Technische Universität München

TUM School of Life Sciences

Improving the Description of Protein-Ligand Flexibility and Ion Interactions in Computer-Aided Enzyme Engineering and Drug Discovery

Okke Melse

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat)

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Dmitrij Frishman
Prüfer der Dissertation:	1.	Prof. Dr. Volker Sieber
	2.	Prof. Dr. Ville R. I. Kaila

Die Dissertation wurde am 01.07.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 08.11.2022 angenommen.

Table of Contents

i. SUMMARY

Biomolecular simulations continue to evolve and are an integral part in numerous scientific fields, including protein/enzyme engineering and (computer-aided) drug design. These simulations can support the design of enzyme engineering and drug discovery studies, guide further lead optimization, and rationalize subsequent experimental findings. Two of the major challenges in this field remain the inclusion of protein flexibility in biomolecular simulations (*e.g.* during molecular docking and binding site identification simulations), and the accurate description of metal ions in a biomolecular environment. However, exactly these two topics are of great importance for enzyme engineering and drug design, since enzyme active sites as well as most drug targets are highly flexible and often contain a metal ion as cofactor. This dissertation addresses these issues and describes the development of two new algorithms *EnzymeMatch* and *DynaBiS*, a benchmarking study to guide the design of biomolecular simulations containing metal ions, and application studies in both enzyme engineering and drug discovery.

The first part of this dissertation describes the development of EnzymeMatch, a bioinformatics algorithm to identify potential biocatalysts. Enzyme engineering approaches such as directed evolution require a natural enzyme as a starting point with at least measurable activity. The developed algorithm predicts these enzymes, and can thereby support enzyme engineering studies. EnzymeMatch only requires the molecular structure of the target substrate as input, automatically predicts the optimal binding site to accommodate binding thereof, and subsequently searches for enzymes containing such a binding site in the Protein Data Bank, while taking both protein- and ligand flexibility into account.

In the second part of this dissertation, the development of the binding site identification algorithm DynaBiS is described. DynaBiS applies soft-core potentials between the ligand and protein to allow for a fully flexible treatment of the entire system, and thereby result in the simulation of conformational adaptation effects. Comprehensive evaluation showed that DynaBiS outperforms traditional binding site identification algorithms, especially in the identification of binding sites for large and flexible ligands, both with the holo or apo structure used as input.

A benchmarking study of biomolecular force field-based Zn^{2+} models constitutes the third part of this dissertation. Large differences in performance between these Zn^{2+} models were observed, and the preferred coordination geometry and type of ligating atoms were determined. This highly valuable information is necessary to design long timescale simulations. These results led to the recommendation of suitable simulation protocols for a variety of modelling approaches, and a guide to further develop these Zn^{2+} models.

Finally, application studies are described to showcase the use of biomolecular simulations in both enzyme engineering and drug discovery. This includes a study aiming to modulate catalytic properties of [2Fe-2S]-dependent dehydratases, in which biomolecular simulations guided the identification of mutation hotspots, which were experimentally investigated with site-directed and saturation mutagenesis. Moreover, two drug discovery studies describe the characterization of the amidohydrolase PurAH, which is involved in the biosynthesis of a natural antibiotic, and the rationalization of observed clavulanic acid inhibition in B3-RQK metallo-β-lactamases, which can support the design of new drugs.

The algorithms developed in this thesis, the benchmarking study of biomolecular force field-based Zn^{2+} models, and the application studies in enzyme engineering and drug design open up new strategies to further advance rational enzyme engineering and structure-based drug discovery.

ii. ZUSAMMENFASSUNG (DE)

Biomolekulare Simulationen entwickeln sich stetig weiter und sind ein fester Bestandteil zahlreicher wissenschaftlicher Bereiche, darunter Enzym-Engineering und (computergestütztes) Wirkstoffdesign. Diese Simulationen können das Design von Studien zum Enzym-Engineering und zur Entwicklung von Arzneimitteln unterstützen, Leitstrukturen optimieren und experimentelle Ergebnisse rationalisieren. Zwei der größten Herausforderungen in diesem Bereich bleiben die Berücksichtigung der Flexibilität von Proteinen (z. B. bei Simulationen zum molekularen Docking und zur Identifizierung von Bindungsstellen) und die genaue Beschreibung von Metallionen in einer biomolekularen Umgebung. Jedoch sind genau diese beiden Themen für das Enzym-Engineering und das Design von Arzneimitteln von großer Bedeutung, da die aktiven Stellen von Enzymen sowie die meisten Zielstrukturen von Arzneimitteln sehr flexibel sind und häufig ein Metallion als Cofaktor enthalten. Die vorliegende Dissertation befasst sich mit diesen Themen und beschreibt die Entwicklung zweier neuer Algorithmen, EnzymeMatch und DynaBiS, eine Benchmarking-Studie zur Steuerung des Aufbaus von biomolekularen Simulationen von Metalloproteinen sowie Anwendungsstudien in den Bereichen Enzym-Engineering und Wirkstoffdesign.

Der erste Teil dieser Dissertation beschreibt die Entwicklung von EnzymeMatch, einem bioinformatischen Algorithmus zur Identifizierung potenzieller Katalysatoren. Enzym-Engineering-Ansätze wie die gerichtete Evolution erfordern als Ausgangspunkt ein natürliches Enzym mit zumindest messbarer Aktivität. Der entwickelte Algorithmus sagt diese Enzyme voraus und kann so Enzym-Engineering-Studien unterstützen. EnzymeMatch benötigt nur die molekulare Struktur des Zielsubstrats als Eingabe, sagt automatisch die optimale Bindungsstelle für dieses Substrat voraus und sucht anschließend in der Protein Data Bank nach Enzymen, die eine solche Bindungsstelle enthalten, wobei sowohl die Protein- als auch die Ligandenflexibilität berücksichtigt werden.

Im zweiten Teil dieser Dissertation wird die Entwicklung eines Algorithmus zur Identifizierung von Bindungsstellen (DynaBiS) beschrieben. DynaBiS wendet Soft-Core-Potentiale zwischen Ligand und Protein an, um eine vollständig flexible Betrachtung des gesamten Systems zu ermöglichen und dadurch die Simulation von Konformationsanpassungseffekten zu ermöglichen. Eine umfassende Evaluierung hat gezeigt, dass DynaBiS die traditionellen Algorithmen zur Identifizierung von Bindungsstellen übertrifft, insbesondere bei der Identifizierung von Bindungsstellen für große und flexible Liganden, sowohl mit der Holo- als auch mit der Apo-Struktur als Eingabe.

Eine Benchmarking-Studie von biomolekularen Kraftfeld-basierten Zn^{2+}-Modellen bildet den dritten Teil dieser Dissertation. Es wurden große Leistungsunterschiede zwischen diesen Zn^{2+}-Modellen festgestellt und die bevorzugte Koordinationsgeometrie und die Art der ligierenden Atome bestimmt. Dies sind äußerst wertvolle Daten, die für die Entwicklung von Simulationen mit langen Zeitskalen erforderlich sind. Die Ergebnisse führten zur Empfehlung geeigneter Simulationsprotokolle für eine Vielzahl von Modellierungsansätzen und zu einem Leitfaden für die weitere Entwicklung dieser Zn^{2+}-Modelle.

Zum Schluss werden Anwendungsstudien beschrieben, die den Einsatz biomolekularer Simulationen sowohl im Enzym-Engineering als auch in der Arzneimittelforschung veranschaulichen. Dazu gehört eine Studie zur Modulation der katalytischen Eigenschaften [2Fe-2S]-abhängiger Dehydratasen. Biomolekulare Simulationen führten zur Identifizierung von Mutationshotspots, die experimentell mit ortsgerichteter und Sättigungsmutagenese untersucht wurden. Darüber hinaus beschreiben zwei Studien zur Wirkstoffentwicklung die Charakterisierung der Amidohydrolase PurAH, die an der Biosynthese eines natürlichen Antibiotikums beteiligt ist, und die Rationalisierung der beobachteten Clavulansäurehemmung in B3-RQK-Metallo-β-Lactamasen, die die Entwicklung neuer Arzneimittel unterstützen kann.

Die in dieser Arbeit entwickelten Algorithmen, die Benchmarking-Studie von biomolekularen Kraftfeld-basierten Zn^{2+}-Modellen und die Anwendungsstudien im Enzym-Engineering und Wirkstoffdesign eröffnen neue Strategien, um das rationale Enzym-Engineering und das strukturbasierte Wirkstoffdesign weiter voranzutreiben.

iii. SAMENVATTING (NL)

Biomoleculaire simulaties zijn sterk in ontwikkeling en een integraal onderdeel van een tal van disciplines binnen de wetenschap, onder andere bij enzym-engineering en (computergestuurde) medicijnontwikkeling. Twee van de grootste uitdagingen op dit gebied blijven het incorporeren van eiwitplasticiteit (b.v. tijdens docking simulaties en de identificatie van de mogelijke eiwit-ligand bindingsplaatsen), en het nauwkeurig beschrijven van metaalionen in een biomoleculaire omgeving. Echter, juist deze twee onderwerpen zijn van groot belang, aangezien een bindingsplaats in een enzym, maar ook drug targets zeer flexibel zijn, en relatief vaak metaalionen bevatten. Dit proefschrift behandelt onderwerpen en beschrijft de ontwikkeling van twee nieuwe algoritmes EnzymeMatch en DynaBiS, een benchmarkstudie als leidraad voor het ontwerp van biomoleculaire simulaties met metaalionen, en twee toepassingsstudies in zowel de biokatalysator-optimalisatie als medicijnontwikkeling.

In het eerste deel van dit proefschrift wordt de ontwikkeling van EnzymeMatch beschreven. EnzymeMatch is een algoritme binnen de bio-informatica om potentiële biokatalysatoren te identificeren. Experimentele methodes zoals gestuurde evolutie hebben een startenzym nodig met minstens een meetbare activiteit. Dit algoritme voorspelt deze enzymen, en kan daardoor enzyme engieering-studies ondersteunen. EnzymeMatch heeft enkel de moleculaire structuur van een specifiek substraat nodig, en berekent automatisch de meest optimale bindingsplaats om dit substraat te binden. Vervolgens zoekt EnzymeMatch enzymen in de eiwitdatabank met een vergelijkbare bindingsplaats. Daarbij houdt EnzymeMatch rekening met zowel eiwit- als ligand flexibiliteit.

Het tweede deel van dit proefschrift beschrijft de ontwikkeling van het bindingsplaats identificatiealgoritme DynaBiS. DynaBiS gebruikt soft-core potentialen tussen het ligand en het eiwit om een volledig flexibele beschrijving van het gehele systeem mogelijk te maken, en zo te resulteren in de simulatie van conformationele aanpassingseffecten. De evaluatie van DynaBiS toonde aan dat DynaBiS beter presteert dan traditionele bindingsplaats identificatiealgoritmes, vooral bij de identificatie van bindingsplaatsen voor grote en flexibele liganden, zowel als de apo- of holo-structuur gebruikt wordt als input.

Een benchmarkstudie van biomoleculaire krachtveld-gebaseerde Zn^{2+} modellen wordt beschreven in het derde gedeelte van dit proefschrift. In deze studie werden grote prestatieverschillen gevonden tussen deze Zn^{2+} modellen, en werden de voorkeursgeometrieën en voorkeurstype van liganden bepaald. Deze resultaten leidden tot aanbevelingen van geschikte simulatieprotocollen voor een variatie aan modelleringsbenaderingen, en sturen de verdere ontwikkeling van krachtveld-gebaseerde Zn^{2+} modellen.

Tenslotte worden toepassingsstudies beschreven die het gebruik van biomoleculaire simulaties demonstreren in zowel enzym-engineering als medicijnontwikkeling. Dit omvat een studie gericht op het moduleren van de katalytische eigenschappen van [2Fe-2S]-afhankelijke dehydratases, waarin biomoleculaire simulaties de identificatie van mutatie hotspots begeleidden, die experimenteel werden onderzocht met site-directed en saturatiemutagenese. Daarnaast beschrijven twee studies de karakterisering van het amidohydrolase PurAH, dat betrokken is bij de biosynthese van een natuurlijk antibioticum, en de rationalisering van de waargenomen remming van clavulaanzuur in B3-RQK metallo-β-lactamases, wat het ontwerp van nieuwe geneesmiddelen kan ondersteunen.

De nieuw ontwikkelde algoritmes, de benchmarkstudie van krachtveld-gebaseerde Zn^{2+} modellen, en de toepassingsstudies in enzym-engineering en medicijnontwikkeling openen nieuwe strategieën die rationeel enzymengineering en structuur-gebaseerde medicijnontwikkeling kunnen bevorderen.

iv. ACKNOWLEDGEMENTS

First, special thanks go to late Prof. Dr. Iris Antes, who gave me the opportunity to perform my PhD thesis in her group and introduced me to the scientific community. She was also always willing to discuss new ideas and she showed me to always keep the bigger picture in mind, both scientifically, as well as in my personal development. I admire her never-lost positive attitude regarding scientific projects, also if they became much more challenging than expected. I would also like to thank Prof. Dr. Volker Sieber, for the excellent collaborations, but especially for taking over supervision for my PhD thesis after Prof. Antes suddenly passed away, and for his support afterwards. I also thank Prof. Dr. Ville Kaila and Prof. Dr. Martin Zacharias, for their immediate support during the turbulent phase of suddenly having to change groups, for inviting me to their group and group seminars, and for their help and scientific discussions to finish open projects and manuscripts.

I would also like to thank Assist. Prof. Dr. Antoine Marion, from whom I learned a lot in the first year of my PhD. I also thank Prof. Dr. Gerhard Schenk, for the very pleasant collaborations, and his contagious enthusiasm when discussing my results, which gave me a boost to go even further into detail. I would also like to thank Dr. Samuel Sutiono, for the very pleasant and successful collaboration, and our enjoyable discussions. I also learned a lot from him about experimental enzymology, and I am very grateful for his continuous belief in our project, and appreciate his willingness to learn from each-other. I would also like to thank my other collaborators, including Jun. Prof. Dr. Jesko Koehnke and Dr. Asfandyar Sikandar, as well as my students and HiWi's, Konstantin Eckel, Zora Rerop, Martin Gesell, Marvin Thielert, Silvia Bergt, Tongyan Wu, Franziska Totzeck, Woo Young Cho, Sarah Fink, and Sophia Zhou (I hope I listed you all), who always gave me interesting new insights into the projects we worked on. And also a great thank you to all (former) TCB group members, Dr. Antoine Marion, Dr. Ilke Ugur Marion, Dr. Ina Bisha, Martin Zachmann, Manuel Glaser, Markus Schneider, Dr. Chen Zheng, Lukas Wietbrock, Dr. Helmut Lutz, Maximillian Meixner, Simone Göppert, and Martijn Bemelmans, who were the people giving me a good time, and were always available to discuss scientific and technical topics.

Last, but not least, I would also like to thank my parents, my girlfriend, and the rest of my family for their continuous support during my life as a PhD student. I realized that the strong and stable basis given by them is essential to bring this thesis to a good end.

V. PUBLICATIONS LIST

First-author publications included in this dissertation:

Melse, O., W.Y. Cho, T. Wu, I. Antes, V.R.I. Kaila, and V. Sieber, *EnzymeMatch: Identification of Enzymes Capable of Catalyzing Target Reactions using Interaction Pattern Matching.* (Submitted).

Melse, O., S. Hecht, and I. Antes, *DynaBiS: A hierarchical sampling algorithm to identify flexible binding sites for large ligands and peptides.* Proteins: Struct. Funct. Bioinform., 2022. **90**(1): p. 18-32.

Melse, O., I. Antes, V.R.I. Kaila, and M. Zacharias, *Benchmarking of Biomolecular Force Field-Based Zn^{2+} for Mono- and Bimetallic Ligand Binding Sites.* (Submitted). Parts are available as preprint at bioRxiv, 2021: p. 2021.06.28.450184.

Melse, O., S. Sutiono, M. Haslbeck. G. Schenk, I. Antes, and V. Sieber, *Structure-Guided Modulation of the Catalytic Properties of [2Fe–2S]-Dependent Dehydratases.* ChemBioChem, 2022. **23**(10): p. e202200088.

Co-authored publications asserted as relevant work by citation:

Sikandar, A., L. Franz, O. Melse, I. Antes, and J. Koehnke, *Thiazoline-Specific Amidohydrolase PurAH Is the Gatekeeper of Bottromycin Biosynthesis.* J. Am. Chem. Soc., 2019. **141**(25): p. 9748-9752.

Pedroso, M.M., D.W. Waite, O. Melse, L. Wilson, N. Mitić, R.P. McGeary, I. Antes, L.W. Guddat, P. Hugenholtz, G. Schenk, *Broad spectrum antibiotic-degrading metallo-β-lactamases are phylogenetically diverse.* Protein & Cell, 2020. **11**(8): p. 613-617.

Additional publications by this author, but not part of this dissertation:

Genz, M., O. Melse, S. Schmidt, C. Vickers, M. Dörr, T. van den Bergh, H.J. Joosten, U.T. Bornscheuer, *Engineering the Amine Transaminase from Vibrio fluvialis towards Branched-Chain Substrates.* ChemCatChem, 2016. **8**(20): p. 3199-3202.

Schwarte, A., M. Genz, L. Skalden, A. Niboli, C. Vickers, O. Melse, R. Kuipers, H.J. Joosten, J. Stourac, J. Bendl, J. Black, P. Haase, C. Baakman, J. Damborsky, U.T. Bornscheuer, G. Vriend, H. Venselaar, *NewProt – a protein engineering portal.* Prot. Eng. Des. Sel., 2017. **30**(6): p. 441-447.

1 INTRODUCTION

1.1 ENZYMES AS BIOCATALYST

1.1.1 Principles of biocatalysis

Enzymes, most of which are proteins, are able to speed up chemical reactions, an essential process found in nature to allow cells to function under biological conditions. The basic principle underlying the enzymatic catalysis can most often be found in their ability to selectively bind and stabilize the transition state of a specific substrate, a basic principle formulated in 1946 by the double Nobel Prize winner Linus Pauling.[1] Enzymes are able accelerate certain chemical reactions, while still showing excellent chemo- stereo- and regioselectivity, under atmospheric pressure and mild pH and temperature.[2] Leonor Michaelis and Maud Menten proposed the following equation, which lies the basis for enzyme kinetics:

$$E + S \underset{}{\overset{K_d}{\rightleftharpoons}} ES \xrightarrow{K_{cat}} E + P \qquad \textbf{(1.1)}$$

illustrating binding of the enzyme (E) and substrate (S) to form an enzyme-substrate complex (ES), with the dissociation rate constant K_d. This is followed by the chemical catalysis producing the product (P), with a rate constant (or turnover number) k_{cat}.[3, 4] Since the enzyme stabilizes the substrate's transition state in the right-hand side of **eq. (1.1)**, the activation energy/barrier is lowered by the enzyme, making the reaction more favorable.

1.1.2 Industrial applications

Enzymes are recognized as useful biocatalyst for numerous industrial applications in multiple industry sectors, including the food and pharmaceutical industry, but also cosmetics, plastics, and textile industry.[5, 6] For example, enzymes can be used in the (bio)synthesis of pharmaceuticals or production of food, but can also be added to washing detergent to improve their functioning. Furthermore, several enzymatic pathways are developed to efficiently produce biofuels, such as isobutanol.[7-9] The interest of industry in enzymes as biocatalyst is mainly because of their excellent properties regarding their high chemo- stereo- and regioselectivity, and their ability to work under mild conditions. Additionally, enzymes as biocatalyst fit perfectly the global aspiration for a more sustainable world because of their resource efficiency, independence of toxic reagents or solvents (enzymes generally live in an aqueous environment), and biodegradability, which led to a boost of the use of enzymes in industry.[10, 11]

While the enzyme market is already a multi-billion dollar business, only a small percentage of known enzymes are commercially used, and even less are produced on industrial scale.[5, 12] Therefore, much progress in this flourishing field can be expected in the next years.

1.2 ENZYME ENGINEERING AND DESIGN

Not all enzymes are immediately suitable for industrial applications. For example, enzymes can become unstable while used at certain temperatures or pH, or in a different solvent environment required for the industrial process. Moreover, enzymes can be very selective, which allows for very selective conversion, but may limit the application. Finally, the enzyme activity can simply be too low for efficient industrial use.[11] There are several strategies possible to optimize these characteristics. For example, the industrial process can be optimized by changing the solvent or optimizing the reactor.[13] Alternatively, enzyme immobilization is often applied, especially to improve the enzyme stability.[14] However, upscaling production lines containing immobilized enzymes appears to be difficult.[15] Finally, the enzyme itself can be modified via several protein engineering strategies to improve the above-mentioned characteristics.[2, 16] The latter can be subdivided in a directed evolution approach and (semi-)rational protein design.

1.2.1 Directed evolution

To tailor a protein toward desired characteristics, one can look at what happens in nature. In Darwinian evolution, genes are randomly modified followed by selection based on the fitness of the phenotype. This idea can be implemented in the laboratory as well within a high-throughput screening setting. The first approaches included several cycles of random mutagenesis via a suboptimal polymerase chain reaction (PCR), *i.e.* error-prone PCR, followed by molecular cloning, transformation and expression, and manual selection of more “fit” protein variants.[2, 17] An alternative to random mutagenesis is homologous DNA shuffling, in which parent genes are first fragmentized, followed by homologous recombination to reassemble these fragments. Circular permutation and random insertion/deletion mutagenesis are two other methods which can be applied in a directed evolution approach. In the first, the terminals of the genes are first covalently linked, followed by random cleavage somewhere in the gene, resulting to a new gene arrangement. In random insertion/deletion mutagenesis, as the name implies, random DNA sequences are either inserted or deleted, leading to a gene with a different length compared to the initial gene.[17]

1.2.2 Semi-rational protein engineering

In semi-rational protein engineering, small, but smart mutation libraries are designed based on advancing computational and machine-learning methods. For example, a Multiple Sequence Alignment (MSA) of the target sequence with homologous proteins can provide information about the evolutionary variability of certain amino acids to design a smart library.[2] Based on this information, Reetz *et al*. developed the Combinatorial Active-site Saturation Test (CAST) strategy, in which amino acids near the active site are targeted simultaneously with random mutations, which was shown to be an effective method to reduce the library size.[18] Two years later, Reetz and Carballeira combined CAST with Iterative Saturation Mutagenesis (ISM), leading to the more commonly used CAST/ISM method.[19] Several excellent reviews have been published by the same author, clearly describing the potential and success stories of CAST and CAST/ISM.[20, 21] Moreover, a new technique inspired by CAST/ISM has been developed, named Focused Rational Iterative Site-specific Mutagenesis (FRISM).[22] This method reduces the library size even further, by applying site-directed mutagenesis instead of saturation mutagenesis. Here, a single CAST site is mutated to several designed mutations with site-directed mutagenesis, whereas the best variant is used for site-directed mutagenesis at the second CAST site, and so on. To ensure proper sampling of the available mutations, this approach is applied on several pathways, *i.e.* changing the order of the CAST sites. This results in *N*! pathways, where *N* represents the number of CAST positions.

Semi-rational engineering approaches thus rely on computational studies to design smart and small libraries. There are several bioinformatics methods available, such as 3DM, which is a bioinformatics method applying structural superposition, MSA, and literature mining. 3DM compares sequences and available experimental structures within a superfamily, and assigned a 3D number: a number which identifies the position in the protein structure rather than the sequence. [23] This allows the prediction of structure-sequence relationships, residue conservation and correlation matrices. This method has been successfully applied in a semi-rational engineering approach regarding PLP-dependent Amine Transaminases.[24] More recently, 3DM combined with CorNet, a tool to analyze co-evolving residue positions, gave some interesting insights in short-chain dehydrogenases, which are involved in the biosynthesis of rare sugars and glycosides.[25, 26] Other approaches, such as (de)stability estimation of certain mutations by FoldX can further guide the development of small mutation libraries.[27-29] Information from these predictors can be combined with use of so-called “one-stop” portals, such as NewProt, in which predictions retrieved from multiple servers are combined automatically to provide a quick and easy overview of predicted (un)desired effects of certain point mutations.[30]

1.2.3 Rational protein engineering

Due to the increasing amount of structural data for proteins, illustrated by the ever-growing number of entries deposited in the Protein Data Bank[31] (**Figure 1**), rational approaches become more available to guide protein engineering studies. This can be performed via bioinformatics analysis, machine learning or biomolecular simulations, the latter described in more detail in Chapter 1.2.4. In contrast to the random mutations performed in the directed evolution approach, a traditional rational protein engineering process contains a combination of single targeted mutations, usually in or nearby the ligand binding site.[17, 32] When one, or a combination of residue positions have been identified which could play a role in enzyme activity, stability, or selectivity, (so-called "mutation hotspots"), these can be investigated *in vitro* via site-directed or site-saturation mutagenesis. In site-directed mutagenesis, a specific mutation is performed on a residue position by modification of the gene with PCR. Usually, a mutation towards an alanine is performed to evaluate the role of certain physico-chemical properties at this position (*i.e.* alanine scanning), or a mutation towards an amino acid of a different type is performed, for example to reduce the side-chain length to generate space in the binding pocket, or to introduce new interaction partners.[17, 33] With site-saturation mutagenesis, all 20 amino acids, or a subset of them, are evaluated at a certain position. This can be performed with a randomized codon in primers, while based on the chosen codon the likeliness of the presence of certain amino acids can be controlled.[19]

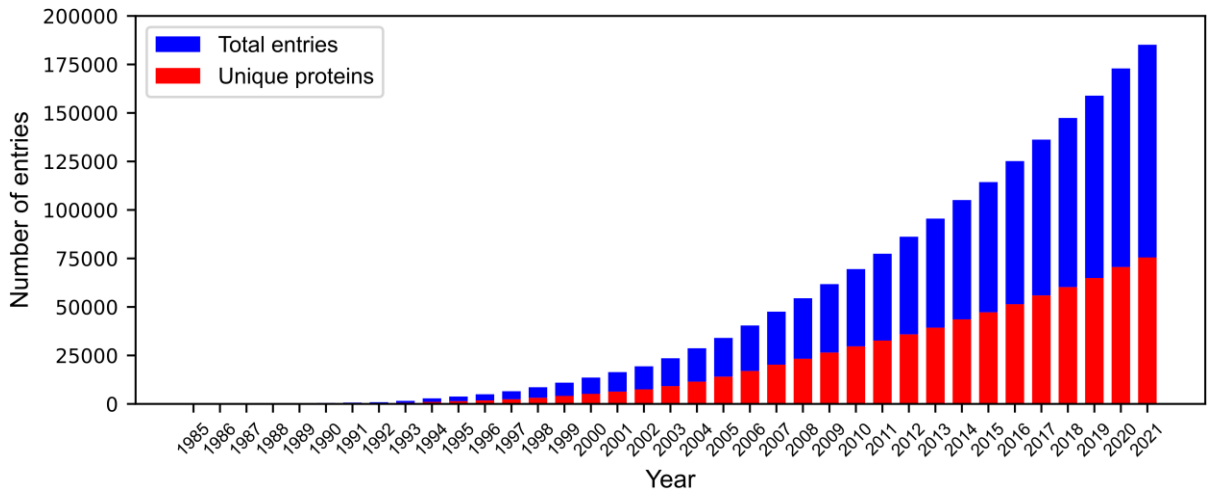


Figure 1. Number of entries in the Protein Data Bank per year. The blue bars indicate the total number of entries in the Protein Data Bank in that year. Red bars indicate the number of unique proteins in the Protein Data Bank, *i.e.* all proteins with 95% sequence similarity merged. Data retrieved from ref [31].

Engineering toward more stable enzymes

Besides all beneficial characteristics of enzymes, such as their substrate selectivity and ability to work under mild conditions, enzymes are not necessarily very stable at conditions required for industrial application. Normally, enzymes get denatured, *i.e.* breaking of weak interactions leading to a change in structure, which can lead to inactivation of enzymes. However, especially in food or textile industry, *e.g.* when enzymes are supplemented to washing detergent, enzymes need to be able to work at high temperatures or high salt concentrations. Also other environmental factors can negatively affect enzyme stability, such as pH or organic solvent.[17]

There are two flavors of enzyme stability: thermodynamic stability and kinetic stability. The first describes the equilibrium between unfolded (denatured) and functional enzyme, while the second describes the free energy barrier between enzymes in their functional folded state and non-functional state.[34] Thus, even if an enzyme is not thermodynamically stable, it can still be active for some time because the barrier between folded and (partially) unfolded enzyme is high, in which case an enzyme is kinetically stable. Most of the bioinformatics analyses described below, as well as the majority of enzyme stability measurements performed in the lab, are however related to thermodynamic stability. There are several protein engineering approaches to improve enzyme stability. For example, a disulfide bond can be introduced between two oxidized cysteine residues to stabilize the folded conformation. Furthermore, glycine residues, which show a high amount of conformational entropy of the backbone, can be mutated toward a proline to reduce the protein conformational entropy. Finally, modification of the buried region in the protein by increasing hydrophobicity, or introduction of certain hydrogen bonds or ionic interactions at the protein surface can stabilize the protein.[17]

Computational studies can help to predict the effect of certain mutations on the protein stability. Molecular dynamics simulations have shown to be helpful to analyze these effects, but these simulations are rather computationally expensive.[35] Therefore, there are several (much more efficient) bioinformatics tools available to predict the effect of specific mutations on protein stability. This can be helpful to guide engineering approaches to improve protein stability, but can also be useful to check if a planned mutation to modify another property, such as enzyme activity or specificity, might have negative influence on the protein stability. Some common bioinformatics tool related to protein stability are listed below:

- PoPMuSiC: focused on single-site mutations in proteins and peptides. All possible mutations in a provided region are performed *in silico*, followed by solvent accessibility calculations which are

used as coefficients in a linear combination of database-derived potentials to predict the effect of each mutation on protein stability.[36]

- FoldX: an empirical force field, which can be applied to predict free energy differences between wild-type and variant to evaluate the effect of a mutation on protein (or nucleic acid) stability, folding and dynamics.[27]
- CUPSAT: predicts the free energy of unfolding between wild-type and single-mutant variants, applying specific atom and torsion angle potentials. A PDB structure of the protein is required.[37]
- I-Mutant: a machine-learning algorithm applying support-vector networks, predicting if a point mutation is stabilizing, destabilizing or neutral. Both a sequence-based and structure-based prediction is possible.[38]

Engineering to modify enzyme-substrate specificity

Enzyme-substrate specificity is one of the main advantages of enzymes as biocatalysts compared to traditional chemical catalysts. However, a too specific enzyme may also limit an industrial application, thus protein engineering can be applied to change and/or broaden an enzyme's substrate scope.[11] The substrate specificity of an enzyme is the result of a very specific binding orientation of the substrate required by the enzyme to make the reaction possible. In other words, an enzyme can catalyze a certain reaction if the substrate is able to bind in the preferred conformation in the active site, by which all active site residues are oriented properly to catalyze the reaction and stabilize the transition state.[17] Thus, in order to modify the enzyme-substrate specificity, one needs to properly understand the binding of the target substrate in the enzyme. Therefore, many rational engineering approaches rely on molecular docking simulations, in which the binding orientation of the substrate (or transition state) in the enzyme's active site is predicted. Furthermore, the active site conformation may adopt upon substrate binding, which can either be simulated with subsequent molecular dynamics simulations, or directly during the molecular docking simulations with more advanced algorithms explicitly allowing for both ligand- and protein flexibility (more details about molecular docking simulations can be found in Chapter 1.5).[39] Based on the predicted enzyme-substrate complex, mutation hotspots can be defined, and the optimal amino acids at these positions can be predicted, which can be incorporated via site-directed mutagenesis.[17]

Numerous engineering strategies based on rational design have been applied. For example, Genz *et al*. combined information from molecular docking and molecular dynamics simulations, together with bioinformatics analysis with 3DM, to predict mutation hotspots in the active site of the amine

transaminase from *Vibrio fluvialis*. Their approach was based on increasing the binding site volume in order to allow production of much bulkier amines. Therefore, they designed a seven-site mutant library, and thereby observed three variants able to synthesize 2,2-dimethyl-1-phenylpropan-1-amine with an enantiomeric excess of >99%, which was not possible with any wild-type amine transaminase until then.[24] Alternative approaches include reduction of the binding site volume, and adjusting the binding site by introduction or deletion of certain physico-chemical properties to enhance binding of the target substrate.[17] The latter approach, together with enlarging binding site volume has been successfully applied to engineer dihydroxy-acid dehydratases, which is described in one of the studies described in this dissertation.[40]

Improving enzyme activity

Enzyme activity describes the rate of the entire process including substrate binding, catalysis, and substrate release, and thus directly correlates with the production rate of the desired product. Therefore, a highly active biocatalyst is generally beneficial, since less enzyme needs to be produced or purchased, and problems related to high protein concentrations in bioreactors can be avoided.[11, 15]

Rational engineering to improve enzyme activity however remains challenging. In order to reduce the activation barrier of the reaction, the transition state should be further stabilized. This can be performed with similar computational strategies as described above), thus a combination of molecular docking and molecular dynamics simulations, followed by hotspot identification and site-directed mutagenesis.[17, 41] OptZyme is a computational procedure applying such an approach aiming to improve enzyme activity.[42] OptZyme optimizes the binding between the enzyme and a transition state analogue (*i.e.* an inhibitor closely similar to the transition state of an enzymatic reaction), rather than the target substrate. These transition state analogues are namely known for most enzymatic reactions. Mutations are predicted affecting the interaction with this transition state analogue, which likely will affect the enzyme activity toward the target substrate as well.

An alternative approach is via a combination of sequence and structural information with literature mining. For example, a core alignment with structurally similar enzymes can be generated with 3DM, which can be further expanded with other sequences belonging to the same superfamily applying MSA. Subsequently, 3DM can perform an automatic data mining approach, searching for literature in which residues from homologous proteins at interesting structural positions (*e.g.* active site residues) were mentioned, together with certain keywords such as “activity”. If a certain structural position is often correlated with activity change, this position may be an interesting mutation hotspot.[23] This approach

has been used successfully to engineer numerous enzymes. For example, the activity of an esterase from *Pseudomonas fluorescens* was improved up to 240-fold with only four mutations of residues near the active site.[43]

De novo enzyme design

In *de novo* design studies, an enzyme is developed catalyzing a desired reaction, which is not related to any enzyme found in nature. This field is still under development, but there are some examples in which this approach has been used successfully.[44-46] Successful *de novo* enzyme design studies are however rare and require a lot of man hours, illustrated by the long author lists and high-impact journals (Nature, Science etc.) these studies are published in.

The main task of *de novo* protein design studies can be subdivided in four steps: (I) prediction of the optimal backbone conformation for the desired role (*e.g.* catalysis of certain reaction), (II) prediction of a sequence which folds into this conformation, (III) scoring all proposed solutions and (IV) design of the functional/active site.[47] Although *de novo* protein design results in proteins with previously unknown sequences, the design procedure still relies on principles retrieved from the Protein Data Bank. For example, fragments which are known to fold in a certain motif or scaffold can be collected from the Protein Data Bank, which can subsequently be combined to predict a new fold. Other strategies include design of protein fragments of a certain fold or motif, *e.g.* by leucine-rich repeats, combined with helical building blocks and loop fragments to predict a desired fold.[47]

1.2.4 Biomolecular simulations in biocatalyst development

Biomolecular simulations are becoming an integral part of numerous research fields, including (structure-based) drug design and protein engineering.[35, 48] The history of biomolecular simulations dates back to the late 1970s, when McCammon, Gelin and Karplus, the latter a Nobel Prize winner of 2013, illustrated the dynamics of proteins by a simulation of bovine pancreatic trypsin inhibitor.[49] They performed this 8.8 ps simulation in vacuum, by solving the equations of motions applying an empirical potential energy function. The authors described the “fluid-like” behavior of proteins, which can be seen as a tripping point in computational chemistry from which on proteins were seen as dynamic structures, thereby initiating the research field of biomolecular simulations. Great progress has been made in this field since then, both regarding the time-scale of the simulations, as well as the accuracy of the respective models.[50, 51]

In silico studies show several advantages over experiments, as (I) no (potentially toxic) reagents are required, (II) events can be studied at a molecular level, and (III) a much larger library can be screened.

