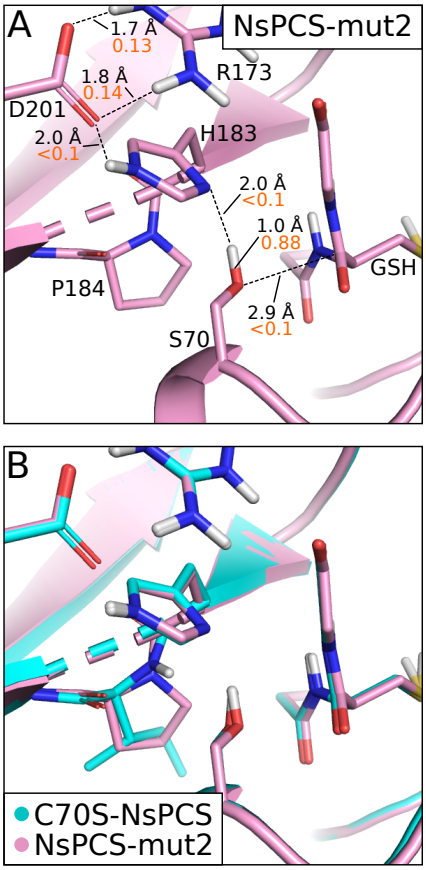


Figure S5: (A) QM/MM optimized active site structure of the NsPCS-mut2 mutant (pink), with relevant distances (black numbers) and Mayer bond orders (orange numbers). (B) Structural comparison of NsPCS-mut2 with the serine mutant C70S-NsPCS (cyan). The mutation of Ile184 to proline eliminates the interaction of the amide hydrogen of the peptide bond of Ile184 and the nucleophile. In addition the C_δ of the sidechain of prolin pushes the serine closer to the substrate.

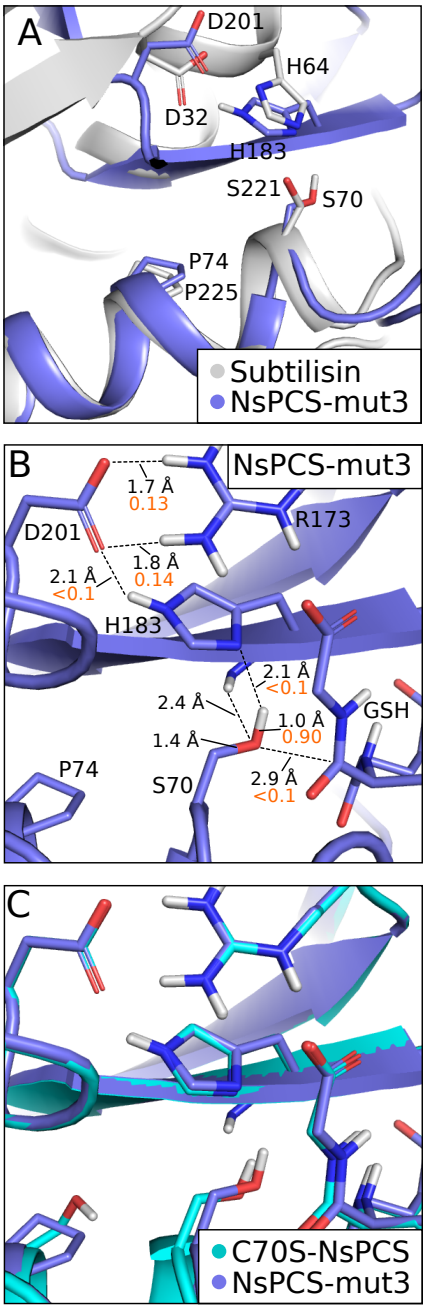


Figure S6: (A) Superposition of the QM/MM optimized structure of NsPCS-mut3 (violet blue) and crystal structure of subtilisin (white) (PDB-ID 1SBC). Subtilisin naturally contains a proline inside its active site α-helix. At the same geometrical position a prolin is introduced in NsPCS to decrease the active site distances. (B) QM/MM optimized structure of NsPCS-mut3, with relevant distances (black numbers) and Mayer bond orders (orange numbers). (C) Geometrical comparison with the serine mutant C70S-NsPCS (cyan), showing the spacial effect of the S74P mutation.

S4 Supplementary Movies

Supplementary movies are provided showing the catalytic mechanisms of trypsin, papain and NsPCS.

References

[1] Bombarda, E.; Ullmann, G. M. pH-Dependent pKaValues in Proteins—A Theoretical Analysis of Protonation Energies with Practical Consequences for Enzymatic Reactions. *J. Phys. Chem. B* **2010**, *114*, 1994–2003.

Publication List

- Martin Culka*, **Florian J. Gisdon***, G. Matthias Ullmann (2017): Computational biochemistry – enzyme mechanisms explored. *Adv. Protein Chem. Struct. Biol.*, 109, 77-112
  DOI: 10.1016/bs.apcsb.2017.04.004

- **Florian J. Gisdon***, Martin Culka*, G. Matthias Ullmann (2016): PyCPR – a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures. *J. Mol. Model.*, 22, 242
  DOI: 10.1007/s00894-016-3116-8

- **Florian J. Gisdon**, Christian G. Feiler, Oxana Kempf, Johannes M. Foerster, Jonathan Haiss, Wulf Blankenfeld, G. Matthias Ullmann, Elisa Bombarda (2022): Structural and biophysical analysis of the phytochelatin-synthase-like enzyme from *Nostoc* sp. shows that its protease activity is sensitive to the redox state of the substrate. *ACS Chem. Biol.*, 17, 4, 883–897
  DOI: 10.1021/acschembio.1c00941

- **Florian J. Gisdon**, Elisa Bombarda, G. Matthias Ullmann (2022): Serine and cysteine peptidases – so similar, yet different. How the active-site electrostatics facilitates different reaction mechanisms. *J. Phys. Chem. B*, 126, 22, 4035–4048 DOI: 10.1021/acs.jpcb.2c01484

* These authors contributed equally to the manuscript

(Eidesstattliche) Versicherungen und Erklärungen

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit versichere ich eidesstattlich, dass ich die Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe (vgl. Art. 64 Abs. 1 Satz 6 BayHSchG).

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit erkläre ich, dass ich die Dissertation nicht bereits zur Erlangung eines akademischen Grades eingereicht habe und dass ich nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden habe.

(§ 8 Satz 2 Nr. 4 PromO Fakultät)

Hiermit erkläre ich, dass ich Hilfe von gewerblichen Promotionsberatern bzw. –vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe noch künftig in Anspruch nehmen werde.

(§ 8 Satz 2 Nr. 7 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass die elektronische Fassung der Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung unterzogen werden kann.

(§ 8 Satz 2 Nr. 8 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass bei Verdacht wissenschaftlichen Fehlverhaltens Ermittlungen durch universitätsinterne Organe der wissenschaftlichen Selbstkontrolle stattfinden können.

. . . . . .

Ort, Datum, Unterschrift