Biomolecular simulations can be applied to rationalize experimental findings and thereby contribute to the understanding of biochemical processes, such as protein-ligand binding events, receptor activation and (enzymatic) catalysis. Because of the increasing accuracy of computational models, together with the vastly rising efficiency and availability of computational resources, biomolecular simulations can also be applied to make predictions to guide drug design-, protein engineering- and *de novo* protein (*e.g.* biocatalyst) development studies.[47, 50-52]

In this thesis, the focus mainly lies on the development and application of biomolecular simulations for enzyme engineering and biocatalyst design. Nevertheless, the majority of the methods described here can be applied for structure-based drug design as well. For example, molecular docking simulations can be conducted to study the binding pose of a ligand in a binding site, after which either the ligand or the protein can be modified to enhance ligand/substrate binding. The first being a classical example for structure-based drug design, and the latter for protein engineering.

1.3 STRUCTURE PREDICTION

Many biomolecular simulation methods, such as molecular docking and molecular dynamics simulations, require a protein structure as input. When no experimental structure of the target protein is available in the Protein Data Bank, a model of the protein structure needs to be generated. The most common strategy to retrieve a model of the protein structure is homology modeling, but recent advances also allow machine-learning methods to predict the 3-dimensional protein structure.

1.3.1 Homology Modelling

Homology modelling, also known as comparative modelling, aims to predict the tertiary structure of the target protein (*i.e.* the protein with unknown structure). Structural information of a homologous protein with a resolved experimental structure is mapped on the protein sequence of the target protein, resulting in a predicted protein model.[53-56] It has been shown that for nearly all proteins a homologous protein is available in the Protein Data Bank, making homology modelling an applicable method for nearly all proteins.[57] Homology modelling relies on the sequence-structure relationship, which states that proteins sharing a high degree of amino-acid sequence similarity generally also share a similar fold.[53] Therefore, the quality of the resulting model for the target protein highly depends on the availability of a template with a high sequence identity. A homology modelling study thus starts with identification of suitable template structures, ideally sharing a high degree of sequence similarity with the target protein. When a suitable template has been found, the position of the backbone atoms of the secondary structure

elements can be predicted. Subsequently, the side-chain positions and the orientation of loop regions are modelled. Dependent on the quality of the resulting model, the sequence similarity between the target- and template structure, and the model accuracy required by the scientist, the homology model can be further refined, which is mostly performed with molecular dynamics simulations. Finally, the models are scored to provide an estimate of the quality of the homology model, and potentially to discriminate between multiple generated homology models.[53-55]

Homology modelling can easily be performed via the online modelling platform SWISS-MODEL, which guides the user with a graphical user interface trough all required steps, and automatically builds the homology model.[55, 58] Despite being an established webserver for homology modelling, alternative modelling algorithms can be beneficial over SWISS-MODEL in certain cases allowing more interference by the user. Especially when only low sequence-identity templates are available (in this case homology modelling should ideally be performed applying multiple templates simultaneously), or when co-factors and/or ligands should be included in the modelling, these alternative modelling algorithms are often preferred. These modelling algorithms allow for full control of all modelling steps by the user, as well as more advanced loop modelling in the predicted protein structures.[59] Examples of homology modelling algorithms include RosettaCM[60], I-TASSER[61] and MODELLER[62], the latter being the algorithm applied in the homology modelling studies performed in this work.

Template search and selection

To identify potential template structures, the target sequence is aligned to all sequences from systems with a solved experimental structure in the Protein Data Bank, which is commonly performed via a BLAST search.[63] This results in a list of potential template structures, together with the sequence identity to the template sequence and some other scores describing the quality of the alignment. This information can be extended with information retrieved from other databases, such as UniProt and the SWISS-MODEL Template Library.[58, 64] As a rule of thumb, 30% sequence identity between two protein sequences generally suggests a common fold, while some studies even define this threshold at 20%.[65, 66] It remains however important to critically examine the templates manually, especially when the sequence identity is at the lower range. For example, Alexander *et al.* illustrated that two proteins sharing 88% sequence identity can end up in a completely different fold, here a 3-α helix fold and an α/β fold.[67] It needs to be noted here that the authors deliberately designed these two proteins to result in a high as possible sequence identity, but still resulting in a different fold. Nevertheless, this illustrates the importance of manual examination of the predicted templates. For example, the protein family of the

target and template proteins should be examined, as the fold of proteins within the same family is more conserved than their sequences.[54, 68] Furthermore, a closer analysis of the actual sequence alignment between the target and template sequence can often result in useful information: a higher local sequence similarity in the protein's interior or the ligand binding site or other relevant regions generally results in more accurate protein models.[56] Other features which should be involved in the examination of protein templates are the quality of the experimental structure (*e.g.* X-Ray resolution), and quaternary structure features.[55]

Model building

Once one ore multiple templates are identified, the model for the target protein can be build. In the classical rigid body approach, the position of the Cα-atoms of structurally conserved regions can be predicted using the sequence alignment between the target and template(s). In the case when multiple templates are used, the templates are first superposed, and the average position of the Cα-atoms are applied for the homology model. This results in a so-called "framework". In most homology modelling algorithms, small deletions are subsequently resolved by relaxation of neighboring residues, after which larger deletions, insertions, and loop regions are handled via the loop prediction pipeline, as described below.[62] A similar approach is applied by ProMod3, *i.e.* the modelling algorithm behind SWISS-MODEL.[55, 69]

An alternative approach, as applied by MODELLER, relies on modelling by satisfaction of spatial restraints.[62, 70] Here, constraints or restraints are defined based on the target-template alignment and the template structure. These restraints include Cα-Cα distances, backbone N-O distances and dihedral angles of the backbone framework and side-chains.[70] These restraints are often supplemented by restraints describing stereo-chemical properties retrieved from a molecular mechanics force field, such as bond lengths, angles, dihedral angles, and non-bonded interactions. These restraints are described by probability density functions (pdfs). The model is then generated using the distance and dihedral angle restraints, and subsequently refined by minimizing the violations of the above-described pdfs.[54, 62]

Loop modelling

Predicting the orientation of non-conserved loop regions and side-chains is more challenging than the prediction of the framework because of the relatively high allowed flexibility in these regions. However, the conformation of the loop regions is highly important, as loops often play an important role in the protein's function and the formation of binding sites.[55, 62] Loop modelling is especially challenging if the sequence identity between the target and template sequence is below 50%, because in these cases

the core regions are often well conserved and aligned, but the loop regions vary.[62] Therefore, loop modelling strategies were developed, aiming to predict the most likely conformation of a small protein sequence, which are anchored in certain positions in Cartesian space from the framework. Besides the anchor positions, the number of residues in the loop, the neighboring residues and the loop's surrounding all affect the loop conformation.[71]

Available loop modelling algorithms can be classified in template-based and template-free methods.[56, 71] Template-based methods search in a database for loops with known conformation with a similar sequence, and ideally within the same class (*e.g.* β-hairpin). These methods can be quite accurate for small loops, but the performance quickly decreases for longer loops, since the possible loop conformations grows exponentially with the loop length while the number of available templates decrease.[62, 71] This approach is however still often applied and further developed, as for example by ProMod3 in SWISS-MODEL.[69] In ProMod3, matching loops are searched in their own StructureDB database, containing $4.5 \cdot 10^6$ loop structures, using a query containing the number of loop residues, the distance between the anchor residues, and the geometric orientation of the two anchor regions described via four defined angles. Subsequently, the matched loops are scored based on summation of several backbone-related scores. If no suitable template was found, or if the number of residues exceed 12 (in the case of ProMod3), a Monte Carlo sampling is applied instead to predict the loop conformation.[56]

Template free methods can be seen as a "mini protein folding problem".[71] Template-free loop modelling methods differ in their energy function, conformational sampling approach and the strategy to tackle the "loop closure" problem, *i.e.* ensuring that the loop termini match the anchor position in the framework. The majority of template-free loop modelling algorithms apply the following pipeline: first a large number of possible loop conformations are sampled applying a coarse-grained sampling method, supplemented with a knowledge- or physics-based energy function. Multiple approaches have been applied to perform this initial sampling, including a random buildup approach[62] (as applied by MODELLER), inverse kinematics[72], Monte Carlo simulated annealing[73], molecular dynamics[74] and Markkov Chain Monte Carlo[73]. The resulting loop conformations are filtered and/or clustered to reduce the set size and redundancy, followed by refinement of these loop conformations with more accurate sampling algorithms and energy functions, *e.g.* full-atom simulations. All these simulations are performed under geometric restraints, such that the generated loops fit to the anchor regions of the framework. Finally, the refined loops are ranked based on a scoring function.[71]

The protein side-chain packing problem

Side-chain placement consists of three steps: definition of possible side-chain conformations, sampling of possible protein side-chain packing (PSCP) combinations and scoring of these PSCPs.[75-77] All PSCP algorithms perform the first step via rotamer libraries, with the sole exception of the Grow-to-Fit molecular dynamics method.[76, 78] A rotamer, short for "rotational isomer", is a single side-chain conformation of a certain residue defined by a set of dihedral angles, while a rotamer library is a collection of rotamers supplemented with probability values.[79] Numerous rotamer libraries are available, such as the dynameomics rotamer library of Dagett[80], the Richardson and Lovell rotamer library[81], and Dunbrack's rotamer library[79], the latter arguable the most known. Rotamer libraries can be subdivided into backbone-independent and backbone-dependent libraries, where in the latter the probabilities for certain rotamers are dependent on the backbone *phi* and *psi* angles. Nearly all rotamer libraries are generated by statistical analysis of structures retrieved from the Protein Data Bank to retrieve probabilities of rotamers, while some more recent attempts to update the rotamer libraries additionally apply molecular dynamics simulations.[79, 80, 82]

Known state-of-the-art PSCP methods include the established method from the Dunbrack's lab SCWRL4[83], a simulated annealing-based method OPUS-Rota[84] and OPUS-Rota2[85]. A method developed by Hartmann, Antes and Lengauer named IRECS[86] is able to select more than one rotamer per side-chain, and a very recently published method FASPR[75], developed in the Zhang Lab was shown to be significantly faster than the other methods. In these PSCP methods, multiple combinations of PSCPs are iteratively optimized by minimizing the amount of clashes and optimizing the favorable interactions between side-chains. Furthermore, it has been shown that the state-of-the-art sampling algorithms in those PSCP methods are able to sample the correct PSCP, but that the main limitation lies in the scoring functions, which is valid for more topics in computational chemistry.[76]

Quality assessment

To estimate if the produced homology model is reasonable, scoring functions have been developed applying statistical potentials to compare the modelled structure to native high-resolution structures. These methods analyze the environment of each residue, and compare this to results obtained from a similar analysis in experimental structures.[62] Most of these scores are represented as a Z-score, which describes the number of standard deviations the model's score is away from the expected value. Thus, a Z-score is the score normalized to the mean 0 and standard deviation 1.[87]

The DOPE score (Discrete Optimized Protein Energy) is based on a statistical potential regarding atomic distances retrieved from the Protein Data Bank.[88] The statistical potential is represented as a pdf based on inter-atomic distances of different residue types. Another scoring function, QMEAN (Qualitative Model Energy Analysis), is a composite scoring function, meaning that multiple scoring terms are combined to the final score.[87, 89] There are two flavors, namely QMEAN4 and QMEAN6, which contain respective four and six statistical potential terms. QMEAN4 contains two potentials describing long-range interactions via potentials of mean force measured between Cβ atoms, a torsion angle potential describing backbone geometry, and a solvation potential.[89] QMEAN6 additionally considers two potentials describing the agreement between calculated and sequence-based predicted secondary structure and solvent accessibility properties.[87] Finally, QMEANDisCo is an extension of the QMEAN scores, especially improving per-residue scores by assessing consistency of residue-residue interatomic distances in the generated homology model, and those observed in close homologues. In QMEANDisCo, the residue-residue interatomic distances of multiple homologues are represented as distance constraints (DisCo), while a neural network defines weights for the individual homologues.[55, 90] Note that this score can only be calculated if homologues are found. All the above-described scores can be derived globally (*i.e.* for the entire structure), or locally (*i.e.* per residue).[87-89] For QMEANDisCo, the global score is the global QMEAN score combined with the average QMEANDisCo score over all residues.[90] Finally, one can also evaluate the stereo-chemical quality of the model, such as bond length, angle, and backbone torsions with algorithms as PROCHECK.[91]

Model refinement

The resulting homology model can be refined by adding potentially missing co-factors and/or ligands, either by superposition to homologues with known experimental structures in their holo form, or via molecular docking simulations (see Chapter 1.5 for a more detailed discussion about molecular docking simulations). Further refinement of the homology models can be performed by (restrained) geometry optimizations or short molecular dynamics simulations. A (very) short geometry optimization is already included in most of the homology modelling algorithms.[62] However, the performance of molecular dynamics simulations to improve the fold of inaccurate homology models have shown to be limited.[92, 93] Finally, it should also be mentioned here that proteins can exist in an equilibrium between multiple conformations, meaning that one cannot always speak of the “correct” structure, and should always keep the natural dynamics of a biological system in mind.[65]

1.3.2 AlphaFold

CASP14 and the protein folding problem

Critical Assessment of Structure Prediction (CASP) is an organization interested in strategies to improve protein structure prediction, sometimes referred to as "solving the protein folding problem". They organize a challenge every two years since 1994, where they provide the participating groups with several modeling targets, which structure was recently solved, but not yet published. The participants receive the sequences, and are asked to predict the protein fold for these proteins, typically within three weeks.[94] Numerous research- and industry groups joined this challenge, often using information from homologous proteins together with physics-based, and later machine learning-based methods to predict the fold of the target proteins. However, predicting the fold based on sequence-only information appeared to be really hard, thus the "protein folding problem" (I personally, among others, prefer "protein structure prediction problem") was often seen as one of biology's grand challenges.[95] However, CASP13, held from May to July 2018, made the news in mainstream media because of the excellent performance by one of the participants, the company DeepMind, with their deep learning-based algorithm "AlphaFold".[96] One edition later, CASP14, held from May to August 2020, CASP received again media attention because of the outstanding performance of DeepMind with AlphaFold2 (hereafter called AlphaFold), who were now able to predict the protein structure reaching a median score of 92.4.[97] It is important to note here that a score above 90 is considered to represent an accuracy similar to experimentally resolved structures.[95] DeepMind recently published their method, and made the source-code publicly available.[98] Because of the expected impact of AlphaFold in protein science, structure-based drug discovery and protein engineering and design, this Chapter briefly introduces the methods applied by AlphaFold, and it's potential.

The algorithm behind AlphaFold

The algorithm behind AlphaFold can be classified into three parts: preprocessing section, the Evoformer block and the structure module.[98] An important aspect of AlphaFold is the repetition of these blocks (called "recycling" by the AlphaFold developers), in which the output of the structure block returns in the Evoformer block. This allows for incremental optimization/refinement of the structure, which was shown to strongly improve the final structure prediction. Finally, several innovative tricks during the training, such as MSA masking, application of FAPE loss and self-distillation training further improved the performance of AllphaFold2[98-101] The individual parts of the AlphaFold algorithm are described below.

Preprocessing and template identification

The preprocessing block takes the input sequence, and searches for related sequences to construct a Multiple Sequence Alignment (MSA). In the MSA, all related sequences are aligned to each-other, resulting in a 2D N_{res} x N_{seq} matrix, where N_{res} and N_{seq} represent the number of residues (or better: alignment length) and the number of sequences, respectively. The MSA contains evolutionary information about the target protein, as residues that are not important for the structure and/or function are expected to mutate with a higher rate during evolution of a species compared to residues which encode important structural properties. In other words, residues that are important for the structure of the protein are expected to be conserved in the MSA, or only mutate to residues with the same type (*e.g.* positively/negatively charged, hydrophilic, hydrophobic, aromatic etc.). Moreover, residues which are close to each-other (note that they may though be far away from each other in the protein sequence) are expected to mutate simultaneously, or at least closely during evolution, due to a process known as "evolutionary pressure". For example, imagine a positively charged residue (*e.g.* lysine) which interacts with a negatively charged residue (*e.g.* aspartate). If the first residue mutates from a lysine to a negatively charged glutamate, this results in a repulsive force with the nearby aspartate. Therefore, the aspartate is expected to mutate toward a positively charged residue, to restore this interaction. Otherwise, the fold may be disrupted, and the species or subfamily carrying this mutation may become extinct. This evolutionary pressure, among other information, can be retrieved from the MSA providing valuable information for the structure determination in later stages.[102]

Besides these MSAs, AlphaFold generates a pair representation, represented by a 2D N_{res} x N_{res} matrix. This pair representation contains predicted distances between the respective residues. In order to predict these inter-residue distances, AlphaFold searches for structural templates, using a similar approach as described in Chapter 1.3.1 in the context of homology modelling. The sequence differences and structure similarities are assessed, thereby retrieving sequence-structure relationships and identify conserved structure fragments. This information is used to generate an initial pair representation.

So far there is not much new compared to existing methods, as these methods are also applied in a similar way in homology modelling (see Chapter 1.3.1), and other protein folding algorithms.[103, 104]

Evoformer block

The main aim of the Evoformer block is to retrieve structural information from the pre-processed data.[98, 102] The main innovations of AlphaFold are that the Evoformer block does not only use MSA information, but mixes the information from the MSA and pair representations during the entire optimization

procedure to predict structural features, as well as optimizes the MSA and pair representations. By feeding this output (optimized MSA and pair representation) back into the Evoformer block, the structural features can additionally be used to further refine the output, and so on.[98, 102]

The Evoformer applies a deep learning architecture named “transformer”.[105] A transformer works with the principle of “attention”, representing regions of high importance.[98, 101] The Evoformer contains two transformers, one for the MSA representation, and one for the pair representation. The transformer for the MSA representation first analyzes the matrix row-wise (*i.e.* over the protein the sequence), identifying parts of the sequences which contain the most valuable information regarding the protein fold. Subsequently, the transformer analyzes the matrix column-wise, thus analyzing which sequences contain most valuable information. The transformer additionally uses information from the pair representation in this analysis.[98, 101, 102] The transformer for the pair representation works slightly different. A main task of this transformer is to enforce the triangle inequality, which states that the sum of the length of two triangle edges must be equal or larger than the remaining side.[106] In other words, it can be that the distances in the pair representation do not sum up to a clear position in Cartesian coordinates for all atoms. For example, the position of two atoms need to be at a certain position to fulfill the majority of the distances in the pair representation, but still violates some of the other distances. In this case, the distances in the pair representation need to be optimized, which is performed by the latter transformer in the Evoformer block.[98]

Structure module

Finally, the structure module uses the refined MSA and pair representations resulting from the Evoformer block to predict the structure. Internally, the structure module does not describe the structure via Cartesian coordinates, but uses a “residue gas” representation, *i.e.* a triangle representation for the backbone atoms and torsion angles for the side-chains, making rotation and translation modifications easier.[98] The structure module starts with the “black hole initialization”, meaning that all residues are located at the origin, having the same orientation. This representation is optimized via translations and rotations to finally lead to the predicted protein structure via a so-called “Invariant Point Attention” method.[98] An important aspect of the structure module is that there are no constraints applied to keep the backbone atoms together (chain constraint). The AlphaFold developers also observed that this chain constraint is often violated by the neural network, such that the neural network can first optimize substructures without having to fix complex loop closure. The peptide bond geometry is only restrained during the final fine-tuning of the structure, by which most introduced gaps are resolved.[98] To finally

resolve all unphysical conformations, backbone gaps and stereochemical violations, the structure is relaxed applying the Amber ff99SB force field[107] under certain constraints.

Impact of AlphaFold, is the protein folding problem solved?

Now AlphaFold is available, is the protein folding problem solved? The short answer is probably: "no", but the development of Alphafold has still been an outstanding performance which can certainly provide a boost to structural biology, and will come in very handy in numerous fields related to protein science. This Chapter may be a bit more of an opinion rather than a description of scientific facts, since the future, and thus the impact of AlphaFold is hard to predict. Moreover, it is important to briefly highlight the difference between the "protein folding problem" and "protein structure prediction problem", already briefly mentioned in above. Where "protein structure prediction" refers to predicting the protein structure and fold based on sequence data, which was successfully addressed by AlphaFold, the "protein folding problem" includes the folding process, which is not addressed at all by AlphaFold, as no information about the folding pathway is provided. Furthermore, protein dynamics are essential for proteins to function, *e.g.* due to domain movements or conformational adaptations of the binding site, which is also not predicted by AlphaFold.[108] This should however not take the shine off DeepMind's work, which still has a lot of applications. For example, protein structures predicted by AlphaFold can be used as search model to solve X-Ray structures, or by computational biologists/chemists to accelerate drug design and protein (*e.g.* biocatalyst) design.[109] Furthermore, it has also already been suggested that AlphaFold can help the way out of the covid-19 pandemic, *e.g.* via the development of (improved) vaccines.[110] Furthermore it should be mentioned that AlphaFold is not completely by its own, as RoseTTAFold approaches the accuracy of AlphaFold, which is also publicly available.[103] Finally, several important questions in structural biology remain unsolved, such as the prediction of important protein-protein or protein-ligand interactions, prediction of protein function, and predictions regarding the presence and importance of different conformational states.[101, 109]

1.4 BINDING SITE IDENTIFICATION

In order to apply biomolecular simulations in structure-based drug design or enzyme engineering, identification of the ligand binding site is essential. This information could be extracted from experiments, for example from an X-Ray structure in which the ligand was co-crystallized. However, this data is often not available, and experimentally still difficult to retrieve. Therefore, multiple computational approaches to identify the binding site in proteins have been presented.[111] Sequence-based methods solely use the

protein sequence to suggest potential ligand binding sites, applying methods such as MSA and Position Specific Scoring Matrices.[112, 113] However, since biomolecular simulations generally require the availability of a protein structure, either experimentally solved or predicted via computational methods (see Chapter 1.3), the structure can also be used to improve the accuracy of the binding site identification methods. Numerous approaches have been proposed, all with their advantages and disadvantages. The most applied methods are briefly described below.

1.4.1 Surface-scanning algorithms

One of the most popular approaches applied in structure-based binding site identification algorithms is scanning the protein surface to search for potential binding sites. These methods are based on the observation that ligand binding sites are often located in cavities in the protein surface.[111, 114] For example, "Putative Active Site with Spheres" (PASS), is a binding site identification algorithm which completely relies on geometric properties of the protein.[115] First, probe atoms are placed around the protein, after which the probe atoms are removed if they clash too much with the protein, if the probe is not sufficiently buried in the protein, or if another probe nearby is more buried. This procedure then continues with smaller probes in the predicted regions, until all probes are removed.[114, 115] Other binding site identification algorithms relying on a similar geometric approach include POCKET[116], LIGSite[117], CAST[118] and LigandFit[119]. In these methods, the largest cavity found is often suggested as the ligand binding site.[114] These methods do however not account for the ability of a pocket to be a ligand binding site. Therefore, other methods have been proposed using probe atoms of different types, in order to additionally scan for physico-chemical properties of the predicted sites and predict interaction forces with the probe atoms.[111, 114] This approach is applied in webservers such as Q-SiteFinder[120], SITEHOUND[121], FTSite[122] and SiteComp[123].

1.4.2 Cavity mapping

The methods above rely on the rigid protein approximation (*i.e.* the assumption that the protein conformation does not change upon ligand binding) as they apply their predictions on a static structure. However, cavities in proteins (and thus ligand binding sites) can split into multiple cavities, merge again into a single cavity, or even completely disappear due to the protein dynamics.[124] POVME[125] and mkgridXf[126] are binding site identification algorithms which map cavities and measure their volume to find cavities large enough to accommodate ligand binding. Both these methods can work on an ensemble of proteins retrieved from a molecular dynamics simulation, thereby partially considering the effects described above. However, the molecular dynamics simulations should be rather long in order to observe

cavity appearing/disappearing effects, making these methods computationally expensive. Furthermore, ligand-induced effects are not considered by these methods, as they rely on structure ensembles retrieved from a molecular dynamics simulation of the apo structure.

1.4.3 Template-based approaches

Another strategy to identify the binding site in proteins is via a comparison of the target protein with proteins with known binding sites. In these methods, a structure alignment is performed on the target protein with a database of proteins with known binding sites, to search for structurally similar protein templates. Subsequently, the information of the binding sites of the templates is transferred to the target protein, generally after performing a cluster analysis. The accuracy of these methods is highly dependent on the availability of structurally similar protein templates.[111] Binding site identification algorithms applying this approach include FINDSITE[127], FunFold[128] and 3DLigandSite[129].

If the structure is not known, template-based approaches can also be performed on protein sequences instead. Here, protein templates are identified by their similarity in the protein sequence, followed by a sequence alignment of the target protein with the identified templates. The amino acids which align in the binding site region of the target proteins are then assumed to form the ligand binding site in the target structure.[111] S-Site[130] is a known binding site identification algorithm applying this approach, using the BioLip Database[131] which contains annotated information of the binding sites of all proteins in the Protein Data Bank. Binding site identification algorithms applying a mixture of these sequence-based and structure-based methods (so-called “hybrid methods”) have also been presented, such as COFACTOR[132] and TM-SITE[130].

1.4.4 Affinity grid representations

AutoLigand[133] and AutoSite[134] are commonly applied binding site identification algorithms, both relying on a principle called “affinity grids”. In these methods, a grid is placed over the entire protein, where each grid point represents a possible position for an atom. The affinity of an atom at each grid point is often pre-calculated, which speeds up the subsequent binding site identification. [133-135] AutoLigand applies the following strategy to identify ligand binding sites in a target protein: first cavities in the target protein (identified via a random set of grid points) are flooded with grid points, followed by a “migration” step, in which low-affinity points are deleted and high-affinity neighboring points are added to the flood. Furthermore, an approach called “ray-casting” is applied to identify potential new pockets at greater distances from the pocket.[133] AutoSite applies the same idea of affinity grids, but calculates the affinity of grid points for multiple atom types, namely for a hydrophobic atom (carbon), a hydrogen bond donor,

and a hydrogen bond acceptor. The high-affinity points of these different types are collected and clustered, which results in a set of predicted binding sites, together with a representation of the physico-chemical properties of the respective binding site.[134]

1.4.5 Blind docking

Blind docking is a docking approach (see Chapter 1.5) without pre-knowledge of the binding site.[39] Thus the ligand is docked all around the protein surface, and a score is calculated for each docked solution, often representing the binding affinity. The top-scored poses finally represent the suggested binding sites of the proteins.[136, 137] Despite being a rather straightforward approach, this method can result in very reliable predictions. Moreover, this method searches for a binding site specifically for the input ligand, while the majority of the methods described above predict general ligand binding sites. However, because of the high number of docked poses required to rigorously scan the protein surface, this method can become computationally expensive.[136-138] Hitényi *et al*. developed such a blind docking approach applying AutoDock, in which they successfully predicted the binding site for 34 of the 43 proteins present in the evaluation set.[137] PEP-SiteFinder[139] is another blind docking approach designed for the prediction of binding sites for peptides. PEP-SiteFinder uses the protein-protein ATTRACT force field and docking algorithm.[140] First, several peptide conformations are predicted based on the peptide sequence, followed by a fast blind docking approach on the entire protein surface. The authors evaluated this approach using an evaluation set consisting of 41 protein systems. Considering the 10 top-scored poses, the binding site residues predicted by PEP-SiteFinder overlapped for almost 90% of the poses with the known (experimentally-resolved) binding site.[139]

1.5 PROTEIN-LIGAND COMPLEX PREDICTION

In both rational protein engineering and structure-based drug design, it is essential to gather knowledge about ligand binding. Based on the structure of a protein-ligand complex, one could engineer the protein (in rational protein engineering) or modify the ligand (in structure-based drug design) to alter properties of interest, such as ligand binding or enzyme activity. For most applications, molecular docking simulations are the best choice to efficiently predict the structure of a protein-ligand complex.[39] A molecular docking simulation consists of two parts: sampling of possible ligand conformations in the binding site, followed by ranking of the generated poses with a scoring function. Within these two steps, multiple challenges need to be addressed, such as consideration of ligand flexibility, description of protein-ligand interactions, design of the scoring function, and potentially consideration of protein flexibility.[39]

Furthermore, there are many degrees of freedom in molecular docking simulations, including the translational and rotational degrees of freedom and the conformational degrees of freedom. Numerous molecular docking algorithms have been proposed, and are still being developed, which all differ in the number of degrees of freedom they ignore, and in the strategy to reliably, but also efficiently perform the sampling and scoring steps, and.[39, 141]

1.5.1 Approaches relying on the rigid protein approximation

The majority of available molecular docking algorithms rely on the rigid protein approximation, meaning that they assume that the protein conformation will not change upon ligand binding. This strongly simplifies the docking problem and allows for much more efficient algorithms, but therefore lose in their accuracy. However, for many proteins this approximation can be applied successfully.[39, 141, 142] All rigid molecular docking algorithms, including the algorithms only considering ligand flexibility, rely on the "lock-and-key" model (**Figure 2A**). This traditional idea, introduced by Emil Fischer in 1894 describes that the function of an enzyme can be explained by the analogy of a lock and a key: the lock (*i.e.* enzyme) and key (*i.e.* ligand/substrate) should fit properly together, as a requisite for a reaction to occur.[143] Thus, a good substrate should geometrically match the binding site.

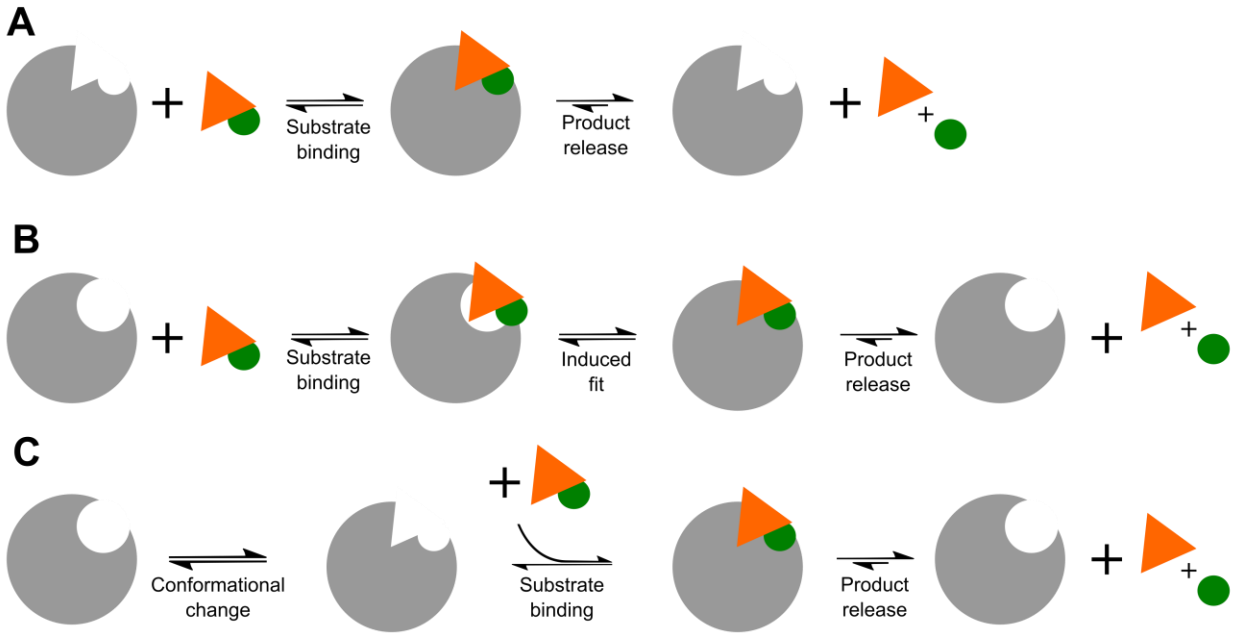


Figure 2. Schematic overview of the lock-and-key model (A), the induced-fit model (B) and the conformational selection model (C), illustrated for a hypothetical system representing an enzyme cleaving a substrate, with the equilibrium lying on the product side. The gray shape represents the hypothetical enzyme, and the cove represents the binding site. The orange and green shape represent a substrate. A detailed description of these models is provided in the main text.

Shape complementary and incremental construction algorithms

The simplest kind of docking methods include the algorithms ignoring all conformational degrees of freedom, as both the protein and ligand are considered as rigid bodies. These methods rely on shape complementary between the ligand and the binding site.[141, 144] DOCK, one of the first molecular docking algorithms, relies on this shape complementary principle.[145] In this algorithm, the binding site cavity is represented by a collection of overlapping spheres touching the molecular surface of the protein, which are subsequently mapped on a sphere representation of the ligand.[141, 145] These methods, ignoring both ligand- and protein flexibility, are nowadays not regularly applied anymore due to increased availability of computation power and the development of new algorithms which take more degrees of freedom into account. Most of the molecular docking methods only consider ligand flexibility, as this is computationally much easier to handle than protein flexibility, and can be implemented in numerous ways.[39, 141, 144, 146, 147] One possibility is an incremental construction of the ligand inside the binding site, as applied by the docking algorithms Hammerhead[148], DOCK[149] (a variant of the version described above), and FlexX[150]. In this method, the ligand is first fragmentized into small rigid fragments, followed by the definition of a base fragment, typically a rather rigid region of the ligand, which is placed inside the binding site. All the different placements of the base fragment act as starting point for the incremental construction algorithm, which iteratively place the remaining fragments of the ligand. The individual molecular docking algorithms applying this approach can differ in their fragmentation approach, base placement strategy, incremental build-up algorithm and scoring function.[39, 141, 145, 148, 149, 151]

Genetic algorithms

Another popular way to implement ligand flexibility is the application of a genetic algorithm, which encode dihedral angles of all rotatable bonds in the ligand, as well as the orientation of the ligand inside the binding site.[39, 141] In a genetic algorithm, molecular properties (here: dihedral angles and ligand orientation) can be encoded in a bit string, or a “chromosome” in the context of a genetic algorithm. Every chromosome is assigned a fitness score, which scores the quality of the docked pose. Often the internal energy calculated via a Molecular Mechanics force field or a score retrieved from a scoring function is used for this fitness score. A first population of chromosomes (“parents”) are randomly generated, followed by generation of new chromosomes (“children”) by genetic operators such as crossover and mutation, which randomly pick parents with a slight bias toward the more “fit” chromosomes. The crossover operator, as inspired by chromosomal crossover during sexual reproduction, combines properties from both parents to generate children, which are again assigned a fitness score. The mutation

operator randomly mutates the chromosome to aim for improved fitness, as inspired by principles from evolution. By altering parameters such as crossover- and mutation rates, population sizes and number of allowed generations, the performance of the genetic algorithm can be modified.[39, 141, 152] Molecular docking algorithms applying this idea include AutoDock[135], AutoDock Vina[153] and Gold[154].

1.5.2 Consideration of protein flexibility in molecular docking simulations

The "lock-and-key" model, which serves as basis for the rigid molecular docking algorithms, is however not always valid. Numerous studies have shown that the conformation of the binding site can adapt on the presence of a ligand.[124, 155-158] Therefore modifications of the "lock-and-key" model have been proposed, such as the "induced fit" model[159], later followed by the "conformational selection" model[160] (**Figure 2B,C**). In the first, the binding site conformation is induced by the ligand, thus the conformational change happens during the binding event, while in the latter, the binding site conformation changes already before the binding event, after which the ligand finds an enzyme/receptor with the "correct" binding site conformation. The question regarding which of the two is correct has been a topic of discussion for many years.[124, 158, 161-163] Regardless of which of the two models is more accurate, the requirement for such a model illustrates the importance to consider protein flexibility in molecular docking simulations.

Ensemble docking and side-chain flexibility

A straightforward way to include information about protein flexibility in molecular docking simulations is to perform rigid docking simulations into multiple protein conformations. These conformations can either be retrieved from experimental data, such as X-Ray or NMR crystallography, or via computational sampling methods such as molecular dynamics or Monte Carlo simulations. This approach is often referred to as "ensemble docking".[164, 165] This ensemble docking approach has been applied frequently, such that even automated pipelines were developed for AutoDock[166] and FlexX (named: FlexE)[167]. However, experimental structures of multiple relevant protein conformations are often not available, while very extensive sampling needs to be performed to retrieve these conformations computationally due to the large phase space which needs to be sampled, making the latter very computationally expensive. Therefore, the application for this ensemble docking method is limited.[142, 165]

However, by limiting the protein flexibility to the side-chains, one can overcome the unfeasible computations required to generate protein conformations.[39, 165] Methods which only consider side-chain flexibility typically rotate small organic groups, or apply rotamer libraries to predict the most likely conformations of the side-chains (see Chapter 1.3.1 for a more extensive discussion about rotamer

libraries).[39] For example, in the docking algorithm GOLD[152, 154], hydrogen atoms can rotate to optimize hydrogen bonds, while in AutoDock[135], protein conformations can be pre-sampled by modifying the side-chain conformations, followed by docking of the ligand using an ensemble docking approach.

Ligand-induced protein flexibility

The above-mentioned approaches to consider protein flexibility do not account for conformational changes induced by the ligand, *i.e.* induced-fit effects. Even if one rather follows the conformational selection model, it is rather unlikely that large conformational changes, such as loop movement events, are properly sampled during pre-sampling. Therefore, several docking algorithms have been presented, which explicitly account for full protein (thus including backbone) flexibility during the docking process. For example, RosettaLigand[168] allows for full protein flexibility, combining a coarse-grained stage, and a Monte Carlo minimization stage, followed by a geometry optimization to relax protein-ligand interactions. DynaDock[169] is another molecular docking algorithm which accounts for full protein flexibility. The DynaDock algorithm consists of two stages: broad sampling and Optimized Potential Molecular Dynamics (OPMD) refinement. In the first broad sampling step, random ligand conformations are sampled within the binding site, allowing a certain protein-ligand overlap. For a selection of these broad sampled poses, an OPMD simulation is performed, during which the protein-ligand overlap is slowly resolved by slight movements of both the ligand and the protein.[169]

1.5.3 Evaluation of docked complexes

After generating possible ligand orientations in the binding site (*i.e.* docked poses) during the sampling stage, the docked poses are ranked with help of a scoring function. Scoring functions aim to rank the docked poses by their free energy of binding, which is given by:

$$\Delta G = \Delta H - T\Delta \mathrm{S} \qquad \textbf{(1.2)}$$

where ΔG represents the Gibbs free energy of binding, ΔH the enthalpy change, T the temperature (in Kelvin) and ΔS the change in entropy. The relation between the free energy of binding and the binding constant K_i is given by:

$$\Delta G = -RT\ ln(K_i) \qquad \textbf{(1.3)}$$

where R represents the gas constant.[141, 158]

Scoring functions need to be efficient because a large number of poses need to be scored. Therefore, the majority of the scoring functions do not aim to accurately reproduce the absolute value of the free energy of binding, but produce a unitless score, which does not necessarily represent any physico-chemical property, but only intends to identify the best generated solution.[141, 158] The majority of available scoring functions can be classified into physics-based, empirical, knowledge-based and machine learning-based scoring functions.[39, 170] Empirical, knowledge-based and machine learning-based scoring functions are only introduced briefly, as mainly physics-based scoring functions were applied in this dissertation.

Physics-based scoring functions

Physics-based scoring functions are one of the best known scoring functions, existing in multiple flavors. The classical physics-based scoring functions are the force field-based scoring functions, which are based on the following terms:

$$\Delta G_{bind} = E_{vdw} + E_{elec} + \Delta G_{solv} \quad \textbf{(1.4)}$$

ΔG_{bind} is the free energy of binding, and E_{vdw} and E_{elec} are the Van der Waals and electrostatic interactions between the protein and ligand, respectively. ΔG_{solv} is the solvation free energy, typically calculated with an implicit solvent model. Not all force field-based scoring functions contain the solvation term, as this is harder to calculate than the other terms, and thereby making the scoring function more computationally expensive.[39, 170] One well-known physics-based scoring function is MM(GB/PB)SA (Molecular Mechanics Generalized Born/Poisson-Boltzmann Surface Area), in which the free energy of binding (ΔG_{bind}) is estimated as the difference in free energy between the bound state ($G_{complex}$) and the unbound states ($G_{receptor}$ and G_{ligand}):

$$\Delta G_{bind} = G_{complex} - \left(G_{receptor} + G_{ligand}\right) \quad \textbf{(1.5)}$$

The individual contributions of the bound and unbound states are either calculated based on a single frame (*e.g.* X-Ray structure of energy-minimized structure), or as an ensemble-average retrieved from (a) molecular dynamics simulation(s). For the latter, molecular dynamics simulations can either be performed for the complex only (single-trajectory protocol), or for multiple states (multiple-trajectory protocol). In a single-trajectory protocol, the energy of the receptor and ligand state are simply retrieved from the complex simulation.[171] However, in this protocol, it is assumed that the receptor and ligand sample similar conformations in the bound and free states, which is questionable. In order to address this so-

called adaptation free energy (*i.e.* free energy associated with conformational adaption of receptor and ligand upon binding), a multi-trajectory protocol can be performed. Here, the free states are simulated as well, either all three states in a 3-trajectory protocol, or only the complex state and the unbound ligand state in a 2-trajectory protocol. The multi-trajectory protocol often introduces a lot of noise, and requires much more simulations, which is the reason that the single-trajectory protocol is most commonly used.[171, 172]

In MM(GB/PB)SA, the free energy of binding is estimated as:

$$\begin{aligned} \Delta G_{bind} &= \Delta H - T\Delta S \\ &\approx \Delta E_{gas} + \Delta G_{solv,pol} + \Delta G_{solv,np} - T\Delta S \end{aligned} \quad \textbf{(1.6)}$$

ΔE_{gas} is the gas-phase (vacuum) interaction energy term, *i.e.* the sum of the change in internal energies (bond, angle, and dihedral energies, which are cancelled out in a single-trajectory protocol) and the Van der Waals and electrostatic interactions. The solvation free energy is divided in two terms, the polar ($\Delta E_{solv,pol}$) and nonpolar ($\Delta E_{solv,np}$) contribution. The polar solvation term is calculated in an electrostatic continuum, calculated applying a Generalized Born (GB) model, or by solving the Poisson-Boltzmann (PB) equation. The nonpolar part is calculated as a function of the surface accessible surface area (SASA). The entropy term ($T\Delta S$) can be approximated by a normal mode analysis[173], but often this term is neglected because of the rather high computational costs, and the contribution often rather small, especially when only relative binding free energies are of interest.[172]

Alternative scoring approaches

Empirical scoring functions sum up potential terms such as the Van der Waals protein-ligand interaction term, a hydrogen-bond term, a term accounting for loss of entropy and a term describing hydrophobic clashes. All terms are weighted by a weighting term, which results from a training of protein-ligand interactions with available experimental binding affinity data.[170, 174] Examples of empirical scoring functions are X-score[175, 176] and the FlexX scoring function[150, 177].

In knowledge-based scoring functions, distance-based potentials (more exact: potential of mean force) are generated for all possible combinations of protein-ligand interactions, based on available structural information (usually the Protein Data Bank). In other words, if a ligand atom and a protein atom are often found at a certain distance from each-other, it is assumed that this is the optimal distance for these atoms and should retrieve the best score. [39, 170]

Finally, machine learning-based scoring functions are the newest type of scoring functions, which depend on a non-linear learning algorithm. Machine learning-based scoring functions can apply numerous machine learning algorithms, such as random forest, deep-learning, or neural networks. These types of scoring functions have shown to be able to outperform the traditional scoring functions already. However, so far these type of scoring functions are barely implemented in any docking algorithm because of their dependence on the training set they were developed with.[170]

1.5.4 Model refinement and analysis

Molecular docking simulations typically result in multiple predicted protein-ligand complex solutions, which are scored based on their binding affinity to the protein. However, there may be more information about the specific target system available, which can be used to further filter down the predicted solutions. This filtering is generally performed with the use of pharmacophore constraints. Moreover, the majority of the molecular docking algorithms result in a static representation of the protein-ligand complex, thus subsequent dynamics simulation (*e.g.* molecular dynamics, Monte Carlo simulations or geometry optimizations) can provide further information about the stability of this complex, and the importance of individual protein-ligand interactions. Finally, if one is interested in more complex electronic properties within the binding site, such as (transition) metal coordination effects or the reaction mechanism, subsequent Quantum Mechanics (QM) or hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) calculations can provide useful insights.

Pharmacophore constraints

When investigating the binding of a certain ligand to a protein, one may already have some knowledge about how the ligand, or a part of the ligand, will bind to the protein. This information can either be retrieved from binding affinity data and/or crystallography data of a similar ligand or homologous protein, via mutation data of binding site residues, or simply because one knows the catalytic site (*e.g.* catalytic residue or triad, iron-sulfur cluster, or metal ion). This information can be used to filter out docked poses that do not fulfill these conditions with the use of pharmacophore constraints.[141, 178, 179]

A pharmacophore is defined as an ensemble of physico-chemical features allowing intermolecular interactions, leading to activation or prevention of a biological response. Thus, a pharmacophore is neither a ligand, nor a functional group, but an abstract moiety which can interact with a biological target. A pharmacophore is associated with a certain pharmacophoric descriptor, such as hydrophobic, aromatic, positively/negatively charged or hydrogen-bond donor/acceptor.[180] One could define such pharmacophore constraints during or after a molecular docking simulation, which enforces a certain

pharmacophore in a pre-defined region in three-dimensional space.[178, 179] This strategy is often applied in virtual screening approaches in computer-aided drug design, but can also be used in molecular docking studies of an enzyme-substrate complex. For example, if a catalytic triad has been identified in an enzyme, one knows where to expect a certain pharmacophore of the ligand in the binding site, which can guide the placement of a pharmacophore constraint. This approach has been applied frequently, and can help to automatically filter out incorrectly docked poses.[141, 178, 179]

Pose refinement by molecular dynamics simulations

In order to obtain more information about the dynamics and stability of the predicted protein-ligand complex, subsequent molecular dynamics simulations can be performed. The stability of the ligand pose during a molecular dynamics simulation can also provide information regarding the quality of the docked pose, since only a decently docked pose will occupy a local minimum allowing equilibrium molecular dynamics simulations.[169, 181] Thus, poses which result in an unstable molecular dynamics trajectory, *i.e.* when the predicted protein-ligand interactions are not present during the majority of the simulation, can be assumed to be incorrectly placed, and thus rejected as protein-ligand complex solution.[169, 181, 182] Besides filtering out incorrectly docked poses, several studies showed that these molecular dynamics pose refinement simulations are also able to discriminate between binders and non-binders, and not predicted protein-ligand interactions can be recovered, especially when small conformational changes of the binding site upon binding is expected.[183-185]

Application of Quantum Mechanical calculations in molecular docking simulations

Finally, pure Quantum Mechanics (QM) or hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) simulations can help to refine and/or to rescore docked poses. This is especially helpful if the binding site contains a complex electronic structure (*e.g.* a highly polarized binding site, or presence of one or multiple transition metal ions) which are hard to describe with Molecular Mechanics (MM).[186, 187] The major bottleneck of the application of QM methods in refinement simulations or rescoring approaches are their rather high computational expenses. Therefore, the region of the system that is described via a QM potential can be reduced, as performed in a QM-cluster approach (sometimes referred to as “QM-only”). In this approach, only the relevant region of the system (generally the ligand and binding site residues) is described via a QM potential, and the remaining atoms are removed. The electrostatics of the protein environment can be mimicked by a homogenous polarizable continuum using a dielectric constant.[188] An alternative to the QM-cluster approach is a hybrid QM/MM potential, in which only the most relevant region is simulated by a QM potential, and the remaining via a MM potential. While all interactions within

the QM- and MM-region are handled by the QM- and MM-code respectively, a coupling scheme is required to account for the interactions between those two regions. This can either be conducted via an additive QM/MM coupling scheme (often referred to as the "QM/MM scheme"), or a subtractive scheme, as applied in ONIOM (Our own N-layered Integrated Molecular Orbital + Molecular Mechanics).[188, 189] See Chapter 2.1.2 in the Theory section for a more detailed discussion on this topic.

Due to the increasingly rising computing power, it became nowadays possible to apply QM-calculations efficiently directly during the molecular docking simulations, or in scoring functions. Numerous QM-based docking algorithms and scoring functions have been developed in the last years, applying different approaches.[190-193] For example, QM calculations can be used to derive system specific partial charges (so-called "polarized protein-specific charges") for the ligand and potentially the binding site.[194, 195] Other methods include the "on-the-fly" scheme, where a QM/MM geometry optimization is included in the docking algorithm.[196] Moreover, several QM-based scoring functions have been developed, which mostly rely on semi-empirical QM potentials due to their reduced computational cost. For example, QMScore is a QM-based scoring function developed by Raha and Merz, combining the AM1 semi-empirical Hamiltonian to calculate gas phase interaction energies, a MM potential for the nonpolar interactions, while using the Poisson-Boltzmann implicit solvent model for the solvent contribution. This method was successfully applied for 23 ligands binding to the zinc metalloenzymes carbonic anhydrase and carboxypeptidase A. [197] More recently, several SQM/COSMO scoring functions have been developed, which combine a semi-empirical QM potential for the enthalpic contribution and COSMO, a QM-based solvent model, for the solvation term. Pecina *et al*. published many variants of this SQM/COSMO scoring function applying different semi-empirical Hamiltonians and correction terms for dispersion, hydrogen-bonding and halogen-bonding, and showed its ability to outperform classical scoring functions for several metalloenzymes.[191, 198-200] Finally, Cavasotto and Aucar recently developed a QM-based scoring function (PM7/COSMO), which is according to the authors suitable for high-throughput docking simulations, as it is only 10 times slower compared to MM-based scoring functions. They evaluated the scoring function on 10 diverse protein systems, showing excellent results, even without applied geometry optimizations of the docked complexes.[192]

1.6 METALLOENZYMES

1.6.1 Dihydroxyacid Dehydratases

Dihydroxy-acid dehydratases (DHADs) and other sugar-acid dehydratases (DHTs) are hydro-lyases (EC 4.2.1) acting on carbon-oxygen bonds. They catalyze the dehydration of sugar-acids (or dihydroxy-acids), producing a C-O double bond in the substrate (**Figure 3**). These enzymes belong to the ilvD/EDD superfamily, which are suggested to contain an iron-sulfur cluster in the active site, either [2Fe-2S] or [4Fe-4S], where enzymes carrying the latter are often unstable in aerobic conditions.[201] These enzymes have gained increased attention due to their involvement in a large variety of biosynthetic and carbohydrate metabolic pathways. For example, DHADs are involved in branched-chain amino acid biosynthesis via the Ehrlich pathway.[202] Moreover, via a modified Ehrlich pathway, DHAD is involved in the production of higher-chain alcohols, which show high potential as biofuel, and can be produced starting from glucose.[9, 203] As an alternative to the *in vivo* approach (by the modified Ehrlich pathway) to produce higher chain alcohols, an *in vitro* pathway has been developed consisting of only eight enzymes to produce isobutanol from D-glucose.[8] The limiting factor in the pathway is the non-natural reaction of D-glycerate to pyruvate by a DHAD from *Saccharolobus solfataricus* (*Ss*DHAD).[8, 204] Other members of the ilvD/EDD family include D-xylonate dehydratase (XyDHT), L-arabinonate dehydratatse (ArDHT), and 6-phosphogluconate dehydratase (6PGDHT), where the first two catalyze the dehydration of D-xylonate and L-arabinonate, respectively.[205, 206] The latter is involved in glucose metabolism via the Entner-Doudoroff (ED) pathway.[207]

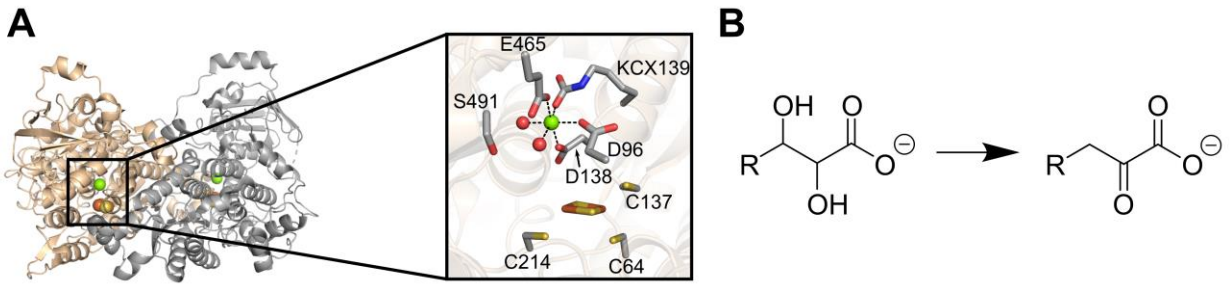


Figure 3. Structure of DHADs, illustrating the active site at the dimer interface and active site composition (A) and chemical reaction catalyzed by DHADs (B). The DHAD dimer is shown in cartoon representation, with both monomers colored in orange and gray, respectively. The inlet represents the active site, with important residues and the [2Fe-2S] cluster shown as sticks. The green and red spheres represent the Mg^{2+} ion and two crystallized water molecules, respectively, which are replaced upon substrate binding. The dotted lines represent coordinate bonds. The residues are indicated by their one-letter abbreviation. The protein structure shown here is from the DHAD from *Mycobacterium tuberculosis* (*Mt*DHAD; PDB-ID: 6ovt).

Despite the promising role of DHADs in biotechnology, there are not many resolved structures of DHADs or other [2Fe-2S]-containing dehydratases belonging to the ilvD/EDD superfamily. However, it is known that the active site is located in a dimer interface, and most DHADs are present in a tetrameric oligomerization (here: dimer of dimers). The active form contains an iron-sulfur cluster, a divalent metal ion, and an essential serine residue, which are all assumed to play a role in substrate binding and/or catalysis.[208-213] There are two reaction mechanisms proposed for DHADs, where the one originally suggested by Pirrung *et al.[208]*, and later by Rahman *et al.[209],* is most often applied, and also fits best to the simulations performed in this work.[209] The reaction starts with proton abstraction from the substrate's C2 atom by a deprotonated serine residue in the active site (*i.e.* the catalytic serine). The resulted carbanion is stabilized by Mg^{2+}, such that Fe2 of the [2Fe-2S] act as Lewis acid accepting the hydroxyl group from C3, followed by tautomerization to the product in keto form.

1.6.2 Metallo-β-Lactamases

β-lactamases (EC 3.5.2.6) are enzymes hydrolyzing substrates carrying a four-membered β-lactam ring (**Figure 4**), which are often found in antibiotics. These enzymes are produced by both Gram-positive and Gram-negative bacteria, thereby making them resistant against a large range of antibiotics.[214, 215] β-lactamases can be classified with the Ambler classification scheme into class A, B, C and D β-lactamases on the basis of their amino acid sequence.[216] Class A, C and D are serine-β-lactamases, which contain a (catalytic) serine residue in their active site, which is involved in the hydrolysis reaction of β-lactams. Class B β-lactamases contain one or two Zn^{2+} ions in their active site activating a hydroxide ion, and are therefore also named metallo-β-lactamases (MBL). The MBLs can be further divided into three subclasses: B1, B2 and B3, based on their sequence similarity and substrate specificity. The B1 and B3 MBLs contain two Zn^{2+} ions in the active site, while B2 MBLs are active in their mono-metallic form, and are even inhibited upon binding of a second Zn^{2+} ion.[217, 218] B3 MBLs show a low sequence similarity to the other two classes, but their substrate scope is similar to B1 MBLs.[219] A B4 class has also been suggested, in which a second Zn^{2+} ion binds when the substrate is already present.[220] This class can however also be classified as a B3 MBL with a different active site motif.[221] The traditional composition of the two Zn^{2+} sites (denoted as α- and β-sites) observed in B3 MBLs is namely: HHH/DHH, for the α- and β-site respectively. However, in a MBL from *Elizabethkingia meningoseptica* (GOB-1/18), Gln116 was observed in the α-site, leading to the sequence motif **Q**HH/DHH (residue varying to common motif indicated in bold).[222] Moreover, Vella *et al*. found via a database search a MBL from *Serratia proteamaculans* (SPR-1), which shows even more variations in the active site composition leading to the sequence motif H**R**H/D**QK**. In this MBL, only one Zn^{2+} ion was observed in the resting state, but it operates in its di-Zn^{2+}

state.[223] These two enzymes can be named B3-Q and B3-RQK, respectively, based on their active site motif.

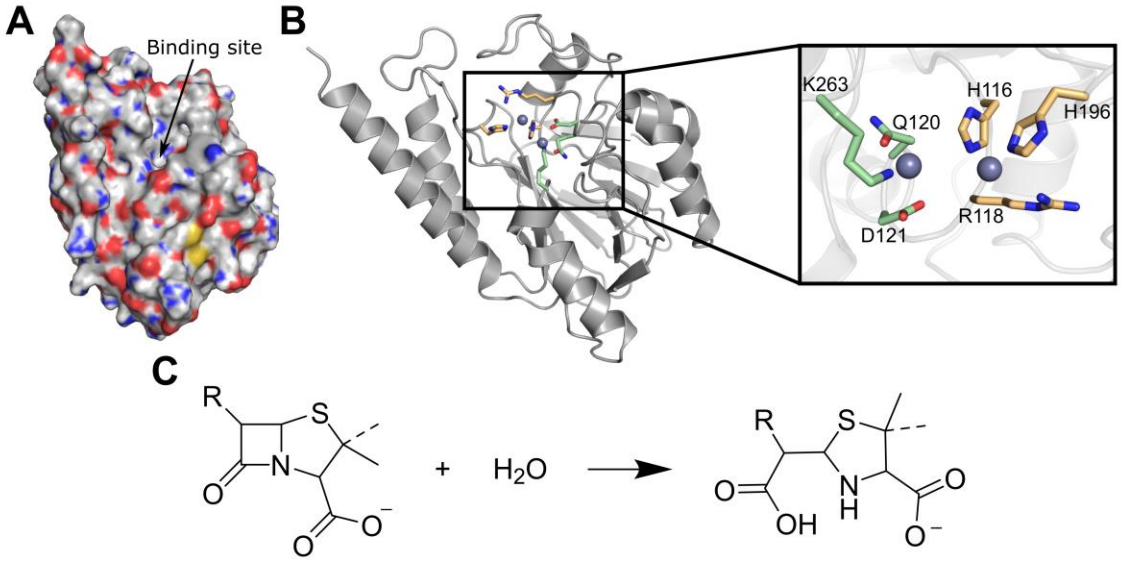


Figure 4. Structural representation of the B3-RQK member CSR-1 MBL (PDB-ID: 6qd2), illustrated as (A) surface representation, and (B) cartoon representation with an inlet illustrating the active site, and (C) the reaction catalyzed by MBLs. The protein is shown as gray surface/cartoon, and the residues coordinating the Zn^{2+} ions in the α-site and β-site are shown as respective orange and green sticks. The Zn^{2+} ions are shown as gray spheres, and are manually modelled by superposition with CSR-1 triple mutant (R118H, Q121H, K263H; PDB-ID: 6dr8). The residues are indicated by their one-letter abbreviation, except KCX, which stands for carboxylated lysine.

1.6.3 Carbonic Anhydrase

Carbonic anhydrases (CAs; EC 4.2.1.1; **Figure 5A**) are enzymes catalyzing the conversion between carbon dioxide (CO_2) and bicarbonate (HCO_3^-), and is widespread in nature.[224, 225] CAs contain a Zn^{2+} ion in the active site, which is coordinated in a tetrahedral geometry by three histidine residues and a water molecule (or hydroxide ion), which is the most common coordination environment for Zn^{2+} in proteins.[226] CAs are popular model systems in numerous research fields, including biophysics and medicinal chemistry, especially for protein-ligand binding studies. This has several reasons, for example, (I) CAs are monomeric and of intermediate size (~30 kDa), (II) they are widely available and inexpensive, (III) a large number of inhibitors have been identified, and (IV) they are well studied. Thus, a lot of data about its structure, catalysis mechanism, internal interaction (*e.g.* hydrogen-bond) networks, and catalytic activity is available and well described.[225] Moreover, there are many experimental structures available: there are 1255 CA structures in the Protein Data Bank to date (checked at: Dec 21th, 2021).

In this work, carbonic anhydrase II (CAII) was one of the model systems in the benchmarking study of Molecular Mechanics-based Zn^{2+} models.[227] CAII proved to be an ideal model system, because of the

common coordination geometry of the Zn^{2+} ion, and its small size, which allowed for long-range molecular dynamics simulations. Furthermore, multiple X-Ray structures were available of CAII in complex with multiple inhibitors, all with available affinity data, allowing for an in-depth analysis of the sampling ability of the evaluated models in metalloproteins.[200] Finally, the binding of sulfonamide inhibitors to CAII has been studied in detail, including experimental deduction of the inhibitor's protonation states, together with important hydrogen-bond networks, which allowed for highly accurate quality assessment of the performed simulations.[224, 228]

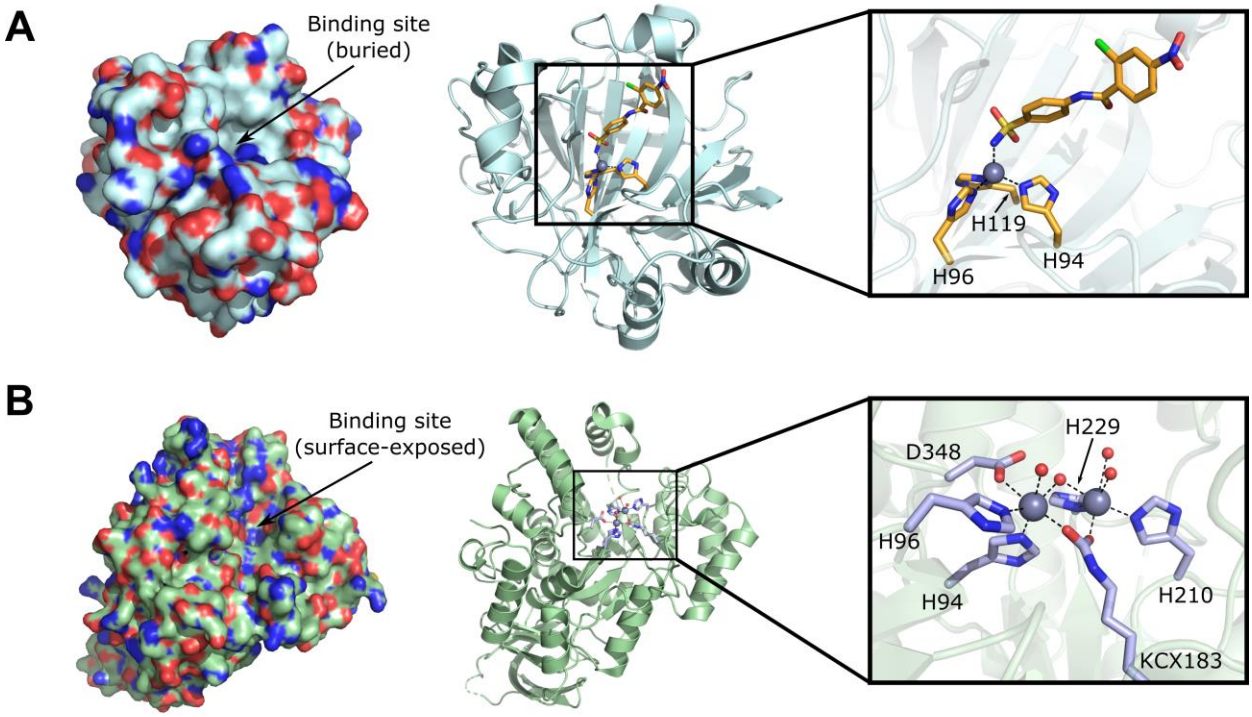


Figure 5. Structural representation of (A) CAII (PDB-ID: 5nxg) and (B) PurAH (PDB-ID: 6i5s). The protein structure is visualized in a surface representation (left) and cartoon representation (center). The inlet shows the active site with the Zn^{2+} ions and crystallized water molecules shown as respective gray and red spheres, and coordinating residues and the CAII inhibitor are shown as sticks. The residues are indicated by their one-letter abbreviation, except KCX, which stands for carboxylated lysine.

1.6.4 Amidohydrolases

The last metalloenzyme studied in this work is an amidohydrolase from *Streptomyces purpureus* (PurAH; **Figure 5B**).[229] This enzyme is involved in the biosynthesis of bottromycin, which is a natural product with antibiotic activity, discovered in 1957.[230] Bottromycins are especially worth studying because of their antibiotic activity against methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *enterococci* (VRE), which are multidrug resistant strains, and therefore play an important role in the antibacterial resistance problem. Bottromycins are macrocyclic peptides, belonging

to the ribosomally synthesized and post-translationally modified peptide (RiPP) family. They contain several non-natural amino acids, such as thiazolines (formed from a cysteine residue), as well as several amino acid analogues. The tetrapeptidic macrocycle and several amino acid substituents have shown to be essential for its antibacterial activity.[231] The biosynthesis of bottromycins, and other RiPPs, have gained increased attention in the last years.[229, 231, 232]

PurAH, an enzyme studied and characterized in this work, is involved in the macroamidine formation. In fact, in this study we could elucidate PurAH as the "gatekeeper" of the macroamidine formation during bottromycin biosynthesis, since PurAH irreversibly cleaves the follower peptide after a thiazoline residue, resulting in a bottromycin precursor.[229] PurAH contains two divalent metal ions in the active site. While the X-Ray structure show two Zn^{2+} ions, enzyme activity could also be observed with Co^{2+} and Mn^{2+}. Based on molecular docking (DynaDock), QM/MM calculations and molecular dynamics simulations, we suggested a possible binding mode of the peptide in the surface-exposed and highly flexible active site of PurAH. These results allowed us to suggest a reaction mechanism, in which an activated (by Zn^{2+}) water molecule or hydroxide ion hydrolyses the peptide bond between the thiazoline and the follower peptide, while D348 acts as a base accepting the remaining proton. Further mutagenesis studies supported the proposed binding mode, as well as the proposed reaction mechanism.

1.7 AIMS OF THIS WORK

1.7.1 Development of an efficient screening algorithm identifying enzymes capable of catalyzing target reactions

Biocatalysts show numerous advantages over traditional chemical catalysts, as extensively introduced above. Moreover, enzyme engineering studies have shown to be able to broaden the substrate scope of enzymes, improve enzyme activity, improve enzyme stability, and much more. However, one first needs to identify a suitable enzyme as starting point in order to develop a biocatalyst for a new target reaction, ideally with at least measurable activity for either the substrate of interest, or a structurally similar substrate. Because of the growing amount of structural data, the aim was to develop a bioinformatics algorithm to screen characterized enzymes on their ability of catalyzing a target reaction. The developed algorithm should fulfill several conditions: (I) the algorithm should be efficient in order to quickly evaluate a large number of potential enzymes, (II) the algorithm should be easy to install and use, also by non-expert users, (III) the user should be able to modify and fine-tune the search algorithm without having to modify the code, and (IV) the output should be easy to understand, contain information relevant for

experimental scientists, and allow for further (either automatic or human) assessment of an enzyme's ability to become a useful biocatalyst.

1.7.2 Development of a binding site identification algorithm with explicit consideration of protein- and ligand flexibility

In order to perform biomolecular simulations in proteins (enzymes in the case of biocatalyst development, or receptors, ion-channels, or other proteins for structure-based drug discovery), it is often essential to know the location of the binding site. However, this information is not necessarily known, and currently available computational tools to predict these binding sites often ignore protein flexibility. Accurate description of this flexibility is however especially important for binding sites for large and flexible ligands, such as peptides. Therefore, the second aim in this dissertation is to develop an algorithm which is able to predict binding sites for large and highly flexible ligands, with explicit consideration of protein- and ligand flexibility. The inspiration for the new binding site identification algorithm is retrieved from the principles of DynaDock[169], in which soft-core potentials are used to allow a certain amount of protein-ligand overlap, which is subsequently resolved to simulate an induced fit effect. However, instead of OPMD, a more efficient algorithm is required because of the large number of potential binding sites that need to be studied. Therefore, an alternative sampling method needs to be selected to explore the potential binding pockets. The performance is compared to existing binding site identification algorithms, such as AutoSite, a blind docking algorithm with AutoDock, and PEP-SiteFinder.

1.7.3 Benchmarking Biomolecular force field-based strategies to simulate Zn^{2+} containing metalloproteins

Biomolecular simulations are being improved year by year, and are already able to accurate simulate the majority of biomolecules. However, interactions with metal ions are still challenging to simulate, especially in a protein environment. Accurate simulations of metalloproteins are however of particular interest because a large number of proteins, including enzymes, rely on a metal ion (often Zn^{2+}) for their activity. These simulations can be accurately performed with QM or QM/MM simulations, but these type of simulations are still too computationally expensive for every-day use, and not suitable for long timescale (molecular dynamics) simulations. Therefore, the third aim of this work is to benchmark and assess available force field-based Zn^{2+} models on their performance in sampling Zn^{2+} ions in ligand binding sites. Besides the stability of the simulations, special focus should be given to the simulated coordination geometries of the Zn^{2+} ions, as well as the preferred type of interactions.

1.7.4 Application of multiscale modelling techniques to engineer [2Fe-2S]-dependent dehydratases

The final aim of this dissertation is to apply a wide range of biomolecular simulations and other bioinformatics tools to guide a rational enzyme engineering project to derive sequence, structure, and activity relationships of DHADs. Because of the relevance of DHADs in the biosynthesis of biofuels, antibiotics and other fine chemicals, it is worth studying which factors determine the difference in substrate specificity of known DHADs. This information can be used to engineer DHADs to alter their substrate specificity, ideally toward D-glycerate, as no suitable DHAD has yet been identified showing high activity on this substrate. Since there are no X-Ray structures available for the target DHADs, homology models need to be generated for all target DHADs. Because of the low sequence-similarity with potential templates, the produced models should be carefully evaluated and refined, *e.g.* with molecular dynamics simulations and other refinement techniques. Furthermore, a DHAD-specific simulation procedure and parameter set should be designed to accurately simulate the iron-sulfur cluster, based on the results from the study described in Chapter 1.7.3. Traditional bonded models may namely fail because of the vacant coordination position for one of the irons, and the highly polarized binding site due to a Mg^{2+} ion nearby. Based on bioinformatics analysis, followed by molecular docking and molecular dynamics simulations, residue positions that are likely to play a role in substrate specificity (so-called engineering hotspots) should be identified. These positions can be further evaluated experimentally with site-directed and saturation mutagenesis, performed by experimental collaborators. Finally, the above-mentioned simulations can be used to further rationalize the experimental findings.

This page was left blank intentionally.

2 THEORY AND METHODS

2.1 POTENTIAL ENERGY FUNCTIONS

2.1.1 Molecular Mechanics

In order to describe (bio)molecular structures, we need a model with which we can calculate physical properties of molecular systems at an atomic level. We could use Quantum Mechanics (QM), which relies on quantum theory to describe the electronic structure of molecular systems using the Schrödinger's wave function. While QM simulations can be highly accurate, they are computationally expensive because of the high number of electrons that need to be considered, and the iterative guessing of the approximate Hamiltonian, to solve, or approximate, the Schrödinger equation.[233] An alternative are classical simulations relying on Molecular Mechanics (MM), in which the electronic motions are ignored, and thus calculate the energy of a system as a function of atom's nuclear positions. These simulations can still be rather accurate thanks to the Born-Oppenheimer approximation, which describes that the electronic and nuclear motion can be decoupled because of their significant difference in mass, and thus momentum.[141, 233] Thus, MM allows for simulations consisting of much more particles (atoms), and is thus often the method of choice for simulations of large complexes, such as proteins.

Force fields

In order to calculate the potential energy of a system with MM, one needs a collection of functions and corresponding (empirically-derived) parameters, which are collected in a so called "force field". A variety of force field flavors are available, including all-atom, united atom and coarse-grained force fields that differ in the proportion of simplification applied. Within these classes, numerous force fields have been developed, applying slightly different functional forms and parameters.[141, 233] Furthermore, polarizable force fields are (being) developed, which explicitly account for electronic polarization effects, and thus additionally contain polarizability parameters.[234] Examples of regularly applied protein force fields are the AMBER, CHARMM, GROMOS, and OPLS force fields.[235-238] Force fields contain both bonded and non-bonded potentials, which will be introduced briefly.

The first potential energy function describes bond stretching, and is generally defined as

$$V_{bond}(r_{ij}) = \sum_{bonds} \frac{1}{2} k_{ij}(r_{ij} - r_{eq})^2 \quad \textbf{(2.1)}$$

where k_{ij} is the force constant, and r_{ij} and r_{eq} are the bond length between atom *i* and *j*, and the equilibrium bond length, respectively.

The potential energy function describing angle bending is defined as

$$V_{angle}(\theta_{ijk}) = \sum_{angles} \frac{1}{2} k_{ijk}(\theta_{ijk} - \theta_{eq})^2 \quad \textbf{(2.2)}$$

where k_{ijk}, θ_{ijk} and θ_{eq} are the force constant, the actual angle defined by atom *i*, *j*, and *k*, and the equilibrium bond angle, respectively. Thus, both bond stretching and bond angle is approximated via a harmonic potential, which can quite accurately describe these potentials for bond lengths and angles around the equilibrium values.

The torsional potential energy is given by

$$V_{tors}(\phi) = \sum_{torsions} \frac{1}{2} V_n[1 + \cos(n\phi - \gamma)] \quad \textbf{(2.3)}$$

where V_n is the potential height, n is the periodicity, ϕ is the dihedral angle and γ is the phase shift.

The non-bonded potentials describe the Van der Waals and electrostatic interaction. The Van der Waals interaction is described by the Lennard-Jones (LJ) potential between atom *i* and *j*:

$$V_{LJ}(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \quad \textbf{(2.4)}$$

where r_{ij} is the distance between atom *i* and *j*, $R_{min,ij}$ is the distance where the LJ-potential reaches its minimum, and ε_{ij} is the LJ-well depth.

The electrostatic potential is described by the Coulomb term:

$$V_{Coul}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \qquad \textbf{(2.5)}$$

where r_{ij} is the distance between atom i and j, ε_0 is the relative permittivity, and q_i and q_j are the partial charges of atom i and j, respectively.

Interactions with metal ions

A classical description applying the coulomb and LJ 12-6 potential for non-bonded interactions is often sufficient. However, limitations of this model, especially the LJ 12-6 term, becomes evident in simulations with (multivalent) metal ions.[239] This is mainly due to the complex nature of these metal ions: the high charge induces strong polarization and charge-transfer effects, and its electron configuration causes the metal-coordinating atoms to orient themselves in certain coordination geometries. Since a metal ion only consists of a single atom, all these properties need to be described by the limited non-bonded parameters of this ion. As the charge of the ion is fixed, the only remaining adjustable parameters are the R_{min} and the well-depth (ε) of the Lennard-Jones potential. There are numerous parameter sets available for these metal ions, as well as alternative (mathematical) constructs to improve the description with metal ions in a biomolecular force field, which were benchmarked in this dissertation (Chapter 3.3).

Li *et al*.[240] realized the relevance of charge-induced dipole interactions in non-bonded interactions with metal ions. This effect is however not described by the classical LJ 12-6 potential, thus including this interaction could improve the description of non-bonded interactions with metal ions (**Figure 6A**). Therefore, Li *et al*. introduced a $1/r^4$ term into the classical LJ 12-6 potential, leading to the so-called 12-6-4 LJ-type model:

$$V_{LJ}(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6} - 2\kappa {R_{min,ij}}^2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{4}\right] \qquad \textbf{(2.6)}$$

introducing κ as a scaling factor with unit $Å^{-2}$, which was parameterized for a large range of multivalent metal ions.[240, 241]

An alternative strategy to improve the description with metal ions in a biomolecular force field is via the use of cationic dummy-atom models (**Figure 6B**), also called multiscale models. The first dummy-atom model was developed for the Mn^{2+} ion in 1990 by Åqvist and Warshel.[242] Nowadays, this model has been parameterized for a large range of multivalent ions, including Mg^{2+}, Fe^{2+}, Cu^{2+}, Ni^{2+} and Zn^{2+}.[243-

246] In the dummy-atom model, non-interaction atoms (*i.e.* dummy-atoms) are placed around the metal ion in a certain geometry matching the coordination geometry which should be sampled, typically tetrahedral or octahedral. The dummy-atoms are kept in place by covalent bonds (see **eq. (2.1)**), but can still rotate freely. The charge of the ion is then distributed among the metal ion and the dummy atoms, thereby introducing a charge delocalization. This charge-delocalization supports the ligating atoms to coordinate the metal ion in a certain coordination geometry, but still allows exchange of metal-ligating atoms.[239, 247, 248]

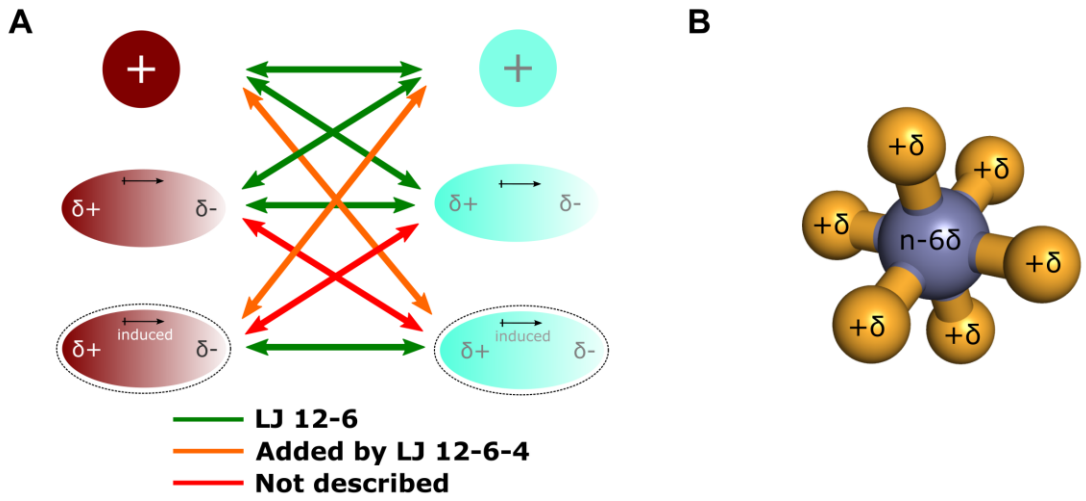


Figure 6. Illustrations of metal ion models. Intermolecular interactions described by (A) the LJ 12-6 and 12-6-4 LJ-type potential, spheres with a "+" represent a charge, the ellipse with a gradient represents a permanent dipole, and an ellipse with dotted border represents an induced dipole. A representation of an (B) octahedral dummy-atom model is shown, with the metal ion in gray, and the dummy-atoms in orange. The δ represents a partial charge, and *n* the charge of the ion.

Another alternative to simulate metal ions in biomolecular simulations is via a 'bonded model'. Here, explicit bonds are placed between the metal ion and its ligating atoms, to enforce a certain coordination geometry.[239, 249, 250] The bonded model however requires a system-specific parameterization for the newly introduced bonded potentials. Multiple strategies to parameterize the bonded parameters have been proposed, with the Seminario method being the most commonly applied method, in which the bonded parameters are derived via the Cartesian Hessian matrix.[251] For metal centers including a Zn^{2+} ion, the parameterization can also be performed empirically applying the Extended Zinc AMBER Force Field (EZAFF).[250] During the design of this force field, several typical zinc coordination models were parameterized, and the resulting parameters are listed in the force field. If the Zn^{2+} center for a certain application is similar to one of the parameterized coordination models, these pre-parameterized parameters can be applied.

Soft-core potentials

In contrast to experiments, molecular simulations can also be applied to study biomolecules in unphysical states. For example, these unphysical simulations can be applied to study the behavior of atoms or molecules in extreme conditions (*e.g.* concentrations above solubility limit), or at very close interatomic/intermolecular distance. These unphysical simulations can also be applied to calculate the free energy of binding (*e.g.* Thermodynamic Integration or Free Energy Perturbation), improve sampling around transition states or other barrier regions (*e.g.* umbrella sampling), or to accelerate ligand-induced conformational changes (*e.g.* application of soft-core potentials).[141] The latter is performed in DynaDock and DynaBiS, where the Lennard Jones and Coulomb potentials (**eq. (2.4) and (2.5)**) are modified to allow atoms at unphysical interatomic distances. By placing a ligand in the binding site allowing certain protein-ligand overlap, followed by gradually re-introducing the normal (*i.e.* non-softened) potentials during a molecular dynamics simulation, an induced fit effect can be simulated which would normally require much longer sampling.[169, 252]

There are several functional forms of soft-core potentials, where the following functional form initially proposed by Taylor *et al*.[253] is used in this study:

$$V_{nb}(r_{ij}) = V_{LJ}(r_{ij}) + V_{Coul}(r_{ij})$$

$$= 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{\left(\alpha^m\sigma_{ij}^6+r_{ij}^6\right)^2}\right)-\left(\frac{\sigma_{ij}^6}{\left(\alpha^m\sigma_{ij}^6+r_{ij}^6\right)}\right)\right] + \frac{(1-\propto)^n q_i q_j}{4\pi\varepsilon_0\sqrt{\alpha + r_{ij}^2}} \qquad \textbf{(2.7)}$$

where σ_{ij} is the distance at which the potential reaches zero, ε_{ij} is the well depth, and α is the soft-core scaling factor, which takes values ranging from 0 to 1. If α is 0, the normal potentials are returned, while the potentials reach (almost) zero if α is 1. Any α value in between 0 and 1 thus returns a "softened" potential, which allows for very small interatomic distances during biomolecular simulations (**Figure 7**). Furthermore, the behavior of the soft-core potential can be further modified via the exponents m and n, which were set in this study to 3 and 6, respectively, as this showed best results in original study from Taylor *et al*.

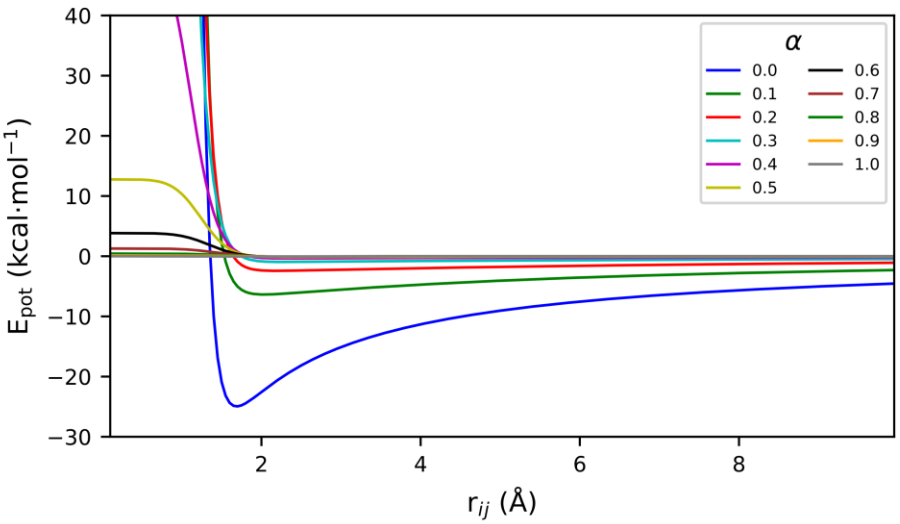


Figure 7. Combined LJ and Coulomb soft-core potential for different values of the soft-core scaling parameter α. The potential is calculated for a C-H pair.

2.1.2 Hybrid Quantum Mechanics/Molecular Mechanics simulations

Classical simulations applying Molecular Mechanics can be combined with Quantum Mechanics to explicitly account for electronic effects, leading to a more accurate description of (a part of) the system. In these "hybrid" simulations, a small part of the system (*e.g.* the active/binding site, or entire solute) is described by a QM potential, while the remainder of the system is described by a classical MM force field. Alternatively, the system can be truncated to a small QM-only system (*i.e.* QM-cluster method), containing solely relevant atoms, for example active site atoms which play a role in the catalyzed chemical reaction. The removed part of the system is either completely ignored (*i.e.* vacuum) or the electrostatic effects is modelled assuming a homogeneous polarizable continuum model (*i.e.* PCM). Depending on the application, the QM-region can be described with the Hartree-Fock method (HF), post-HF methods, Density Functional Theory (DFT), or Semi-Empirical (SE) methods, among others. In the first, the many-electron wave function is described with a Slater determinant, *i.e.* a determinant of single-electron wave functions. However, the Hartree-Fock method does not accurately describe electron correlation, therefore several post Hartree-Fock methods have been developed aiming to tackle this problem. Known post Hartree-Fock methods are the Configuration Interaction and Møller-Plesset method. The Density Functional Theory applies a different approach, as the energy is calculated as a function of electron density instead. While an energy function correlating electron density to the system's energy must exist according to the Hohenberg-Kohn theorems, this function is not known. Therefore, numerous approximate functionals are proposed, which have shown to be able to lead to very accurate results, making this a

popular QM method at the moment. Semi-Emprical methods are typically Hartree-Fock based methods, in which numerous approximations were made based on empirically-derived information, making the calculations much less computationally demanding.[141, 233] A further detailed description of these "pure" QM methods, derivations, postulates and theorems is out of scope for this Chapter, but extensively described in excellent books, reviews, and further literature.[141, 189, 233] In the remainder of this Chapter, the focus lies on combining QM and MM methods in a single simulation.

QM/MM interface

One of the main challenges in QM/MM simulations is a proper description of the electrostatic coupling (polarization) between the QM- and MM-region. This is handled by the so-called "embedding scheme". Three embedding schemes are commonly used: mechanical embedding, electrostatic embedding and polarized embedding.[141, 188, 189, 233] In the first, the electrostatic coupling is handled at the MM-level, by adding point charges (or higher order multipoles) of the QM-atoms in the MM-system. In this way, the MM-atoms are polarized by the QM-region, but not vice-versa. Therefore, in the electrostatic embedding scheme, the MM point charges are included in the QM calculations, such that the QM-region is polarized by the MM-region, and thus can adapt to changes in the electronic structure of the entire system. This embedding scheme is most often applied in biomolecular simulations, also in the calculations performed in this dissertation. Finally, a polarized embedding scheme has been suggested, in which a polarizable force field is applied for the MM-region, such that both the QM- and MM-region polarize each-other.

The second consideration is how to treat bonds crossing the QM/MM interface. The easiest would be to avoid these bonds crossing the interface, but this is not often practically possible, for example in simulations of proteins, with the active site in the QM-region. The most common treatment (*i.e.* boundary scheme) of these interface-crossing bonds is via the use of link atoms. Link atoms are usually hydrogen atoms, present in the QM-region at the end of the interface-crossing bond to saturate the free valence atoms, and thus caps the dangling bond. The link atom can be constrained to avoid addition of degrees of freedom. Furthermore, the MM-atoms close to the boundary may over-polarize the QM-system due to the close link atom in an electrostatic or polarized embedding scheme. This over-polarization can be reduced in multiple ways, while the most common way is to shift charges to neighboring MM-atoms. An alternative to the link atom scheme is the localized-orbital scheme, in which frozen hybrid orbitals replace the interface-crossing bond, capping the QM-region.[188, 189] Finally, besides the choice of the embedding and boundary scheme, the location of the QM/MM boundary is highly important. To avoid

artefacts, the interface should ideally not cut highly polarized bonds (*e.g.* peptide bonds) and be located as far as possible from where the studied event takes place.

Additive and subtractive schemes

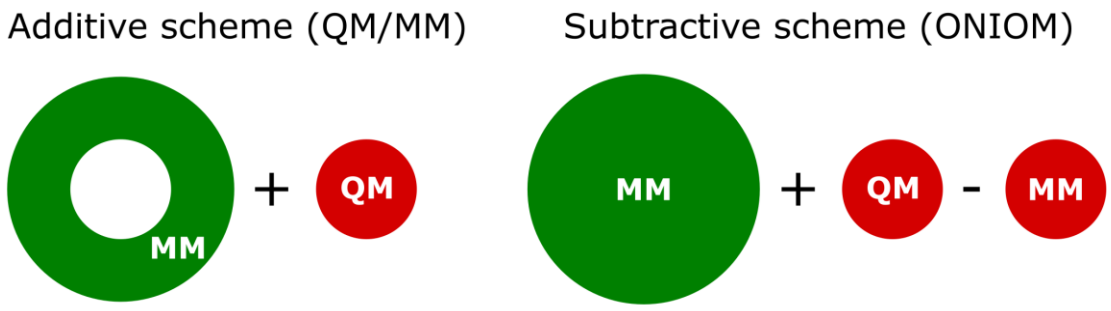


Figure 8. Illustration of the additive and subtractive scheme applied in QM/MM and ONIOM calculations.

In a hybrid QM/MM simulation, the energy of the QM and MM regions are calculated separately, often by two different programs. The question remains how to use these simulations to calculate the total energy of the entire system. Two schemes are available: the additive and subtractive scheme. In the additive scheme, often also named "QM/MM scheme", the total energy is calculated as follows:

$$E_{QM/MM} = E_{MM}^{outer} + E_{QM}^{inner} + E_{QM/MM}^{boundary} \qquad \textbf{(2.8)}$$

Thus, the MM calculation is performed on the outer region only, the QM calculation on the inner region only, and the boundary effects are included via an additional QM/MM coupling term (**Figure 8**). The coupling term contains the bonded and non-bonded interactions between both subsystems. This additive scheme has the advantage that no MM-parameters are required for the inner region, and that no region is simulated twice (by both the QM and MM code).

The subtractive scheme is defined as follows:

$$E_{QM/MM} = E_{MM}^{all} + E_{QM}^{inner} - E_{MM}^{inner} \qquad \textbf{(2.9)}$$

Thus, three simulations are required: one MM calculation of the entire system, one QM calculation of the inner system, as well as a MM calculation of the inner system to avoid double counting (**Figure 8**). This also adds the drawback that MM-parameters need to be available for the inner region, in contrast to the additive scheme. However, no special boundary term is required anymore, and the MM-system does not need to be truncated, such that these calculations can be performed by any MM-code.[188, 189] One of the most widely used subtractive scheme is ONIOM: Our own N-layered Integrated Molecular Orbital +

Molecular Mechanics. ONIOM also allows for hybrid simulations with more than two layers, as in ONIOM3(MO:MO:MO) or ONIOM3(MO:MO:MM), where the first contains three QM layers (often with different functionals, *e.g.* DFT and SE), and the latter contains two QM layers, and one MM layer.[254]

2.2 Applied methodology

This section briefly describes the most relevant simulation procedures applied in this this work. Details in simulation procedure may differ between the different studies described in this dissertation, thus please refer to the respective publication for full details. **Table 1** lists the most relevant software applied in this work.

Table 1. Overview of regularly applied scientific software and algorithms in this work.

Software/Algorithm	Version/release	Purpose
Amber and AmberTools	Amber14/AmberTools14, Amber16/AmberTools16 and Amber18/AmberTools18	Molecular Dynamics, Geometry Optimization, SQM
Gromacs	Version 4.5.6	Molecular Dynamics (system preparation for DynaBiS)
Gaussian	Gaussian09	QM calculations: geometry optimization, single point calculations, RESP fitting, interaction energy scanning
TURBOMOLE	Version 7.1	QM calculations: geometry optimization and single point calculations
Chemshell	Version 3.4 (Tcl-version)	Hybrid QM/MM calculations
AutoDock & AutoGrid	Version 4.2.6	Molecular docking
FlexX (LeadIT)	Version 2.3.2	Molecular docking
DynaDock	Release 604 and older	Molecular docking
AutoSite	Version 2.0.3	Binding site identification
DynaBiS	Original release	Binding site identification
EnzymeMatch	Original release	Prediction of enzymes catalyzing target reaction
FitFF	Original release	Derivation of Amber-compatible force field parameters from QM interaction energy scans
VMD	Version 1.9.3	Visualization
PyMol	Version 1.4	Molecular visualization and *in silico* mutations (using Dunbrack's rotamer library)
Chemdraw	Version 20.0	Molecular editing
Matplotlib	Version 3.3.1	Plotting
PyCharm	Multiple versions	Programming IDE
Vim	7.4	Text editing and programming IDE
PEP-SiteFinder	1.0 (after March 2014 update)	Identify peptide binding sites

2.2.1 Homology modelling

Potential templates were identified based on a BLAST[255] search, and further manually filtered based on their characterized substrate scope. Pairwise and multiple sequence alignments were performed with Clustal Omega[256], except stated otherwise, applying the default parameters. Homology models were generated with Modeller[62] applying the *salign* module, either on single- or multiple (structurally pre-aligned) templates. Homology models of the homo-dimers were generated, and potentially missing ions and cofactors were placed using structural information from homologues with resolved experimental structures, or docked using a molecular docking algorithm (see Chapter 1.5). The DOPE[88] and QMEAN[87, 89] scores were calculated, and Ramachandran plots were generated. All this information was combined to identify the model with the highest quality.

2.2.2 Molecular docking simulations

AutoDock

Molecular docking simulations with AutoDock4.2 were generally performed with a 60x60x60 points grid (size may differ between applications), with 0.375 Å spacing generated with AutoGrid. ≥150 poses (GA runs) were generated, with a population size of 150, allowing a maximum of $2.5 \cdot 10^6$ energy evaluations with AutoDock. The poses were ranked with the AutoDock scoring function, and clustered with a RMSD tolerance of 1.0 Å.

FlexX

Docking with FlexX was performed within the LeadIT platform (https://www.biosolveit.de/LeadIT/). The binding site was defined as all residues within 10 Å of the binding site center (size may differ between applications). The 'Enthalpy and Entropy' (*i.e.* Hybrid) approach was applied for initial base placement, a clash factor of 0.6 was applied, and the maximum allowed overlap volume was set to 3.5 $Å^3$.

DynaDock

Prior to application of DynaDock, the ligands were parameterized, and the system was prepared, heated and equilibrated as described in Chapter 2.2.3. ≥200 random ligand conformations were generated during the broadsampling step in the equilibrated protein structure, allowing 60-80% protein-ligand overlap, and 40-75% intra-ligand overlap. The broad sampling was restricted to a sphere of 8 Å around the center of the binding site (see individual studies for exact sphere size and definition of binding site). The generated poses were clustered with cpptraj from AmberTools, applying the hieragglo clustering algorithm. The centroids of each cluster continued to the next step, *i.e.* the OPMD simulations. In this

step, the soft-core parameter α is optimized using a steepest descent energy minimization. The OPMD simulation was performed at 300 K with an integration step of 1 fs, and the langevin thermostat[257] was used with a collision frequency of 1 ps^{-1}. As the simulation was performed in vacuum, all backbone atoms which were not within a sphere of 1.5 nm from any ligand atom were restrained with a force constant of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (identical to the default restraint force constant in Gromacs). To avoid too fast decrease of the soft-core parameter α, the soft-core parameter was kept fixed for at least 600 steps after α was optimized. Moreover, all soft-core parameters (*i.e.* for Lennard-Jones, Coulomb, and dihedral terms, if applied) were enforced to be equal. Finally, the simulation continued for 500 ps with standard force field potentials (*i.e.* non-softened potentials) after α reached zero.

2.2.3 Molecular dynamics simulations

System and ligand preparation, parameterization, definition of protonation states

For each biomolecular system, the first occurrence of any multi-resolved residue was selected, and the protonation state of all ionizable protein residues was determined at pH 7.0, except stated otherwise, with the PROPKA3.0 software package[258, 259]. The protonation state was checked visually, especially for the metal centers. The AMBER ff14SB[238] force field was applied for all standard protein residues. The system was solvated in a rectangular box or a truncated octahedron filled with TIP3P[260] water, applying a 12 Å buffer region, and counterions (Na^+ or Cl^-) were added to neutralize the system. This was performed using the *tleap* module of the AmberTools software package. The metal ion parameters applied differs for the individual studies, and is described in the Methods sections of the respective publications.

Bonded and Van der Waals force field parameters for all ligands were retrieved from the General Amber Force Field (GAFF)[261], and charges were derived applying the RESP procedure based on a QM-geometry optimized structure performed at the HF/6-31G(d) level of theory. The QM calculations were performed with Gaussian09 (revision E.01)[262]. In this dissertation, new force field parameters were also derived for a carboxylated lysine, as well as a thiazoline residue, as there exist no parameters for these residues in the ff14SB force field. For the thiazoline residue parameterization, an aspartate and thiazoline residue were treated as one residue, as there is no peptide bond between them. For all atoms within the aspartate residue not directly bound to the thiazoline residue, the standard ff14SB force field bonded parameters were used. GAFF parameters were selected for the bonded and Van der Waals parameters for the thiazoline residue, the remaining aspartate residue, and the carboxylated lysine. For the charge derivation, the residues were capped with an N-terminal methyl group and a C-terminal acetyl group to

avoid terminal charges during the geometry optimization. A QM-geometry optimization at the HF/6-31G(d) level of theory was performed, which served as basis for the RESP fitting, using an in-house script to keep the charges of the capping groups fixed to the values from the ff14SB force field.

Simulation protocol

The procedure for minimization and heat up was derived from the method described in Duell *et al.*[263] During the minimization (XMIN method; ntmin=3), the box size was adjusted sequentially in steps of 0.02 g/cm^3 to bring the density from 0.8 g/cm^3 to 1.0 g/cm^3, using *sander* from the Amber/AmberTools software package. In each sequential minimization step, 3.0 kcal·mol^{-1}·Å^{-2} restraints were applied to all protein atoms. Once the target density was reached, one additional minimization step was performed without application of any restraints. In the heat up procedure, the system was equilibrated for 10 ps at 0 K and afterwards heated to 20 K following a *NVT* ensemble, applying the Langevin thermostat[257] with a collision frequency of 4.0 ps^{-1}. 3.0 kcal mol^{-1}·Å^{-2} restraints were applied to the position of all atoms in the following sequence: 0-5 K, 5-10 K, and 10-20 K, each for 50 ps. For the heat up to 200 K, the restraint was only applied to the backbone atoms, as well as potential metal ions and hydroxide ions. This heat up was performed in the following sequence: 20-50 K, 50-100 K, and 100-200 K in 50 ps, 100 ps and 100 ps, respectively. Afterwards, the system was equilibrated at 200 K for 200 ps without any positional restraints. Finally, the system was heated to 300 K in 400 ps and equilibrated for another 500 ps at the target temperature in a *NPT* ensemble, applying the Berendsen barostat[247] with a relaxation time of 1 ps and compressibility of 44.6 x 10^{-6} bar^{-1} to keep the pressure constant at 1 bar. During the heat up simulations, periodic boundary conditions were applied and the SHAKE algorithm[248] was used on all bonds involving hydrogen atoms. A nonbonded cut-off of 12 Å was used, the particle mesh Ewald method[264] was applied for the long-range electrostatics and an integration step of 1 fs was used. The heat-up and production simulations were conducted in three replicas with the *pmemd.cuda* MD engine from the Amber/AmberTools software package. The post-processing of the simulations strongly differs between all studies, and is accurately described in the respective publications.

2.2.4 Hybrid QM/MM simulations

For the QM/MM calculations described in this dissertation, the systems were prepared as follows. The proteins were protonated with the *tleap* module from the Amber/AmberTools software package. The protonation state of all QM-residues was visually inspected, especially for residues coordinating metal ions. Force field parameters were assigned as described in Chapter 2.2.3, and Amber topologies and input coordinate files were prepared. The definition of the QM-region is system-specific, and described in detail

in the respective publications. QM/MM-boundaries were only placed on non-polar bonds, mostly on the Cα-Cβ bond. If multiple consecutive residues were included in the QM-region, the atoms of these residues were also included, and the QM/MM-boundary was placed at the N-Cα and Cα-C' bonds instead. The QM/MM-boundary was described with an electrostatic embedding scheme, with link atoms placed at the boundary (hydrogen atoms), and the charges at the QM/MM-boundary were shifted away by a dipole on the recipient atom, as implemented in the *shift* scheme in ChemShell.

The QM-atoms were described with the TPSS meta-GGA functional[265], applying RI-J approximation[266] on a m4 multigrid[267], and Grimme's D3 dispersion correction[268]. The def2-SVP[267, 269] basis set was applied for all atoms except the metal ions, which were described with the def2-TZVP basis set[267, 270]. The SCF convergence criteria was set to 10^{-7} au.

QM/MM calculations were performed using the ChemShell suite.[271] The DL_POLY module implemented in ChemShell was used for the MM-calculations, and an interface with Turbomole[272] was applied for the QM-calculations. The geometry optimizations of all QM-atoms and the neighboring MM-atoms were performed using the DL-Find[273] optimizer.

3 RESULTS

3.1 ENZYMEMATCH: IDENTIFICATION OF ENZYMES CAPABLE OF CATALYZING TARGET REACTIONS USING INTERACTION PATTERN MATCHING

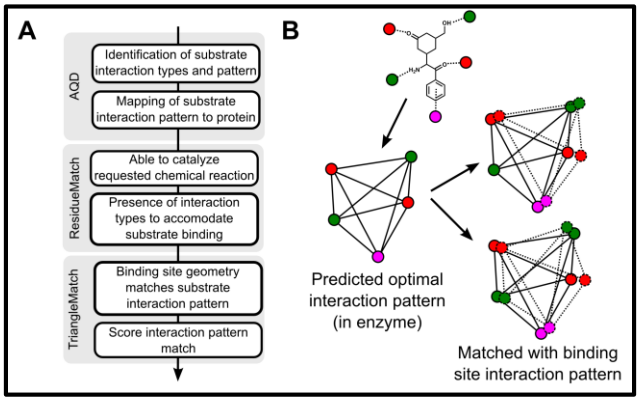
Enzymes gain increasing attention for their role as biocatalyst because they deliver tools for green chemistry, addressing global challenges toward the Bioeconomy. Enzymes can work at mild conditions and low temperatures, and form products with high chemo-, regio-, and stereoselectivity, making them valuable tools for the food and pharma industry.[11] Moreover, enzyme engineering has shown to be a valuable tool to modify certain properties of enzymes, including enzyme activity, selectivity and stability.[16, 274] However, in order to optimize an enzyme towards an industrially applicable biocatalyst, an enzyme with measurable desired activity needs to be available to function as starting point for engineering studies. Therefore, EnzymeMatch was developed in this dissertation: a bioinformatics algorithm to identify enzymes capable of catalyzing a target reaction using interaction pattern matching. First, an interaction pattern is automatically generated based on the structure of the target substrate in the "Automatic Query Design" (AQD) module, in which substrate flexibility is explicitly considered. Subsequently, the "ResidueMatch" module compares the available physico-chemical properties required to bind the target substrate to those found in characterized enzymes. This mode can further filter for a certain required chemical reaction that needs to be catalyzed. Finally, the "TriangleMatch" module structurally matches the interaction patterns found in the enzymes identified by ResidueMatch to the optimal interaction pattern of the substrate, applying a newly designed graph-theoretical interaction pattern matching approach. The required structures from potential enzymes are automatically retrieved from the Protein Data Bank, and annotation about the binding site is retrieved from the BioLiP database.[131] EnzymeMatch was extensively evaluated in both UNIX-based and Windows operating systems, and significant effort was spent on a user-friendly user interface. An extensive manual including several tutorials was written as well, describing all functionalities of this algorithm.

Okke Melse, Iris Antes and Volker Sieber devised the study, and Okke Melse with support from Iris Antes translated these ideas to an actual programmable algorithm. Woo Young Cho and Tongyan Wu supported in the programming and evaluation of EnzymeMatch. Okke Melse wrote the Applications Note with support from Ville R. I. Kaila and Volker Sieber.

EnzymeMatch: Identification of Enzymes Capable of Catalyzing Target Reactions using Interaction Pattern Matching

Okke Melse, Woo Young Cho, Tongyan Wu, Iris Antes, Ville R. I. Kaila, Volker Sieber

Submitted



Source code freely available from GitHub:

https://github.com/MelseO24/EnzymeMatch

distributed under the terms of the GNU General Public License.

3.2 DynaBiS: A Hierarchical Sampling Algorithm to Identify Flexible Binding Sites for Large Ligands and Peptides

Knowledge about the location of a ligand binding site is highly important in both drug discovery and protein engineering, and numerous computational algorithms rely on this information. Experimental elucidation of the binding site location is however still difficult. A large variety of binding site identification algorithms have been presented, but the majority relies on the rigid protein approximation. However, typically only an apo structure is available when a binding site needs to be predicted, in which the binding site is not adopted yet to the presence of a ligand. Therefore, this publication describes the development of DynaBiS: a binding site identification algorithm that explicitly accounts for both protein- and ligand flexibility, in order to find flexible binding sites for large ligands, such as peptides.

The DynaBiS algorithm consists of two steps: “surface screening” and “pocket sampling”. In the first, random ligand conformations are sampled around the protein surface, allowing a certain amount of protein-ligand overlap. After clustering and selection of the most promising binding sites, the pockets are further screened during the pocket sampling step. Here, Monte Carlo/Simulated Annealing (MC/SA) simulations are performed applying a soft-core potential, as inspired by the flexible docking algorithm DynaDock. During this step, the binding pocket is analyzed in more detail to investigate if the target ligand could bind in this respective pocket. At the end of the MC/SA simulations, the overlap is resolved to simulate the induced fit/conformational selection effect. The performance of DynaBiS was evaluated against a diverse evaluation set, consisting of both peptide- and small-ligand binding sites. Additionally, the apo structure was included for the majority of the evaluation systems as well to analyze how well protein flexibility is simulated. This evaluation showed that DynaBiS was able to identify all binding sites from the evaluation set as potential binding sites. Furthermore, DynaBiS predicted the correct binding site in the top-5 ranked binding site predictions for all but one system, and 19 out of 26 binding sites were predicted as the top-ranked binding site. With this outstanding performance, DynaBiS outperformed other commonly used binding site identification algorithms. The major improvement was observed in the identification of peptide binding sites using an apo structure as input, but still performing well with rigid small-ligand binding sites.

Iris Antes initially designed the project, and the initial coding was performed by Sabrina Hecht, and revised by Okke Melse, who also designed the evaluation, and performed all calculations with DynaBiS and the other binding site algorithms. The manuscript was written by Okke Melse and Iris Antes.

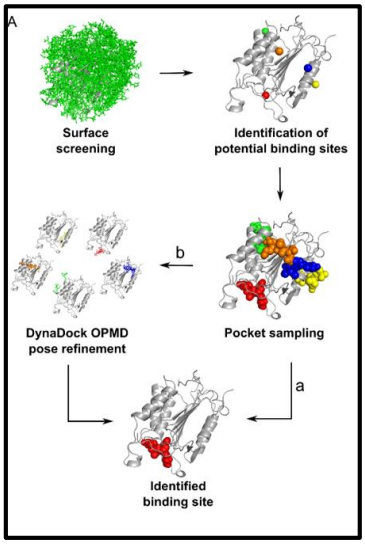
DynaBiS: A Hierarchical Sampling Algorithm to Identify Flexible Binding Sites for Large Ligands and Peptides

Okke Melse, Sabrina Hecht, Iris Antes

Proteins: Structure, Function, and Bioinformatics

2022, **90**(1):18-32

https://doi.org/10.1002/prot.26182



Received: 15 December 2020 | Revised: 24 June 2021 | Accepted: 11 July 2021
DOI: 10.1002/prot.26182

RESEARCH ARTICLE

PROTEINS WILEY

DynaBiS: A hierarchical sampling algorithm to identify flexible binding sites for large ligands and peptides

Okke Melse[1] | Sabrina Hecht[1,2] | Iris Antes[1]

[1]TUM Center for Functional Protein Assemblies and TUM School of Life Sciences, Technische Universität München, Freising, Germany

[2]Quattro Research, Planegg, Germany

Correspondence
Iris Antes, TUM Center for Functional Protein Assemblies and TUM School of Life Sciences, Technische Universität München, Emil-Erlenmeyer-Forum 8, 85354 Freising, Germany.
Email: antes@tum.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: SFB 1035, Project A10, SFB749, project C08, CIPSM

ABSTRACT

Knowing the ligand or peptide binding site in proteins is highly important to guide drug discovery, but experimental elucidation of the binding site is difficult. Therefore, various computational approaches have been developed to identify potential binding sites in protein structures. However, protein and ligand flexibility are often neglected in these methods due to efficiency considerations despite the recognition that protein–ligand interactions can be strongly affected by mutual structural adaptations. This is particularly true if the binding site is unknown, as the screening will typically be performed based on an unbound protein structure. Herein we present DynaBiS, a hierarchical sampling algorithm to identify flexible binding sites for a target ligand with explicit consideration of protein and ligand flexibility, inspired by our previously presented flexible docking algorithm DynaDock. DynaBiS applies soft-core potentials between the ligand and the protein, thereby allowing a certain protein–ligand overlap resulting in efficient sampling of conformational adaptation effects. We evaluated DynaBiS and other commonly used binding site identification algorithms against a diverse evaluation set consisting of 26 proteins featuring peptide as well as small ligand binding sites. We show that DynaBiS outperforms the other evaluated methods for the identification of protein binding sites for large and highly flexible ligands such as peptides, both with a holo or apo structure used as input.

KEYWORDS
algorithms, binding site identification, drug design, DynaDock, molecular docking, peptides, protein flexibility

1 | INTRODUCTION

The identification of protein binding sites is crucial for a variety of research fields, for example, protein engineering or drug discovery. It remains, however, difficult to retrieve this information experimentally, therefore multiple computational approaches have been proposed to fulfill this task.[1] Several strategies have been applied, starting from sequence-based approaches applying methods such as Multiple Sequence Alignment and Position Specific Scoring Matrices to detect hot spots in the sequence, and thereby predict the binding site.[2,3] Alternatively, a variety of structure-based approaches have been presented, aiming to identify ligand binding sites. A typical approach is scanning the protein surface with probe atoms to identify high-affinity sites, which can be further differentiated for their ligand preference by the use of different probe types. This technique is applied in the webservers Q-SiteFinder[4] and SITEHOUND.[5]

Other binding site identification methods, such as AutoLigand[6] and AutoSite,[7] apply affinity grids in a different way. AutoLigand aims to find the binding sites by first flooding cavities in the protein with grid points, after which the site is further explored via migration of lower-affinity points and ray-casting to identify potential new high-affinity pockets. AutoSite collects and clusters high-affinity points of different types obtained from AutoGrid[8] to identify potential binding sites. Another strategy for protein binding site identification is comparison of the target protein to structures with known binding sites, as performed in 3DLigandSite,[9] or mapping cavities large enough to accommodate a ligand, as applied in mkgridXf[10] and POVME.[11] Blind docking, in which docking simulations are performed without prior knowledge about the location of the binding site is another commonly used strategy. Hetényi et al. developed a blind docking approach using AutoDock, which they successfully applied to identify binding sites for a large set of proteins, including several in their apo state.[12,13] PEP-SiteFinder applies a similar blind docking approach in which pre-generated peptide conformations are docked in a rigid blind docking approach with the ATTRACT docking protocol and force field. The authors showed that while considering the 10 best poses, this method was able to match any of the protein residues interacting with the ligand for 90% of the binding sites from an evaluation set consisting of 41 systems.[14]

Most of these structure-based approaches are based on the rigid protein approximation and thereby neglect the protein flexibility. However, binding site dynamics, such as an induced fit upon binding, is a crucial effect strongly affecting ligand binding and affinity, as has been discussed frequently.[12,15-17] Accurate sampling of these effects is especially valuable if the binding site is unknown, because the structure of the protein will typically only be available as an apo structure. Conclusive and efficient modeling of this structural flexibility however remains a challenge for computational approaches. A straightforward approach to include some of the protein's flexibility in the predictions made, is the use of an ensemble of protein conformations, either retrieved from experiment or from sampling techniques such as molecular dynamics (MD) simulations. However, the availability of sufficient experimental structures is low, while the latter is computationally expensive, especially if large structural changes are to be expected. Furthermore, ligand induced structural changes are often not sampled extensively with this method.[18]

Here we present DynaBiS, a hierarchical sampling algorithm developed to identify (flexible) binding sites with explicit consideration of both protein and ligand flexibility. DynaBiS was developed to improve the identification of binding sites for large and highly flexible ligands such as peptides, based on protein structures in their apo state. The binding site identification is ligand-specific, meaning that only (potentially allosteric) binding sites are considered which might be able to accommodate the target ligand. In this way, binding sites, which are too small (e.g., metal ion binding sites) are not considered. The procedure to simulate protein and ligand flexibility applied in DynaBiS is based on our previously presented flexible docking algorithm DynaDock,[16] in which protein–ligand poses containing a defined amount of overlap are refined by soft-core potential-based MD simulations. During these simulations, DynaDock resolves the overlap by a specially designed procedure called optimized potential molecular dynamics (OPMD), which allows for efficient and accurate sampling of conformational adaptation effects (e.g., induced fit) of both the ligand and the protein. In DynaBiS, we adapted and optimized this concept for the identification of flexible ligand binding sites. This more efficient implementation of overlap refinement is justified, since DynaBiS was developed to identify binding sites in apo structures overcoming the rigid protein approximation by considering an intrinsic flexibility of the entire protein binding site during the search procedure, rather than to accurately reproduce the holo structure or conformational adaptation effects. In this way, the computation time of a typical DynaBiS simulation is in a similar range as an AutoDock blind docking simulation (Table S5). Furthermore, we analyzed the necessity of allowing protein–ligand overlap for the identification of such a binding site. We extensively evaluated the performance of the individual steps of the DynaBiS method using a diverse evaluation set consisting of both small ligand and peptide binding proteins, resolved either in the holo or apo state. Furthermore, we compared the DynaBiS performance to other commonly used binding site identification algorithms thereby illustrating the improved sampling capacity of the new DynaBiS algorithm. We show that DynaBiS is able to simulate all evaluated binding sites correctly and that the correct binding site was found in the top-5 ranked binding site predictions for all but one of the evaluation systems. While strongly improving the binding site identification in peptide-binding proteins, DynaBiS shows equal performance as existing approaches regarding small ligand binding systems.

2 | METHODS

2.1 | General algorithm

The DynaBiS binding site identification algorithm consists of two sampling steps, as illustrated in Figure 1(A). First, random sampling of potential ligand conformations covering the whole protein surface is performed, named "surface screening." During this step, a user-defined amount of overlap (in percent) between the ligand and the protein is allowed, providing a simple approximation to protein and ligand flexibility upon ligand binding. This results in the identification of several potential binding sites spatially distributed over the protein. These potential binding sites are further investigated in the subsequent step, named "pocket sampling." In contrast to the previously presented DynaDock approach, in this second step Monte Carlo/Simulated Annealing (MC/SA) simulations are performed instead of OPMD, as the former are less resource intense. Furthermore, because the aim of DynaBiS is solely the identification of the binding site, a more extensive and less accurate optimization procedure will suffice. During this step, a defined protein–ligand overlap is still allowed, but is resolved in the final binding sites through a subsequent geometry optimization with standard potentials. Since MC/SA optimizes the potential energy of the system, the simulation will be directed to the energetically favorable state,

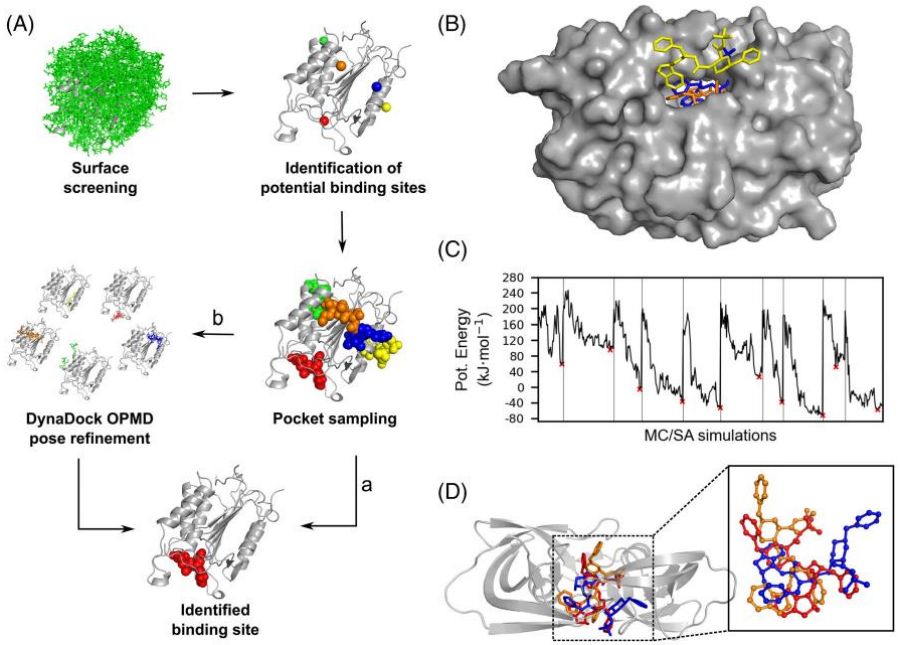


FIGURE 1 Overview of the DynaBiS procedure for binding site identification. In the flowchart (A), both the standard DynaBiS (a) and combined DynaBiS & DynaDock (b) pipelines are shown. A more detailed overview of these steps in the DynaBiS algorithm is shown in the Figure S2. In (B), the positions of the best pose after surface screening (yellow licorice), the top-ranked pose after pocket sampling (blue licorice), and the reference ligand (orange licorice) are shown in a surface representation of one of the evaluation systems (PDB-ID: 1hsh). The line plot (C) describes the potential energy optimization during 10 MC/SA simulations in the pocket sampling step: the vertical lines indicate the start of a new MC/SA simulation and the red crosses indicate the lowest energy pose. This low energy pose undergoes a geometry optimization with standard potentials to resolve the overlap and results in one of the final poses from the DynaBiS procedure. In (D), the DynaDock pose refinement from the combined DynaBiS and DynaDock approach is visualized in a cartoon representation for one of the evaluation systems (PDB-ID: 1hsh). The pose resulting from DynaBiS, the DynaDock optimized pose and the reference pose are shown as respective blue, red and orange licorice

that is, the local minimum of the ligand in the sampled binding site, and thereby improve the localization of the binding site. This is illustrated in Figure 1(B), where a surface screening result with a GC_{offset} of 13.12 Å was refined during pocket sampling to a GC_{offset} of 3.39 Å. All final poses from the MC/SA step can directly be scored and ranked using the "pepscore" scoring function. During both the surface screening and the MC/SA simulations in the pocket sampling, binding site flexibility is approximated by allowing protein–ligand overlap. The binding site (see below for definition) is allowed to move only during the final energy minimization to resolve the overlap. Alternatively, an additional DynaDock molecular docking simulation can be performed in the predicted binding site, which leads to more accurate docking poses through a more intense sampling of each identified binding site. A detailed schematic overview of the individual steps of DynaBiS is provided in the Figure S2.

Since DynaBiS is implemented in the "in-house" program DynaCell, the source code is not yet publicly available, but we plan to distribute the code in the near future. Until publication of DynaCell, we would be glad to share the compiled code to other computational research groups on request.

2.2 | Soft-core potentials

A soft-core potential-based description of overlapping atoms can improve the sampling of conformational adaptation considering both ligand and protein flexibility considerably.[16] To this end, we apply soft-core potentials for the nonbonded interactions between ligand and protein in DynaBiS. This allows us to perform the sampling with a certain amount of protein–ligand overlap during both the surface screening as well as the pocket sampling steps. The implementation of the soft-core potentials in the "in-house" program DynaCell is based on the proposed mathematical form of Taylor et al.[19]:

$$V_{nb}(r) = V_{LJ}(r) + V_{Coul}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{\left(\alpha^m\sigma_{ij}^6 + r_{ij}^6\right)^2}\right) - \left(\frac{\sigma_{ij}^6}{\left(\alpha^m\sigma_{ij}^6 + r_{ij}^6\right)^2}\right)\right] + \frac{(1-\alpha)^n q_i q_j}{4\pi\varepsilon_0\sqrt{\alpha + r_{ij}^2}} \quad (1)$$

The soft-core scaling parameter α varies from 0 to 1, with the value $\alpha = 0$ returning the unmodified nonbonded interactions while

$\alpha = 1$ scales these potentials down to nearly 0. Therefore, this parameter can be used to scale the nonbonded interactions such that the Van der Waals radii of two atoms (e.g., protein and ligand atoms) can overlap without the corresponding interaction energy reaching infinite values. The additional scaling exponents *m* and *n* allow further modification of the softening character of the nonbonded potentials and were set to $m = 3$ and $n = 6$ in this study. In the DynaDock flexible docking method, this overlap is gradually removed during an OPMD simulation, in which the α scaling parameter is optimized to finally reach zero (i.e., normal potentials) and the induced fit of the protein is modeled explicitly.[16] In the DynaBiS approach, a less resource intensive geometry optimization is applied instead to reduce this overlap (after pocket sampling), since a large number of structures need to be refined in order to analyze multiple binding sites, as described in more detail below.

2.3 | Surface screening

The first step in the DynaBiS procedure consists of a screening of the whole protein surface to identify potential binding sites. In this step, random ligand conformations are generated by applying random translations and rotations of the entire ligand, as well as rotations over a random selection of rotatable bonds. Poses with a geometric center farther from the protein surface than a user-defined distance (default 5 Å) are rejected. To ensure that only physically relevant and high-affinity poses are considered, any newly generated pose has to fulfill the following three acceptance criteria: (1) the overlap between the Van der Waals radii of any protein–ligand atom pair should be below a user-defined cutoff, (2) the distance between the geometric center of the ligand and the protein surface should be below a user-defined cutoff, and (3) the non-soft-core coulomb energy component needs to be below a user-defined value. These and all other specific settings applied in this study are listed in the Table S1. The geometry of each accepted pose is optimized using a steepest descent energy minimization, applying soft-core potentials for the nonbonded energies, by default with $\alpha = 0.5$, during which only the ligand is treated flexible. This process is pursued until a user-defined number of poses are accepted.

The minimized structures are ranked according to their potential energy to extract a certain percentage of the top-scoring poses. In addition, all top-scoring poses are filtered, such that only poses with a distance from each other larger than a user defined value are chosen as starting conformations for the subsequent pocket sampling step. This filter is included to avoid selecting multiple poses located in the same potential binding site for the pocket sampling step and thus to ensure subsequent coverage of all energetically favorable binding pockets.

2.4 | Pocket sampling

The identified potential binding sites are analyzed in more detail during the successive pocket sampling step. This step consists of a user-defined number of independent MC/SA simulations for each potential binding site identified during surface screening. Each individual MC/SA simulation will result in one optimized pose, which identifies the potential binding site. This pose should only be considered as a place mark for the binding site. To retrieve more accurate information about the binding mode (docking poses), one should conduct subsequent docking (e.g., DynaDock) calculations. The MC/SA algorithm itself has been documented extensively.[20-22] Our implementation of the MC/SA scheme contains 9 MC temperature cycles with decreasing temperatures, in which the initial MC temperature cycle is performed at 1000 K and each successive MC temperature cycle at 85% of the previous temperature. The ninth and last MC temperature cycle is performed at 300 K. During each MC temperature cycle, the ligand conformation is modified randomly, either by translation or rotation of the entire ligand or by rotation of a rotatable bond within the ligand. Translation steps are performed at most every 10th step to promote bond rotation modifications over translation. If the energy (containing the soft-core nonbonded potential) of the newly generated conformation is lower than the previous conformation, the modification is automatically accepted, otherwise the acceptance or rejection is determined by the Metropolis criteria.[23] Additionally, the overlap between the Van der Waals radii of any protein–ligand atom pair needs to be below a user-defined value (see Table S1) for the new conformation to be accepted, to avoid sampling of unphysical ligand poses buried inside the protein. Furthermore, ligand displacement during the MC temperature cycle is restricted to ensure local sampling. Thus a newly generated conformation is only accepted if its geometric center is located within a user-defined cutoff from the initial starting conformation of the MC/SA simulation. This cutoff value is decreased by 10% for every MC temperature cycle. The protein is kept fixed during these MC/SA simulations, but binding site flexibility is approximated via the allowed protein–ligand overlap. The first MC temperature cycle is terminated once 200 conformations are accepted, or when 3000 consecutive conformations are rejected. For the following MC temperature cycles, the number of accepted conformations required to terminate each MC temperature cycle is scaled down linearly from 200 conformations for the first temperature cycle to 50 conformations for the 9th MC temperature cycle. The last MC temperature cycle can also be terminated once convergence is reached, that is, if the potential energy of 15 consecutive steps differ less than 10 kJ/mol. The specific MC settings used in this study result from a rough optimization aiming to retrieve diverse poses (data not shown). The conformation with the lowest potential energy of the current MC temperature cycle is used as input conformation for the next MC temperature cycle.

After each MC/SA simulation (each containing the 9 MC temperature cycles as described above), the frame with the lowest potential energy simulated during the MC/SA simulation (Figure 1(C)), undergoes a geometry optimization. In this geometry optimization, both the ligand and the binding site (defined as all residues within 6.5 Å of the ligand) are optimized, thus both the protein and ligand are allowed to move. This simulation is performed without application of soft-cores, thus $\alpha = 0$, in order to resolve any overlap between the

ligand and the binding site. The resulting, optimized protein structure is used as input structure for the next MC/SA simulation.

2.5 | Scoring and ranking

Finally, either a DynaDock simulation can be first performed to additionally retrieve accurate docking poses and simulate larger protein–ligand conformational adaptation effects (Figure 1(A), pathway b), or all binding sites are directly scored (Figure 1(A), pathway a). Scoring was performed with the pepscore scoring function according to Equation (2), which is described in more detail in Antes (2010).[16]

$$E_{pep} = a_1(E_{vdW,P-L} + E_{vdw,L}) + a_2(E_{Coul,P-L} + E_{Coul,L}) + a_3(E_{dih,L}) + a_4(E_{impr,L}) + a_5(E_{1-4vdW,L}) + a_6(E_{1-4Coul,L}) \quad (2)$$

Briefly, the pepscore consists of scaled intramolecular interactions of the ligand (dihedral, improper, and nonbonded 1–4 interaction terms) and the nonbonded interactions between the ligand and the protein. The pepscore was originally developed without a GB solvent term, as it was shown before that this did not improve its scoring ability. The following parameter set was used: $a_1 = 0.87459$, $a_2 = 0.013073$, $a_3 = 0.052852$, $a_4 = 0.39477049$, $a_5 = 0.329739$, and $a_6 = 0.13516631$.

2.6 | Simulation conditions and evaluation specific settings

The performance of DynaBiS, as well as the performance of the individual steps within DynaBiS was evaluated on a diverse evaluation set, and compared with the performance of other binding site identification algorithms. Furthermore, the potential of combining the binding site prediction (performed by DynaBiS) with DynaDock post-refinement simulations in a single pipeline was investigated.

2.7 | Evaluation set

To validate the DynaBiS procedure, a diverse evaluation set was constructed consisting of protein–ligand systems with deep buried binding sites as well as surface-exposed binding sites. In addition, the systems were selected to span a wide range of ligand sizes, balanced over small ligand and peptide systems. We did not add systems from which it was immediately clear that binding site identification algorithms relying on the rigid protein approximation (AutoSite and AutoDock) would fail (e.g., movement of entire protein domains), in order to allow a fair performance comparison. In addition, the aim of DynaBiS is namely not to predict the conformational changes between apo and holo structures, but solely to identify the binding sites using apo structures as input. The evaluation set contains 8 small ligand and 7 peptide protein–ligand complexes (Table 1). For 11 systems, the apo structures were added to the evaluation set as well, to enable performance validation of the binding site identification methods in their ability to find and identify binding sites in structures which were crystallized in absence of a ligand.

2.8 | System preparation

All calculations are based on the PDB structures listed in Table 1. First, all water molecules were removed and missing heavy atoms or non-resolved residues were built using IRECS.[24] Afterwards, the systems were protonated for pH value of 7.4 and the parameters for the small ligands were retrieved from the PRODRG server.[25] The GROMOS43A1 force field[26] parameters were used for all proteins and peptides. The geometry of the structures was optimized using DynaCell applying the steepest descent method with the AGBNP implicit solvent model,[27] and no cutoff for nonbonded interactions. The resulting structure was used as starting structure for the DynaBiS simulations.

In contrast to DynaBiS, DynaDock requires protein structures equilibrated at a certain temperature. Therefore, the ligand-free structures (thus the ligand was removed from the holo structures) were solvated in a SPC water box applying a 1.0 nm buffer region, and counterions were added to neutralize the system. Each system was minimized with GROMACS (http://www.gromacs.org, RRID: SCR_014565) version 4.5.6.[28] using the steepest descent algorithm applying periodic boundary conditions, particle mesh Ewald[29] for the long-range electrostatics and a 1 nm cutoff for the nonbonded interactions until the maximum force in the system was below 1000 kJ/mol/nm. For heat-up, the system was heated in each 50 ps simulations in the following sequence: 0–5 K, 5–10 K, and 10–20 K applying 1000 kJ/mol/nm^2 restraints on all atom positions followed by 50 ps simulations of 20–50 K, 50–100 K, and 100–200 K applying 1000 kJ/mol/nm^2 restraints to the protein backbone only. At 200 K, the system was equilibrated for 150 ps, followed by heating to 300 K in 100 ps, applying 100 kJ/mol/nm^2 positional restraints on the protein backbone atoms. Finally, the system was equilibrated at 300 K for 50 ps without any positional restraints. The heat-up simulations were performed in a NVT ensemble using particle mesh Ewald for the long-range electrostatics and a 1.4 nm cutoff for the nonbonded interactions. A time step of 1 fs was used and the pair-list was updated every 10th integration step. The Berendsen thermostat[30] was applied for temperature coupling using a coupling time constant of 0.1 ps and the bonds were constrained to their equilibrium bond length using the LINCS algorithm.[31] All water molecules and counterions were removed before application to DynaDock simulations.

2.9 | Evaluation specific DynaBiS settings

In this study, the surface screening was performed with $\alpha = 0.5$. Poses were only accepted if (1) the overlap between the van der Waals radii of any protein–ligand atom pair did not exceed a maximum of 35%,

TABLE 1 Evaluation set

PDB ID[a]	Annotation	Ligand name/peptide sequence	Nr. rotatable bonds	Ligand diameter (Å)
Small ligands				
2fm2	HCV-NS3 protease	Org. lig (3 BC)	16	19.7
1hsh (1hsi)	HIV-II protease	Org. lig (MK1)	14	18.1
1bju (1s0q)	Trypsin	Org. lig (GP6)	4	14.4
1efy (2paw)	Poly (ADP-ribose) polymerase	Org. lig (BZC)	3	11.0
1stp (1swb)	Streptavidin	Biotin (BTN)	5	8.3
1j8a (1s0q)	Trypsin	Benzamidine (BEN)	2	7.1
2mcp	Immunoglobulin	Phosphocholine (PC)	3	6.2
1ldm	Lactate dehydrogenase	Oxamic acid (OXM)	1	4.4
Peptides				
1io6 (1gfd)	Grb2[b] SH3 domain	RHYRPLPPLP	33	31.6
1be9 (1bfe)	PSD-95[c] PDZ domain	KQTSV	23	19.9
1awq (2cpl)	Peptidyl-propyl cis-trans isomerase A	HAGPIA	15	15.6
1pau (1qx3)	Caspase-3	DEVD	16	13.7
1a30	HIV-1 protease	EDL	13	10.4
8tln (1fjq)	Thermolysin	VK	10	9.3
2cyh (2cpl)	Peptidyl-propyl cis-trans isomerase A	AP	3	6.7

Note: All ligand structures of the systems in the evaluation set are shown in the Figure S1.
[a]The PDB-ID between brackets indicate the PDB of the apo system.
[b]Growth factor receptor-bound protein 2.
[c]Postsynaptic density protein 95.

(2) the distance between the geometric center of the ligand and the protein surface was below 0.5 nm, and (3) the coulomb term of the protein–ligand interactions for the sampled ligand was below a system-specific value (exact values are provided in the Table S1). These settings result either from optimization (see Section 3), or from in-house experience with DynaDock. The energy threshold (condition 3) was generally set to 0 kJ/mol, and increased to 500 kJ/mol if the lower limit did not result in any poses. For the evaluation of the effect of protein–ligand overlap on the surface screening results, 10 000 poses were generated for each investigated protein–ligand overlap value. For the final performance evaluation and comparison to other binding site identification methods, 1000 poses were generated during the surface screening. To obtain a set of optimally distributed poses, from the top-ranked 2% of the sampled poses, only those were collected with a distance greater than or equal to 1.5 nm between their geometric centers. These poses formed the starting structure set for the pocket sampling step, and thus practically serve as the center of the investigated binding pocket.

Pocket sampling was performed with $\alpha = 0.9$, which was set so high to avoid high penalization of overlapping atoms during the simulation. A maximum allowed protein–ligand overlap was defined system dependent, either 30% or 40% (system specific settings are provided in the Table S1), and the sampling was restricted to 0.2 nm around the starting pose. Each pocket sampling simulation contained 50 MC/SA simulations for the analysis of different protein–ligand overlap values and 10 MC/SA simulations for the final performance comparison to the other methods.

To analyze the performance of DynaBiS, a DynaBiS procedure as described above was performed 10 times initiating from different random seeds, after which the resulting poses were merged. To retrieve spatially different poses, all resulting protein–ligand complexes were clustered with R[32,33] using the Agglomerative Nesting (AGNES) method with average linking, with the cluster tree being cut at the second level. Each resulting cluster was scored and ranked by the average pepscore of all poses inside each cluster and the pose with the best pepscore within each cluster was selected as cluster representative. The order of the final poses results from the order of the ranked clusters, thus the top-ranked pose is the representative pose of the top-ranked cluster, and the second ranked pose is the representative pose of the second ranked cluster, and so on.

2.10 | DynaDock post-refinement

We analyzed if the binding site identification by DynaBiS could be combined in a single pipeline with molecular docking simulations in the high-ranked binding sites using DynaDock. DynaDock uses the same soft-core potentials (Equation (1)) to simulate conformational adaptation effects upon ligand binding.[16] The additional DynaDock docking was performed in the five top-ranked binding sites retrieved from the cluster analysis. During the DynaDock broad sampling, 200 random ligand conformations were generated in the equilibrated protein structures. For peptide and small molecule ligands, a maximum protein–ligand overlap of respective 80% and 60% was allowed.

Furthermore, only 40% and 75% of the ligand atoms were allowed to have any intra-ligand overlap for the systems binding peptides and small molecule ligands, respectively. To limit the sampling to the specific binding site, the broad sampling was restricted to a sphere with an 8 Å radius around the geometric center of the initial ligand conformation.

The introduced overlap was subsequently resolved by an OPMD simulation, during which the soft-core parameter α is optimized with respect to the system's potential energy, using a steepest descent energy minimization at each MD step. The simulations were initialized with maximum soft-score nonbonded interactions ($\alpha = 1$) and after the scaling parameter was optimized to $\alpha = 0$ (minimum soft-score), the simulations were continued for 500 ps with standard force field potentials to equilibrate the obtained protein–ligand complex.

From the resulting ligand conformations, only the poses which geometric center is within 50% of the ligand diameter from the starting pose were further analyzed, to ensure that only poses within the investigated binding site were considered.

2.11 | AutoDock and AutoSite simulations

The AutoDock blind docking simulations were performed according to the Hetényi approach[13] with AutoDock release 4.2.6.[8] All water molecules and ions were removed from the experimental structures. A grid consisting of 126 grid points in each dimension was placed around the entire protein with a spacing of 0.55 Å centered on the geometric center of the system. Two hundred and fifty Lamarckian Genetic Algorithm runs were performed, using a population size of 250, mutation rate 0.02, crossover rate 0.8, and up to 1×10^7 energy evaluations. The docked conformations were clustered with a RMSD tolerance of 2.0 Å. The rotatable bonds were determined using AutoDockTools, with the exception of the bond between the phenyl-group and the hydroxide ion of the phenylalanine in the system with PDB-ID 1io6, which is treated as rigid to stay within the maximum of 32 rotatable bonds allowed in AutoDock.

The AutoSite[7] calculations were performed with AutoSite version 2.0.3 applying the default setup. Each binding site predicted by AutoSite (called "fill") was used as a starting point for subsequent AutoDock docking simulations for the combined AutoSite and AutoDock approach. In the combination approach, each AutoDock simulation was performed in a grid consisting of 60 points in each dimension, separated with a spacing of 0.375 Å, centered on the subsequent fill's geometric center. The docking consisted of 70 Lamarckian Genetic Algorithm runs with a population size of 150, crossover rate 0.8, mutation rate 0.02, and 2.5×10^6 energy evaluations.

2.12 | Performance analysis

To analyze the performance of the different methods in identifying the native binding sites, the absolute distance between the geometric center of the docked pose, or fill in case of AutoSite, and the reference was measured, denoted as GC_{offset}:

$$GC_{offset} = \left| GC_{docked\ pose} - GC_{reference} \right| \quad (3)$$

where, GC represents the geometric center. This measurement was used because solely the translation of the pose is of importance for the identification of the binding site, and not the actual conformation of the ligand inside the pocket. We apply a system-dependent threshold to determine whether a binding site was identified correctly. To account for variations in binding site volume, this GC_{offset} threshold for a binding site to be successfully identified was defined as half the ligand diameter, which was calculated as follows:

$$Cutoff = \frac{\sum_{n=1}^{N} \max\left(\left| x_{i,n} - x_{j,n} \right| \right)}{2N} \quad (4)$$

where, N represents the number of poses generated during the surface screening, and i and j represent heavy-atoms within the ligand. The ligand diameter was measured as the average of the maximum distance between any atom pair in the ligand sampled during the random conformational sampling in the surface screening step. This ligand-size dependent threshold is applied because of the large variety of ligands in the evaluation set: a slightly translated long peptide for example is still assumed to be in the same binding site, while for a small ligand, a translation of the same magnitude moves the ligand to an adjacent (potentially incorrect) binding site. Additionally, since this threshold can be applied on all evaluated methods, the performance of these methods can be accurately compared and assessed.

3 | RESULTS AND DISCUSSION

We developed DynaBiS to account for the flexibility of both protein and ligand during the binding site identification process, thereby approximating conformational adaptation effects. The rigid protein approximation has been proposed to be one of the most important reasons for the often inadequate prediction of ligand binding sites, especially for peptides and other bulky ligands.[34-36] In the DynaBiS binding site identification algorithm we present here, the protein surface is scanned by generation of random ligand poses while allowing a certain amount of protein–ligand overlap, which are further investigated by MC/SA-based pocket sampling applying soft-core potentials (Figure 1), as inspired by our previously presented molecular docking algorithm DynaDock.[16]

3.1 | Effect of protein–ligand overlap on the DynaBiS procedure

The main difference between the DynaBiS procedure and other binding site identification methods results from allowing a certain amount

of protein–ligand overlap for an efficient approximation of conformational adaptation effects. The amount of protein–ligand overlap allowed in these simulations can have a considerable effect on the binding site identification performance. For example, a too large protein–ligand overlap could lead to identification of non-existing pockets, while an overlap which is too small will prohibit the successful identification of flexible binding sites for which conformational adaptation effects are important for pocket formation upon ligand binding. To define optimal default values for the allowed protein–ligand overlap in DynaBiS, we systematically investigated the effect of the applied protein–ligand overlap values on the individual steps in the DynaBiS procedure: surface screening and pocket sampling. This effect was investigated in eight structures of the evaluation set, containing four holo structures and four apo structures.

We observed that the likeliness of sampling a binding site close to the reference (i.e., observed binding site in the crystal structure) increases strongly if a protein–ligand overlap greater than 30% is allowed (Figure 2(A)). As generally a GC_{offset} (see Section 2 for definition) smaller than 3–4 Å means that the ligand is located in the correct binding site, the data show that if the overlap value is set to at least 35%, the final set of binding sites obtained by the surface screening step contains the correct binding site for all systems. From the distribution of the sampled binding sites for the different overlap values (Figure 2(B)), it can be observed that in all cases the sampled binding sites are equally well distributed over the whole protein surface with a slightly smoother distribution and decreased median GC_{offset} of the sampled binding sites for high overlap values (all *p*-values from two-sided *T*-tests were below .01 for all overlap pairs). This shows that while for high overlap values the binding site identification probability increases, the protein surface is still sampled throughout. This is illustrated for one system (PDB-ID: 1be9) in Figure 2(C), which additionally shows a small tendency of the accepted surface screening poses towards loop regions, especially for the low-overlap surface screening simulations. Based on these results we concluded that 35% is generally a good compromise between proper sampling and efficiency, that is, avoiding sampling of too many non-existing binding sites.

Initial control pocket sampling simulations in the known binding site from the crystal structure were performed to investigate the ability of the pocket sampling step to investigate the binding site. All simulations resulted in at least one binding site with a GC_{offset} below 2 Å and average GC_{offset} of all resulting binding sites between 2 and 7 Å (Table S4). This illustrates the thorough sampling and the ability to comprehensively sample the reference binding site in the pocket sampling simulations. Next, we systematically investigated the effect of different protein–ligand overlap values on the performance of the pocket sampling step, independently of the surface screening results. For this analysis, we performed 50 independent pocket sampling simulations for each system. All pocket sampling simulations started in the same binding site, namely the best-sampled binding site (i.e., lowest GC_{offset}) from the respective system, obtained from the surface screening allowing a maximum of 35% overlap. These pocket sampling simulations were performed for a large range of maximum allowed protein–ligand overlap values (Figure 3). To further study if

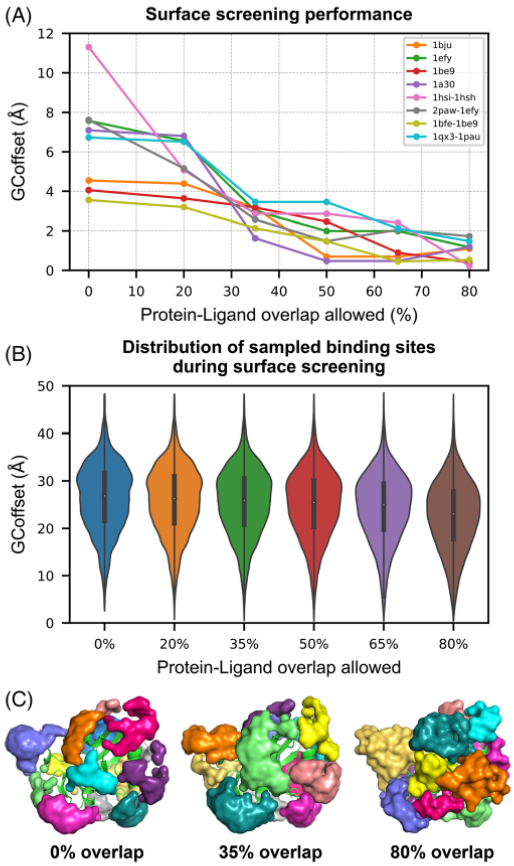


FIGURE 2 The effect of maximum allowed protein–ligand overlap on the performance of surface screening. Surface screening performance is shown as the lowest GC_{offset} (for definition see Methods section) observed for a surface screening of 10 000 poses as a function of the maximum allowed protein–ligand overlap (A). The distribution of the GC_{offset} values of all these generated poses is shown in a violin plot (B). The plotted values result from all poses of all the systems in the subset of the evaluation set. The distributions for the individual systems are provided in the Figure S5. In (C), the distribution of the geometric centers of the poses sampled during the surface screening is visualized as a surface around one of the evaluation systems (PDB-ID: 1be9), in which the different colors represent different clusters, resulting from an average-linkage hierarchical clustering of these generated poses

the pocket sampling is also able to identify a binding site starting from a ligand pose which requires a longer optimization path to find the global minimum in the binding site, the same pocket sampling evaluation was performed starting from a pose with a large GC_{offset} between 5 and 8 Å as well (Figure 3, dashed lines, and Table 2). In addition, as the results differed considerably depending on the type of protein

structure used (either apo- or holo-), we analyzed this step for apo-protein and holo-protein structures separately.

If pocket sampling was performed based on the holo protein-structures, the best results were obtained with a 20% overlap (all GC_{offset} values below 2 Å), even if the pocket sampling was started from a non-optimal pose with a GC_{offset} between 5 and 8 Å (Figure 3(A), dashed lines). Binding site prediction in a holo structure does however not resemble a realistic example, since the protein binding site is already adapted to the presence of the ligand, hence the rigid protein approximation is always valid. Thus, we performed this analysis based on the apo protein-structures as well. Pocket sampling allowing a maximum of 20% protein–ligand overlap in apo structures resulted in much higher GC_{offset} values than observed for the holo structures, as shown in Figure 3(B). For two cases, the initial binding site could not be sampled with the pocket sampling allowing 20% overlap at all, probably due to too large initial overlap of the starting pose. For the apo structures, the best results were obtained for maximum overlap values of 30% and 40%, the latter featuring the best-sampled average GC_{offset} of 1.6 Å (Table 2). In this context it is especially noteworthy that by applying an overlap of 40%, very good GC_{offset} values could even be obtained from the starting pose set with the high GC_{offset} between 5 and 8 Å, as all final values are smaller than 3 Å (Table 2). This illustrates the ability of the pocket sampling step to significantly improve the localization of the binding site after surface screening, even if the initial pose is not optimally placed. Noteworthy are the large RMSD values observed for the poses with the lowest GC_{offset} values (Table 2), that is, the pose which indicates the location of the binding site. This is most likely due to the extensive sampling approach in the pocket sampling step in DynaBiS, which aim is to investigate the binding pocket rather than to reproduce a correct binding mode of the ligand. Since DynaBiS is able to identify most binding sites correctly (as described in more detail below), a low-RMSD pose apparently already scores higher compared with poses in the incorrect binding site, indicating that a binding site identification algorithm does not need to be able to simulate a low-RMSD pose in order to identify the binding site.

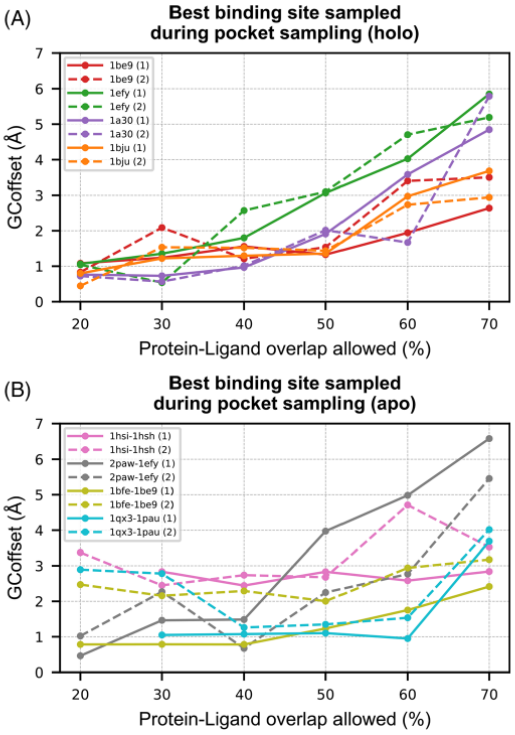


FIGURE 3 The effect of maximum allowed protein–ligand overlap on the performance of pocket sampling, studied in holo (A) and apo (B) structures. The lowest observed GC_{offset} from 50 pocket sampling simulations is plotted against the applied maximum allowed protein–ligand overlap. The pocket sampling simulations started from a pose with the lowest GC_{offset} from the 35% overlap surface screening or from a pose with a GC_{offset} between 5 and 8 Å (Table 2), represented as solid and dashed lines respectively

Finally, we observed that pocket sampling allowing only low protein–ligand overlap supports local sampling, while pocket sampling allowing a large protein–ligand overlap naturally scans a larger area around the binding pocket. This effect is mainly caused by a higher acceptance rate for translation moves of the ligand in the investigated pocket for high overlap values. On the other hand, allowing too much protein–ligand overlap results in buried poses, which lead to a drifting off of the ligand (see high GC_{offset} values in Figure 3(B) for 70% overlap) and the sampling of non-existing pockets, as illustrated for one system (PDB-ID: 1a30) with 70% overlap in Figure S6.

3.2 | Performance evaluation and method comparison

To evaluate the overall performance of DynaBiS, we compared the DynaBiS binding site predictions with several other binding site identification methods. The evaluation was performed on a diverse evaluation set containing 8 small ligand and 7 peptide systems. For 11 of the systems, the evaluation set contained both the holo and the apo structure. We compared the DynaBiS performance with a blind docking approach performed with AutoDock, as described by Hetényi et al.[13] and with the predictions obtained by AutoSite,[7] which uses probe atoms of different types to create three-dimensional volumes, named fills, in the protein.

As shown in Figure 4(A), in which the number of correctly identified binding sites on the basis of the top-ranked pose is provided for the evaluation set, DynaBiS outperforms the other evaluated methods in terms of the number of correctly predicted binding sites. Regarding small ligand binding sites, both DynaBiS and AutoDock performed equally strong in the small ligand—apo category, with AutoDock performing slightly better in the small ligand—holo setting (7 vs. 5). Especially for the identification of peptide binding sites in apo protein-structures (peptide-apo category), DynaBiS outperformed the AutoDock based methods as four binding sites were correctly identified by DynaBiS vs. only one by AutoDock blind docking (Table 3).

TABLE 2 Evaluation of DynaBiS performance

Surface screening (35% max. overlap)						Pocket sampling (40% max. overlap)			
						Best sampled		Average	
System	GC_{offset}	RMSD	Max OL (%)	Avg OL (%)	% atoms OL	GC_{offset}	RMSD	GC_{offset}	RMSD
1be9	3.18	3.98	34.47	3.06	15	1.56	1.69	4.18	5.79
	5.64	8.41	30.10	3.27	13	1.21	5.18	4.94	8.38
1efy	3.01	5.41	32.50	12.08	40	1.80	7.29	6.57	8.76
	5.06	7.41	29.14	11.62	47	2.57	5.02	7.32	9.42
1a30	1.63	5.71	27.82	4.81	32	0.97	5.98	5.24	7.64
	5.99	7.60	31.45	5.88	35	1.02	6.65	7.47	5.63
1bju	3.16	6.97	33.23	10.45	59	1.30	8.87	4.32	9.03
	5.95	7.58	11.20	0.58	9	1.52	5.83	6.15	8.86
1bfe–1be9	2.13	11.97	32.44	6.37	25	0.78	9.45	3.56	9.61
	6.01	11.57	14.17	0.34	6	2.29	3.31	5.94	9.99
1qx3–1pau	3.46	8.97	34.52	12.04	41	1.08	6.36	3.84	7.49
	7.52	9.00	29.77	4.36	18	1.26	4.73	5.78	8.32
1hsi–1hsh	2.86	11.40	27.47	8.12	30	2.44	8.64	5.88	10.80
	6.80	10.32	34.77	6.22	32	2.74	9.03	6.12	10.01
2paw–1efy	2.57	5.36	28.78	4.91	30	1.49	3.53	6.48	8.36
	7.13	7.82	13.43	0.60	3	0.67	7.02	7.40	9.43
Average (1)[a]	2.75	7.47	31.40	7.73	34	1.60	6.25	5.36	8.26
Average (2)[b]	6.26	8.71	24.25	4.11	20	1.66	5.84	6.39	8.75
Overall average	4.51	8.09	27.83	5.92	27	1.63	6.04	5.87	8.51

Note: For each system, two poses resulting from the surface screening performed with a maximum protein–ligand overlap of 35% are provided. These poses were used as starting poses for the respective pocket sampling step. The first entry for each system represents the sampled pose from the surface screening with the lowest GC_{offset}, while the second is a pose with a GC_{offset} between 5 and 8 Å. The second entry is provided to evaluate the pocket sampling ability to optimize less optimal poses in the correct binding site. The protein–ligand overlap (OL) columns represent the maximum overlap of the Van der Waals radii between any protein–ligand atom pair of the respective pose. The pocket sampling column shows the GC_{offset} and heavy-atom RMSD values (with respect to the crystal structure after superposition of the protein) of the best-sampled pose and the average of the resulting poses from 50 pocket sampling simulations, initiated from the respective surface screening pose.

[a]Average of all best-sampled poses from the surface screening (i.e., first entry of each system).

[b]Average of all poses with a GC_{offset} between 5 and 8 Å from the surface screening (i.e., second entry of each system).

Next, we analyzed if the correct binding site was present in the top-5 ranked binding sites (Figure 4(B)), as this is an evaluation setup generally used to analyze enrichment effects in the top-ranked poses. With DynaBiS, the correct binding site was found in the top-5 ranked binding sites for all but one system (PDB-ID: 2mcp) in the small ligand—holo category (Table 3 and Figure S7. Nevertheless, this system was still sampled considerably closer to the reference binding site by DynaBiS compared with AutoDock blind docking (GC_{offset} = 4.02 and 21.90 Å, respectively), but the final pose was still located outside the 50% ligand diameter criterion used in this study to define a binding site as identified. Overall, with AutoDock and AutoSite, the correct binding site was included in the top-5 ranked binding site predictions only for respective 21 and 17 out of 26 systems. The difference in performance is especially striking for the protein–peptide complexes, as the DynaBiS results list the correct binding site in the top-5 ranked binding site predictions for all systems, whereas looking at the AutoDock blind docking results, the correct binding site is only contained in the top-5 ranked binding site predictions for 9 out of 13 peptide systems, and only for 3 out of 6 in the peptide-apo category. This illustrates that the strong performance of DynaBiS is mainly due to the superior identification of peptide binding sites, especially if apo protein structures are used. For small ligands, DynaBiS showed the same performance as AutoDock. This demonstrates that DynaBiS is especially suited for the prediction of peptide binding sites in apo protein structures for which the consideration of protein and ligand flexibility during sampling is especially important, while still performing as well as AutoDock for the prediction of small ligand binding sites.

In our evaluation set, AutoSite was outperformed by both DynaBiS and AutoDock blind docking. We could show that this is mainly due to insufficient sampling of the peptide systems (Figure 4 (C)), as well as due to inaccurate scoring especially of the small ligand–holo systems, as discussed in more detail below. However, it needs to be noted here that AutoSite is much more efficient compared with the AutoDock blind docking approach, since AutoDock blind docking

was performed with exceptionally high docking parameters in this study (Table S5).

3.3 | Role of sampling and scoring in binding site prediction

To investigate if the incorrect prediction of certain binding sites is mainly due to inaccurate sampling or scoring, we also listed the best sampled binding site (i.e., binding site with lowest GC_{offset} from all sampled binding sites) in Figure 4(C) and Figure 5 to allow comparison with the best scored binding sites. It can be observed that the correct binding sites are found in all of the 26 evaluation systems with the sampling method in the DynaBiS algorithm (Figure 4(C)). AutoDock blind docking did not sample the correct binding site for one system (PDB-ID: 2mcp), and AutoSite did not sample the correct binding site for several peptide systems (Table S2).

The DynaBiS scoring of these binding sites with the forcefield-based pepscore works rather well: the correct binding site was listed as top-ranked prediction for 19 out of 26 binding sites, and all correct binding sites but one (PDB-ID: 2mcp) were present in the top-5 ranked binding site predictions. As shown in Table S2, the binding site which was not predicted (PDB-ID: 2mcp) was properly sampled by both AutoSite and DynaBiS (not by AutoDock blind docking), but ranked as a low-affinity binding site by all of these methods. This indicates that this inability to identify the correct binding site in this system is mainly due to a failing scoring function.

The AutoDock scoring function performed well for the small ligand systems, identifying almost all small ligand binding sites correctly as top-ranked binding site. It needs to be noted here that AutoDock applies a semi-empirical scoring function,[37] while DynaBiS applies a force field based pepscore[16] optimized for peptides. AutoDock's scoring performance for peptide systems is much lower: only one of the correctly sampled peptide-apo binding sites was selected as top-ranked binding site for the AutoDock blind docking approach.

3.4 | Effect of binding site size and ligand flexibility on binding site identification performance

Figure 5 shows the relationship between the binding site size or ligand flexibility and the performance of the binding site identification methods. We showed that DynaBiS steadily predicts a binding site in the top-5 ranked binding site predictions with a GC_{offset} below 5 Å, which illustrates that the performance of DynaBiS is barely affected by the size of the binding site or ligand flexibility. Regarding the best sampled binding sites (lowest GC_{offset}), all GC_{offset} values were even below 4 Å for the DynaBiS method. In contrast, both the AutoDock blind docking approach and AutoSite predicted less binding sites correctly in their top-5 ranked binding site predictions with increasing binding site size and ligand flexibility. Interestingly, the largest distribution in the GC_{offset} values (including the highest GC_{offset} values) was found for the small binding sites. DynaBiS performs particularly strong for peptide systems: the average GC_{offset} considering the best binding site in the top-5 ranked binding site predictions for the small ligand systems is 2.3 ± 5.9 Å and 2.2 ± 1.3 Å for the AutoDock blind docking approach and DynaBiS respectively, while the average GC_{offset} values for the peptide systems for these two methods are respective 7.4 ± 6.7 Å and 3.0 ± 1.7 Å. This trend is less apparent in the best-scored

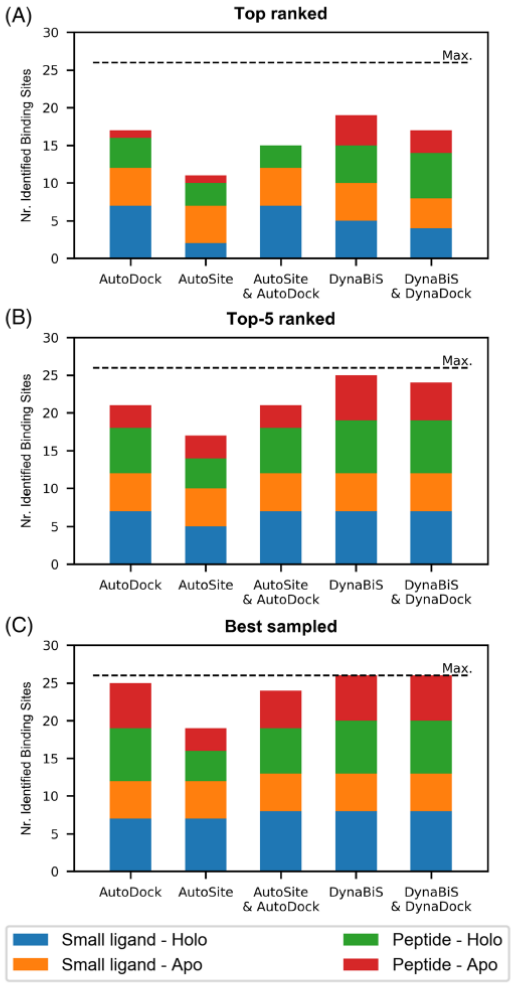


FIGURE 4 Binding site identification performance of evaluated methods, subdivided and color-coded by ligand type (small ligand/peptide) and type of crystal structure used (apo/holo). The dashed line represents the total number of systems in the evaluation set. The stacked bar charts represent the amount of correctly predicted binding sites only considering the top-ranked binding site (A), the best binding site from the top-5 ranked binding site predictions (B) or considering all sampled binding sites (C)

TABLE 3 Overview of the performance evaluation analysis

PDB ID	Nr. rot. Bonds	Ligand diameter (Å)	AutoDock	AutoSite	AutoSite& AutoDock	DynaBiS	DynaBiS& DynaDock
SL-holo							
2fm2	16	19.66	**1.15 (5.54)**	**1.75** (25.30)	**2.09 (3.48)**	**4.99** (26.59)	**4.09 (4.09)**
1hsh	14	18.12	**0.74 (3.00)**	**0.98 (0.98)**	**1.33 (2.00)**	**3.39 (3.39)**	**2.19 (2.19)**
1stp	5	8.31	**0.19 (0.19)**	**0.83 (0.83)**	**0.25 (0.25)**	**0.48 (0.48)**	5.15 (8.50)
1bju	4	14.36	**0.35 (0.35)**	**1.98** (21.54)	**0.62 (0.62)**	**1.38 (2.68)**	**2.85** (20.46)
2mcp	3	6.15	21.90 (29.53)	9.32 (44.94)	21.90 (29.12)	4.02 (4.36)	**1.77 (1.77)**
1efy	3	10.96	**0.95 (2.62)**	8.42 (8.42)	**0.69 (2.95)**	**1.30 (2.56)**	**0.88** (6.04)
1j8a	2	7.05	**0.20 (0.20)**	**1.83** (20.40)	**0.64 (0.73)**	**1.54 (1.54)**	**1.22** (16.30)
1ldm	1	4.35	**0.51 (0.51)**	2.33 (2.33)	**0.52 (0.52)**	**1.67** (7.01)	**1.37 (1.37)**
		NR-BSIdentified:	**7 (7)**	**5 (2)**	**7 (7)**	**7 (5)**	**7 (4)**
SL-apo							
1hsi–1hsh	14	18.09	**1.75 (2.49)**	**8.03 (8.03)**	**1.81 (3.13)**	**2.69 (4.83)**	**3.99 (4.18)**
1swb–1stp	5	10.87	**0.32 (0.32)**	**0.32 (0.32)**	**0.50 (0.50)**	**1.54 (1.54)**	**0.38 (0.65)**
1s0q–1bju	4	14.34	**0.53 (0.53)**	**2.55 (2.55)**	**0.81 (0.81)**	**1.47 (3.09)**	**4.04 (5.53)**
2paw–1efy	3	10.83	**0.89 (0.89)**	**4.47 (4.47)**	**0.46 (3.07)**	**2.15 (2.99)**	**2.23** (6.26)
1s0q–1j8a	2	7.07	**0.16 (2.98)**	**1.46 (1.46)**	**0.65 (3.23)**	**1.54 (2.67)**	**2.50 (2.50)**
		NR-BSIdentified:	**5 (5)**	**5 (5)**	**5 (5)**	**5 (5)**	**5 (4)**
PL-holo							
1io6	33[a]	31.61	**3.71 (2.44)**	18.86 (18.86)	18.08 (18.08)	**1.90 (1.90)**	**2.10 (2.10)**
1be9	23	19.89	**8.54** (25.58)	**6.37 (6.37)**	**4.17** (22.71)	**1.50 (3.81)**	**1.56 (2.04)**
1pau	16	13.65	**2.83 (2.83)**	**2.94 (2.94)**	**0.21 (0.21)**	**2.42 (3.91)**	**3.80 (3.80)**
1awq	15	15.57	**5.23 (5.23)**	**4.46 (4.46)**	**3.31 (3.31)**	**3.06 (6.59)**	**2.13 (4.39)**
1a30	13	10.4	15.01 (16.93)	5.29 (5.29)	**3.34** (16.75)	**2.15 (2.15)**	**3.10 (3.39)**
8tln	10	9.31	**2.85** (8.46)	5.69 (5.69)	**4.48** (8.17)	**3.81** (5.34)	**2.05 (4.15)**
2cyh	3	6.55	**0.44 (0.44)**	**1.51** (9.48)	**0.23 (0.32)**	**1.99** (8.33)	**1.01** (17.64)
		NR-BSIdentified:	**6 (4)**	**4 (3)**	**6 (3)**	**7 (5)**	**7 (6)**
PL-apo							
1gfd–1io6	33[a]	31.61	**3.70 (8.29)**	**13.96** (17.92)	**14.34** (17.33)	**4.97 (8.21)**	**0.78 (8.62)**
1bfe–1be9	23	19.88	**5.89** (14.31)	**7.25 (7.25)**	**6.20** (23.92)	**4.00** (18.48)	**2.41 (9.28)**
1qx3–1pau	16	13.56	25.28 (32.57)	**5.17** (11.23)	29.18 (40.50)	**3.82 (3.82)**	**5.11** (32.05)
2cpl–1awq	15	16.47	9.44 (9.44)	11.06 (11.06)	**5.14** (8.66)	**3.66 (6.33)**	**5.03** (18.66)
1fjq–8tln	10	9.02	**2.59** (7.91)	6.27 (6.27)	5.63 (16.03)	**4.10** (7.77)	**3.55 (3.55)**
2cpl–2cyh	3	6.55	11.20 (11.20)	9.46 (9.65)	29.35 (29.35)	**1.58 (1.58)**	6.54 (29.37)
		NR-BSIdentified:	**3 (1)**	**3 (1)**	**3 (0)**	**6 (4)**	**5 (3)**
		NR-BindSiteIdentified:	21 (17)	17 (11)	21 (15)	25 (19)	24 (17)

Note: The values represent the GC_{offset} of the best binding site found in the top-5 ranked binding site predictions. The values in brackets represent the GC_{offset} of the best-ranked binding site. Values printed in bold are below the binding site identification criteria (50% ligand diameter), and therefore selected as identified.

Abbreviations: PL, peptide ligand; SL, small ligand.

[a]For the AutoDock calculations, the bond between the hydroxyl moiety and the phenyl group was inactivated as AutoDock limits the number of rotatable bonds to 32.

binding sites, indicating that the scoring functions of both AutoDock and DynaBiS are not substantially affected by the binding site size and ligand flexibility.

Based on our observation regarding the strong performance of DynaBiS in peptide-binding proteins, we additionally compared DynaBiS with PEP-SiteFinder,[14] a binding site identification method specifically designed for peptide-systems. The results are provided in the Table S3. PEP-SiteFinder first predicts the peptide's conformation using PEP-FOLD, followed by blind docking applying the ATTRACT docking protocol using the coarse-grained ATTRACT force field.[38]

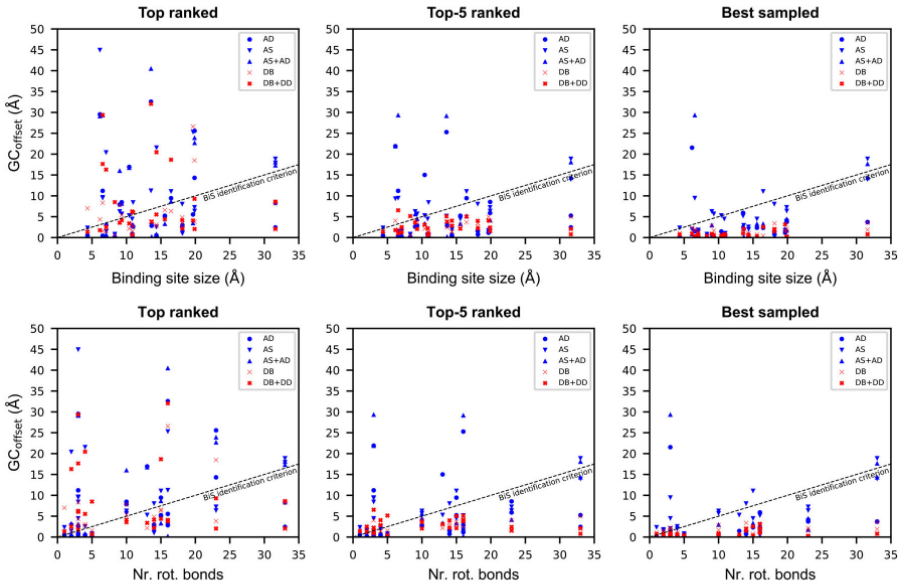


FIGURE 5 The effect of binding site size and ligand flexibility on binding site identification performance in the evaluated methods. AutoDock related methods are represented in blue and DynaBiS related methods in red. The dashed line represents 50% ligand diameter and thereby the cutoff for considering a binding site as identified. The panels show the GC_{offset} for the best-scored binding site (left), the best binding site found in the top-5 scored binding site predictions (center) and the best sampled binding site (i.e., lowest GC_{offset} from all sampled binding sites, right) ordered by binding site size (top) and by number of rotatable bonds (bottom), the latter as measure for ligand flexibility. The ligand diameter was used as measure for the binding site size. AD: AutoDock blind docking, AS: AutoSite, AS + AD: combined AutoSite and AutoDock approach, DB: DynaBiS, DB + DD: combined DynaBiS and DynaDock approach

Since PEP-SiteFinder requires peptides of at least five amino acids, this evaluation could only be performed for three systems of the evaluation set. DynaBiS predicted all binding sites closer to the reference than PEP-SiteFinder. Considering all sampled binding sites in order to compare the sampling algorithm of both methods, PEP-SiteFinder only sampled a binding site closer to the reference for the system with PDB-ID 1io6: $GC_{offset} = 1.15$ and 1.90 Å for PEP-SiteFinder and DynaBiS respectively. However, PEP-SiteFinder did not score this binding site correctly, as it was not present in the top-5 ranked binding site predictions. In both cases, considering only the best scored or the best binding site from the top-5 ranked binding site predictions, the binding sites predicted by DynaBiS were closer to the reference than PEP-SiteFinder, illustrating again the importance of including protein flexibility in the binding site identification process.

We additionally investigated the extend of protein movement simulated by DynaBiS. During the MC/SA simulations in the pocket sampling step, the protein flexibility is approximated by allowing protein–ligand overlap. The protein atoms in the binding site are allowed to move during the subsequent energy minimization, at which the protein–ligand overlap will be resolved. Since only the binding site atoms are treated flexible and no long-term MD simulations are performed, large backbone movements cannot be expected, and were also not observed. However, regarding the pocket sampling simulations in the apo structures, the binding site residues moved up to 3.5 Å away from the input structure (Figure S3). Also the backbone atoms moved a certain amount, illustrating the importance of considering backbone flexibility. The conformation of the holo structures was however not reproduced (Figure S4), as this probably requires long-term MD simulations to obtain.

3.5 | Combining binding site identification with molecular docking simulations

We were interested if DynaBiS binding site identification could be combined with DynaDock molecular docking simulations in the top-ranked binding sites in a single pipeline, to additionally retrieve more accurate docking poses. As DynaDock treats the entire protein as flexible and contains a more extensive OPMD procedure to investigate the binding pose, we also briefly investigated if a blind docking approach achieved by combining DynaBiS and DynaDock (Figure 1(A), pathway b) could potentially further improve sampling and identification of binding sites. For a direct comparison to the combined DynaBiS and DynaDock approach, we combined AutoSite with AutoDock docking calculations, in which the AutoSite fill guided the construction of the grid box. In the AutoDock and AutoSite

combination approach, AutoDock docking was performed in all predicted binding sites from AutoSite, after which all resulting poses were combined and further filtered as described for the AutoDock blind docking approach. In the DynaBiS and DynaDock combination approach, DynaDock docking simulations were only performed in the top-5 ranked binding sites predicted by DynaBiS. The resulting docking poses acted as indicators for the location of the binding site.

As shown in Figure 4(B) and Table 3, the correct binding site was present in the top-5 ranked binding site predictions for four more systems by the combined AutoSite and AutoDock approach compared with the individual AutoSite approach. This improvement is mainly due to improved sampling (Figure 4(C) and Table S2): five more binding sites were correctly sampled, of which four belong to the peptide systems. Sampling of the small ligand systems was only slightly improved, however, pose refinement by AutoDock strongly improved the scoring of the small ligand systems. This is illustrated by the five additional correctly predicted binding sites in the top-ranked sites, which all belong to the small ligand holo set. This post-refinement was only beneficial for the binding site identification of AutoSite, as the combination of DynaBiS and DynaDock resulted in slightly less correctly identified binding sites. All binding sites were still correctly sampled in the DynaBiS and DynaDock combination approach (Figure 4 (C) and Table S2), thus the observed reduced performance compared with the single DynaBiS approach is mainly due to incorrect scoring.

Finally, we briefly investigated if a blind docking pipeline containing of DynaBiS as a binding site identification algorithm followed by DynaDock molecular docking (in which the docking simulations are performed in the identified binding site by DynaBiS) could result in accurate docking poses as well. Because DynaBiS uses ligand poses as binding site location indicators, a direct comparison with the DynaDock docking poses is possible. For 20 of the evaluation systems, this combined pipeline resulted in docking poses with a RMSD below 3 Å and for the remaining 6 systems, poses with a RMSD below 5 Å could be obtained. With DynaBiS alone good poses with a RMSD better than 3 Å could only be found for 14 systems, while 9 systems had RMSD values below 5 Å and 2 poses had even RMSD values larger than 7 Å. This clearly demonstrates the ability and benefit to combine DynaBiS and DynaDock in a single pipeline to additionally retrieve accurate docking poses (Figure S8).

4 | CONCLUSIONS

We developed a binding site identification algorithm DynaBiS, with which we could successfully identify both buried and surface exposed binding sites, binding small and rigid as well as large and flexible ligands. This approach was inspired by the DynaDock docking algorithm, applying soft-core potentials between the ligand and the protein to allow them to partially overlap. With this approach, the limitations of the rigid protein approximation can be overcome. We showed that with DynaBiS all binding sites from the evaluation set were successfully identified as potential binding sites and included in the pocket sampling step. Regarding binding site identification, also for all but one system the correct binding site was present in the top-5 ranked binding site predictions and 19 out of 26 binding sites were predicted correctly as the top-ranked binding site. Comparison with an AutoDock blind docking approach and AutoSite showed that DynaBiS outperforms both methods in the prediction and identification of peptide-binding binding sites, especially if an apo structure is used as input.

ACKNOWLEDGMENTS

The authors acknowledge the Deutsche Forschungsgemeinschaft for financial support via SFB 1035, project A10, SFB749, project C08, and CIPSM. Open Access funding enabled and organized by Projekt DEAL.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26182.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Okke Melse https://orcid.org/0000-0002-0021-3466
Iris Antes https://orcid.org/0000-0002-2241-7187

REFERENCES

1. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. *Comput Struct Biotechnol J*. 2020;18:417-426.
2. Ding Y, Tang J, Guo F. Identification of protein–ligand binding sites by sequence information and ensemble classifier. *J Chem Inf Model*. 2017;57(12):3149-3161.
3. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*. 2001;307(1):447-463.
4. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*. 2005;21 (9):1908-1916.
5. Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res*. 2009;37:W413-W416.
6. Harris R, Olson AJ, Goodsell DS. Automated prediction of ligand-binding sites in proteins. *Proteins: Struct Funct Bioinform*. 2008;70(4): 1506-1517.
7. Ravindranath PA, Sanner MF. AutoSite: an automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics*. 2016;32(20):3142-3149.
8. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30(16):2785-2791.
9. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res*. 2010;38: W469-W473.
10. Monet D, Desdouits N, Nilges M, Blondel A. mkgridXf: consistent identification of plausible binding sites despite the elusive nature of cavities and grooves in protein dynamics. *J Chem Inf Model*. 2019;59 (8):3506-3518.
11. Wagner JR, Sørensen J, Hensley N, et al. POVME 3.0: software for mapping binding pocket flexibility. *J Chem Theory Comput*. 2017;13 (9):4584-4592.

12. Hetényi C, van der Spoel D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett*. 2006;580(5):1447-1450.
13. Hetényi C, van der Spoel D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci*. 2002;11(7):1729-1737.
14. Saladin A, Rey J, Thévenet P, Zacharias M, Moroy G, Tufféry P. PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces. *Nucleic Acids Res*. 2014;42(W1):W221-W226.
15. Stank A, Kokh DB, Fuller JC, Wade RC. Protein Binding Pocket Dynamics. *Acc Chem Res*. 2016;49(5):809-815.
16. Antes I. DynaDock: a new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. *Proteins: Struct Funct Bioinform*. 2010;78(5):1084-1104.
17. Plattner N, Noé F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun*. 2015;6(1):7653.
18. Stefan H, Salo-Ahen OM, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit*. 2010;23(2):209-219.
19. Taylor RD, Jewsbury PJ, Essex JW. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J Comput Chem*. 2003;24(13):1637-1656.
20. Vanderbilt D, Louie SG. A Monte carlo simulated annealing approach to optimization over continuous variables. *J Comput Phys*. 1984;56(2):259-271.
21. Bouzida D, Rejto PA, Arthurs S, et al. Computer simulations of ligand–protein binding with ensembles of protein conformations: a Monte Carlo study of HIV-1 protease binding energy landscapes. *Int J Quantum Chem*. 1999;72(1):73-84.
22. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science*. 1983;220(4598):671-680.
23. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21(6):1087-1092.
24. Hartmann C, Antes I, Lengauer T. IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci*. 2007;16(7):1294-1307.
25. Schuttelkopf AW, van Aalten DMF. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr Sect D Biol Crystallogr*. 2004;60(8):1355-1363.
26. van Gunsteren WF, Billeter SR, Eising AA, et al. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*. Zürich, Switzerland: VdF: Hochschulverlag AG an der ETH Zürich; 1996.
27. Gallicchio E, Levy RM. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J Comput Chem*. 2004;25(4):479-499.
28. van der Spoel D, Lindahl E & Hess B. Gromacs User Manual version 4.5.6, www.gromacs.org (2010).
29. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *J Chem Phys*. 1995;103(19):8577-8593.
30. Berendsen HJC, Postma JPM, Gunsteren WFV, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys*. 1984;81(8):3684-3690.
31. Hess B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J Chem Theory Comput*. 2008;4(1):116-122.
32. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
33. Mächler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. *Cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.4; 2013.
34. Antunes DA, Devaurs D, Kavraki LE. Understanding the challenges of protein flexibility in drug design. *Expert Opin Drug Dis*. 2015;10(12):1301-1313.
35. Satya PG. Protein flexibility: a challenging issue of drug discovery. *Curr Chem Biol*. 2018;12(1):3-13.
36. Ciemny M, Kurcinski M, Kamel K, et al. Protein–peptide docking: opportunities and challenges. *Drug Discov Today*. 2018;23(8):1530-1537.
37. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*. 2007;28(6):1145-1152.
38. Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT. *Proteins: Struct Funct Bioinform*. 2010;78(15):3131-3139.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Melse O, Hecht S, Antes I. DynaBiS: A hierarchical sampling algorithm to identify flexible binding sites for large ligands and peptides. *Proteins*. 2022;90(1):18-32. https://doi.org/10.1002/prot.26182

Supporting Information

DynaBiS: A Hierarchical Sampling Algorithm to Identify Flexible Binding Sites for Large Ligands and Peptides

Okke Melse, Sabrina Hecht, Iris Antes*

TUM Center for Functional Protein Assemblies and TUM School of Life Sciences, Technische Universität München, Emil-Erlenmeyer-Forum 8, 85354 Freising, Germany

Contents

Table S1. DynaBiS settings used for the evaluation set during the method comparison with AutoDock and AutoSite.

PDB ID	Nr. rotatable bonds	Ligand diameter (Å)	Surface screening Energy Threshold† (kJ·mol^{-1})	Pocket sampling Maximum allowed‡ P-L overlap (%)
SL-holo				
2fm2	16	19.66	0	40
1hsh	14	18.12	0	40
1stp	5	8.31	0	30
1bju	4	14.36	500	30
2mcp	3	6.15	500	30
1efy	3	10.96	0	30
1j8a	2	7.05	500	30
1ldm	1	4.35	500	30
SL-apo				
1hsi-1hsh	14	18.09	500	40
1swb-1stp	5	10.87	500	30
1s0q-1bju	4	14.34	500	30
2paw-1efy	3	10.83	500	30
1s0q-1j8a	2	7.07	500	30
PL-holo				
1io6	33	31.61	0	40
1be9	23	19.89	0	40
1pau	16	13.65	500	40
1awq	15	15.57	0	40
1a30	13	10.4	0	40
8tln	10	9.31	0	40
2cyh	3	6.55	0	30
PL-apo				
1gfd-1io6	33	31.61	500	40
1bfe-1be9	23	19.88	500	40
1qx3-1pau	16	13.56	500	40
2cpl-1awq	15	16.47	500	40
1fjq-8tln	10	9.02	500	40
2cpl-2cyh	3	6.55	500	30

†Energy threshold applies to the maximum value of the coulomb term (considering the soft-core potential) of the sampled ligand in order for the generated conformation to be accepted.

‡This value was set to 30% for small and rigid (nr. rot. bonds < 6) ligands and peptides to avoid too much overlap, since this will be hard to refine for these kind of ligands, based on in-house experience with DynaDock.

Table S2: Best sampled binding site for each evaluated method.

PDB ID	Nr. rotat. Bonds	Ligand diameter (Å)	AD	AS	AS+AD	DynaBiS	DB+DD
SL-holo							
2fm2	16	19.66	**1.15**	**1.75**	**2.09**	**1.61**	**3.16**
1hsh	14	18.12	**0.74**	**0.98**	**0.23**	**3.39**	**2.03**
1stp	5	8.31	**0.08**	**0.83**	**0.25**	**0.48**	**0.36**
1bju	4	14.36	**0.16**	**1.98**	**0.62**	**1.09**	**0.35**
2mcp	3	6.15	21.54	**2.04**	**0.85**	**2.28**	**0.96**
1efy	3	10.96	**0.66**	**1.05**	**0.67**	**0.35**	**0.69**
1j8a	2	7.05	**0.08**	**1.83**	**0.64**	**1.54**	**0.67**
1ldm	1	4.35	**0.15**	2.33	**0.52**	**1.06**	**0.73**
		NR-BS-Identified:	**7**	**7**	**8**	**8**	**8**
SL-apo							
1hsi-1hsh	14	18.09	**1.33**	**8.03**	**1.16**	**1.68**	**1.51**
1swb-1stp	5	10.87	**0.08**	**0.32**	**0.50**	**0.58**	**0.22**
1s0q-1bju	4	14.34	**0.16**	**2.55**	**0.81**	**0.60**	**0.61**
2paw-1efy	3	10.83	**0.65**	**4.47**	**0.46**	**0.58**	**0.78**
1s0q-1j8a	2	7.07	**0.03**	**1.46**	**0.65**	**1.12**	**0.14**
		NR-BS-Identified:	**5**	**5**	**5**	**5**	**5**
PL-holo							
1io6	33†	31.61	**3.71**	18.86	17.66	**1.90**	**0.78**
1be9	23	19.89	**3.73**	**6.37**	**1.79**	**1.50**	**0.28**
1pau	16	13.65	**2.09**	**2.94**	**0.21**	**1.35**	**2.08**
1awq	15	15.57	**2.35**	**4.46**	**3.31**	**0.77**	**0.30**
1a30	13	10.4	**1.45**	5.29	**0.66**	**0.80**	**0.43**
8tln	10	9.31	**0.17**	5.69	**0.45**	**1.66**	**0.55**
2cyh	3	6.55	**0.19**	**1.51**	**0.23**	**1.99**	**0.33**
		NR-BS-Identified:	**7**	**4**	**6**	**7**	**7**
PL-apo							
1gfd-1io6	33†	31.61	**3.70**	**13.96**	**14.34**	**3.58**	**0.78**
1bfe-1be9	23	19.88	**4.15**	**7.25**	**4.45**	**2.12**	**0.09**
1qx3-1pau	16	13.56	**5.91**	**5.17**	**2.93**	**1.22**	**0.99**
2cpl-1awq	15	16.47	**2.32**	11.06	**3.29**	**0.43**	**2.55**
1fjq-8tln	10	9.02	**0.69**	6.27	**2.98**	**2.86**	**0.88**
2cpl-2cyh	3	6.55	**1.50**	9.46	29.35	**1.58**	**0.12**
		NR-BS-Identified:	**6**	**3**	**5**	**6**	**6**
		NR-BindSite-Identified:	**25**	**19**	**24**	**26**	**26**

The values represent the GC_{offset} of the best sampled pose. Values printed in bold are below the binding site identification criteria (50% of ligand diameter), and therefore selected as identified. AD: AutoDock blind docking, AS: AutoSite, AS+AD: combined AutoSite & AutoDock approach, DB: DynaBiS, DB+DD: combined DynaBiS & DynaDock approach.

†For the AutoDock calculations, the bond between the hydroxyl moiety and the phenyl group has been inactivated as AutoDock limits the number of rotatable bonds at 32.

Table S3. PepSiteFinder binding site identification performance.

PL-holo	DynaBiS	PepSiteFinder
Best scored		
1io6	**1.90**	3.63
1be9	**3.81**	4.72
1awq	**6.56**	10.77
Top-5		
1io6	**1.90**	3.07
1be9	**1.50**	3.05
1awq	**3.06**	5.48
Best sampled		
1io6	1.90	**1.15**
1be9	**1.50**	4.72
1awq	**0.77**	2.64

Table S4. Control pocket sampling simulations

	Best sampled		Average	
System	**GC_{offset} (Å)**	**RMSD (Å)**	**GC_{offset} (Å)**	**RMSD (Å)**
1be9	1.54	1.66	4.75	5.30
1efy	1.97	3.82	6.03	7.57
1a30	0.78	3.29	5.03	7.13
1bju	1.38	2.35	4.54	5.60
1bfe-1be9	1.79	1.87	3.99	4.53
1qx3-1pau	0.49	2.97	3.78	6.41
1hsh-1hsi	0.00	0.09	2.83	5.26
2paw-1efy	1.57	3.34	5.93	7.85
Average	1.19	2.42	4.61	6.21

The results of pocket sampling simulations (50 MC/SA runs) starting from the reference conformation are listed.

Table S5. CPU time benchmarking

Calculation	**CPU time (sec)**
DynaBiS – Surface screening (1000 poses)	562 ± 287 (9m22s ± 4m47s)
DynaBiS – Pocket sampling (1 binding site, 10 MC runs)	3392 ± 1471 (56m32s ± 24m31s)
DynaBiS – full run† (assuming 3 binding sites)	10738 ± 2564 (2h59m58s ± 42m44s)
AutoDock‡ (250 GA-runs, $10 \cdot 10^6$ ga_num_evals)	23472 ± 12823 (6h31m12s ± 3h33m43s)
AutoSite	2 ± 1

All calculations were performed as a single-core job on an Intel Xeon CPU E3-1270 v5 (3.60 GHz). The provided values are average CPU time for the calculations performed on the subset of the evaluation set, *i.e.* 1be9, 1efy, 1a30, 1bju, 1bfe-1be9, 1qx3-1pau, 1hsh-1hsi, 2paw-1efy.

†Note that in this study, we run 10 replicates of DynaBiS in parallel. Thus the total CPU time for this setup is 10 times higher than the value reported here, but can be ran in parallel on multiple CPU cores.

‡The CPU time of AutoDock is linearly scalable with the number of GA-runs, thus the CPU time can significantly reduce or increase with different values for this setting. In contrast to DynaBiS, the spatial distribution cannot be enforced. Therefore, a large amount of poses (GA-runs) are required to retrieve accurate results, ideally together with a large amount of allowed energy evaluations (ga_num_evals).

Figure S1. List of ligands in the evaluation set. Small ligands (top) and peptides (bottom) included in evaluation set. The value between brackets indicates the calculated ligand diameter of the respective ligand.

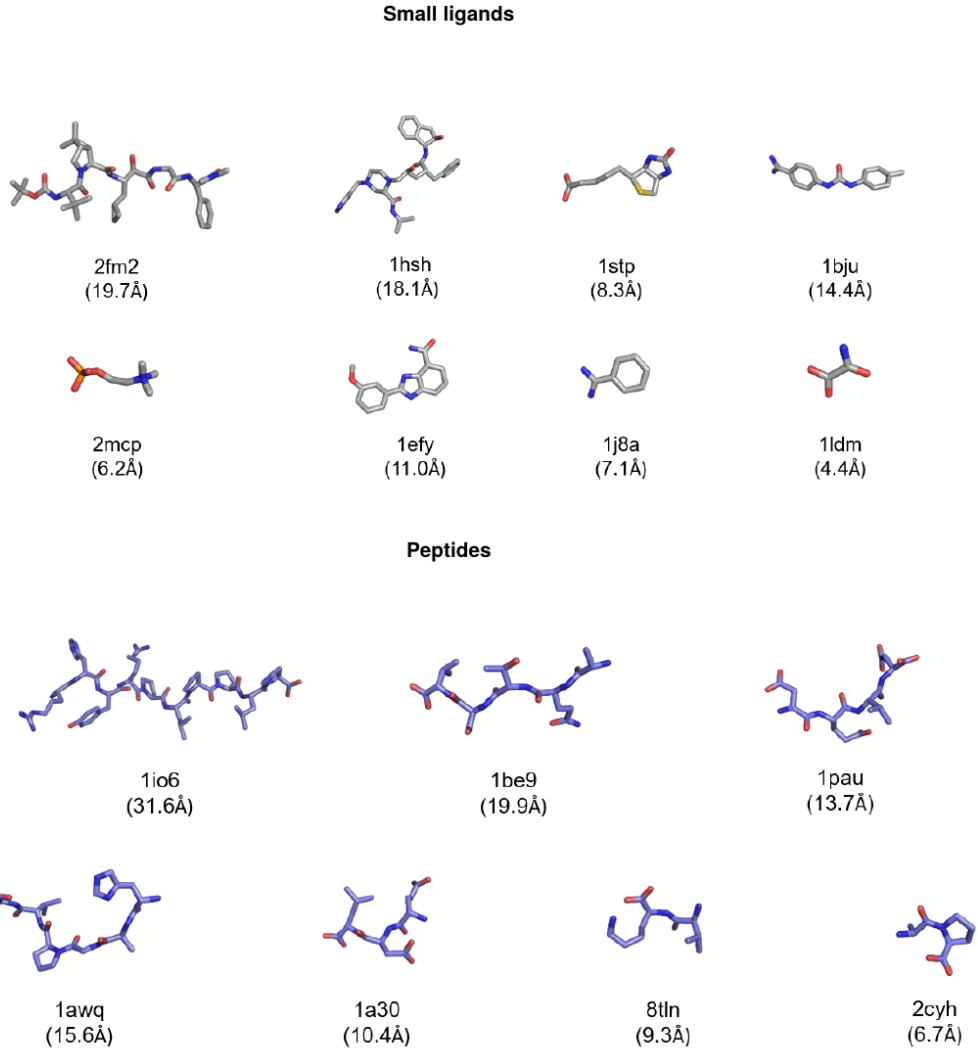


Figure S2. Schematic representation of the DynaBiS algorithm, shown for the situation of four spatially different pockets (a number which is automatically defined by DynaBiS). N represents the number of MC/SA simulations in the pocket sampling, which is set by the user (default: 10).

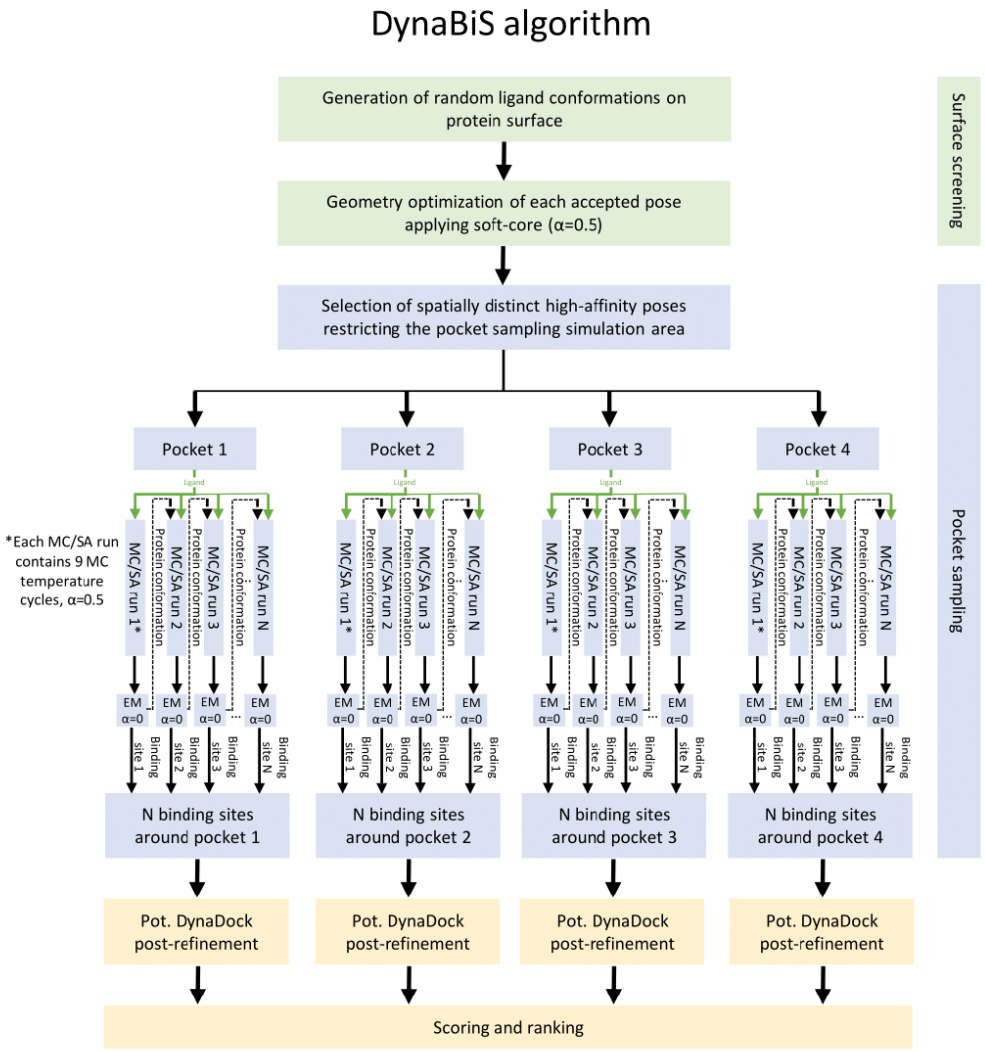


Figure S3. Protein flexibility during pocket sampling simulations in apo-structures. Average per-residue RMSD values of important binding site residues are shown from pocket sampling simulations containing 50 MC/SA runs. The RMSD is measured to the input protein structure (i.e. apo-structure).

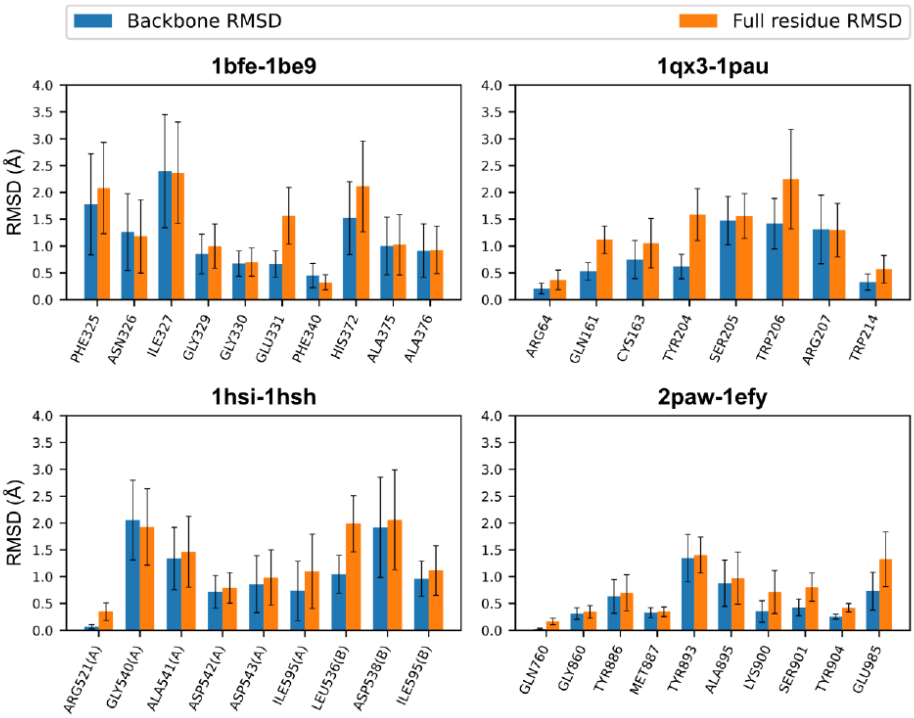


Figure S4. Difference in protein structure between apo- and holo-structure. The red bars indicate the RMSD of binding site residues between the apo- and holo- X-Ray structures, while the green bars represent the RMSD with respect to the holo-structure of all poses sampled during the pocket sampling step (50 MC/SA runs), starting from the crystallized ligand conformation.

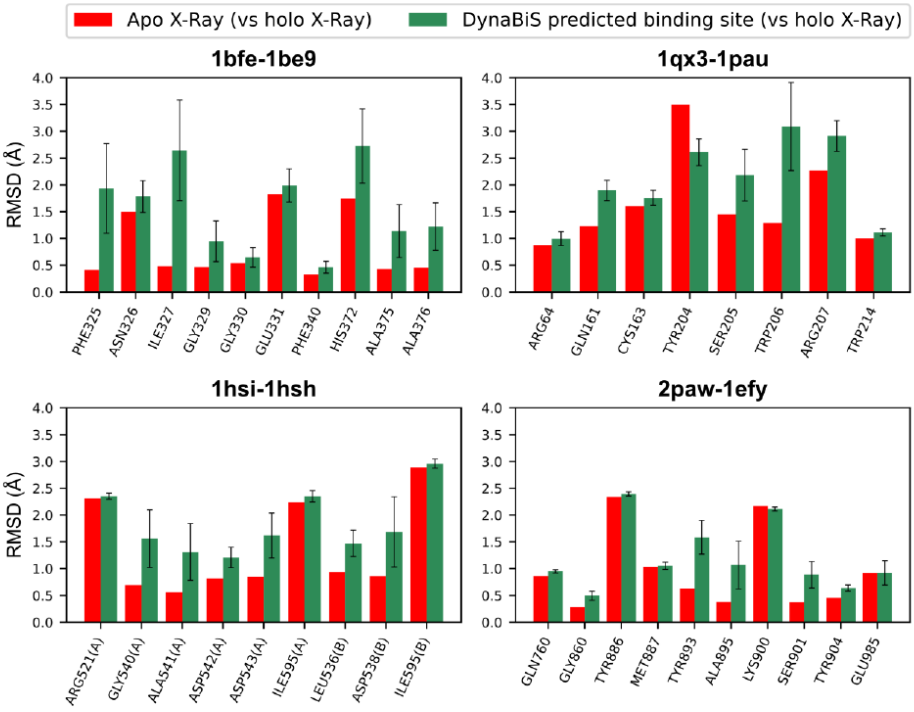
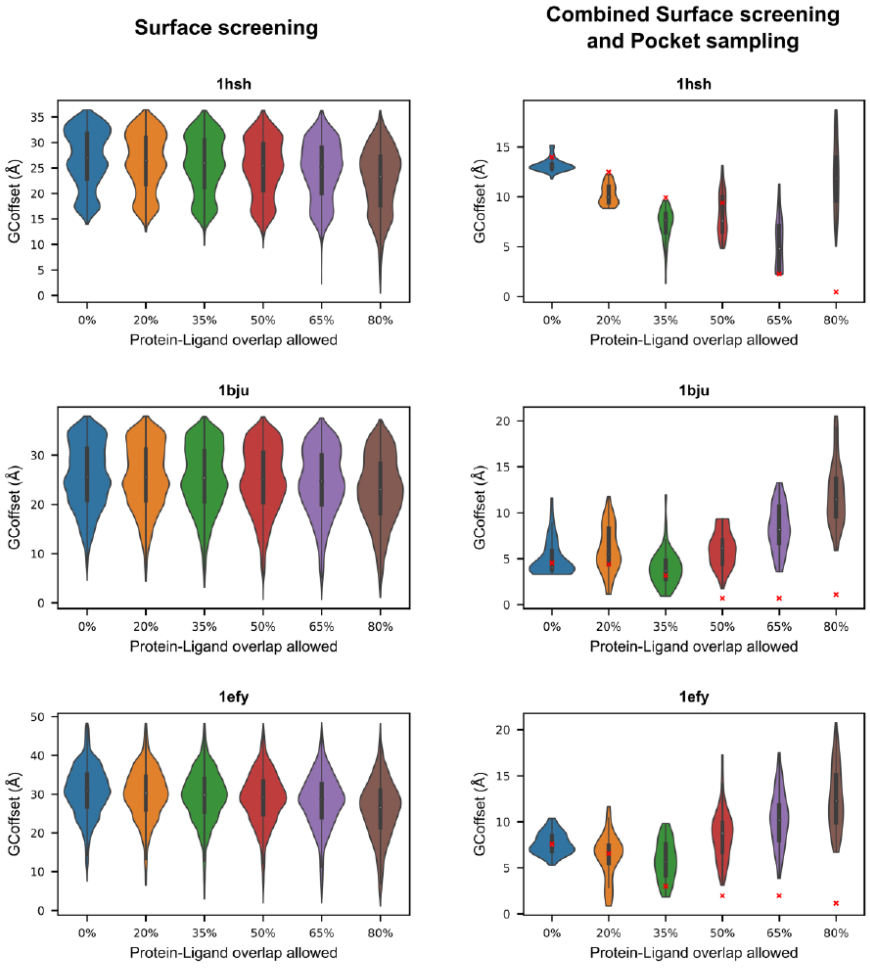
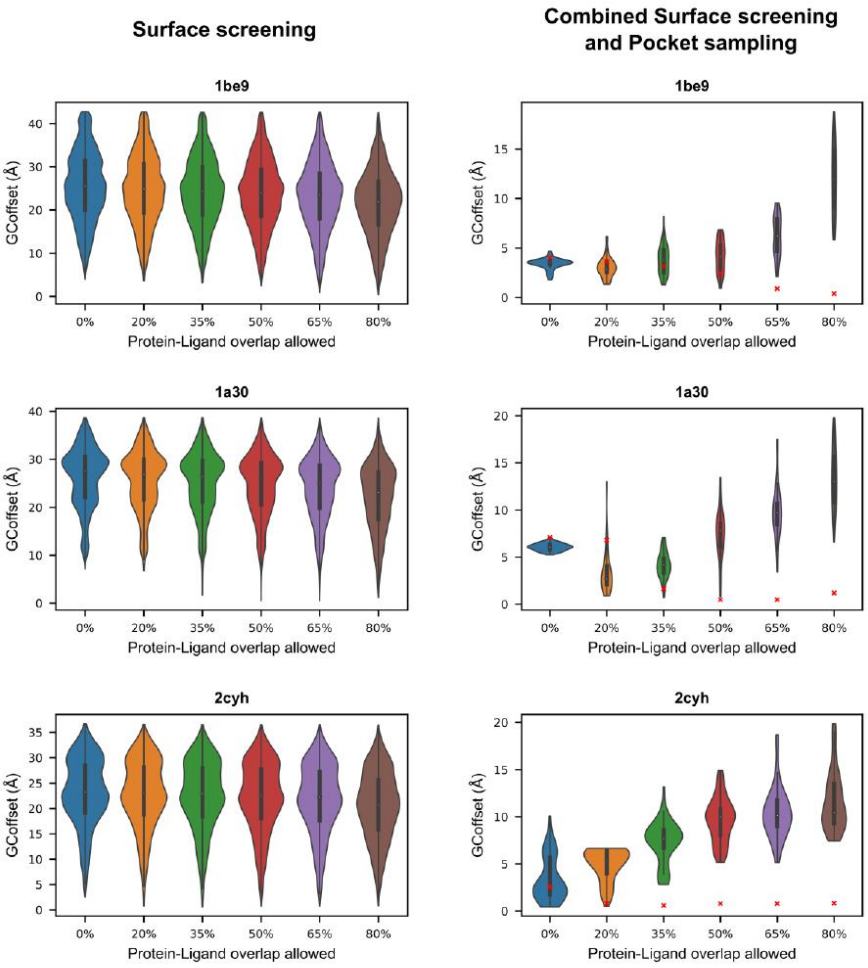
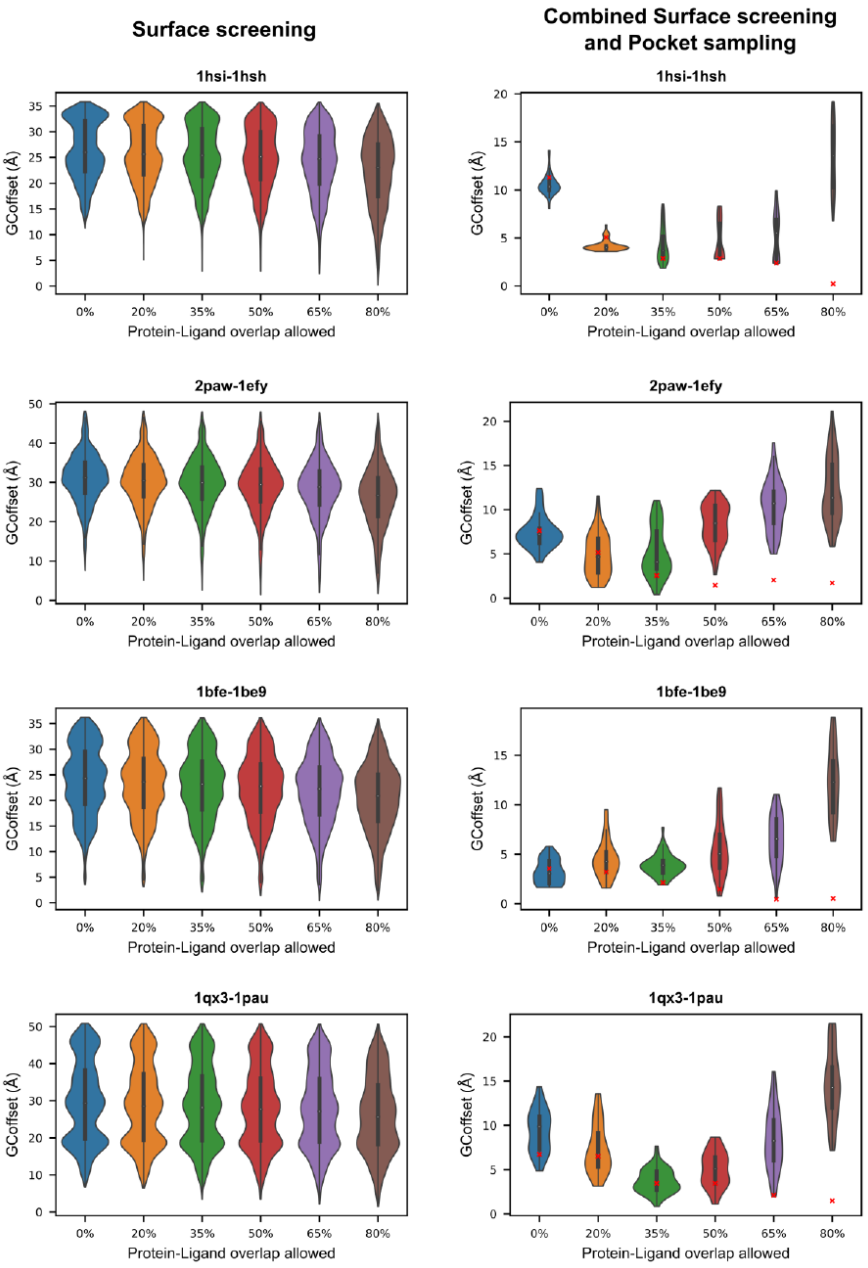


Figure S5. Evaluation of the effect of maximum allowed protein-ligand overlap on the DynaBiS performance for each individual evaluation system. The allowed protein-ligand overlap was identical for the surface screening and pocket sampling simulations: thus all pocket sampling simulations started from the pose with the lowest GC_{offset} resulting from the surface screening with the same maximum allowed protein-ligand overlap. The GC_{offset} of these starting poses are indicated by a red cross.




Surface screening
Combined Surface screening and Pocket sampling
1be9
1be9
1a30
1a30
2cyh
2cyh
GCoffset (Å)
Protein-Ligand overlap allowed
0%
20%
35%
50%
65%
80%


Surface screening
Combined Surface screening and Pocket sampling
1hsi-1hsh
2paw-1efy
1bfe-1be9
1qx3-1pau
GCoffset (Å)
Protein-Ligand overlap allowed
0%
20%
35%
50%
65%
80%

11

Figure S6. Effect of protein-ligand overlap in pocket sampling on resulting poses. The resulting poses of 50 pocket sampling simulations are shown for 3 different maximum allowed protein-ligand overlap values in 2 systems.

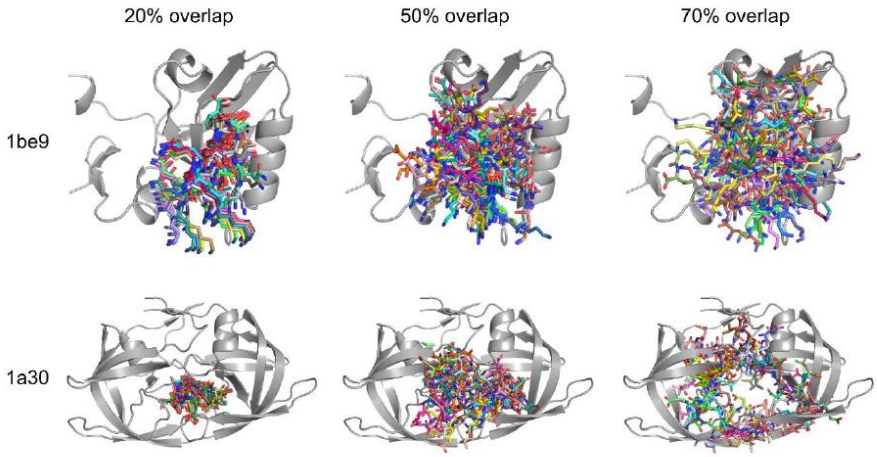


Figure S7. Binding site prediction in 2mcp**,** the top-ranked pose is shown.

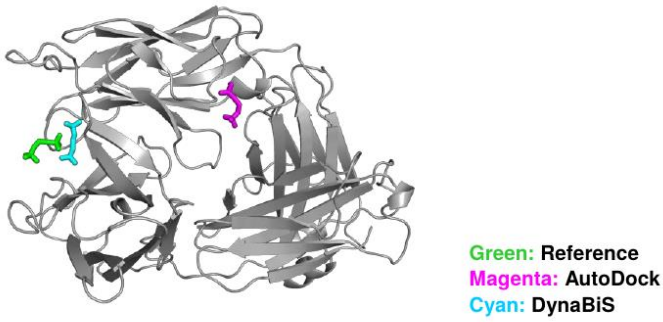
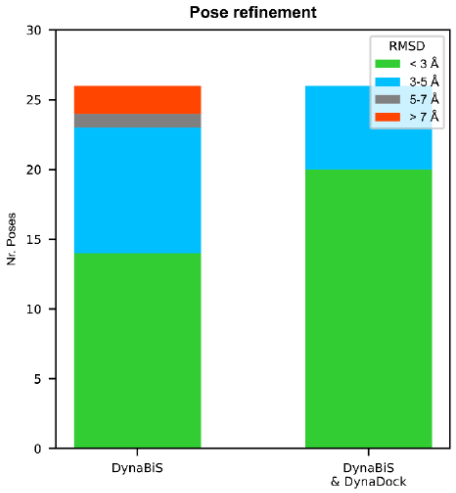


Figure S8. Ability of DynaDock to retrieve a docked pose in the already identified binding pocket (*i.e.* combined DynaBiS & DynaDock approach). All best-sampled poses for each system, either before (left) or after (right) DynaDock refinement are classified based on their RMSD to the reference ligand, *i.e.* the binding conformation found in the crystal structure.



3.3 BENCHMARKING BIOMOLECULAR FORCE FIELD-BASED Zn^{2+} FOR MONO- AND BIMETALLIC LIGAND BINDING SITES

In this study, a large number of commonly-used biomolecular force field-based Zn^{2+} models were benchmarked. An accurate description of Zn^{2+} in biomolecular systems is highly important because Zn^{2+} can play an important role in biocatalysis or support protein folding and -activation.[239] However, modelling of a Zn^{2+} ion applying classical force fields remains challenging, as advanced structural properties of the ion, including strong polarization effects and adaptation of multiple coordination geometries needs to be described solely by the non-bonded interaction terms with a single atom. Various Zn^{2+} models have been proposed, either applying the traditional 12-6 Lennard-Jones (LJ) potential, or a 12-6-4 LJ-type potential to include charge-induced dipole effects.[239, 240, 275, 276] Moreover, several dummy-atom models were included in the benchmarking set, as well as bonded models with various bonded parameters.[245, 246, 249, 250] The performance of these Zn^{2+} models was assessed in challenging environments: in the mono-metallic Carbonic Anhydrase II (CAII) and the bimetallic metallo-β-lactamase VIM-2. These ligand binding sites of both systems represent highly challenging benchmarking environments for the Zn^{2+} models because of the large number of possible Zn^{2+}-ligating atoms and the relatively large flexibility allowed in the pocket. Furthermore, these benchmarking environments also represent actual application-cases for drug design or biocatalyst development. The benchmarking study focused on properties that are important for molecular dynamics simulations, such as a correct and stable description of the coordination geometry and type of ligation. Based on these results, suitable simulation conditions for a variety of modelling approaches were suggested. Additional attempts to develop a new tetrahedral dummy-atom model, as well as a combination of parameters from different Zn^{2+} models provided further insights in promising parameterization strategies to further improve existing Zn^{2+} models.

Okke Melse and Iris Antes initially designed the project, and Ville R. I. Kaila gave useful suggestions for the development attempts of a tetrahedral Zn^{2+} model. Okke Melse designed the benchmarking study, performed all simulations, analyzed the data, and wrote the initial draft of the manuscript. Finally, all authors were involved in writing the final version of the manuscript.

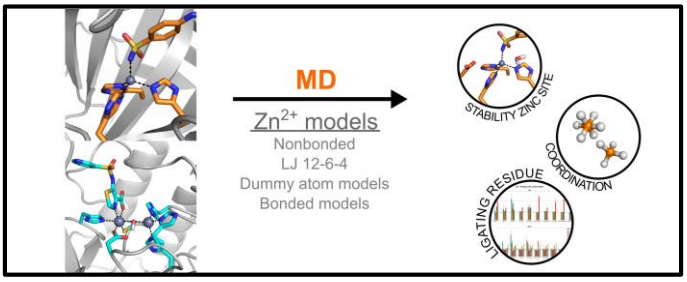
Benchmarking Biomolecular Force Field-Based Zn^{2+} for Mono- and Bimetallic Ligand Binding Sites

Okke Melse, Iris Antes, Ville R. I. Kaila, Martin Zacharias

Submitted

Preprint available at bioRxiv

https://doi.org/10.1101/2021.06.28.450184



3.4 Structure-Guided Modulation of the Catalytic Properties of [2Fe-2S]-Dependent Dehydratases

In this publication, sequence, structure and activity relationships within the [2Fe-2S]-dependent ilvD/EDD superfamily were deduced to understand, and rationally modify the substrate preferences of these different dehydratases. Based on this analysis, a new classification of these enzymes was proposed based on their evolutionary relationships and substrate preference into: sugar acid dehydratase (SADHT), branched-chain acid dehydratase (BCADHT) and promiscuous acid dehydratase (PADHT). Enzymes belonging to the first group, such as the DHT from *Paralcaligenes ureilyticus* (*Pu*DHT), are most active to substrates with longer chain length, such as D-gluconate. BCADHTs are predominantly active toward DHIV, such as the DHAD from *Fontimonas thermophilia* (*Ft*DHAD). The last category, PADHT, show a substrate profile which lies in between of the two other categories, with the DHAD from *Saccharolobus solfataricus* (*Ss*DHAD) as its currently only characterized member. For the rational design, homology models of *Pu*DHT, *Ft*DHAD and *Ss*DHAD were produced. Subsequently, several mutation hotspots were identified via molecular modelling which are likely to play a role in substrate specificity, and/or enzyme activity. These positions were confirmed to be catalytically relevant by *in vitro* site-directed mutagenesis. Further investigations led to several interesting enzyme variants, including a variant (P73G) of *Ft*DHAD with improved substrate promiscuity toward D-gluconate by >10-fold, while retaining a high activity toward DHIV, *i.e.* the biological substrate of the wild-type enzyme. The hypothesis is raised that this effect is due to a slight increase of binding site volume. Moreover, molecular docking and molecular dynamics simulations suggested that the C-terminal histidine in *Pu*DHT plays a role in substrate stabilization. Thus, saturation mutagenesis was performed on this position, which showed that a mutation toward a phenylalanine indeed shifts the substrate preference toward shorter sugar acids, showing around six-fold improved activity toward D-glycerate. In summary, the sequence, structure and activity relationships identified in this work may guide further engineering of DHADs to further improve biocatalysis cascades, leading to more efficient production of fine chemicals.

Okke Melse designed and performed the structural modelling including parameterization of the [2Fe-2S]-containing binding site, molecular docking and molecular dynamics simulations, prediction of mutation hotspots, and suggestion of *in vitro* mutagenesis strategies. Samuel Sutiono designed the *in vitro* experiments, which were conducted together with Magdelena Haslbeck. Okke Melse and Samuel Sutiono wrote the manuscript, with support by Gerhard Schenk and Volker Sieber.

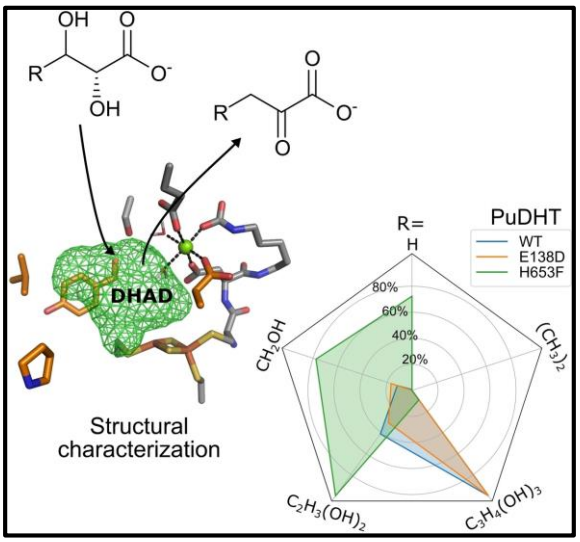
Structure-Guided Modulation of the Catalytic Properties of [2Fe-2S]-Dependent Dehydratases

Okke Melse, Samuel Sutiono, Magdalena Haslbeck, Gerhard Schenk, Iris Antes, Volker Sieber

ChemBioChem

2022. **23**(10): p. e202200088.

https://doi.org/10.1002/cbic.202200088



ChemBioChem

Research Article
doi.org/10.1002/cbic.202200088

Chemistry Europe
European Chemical Societies Publishing

www.chembiochem.org

Structure-Guided Modulation of the Catalytic Properties of [2Fe–2S]-Dependent Dehydratases

Okke Melse[+],[a, b, c] Samuel Sutiono[+],[a, b] Magdalena Haslbeck,[a, b] Gerhard Schenk,[d, e] Iris Antes[#],[b, c] and Volker Sieber*[a, b, d, f]

Dedicated to the memory of Prof. Dr. Iris Antes, a well-respected professor, supervisor, and friend, who sadly passed away shortly before the submission of this work.

The FeS cluster-dependent dihydroxyacid dehydratases (DHADs) and sugar acid-specific dehydratases (DHTs) from the ilvD/EDD superfamily are key enzymes in the bioproduction of a wide variety of chemicals. We analyzed [2Fe–2S]-dependent dehydratases *in silico* and *in vitro*, deduced functionally relevant sequence, structure, and activity relationships within the ilvD/EDD superfamily, and we propose a new classification based on their evolutionary relationships and substrate profiles. *In silico* simulations and analyses identified several key positions for specificity, which were experimentally investigated with site-directed and saturation mutagenesis. We thus increased the promiscuity of DHAD from *Fontimonas thermophila* (*Ft*DHAD), showing >10-fold improved activity toward D-gluconate, and shifted the substrate preference of DHT from *Paralcaligenes ureilyticus* (*Pu*DHT) toward shorter sugar acids (recording a six-fold improved activity toward the non-natural substrate D-glycerate). The successful elucidation of the role of important active site residues of the ilvD/EDD superfamily will further guide developments of this important biocatalyst for industrial applications.

Introduction

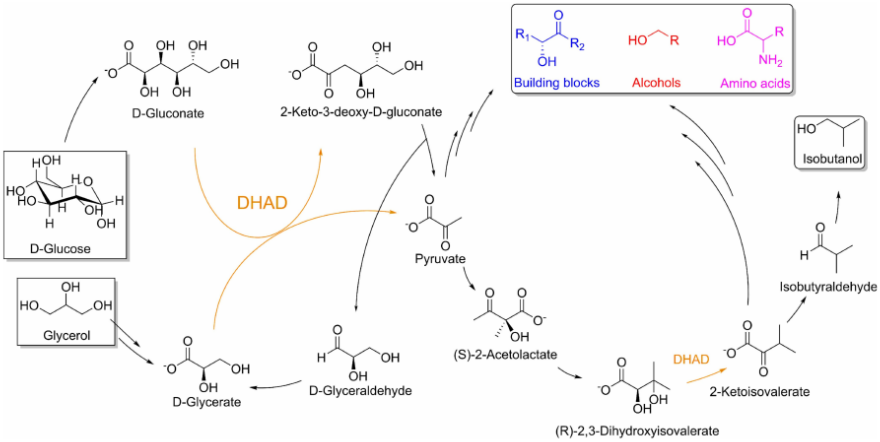
Production of higher chain alcohols, such as isobutanol, has gained major interest in the last decade because of their potential as biofuels with properties similar to gasoline. Isobutanol has a higher energy density and is less hygroscopic than ethanol, the traditional biofuel.[1] Bioproduction of isobutanol from fermentation can be achieved by a modified Ehrlich pathway, *i.e.* catabolism of branched chain amino acids (BCAA).[2] One key enzyme in this pathway is dihydroxyacid dehydratase (DHAD; Scheme 1), which catalyzes the dehydration of (*R*)-2,3-dihydroxyisovalerate (DHIV) to 2-ketoisovalerate (KIV). As an alternative to the fermentation approach, a cell-free system with minimized cofactor and enzyme requirements was developed.[3] This *in vitro* system is able to convert D-glucose to isobutanol utilizing only eight enzymes, in contrast to the 15 enzymes needed in the *in vivo* approach. The key enzyme in this system is a promiscuous DHAD from *Saccharolobus solfataricus* (*Ss*DHAD). In addition to the dehydration of DHIV to KIV, this enzyme also catalyzes the dehydration of the sugar acids D-gluconate to 2-keto-3-deoxy-D-gluconate (KDG) and D-glycerate to pyruvate (Scheme 1).[4] While the first reaction is part of the semi- or non-phosphorylative Entner-Doudoroff (ED) pathway, the latter is a non-natural reaction that proceeds at a very slow rate, thus serving as the major bottleneck in the cell-free system.[3,5] The dehydration of D-glycerate is also the key step in the valorization of glycerol, thus enhancing the significance of DHAD for applications in biotechnology.[6,7]

In contrast to DHADs, which use DHIV as preferred substrate, a closely related group of dehydratases (DHTs) prefer various sugar acids. D-xylonate DHT from *Caulobacter crescentus* (*Cc*XylDHT) and L-arabinonate DHT from *Rhizobium leguminosarum* (*Rl*ArDHT) have recently been characterized; both are candidate alternatives for *Ss*DHAD in the *in vitro* cascade.[8] However, while *Cc*XylDHT and *Rl*ArDHT are reactive toward long sugar acid substrates (*e.g.* D-gluconate) we recently showed that they are practically inactive toward D-glycerate.[9] In that study, several novel dehydratases (DHADs and DHTs) with

[a] O. Melse,[+] Dr. S. Sutiono,[+] M. Haslbeck, Prof. V. Sieber
Chair of Chemistry of Biogenic Resources
Campus Straubing for Biotechnology and Sustainability
Technical University of Munich
Schulgasse 16, 94315 Straubing (Germany)
[b] O. Melse,[+] Dr. S. Sutiono,[+] M. Haslbeck, Prof. I. Antes,[#] Prof. V. Sieber
SynBiofoundry@TUM
Technical University of Munich
Schulgasse 22, 94315 Straubing (Germany)
[c] O. Melse,[+] Prof. I. Antes[#]
TUM Center for Functional Protein Assemblies
Technical University of Munich
Ernst-Otto-Fischer-Straße 8, 85748 Garching (Germany)
[d] Prof. G. Schenk, Prof. V. Sieber
School of Chemistry and Molecular Biosciences
The University of Queensland
68 Cooper Road, 4072 St. Lucia (Australia)
[e] Prof. G. Schenk
Sustainable Minerals Institute
The University of Queensland
47 Staff House Road, 4072 St. Lucia (Australia)
[f] Prof. V. Sieber
Catalytic Research Center
Technical University of Munich
Ernst-Otto-Fischer-Straße 1, 85748 Garching (Germany)
E-mail: sieber@tum.de
[+] These authors contributed equally to this work.
[#] Deceased in 2021.

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cbic.202200088

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing



Scheme 1. Minimized synthetic pathway for the production of isobutanol from D-glucose and glycerol enabled by *Ss*DHAD (orange). Furthermore, the dehydration products of the *Ss*DHAD-catalyzed reaction, *i.e.* pyruvate and 2-ketoisovalerate, can be converted into additional chemical building blocks, alcohols, and amino acids (shown in blue, red and magenta, respectively).

activity toward D-glycerate were described, with the DHAD from *Fontimonas thermophila* (*Ft*DHAD) and the DHT from *Paralcaligenes ureilyticus* (*Pu*DHT) being the most promising. Both showed up to 50-fold higher activity toward D-glycerate when compared to *Ss*DHAD. Further substrate profiling demonstrated however that *Ft*DHAD, in contrast to *Cc*XylDHT and *Rl*ArDHT, is significantly less active toward longer sugar acids (such as D-gluconate), while *Pu*DHT is not active on branched-chain substrates such as DHIV, in contrast to *Ss*DHAD (Table 1).[9]

DHTs and DHADs belong to the ilvD/EDD (**i**soleucine, **l**eucine, **v**aline **D**ehydratase/**E**ntner-**D**oudoroff **D**ehydratase) superfamily of enzymes.[4,8,10] To date, there is no consistent naming system for enzymes from this superfamily, which complicates their comparative analysis. Most of the characterized representatives are tetrameric, containing four identical monomers arranged as a dimer of homodimers. Each dimer contains two active sites, located at the dimer interface (Figure 1). Despite the central role of these enzymes in metabolism and potential in biotechnology, X-ray structures are currently only available for a few of these enzymes including the DHADs from *Arabidopsis thaliana* (*Ath*DHAD; PDB: 5ze4)[11] and *Mycobacterium tuberculosis* (*Mt*DHAD; PDB: 6ovt),[12] and the DHTs *Cc*XylDHT (PDB: 5oyn)[13] and *Rl*ArDHT (PDB: 5j84/5j85).[14] This paucity of available structures limits both in-depth comparative functional studies among members of the ilvD/EDD superfamily and the (semi-)rational engineering of variants with properties suitable for industrial applications. No *in silico* studies of DHADs or DHTs belonging to the ilvD/EDD superfamily have yet been reported, partially due to the limited availability of experimental structures, but also because of the complex nature of the catalytic site, which requires both an FeS cluster ([2Fe–2S] or [4Fe–4S]) and a nearby Mg^{2+} ion (Figure 1).[13,14] The complex electronic properties of the FeS cluster and the highly polarized binding site are difficult to parameterize and simulate, making computational studies challenging

Table 1. Experimental specific activities of DHTs and DHADs toward sugar acid substrates of varying size and the branched chain acid DHIV.

	V (U/mg)[e] D-glycerate	L-threonate	D-xylonate	D-gluconate	DHIV
*Pu*DHT[a]	0.31 ± 0.03	5.71 ± 0.98	19.86 ± 1.06	48.23 ± 1.18	n.a.
*Rl*ArDHT[b]	< 0.01 ± 0.00	n.d.	6.19 ± 0.76	7.65 ± 0.71	n.d.
*Cc*XylDHT[b]	< 0.01 ± 0.00	n.d.	25.65 ± 0.67	47.91 ± 2.86	n.d.
*Ft*DHAD[c]	0.96 ± 0.07	2.23 ± 0.17	0.36 ± 0.01	n.a.	12.92 ± 0.63
*Mt*DHAD[12]	n.d.	n.d.	n.d.	n.d.	1.87 ± 0.06
*Ath*DHAD[17]	n.d.	n.d.	n.d.	n.d.	Active[d]
*Ss*DHAD[c]	0.02 ± 0.00	0.03 ± 0.00	2.18 ± 0.03	0.77 ± 0.04	0.75 ± 0.03

[a] Activity determined in this study at 30 °C. [b] Activity toward D-glycerate was taken from Sutiono *et al.* (2020)[9] and that toward D-xylonate and D-gluconate from Andberg *et al.* (2016).[8] [c] Activities in this study were determined at 50 °C. The loading of [Fe–S] was not determined. Thus, although the values of the activities for the DHADs may be higher than reported here, the substrate preference will remain the same. [d] Activity was determined, but no absolute value provided in cited study. [e] Activity of *Pu*DHT, *Ft*DHAD and *Ss*DHAD was determined with a substrate concentration of 25 mM. n.d. = activity was not determined in the previous studies; n.a. = activity < 0.01 U/mg.

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

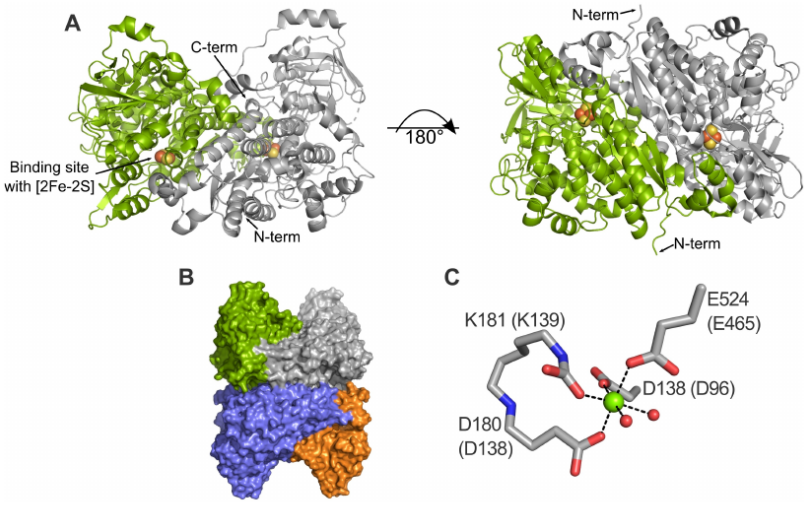


Figure 1. Oligomerization of [2Fe–2S]-dependent dehydratases. The figures were created based on a DHAD from *M. tuberculosis* (*Mt*DHAD, PDB: 6ovt). The single monomers are shown as respective green and grey cartoon, forming a dimer (A). The termini and the binding site are indicated in the figure. Two dimers pack together, forming a tetramer (B). The octahedral coordination of the Mg^{2+} ion is completed by water molecules in the X-ray structure (C). The alignment numbers (see Figure S1) are used for residue numbering; residue numbers from the original sequence of *Mt*DHAD are shown in brackets.

and time-consuming. There are also no bound substrates in any of the available DHAD or DHT X-ray structures. Therefore, molecular docking simulations are required to identify likely binding poses of the substrates, as well as residues that play an important role in substrate binding and catalysis. Such residues may represent mutational hotspots to engineer DHAD/DHT variants with tailored properties.

In this work, we introduce a new classification system, which will allow more straightforward comparison among members of the ilvD/EDD superfamily based on their substrate profiles and evolutionary relationship, in particular the [2Fe–2S]-dependent dehydratases. Furthermore, we used the available crystal structures of DHADs and DHTs to build homology models of *Ss*DHAD, *Ft*DHAD and *Pu*DHT. These models, together with computational simulations, allowed predictions of the origin of the substrate selectivity of these enzymes. We used this insight to rationally design variants of these three enzymes with improved activity toward substrates of interest. The success of our predictions was highlighted in the H653F variant of *Pu*DHT (residue numbering according to the sequence alignment; Figure S1), which has a six-fold higher activity toward D-glycerate than its wild-type counterpart. This study sheds light on the sequence, structure and activity relationship within the ilvD/EDD superfamily and thus facilitates targeted bioengineering studies to design optimized biocatalysts for industrially relevant biotransformations.

Results and Discussion

Substrate profiles and phylogenetic relationship: New classification of DHADs and DHTs

In previous work we identified dehydratases that are active on D-glycerate and partially characterized more than 20 enzymes.[9] We focused primarily on the dehydratases that contain a [2Fe–2S] center because of their relative stability in the presence of oxygen.[15] These enzymes can be grouped into two clusters based on their substrate profiles. The first cluster consists of DHTs, which are active toward longer sugar acids, and the second cluster contains DHADs, which have a preference for branched dihydroxyacids such as DHIV. We used one enzyme from each cluster as model enzyme in this work, namely *Pu*DHT and *Ft*DHAD.[9] We included *Ss*DHAD in our study as this enzyme was reported to be active on both substrate classes.[4] We characterized these enzymes under similar experimental conditions, and compared the results to other known dehydratases (Table 1). In the assays, sugar acids with increasing chain length (*i.e.* from D-glycerate *via* L-threonate and D-xylonate to D-gluconate) and the branched DHIV were used as substrates. The concentration for each substrate was 25 mM, which is well-above the K_M of previously characterized DHADs and DHTs.[4,8,9,12] Furthermore, we focused primarily on relative activities rather than absolute values in order to compare the change in the substrate profiles of the investigated enzymes.

In agreement with our previous study, *Pu*DHT is most active toward D-gluconate, with the activity decreasing as the chain length of the sugar acid substrate decreases; no activity toward

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

DHIV is recorded (Table 1). In contrast, *Ft*DHAD has virtually no activity toward D-gluconate, but is active toward shorter sugar acid substrates, in particular L-threonate. Maximum activity for *Ft*DHAD is recorded with its biological substrate, DHIV, more than five-times faster than that measured with L-threonate. The substrate preference of *Ss*DHAD lies somewhere in between those of *Pu*DHT and *Ft*DHAD with the relatively large sugar acid D-xylonate being the most preferred reactant, but the enzyme displaying significant activity for DHIV as well.[16] To explore the evolutionary relationship of these dehydratases further, we constructed a phylogenetic tree based on the sequences of all reported DHADs and DHTs in the literature (Figure 2). Both *Ft*DHAD and *Ss*DHAD cluster with other DHADs such as the enzymes from *A. thaliana* (*Ath*DHAD) and *M. tuberculosis* (*Mt*DHAD). However, *Ss*DHAD appears to have diverged from other DHADs early, maintaining larger similarities to DHTs. *Pu*DHT clusters closely with other known DHTs, including *Cc*XylDHT and *Rl*ArDHT.

Although DHADs and DHTs all belong to the ilvD/EDD superfamily, only DHADs are associated with a function in BCAA biosynthesis (due to their activity toward DHIV). DHTs, with preference for sugar acids, are key enzymes in the oxidative pentose (OP) pathways (Dahms and Weimberg pathways).[18,19] *Ss*DHAD, on the other hand, appears to occupy a unique evolutionary position, demonstrating significant activity toward both sugar acids and DHIV, and is thus believed to play a role in the BCAA, ED and OP pathways. However, strictly speaking, since all of these substrates contain a hydroxyl group at the α- (C2) and β-positions (C3) of an acid substrate, all of these enzymes are, in fact, dihydroxyacid dehydratases, *i.e.* DHADs. Thus, to make a clearer distinction, we propose another naming scheme, where enzyme homologous to *Pu*DHT, *Cc*XylDHT and *Rl*ArDHT are named **s**ugar **a**cid **deh**ydra**t**ases (SADHTs), and enzymes homologous to *Ft*DHAD, which are predominantly active toward DHIV, are named **b**ranched **c**hain **a**cid **deh**ydra**t**ases (BCADHTs). Finally, enzymes that display comparable activities toward sugar and branched chain acid substrates (such as *Ss*DHAD) are labelled as **p**romiscuous **a**cid **deh**ydra**t**ases (PADHTs). To date, *Ss*DHAD is the only PADHT that has been characterized but the phylogenetic inference (Figure 2) suggests that enzymes such as *Sto*DHAD from *Sulfurisphaera tokodaii* and *Ms*DHAD from *Metallosphaera sedula* should also have a promiscuous substrate profile.

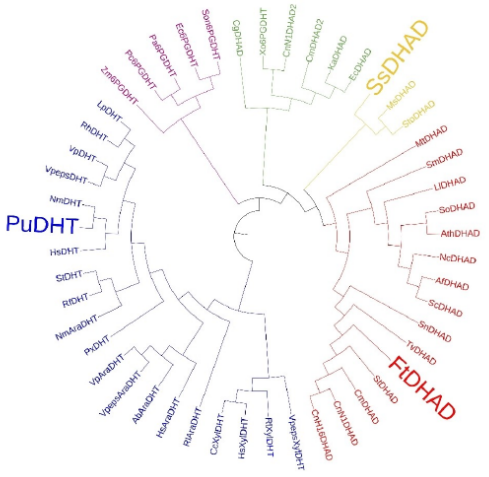


Figure 2. Phylogenetic tree of characterized and selected putative enzymes from the IlvD/EDD superfamily. The blue branch represents sugar acid dehydratases (SADHTs) and the red branch represents DHADs active on branched chain acids (BCADHTs). The yellow branch represents promiscuous DHADs active on sugar acids and branched chain acids (PADHT). *Sto* and *Ms*DHADs have not yet been characterized but were included in the tree to highlight the PADHT clade. The green branch represents DHADs, which contain a [4Fe-4S] cluster, and the purple branch represents 6-phosphogluconate dehydratases. The tree was constructed using MAFFT with the default method and visualized using iTOL.[20,21]

In silico modeling of representative DHADs

To gain structural insights relevant to the engineering of DHADs with optimized catalytic properties, homology models of representative DHADs for each of the three classes were generated (*i.e. Pu*DHT, *Ft*DHAD and *Ss*DHAD). The crystal structures of *Ath*DHAD, *Mt*DHAD, *Cc*XylDHT and *Rl*ArDHT were used as templates. A multiple sequence comparison was performed to align all templates and target proteins (Figure S1). Since residues sharing an alignment position are located in the same three-dimensional space in the protein, the alignment positions allow easier identification and comparison of residues based on their location in the protein. Therefore, we used the alignment positions of amino acids, rather than their position in the sequence throughout the remainder of this study. The correlation between sequence and alignment positions for relevant residues is shown in Table S1. The homology models (Figure S2) scored well during the QMEAN[22] quality estimation analysis, with QMEAN4 Z-scores ranging from −1.34 to −0.62 (Table S2), indicating a high quality for the homology models (see the Computational Biology section in Supporting Information for further details).

Two models for the reaction mechanism employed by DHADs and DHTs were recently proposed. Rahman *et al.*,[14] based on their studies with *Cc*XylDHT, suggested a mechanism for the dehydration reaction where the proton from the C2 atom of the substrate (Scheme 1) is removed by the deprotonated alkoxide side chain of the serine residue in alignment position 552 (see Figure 3 for details). The resulting carbanion is stabilized by the Mg^{2+} in the active site, leading to a weakening of the C3–O bond. Subsequently, one of the iron atoms of the [2Fe–2S] cluster (Fe2) promotes the abstraction of the hydroxyl group from C3, which triggers the tautomerization of the product to its keto form. This mechanistic model is supported by mutagenesis data that demonstrate the essential role of Ser552, as well as Mössbauer and EPR spectra that highlight the significance of the Lewis acid behavior of the [2Fe–2S] cluster in the catalytic cycle.[11–13] More recently, an alternative reaction mechanism was proposed for the DHIV dehydration catalyzed

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

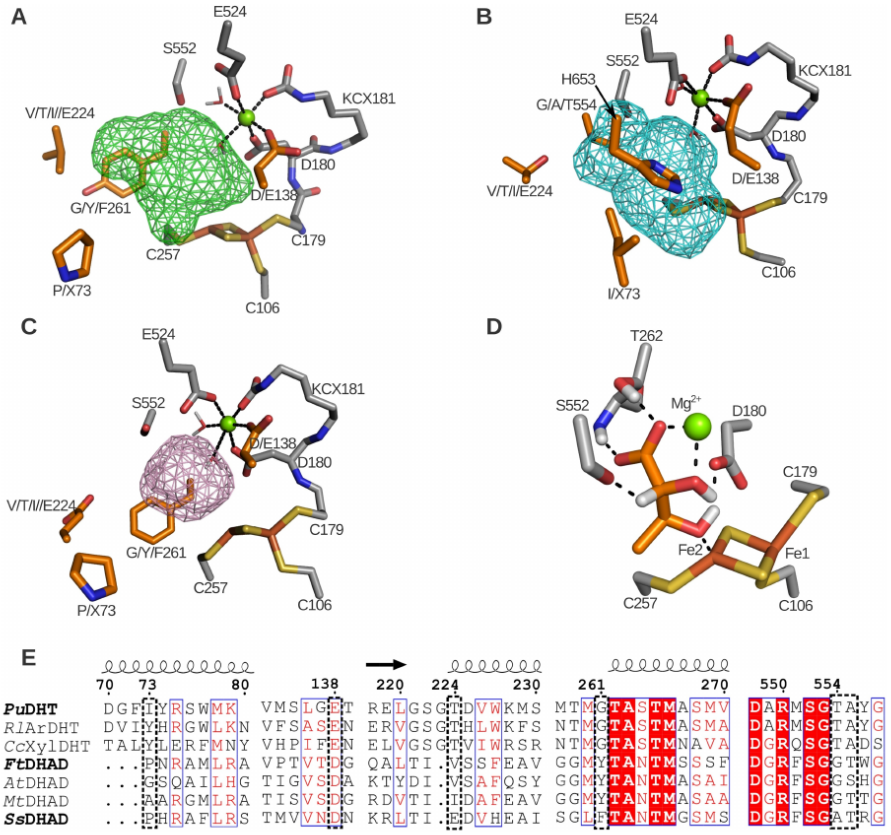


Figure 3. Structural representation of the binding site of *Ft*DHAD (A), *Pu*DHT (B) and *Ss*DHAD (C). The binding pocket is visualized by the mesh surface and residues identified as mutation hotspots and other relevant binding site residues are shown as orange and grey sticks, respectively. DHIV docked in *Ft*DHAD is used to visualize the conserved binding mode of dihydroxyacids in DHADs (D). Alignment numbers are used for the residue numbering, and amino acid side chains are indicated using their single letter abbreviation, except KCX, which stands for carboxylated lysine; a singular X indicates any amino acid. Note that the amino acid at alignment position 73 originates from the other monomer. The sequence alignment of the hotspot regions is shown (E), where the numbers above the sequences indicate the alignment numbers, and helices and arrows indicate the secondary structure of that corresponding sequence (alpha-helix or beta-sheet, respectively). The dotted boxes highlight the mutation hotspots.

by a DHAD from the cyanobacterium *Synechocystis* sp. PCC 6803 (*Sn*DHAD).[23] In this model, the C2 hydrogen is abstracted by a base (possibly a water molecule coordinated by Asp180). Subsequently, Ser552 donates a proton to 3-OH and thereby facilitates elimination of this hydroxyl group, releasing a water molecule. We performed docking simulations (see below) of relevant substrates in the homology models of *Ss*DHAD, *Ft*DHAD and *Pu*DHT, which largely support the reaction mechanism proposed by Rahman and colleagues. All of our docking simulations indicate that Ser552 is in close proximity to the C2 hydrogen, Fe2 is near C3–OH and the Mg^{2+} ion is close to C2–OH, where it can stabilize the negative charge of the carbanion (Figure 3).

Hotspot identification

From the multiple sequence alignment (Figure S1), a number of conserved residues were identified within each of the three DHT classes (*i.e.* SADHT, BCADHT and PADHT), including the three cysteine residues that coordinate the two iron atoms of the FeS cluster (Cys106, Cys179 and Cys257), and the catalytic serine residue (Ser552), but there are some interesting variations. For example, the motif around the second cysteine ligand of the [2Fe–2S] cluster (Cys179; underlined) is identical for SA- and PADHTs (*i.e.* GCDKTT; Figure S1) but is replaced by GCDKNM in BCADHTs. However, in the motif surrounding the third cysteine ligand of the FeS cluster (Cys257), PA- and BCADHTs are well conserved (GTCSG *vs* GACGG), whereas the

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

SADHTs are less so (GHCM/NT). Since representatives from the three DHT classes differ significantly in their substrate profiles while maintaining a similar overall structure and mechanism (Table 1), residues with high intra-class but low inter-class conservation are expected to be major determining factors (*i.e.* "hotspots") that control substrate specificity in these enzymes. Using the homology models of *Pu*DHT, *Ft*DHAD and *Ss*DHAD (Figure S2), an amino acid sequence comparison (Figure S1) and *in silico* substrate docking (Figure 3D), we identified several hotspots. Only residues in or within proximity of the active site were considered. The identified hotspots are (alignment positions): 73, 138, 224, 261, 554–555 (SGXX motif) and the C-terminal histidine of SADHTs (Figure 3A–C, E). We performed site-directed mutagenesis or saturation mutagenesis of these hotspots to validate our predictions, and to define the contribution of each of these hotspots on the substrate profile and enzyme activity (Table 2).

Influence of alignment position 73 on substrate specificity and catalytic efficiency

The residue located in alignment position 73 is on an N-terminal helix of a monomer of the DHAD/DHT homodimer (Figure S2); however, this residue forms part of the active site of the other monomer (Figures 3 and S2). Diverse residues of varying size are found at this position, ranging from proline in *Ss*DHAD and *Ft*DHAD, to the polar and bulky tyrosine in *Rl*ArDHT and *Cc*XyDHT (Figure 3). In *Pu*DHT an isoleucine is found in this position, and saturation mutagenesis did not result in variants with a significant change in specificity toward D-gluconate, L-threonate and DHIV (Figure S3). This observation stands in contrast to an earlier study on the structure of *Cc*XylDHT, which predicted that the residue in this position is important for the substrate preference of SADHTs.[13] The activity landscape of the *Ft*DHAD P73X library toward DHIV indicates that >70% of the variants show a significant decrease in activity in comparison to the wild-type enzyme (Figure S3). Decreased activity was also observed toward L-threonate. However, several variants in this library were observed that have almost >three-fold higher activity toward D-gluconate than wild-type *Ft*DHAD. Importantly, and in contrast to the wild-type enzyme, three of them display comparable activity toward DHIV and L-threonate (Figure S3). Since our main goal is to find enzymes which show enhanced substrate promiscuity, we selected one of these three variants for further studies. It contains the proline to glycine substitution at position 73, and has a >10-fold higher activity toward D-gluconate, while retaining ~64% of the activity toward DHIV when compared to wild-type *Ft*DHAD (Table 2 and Figure 4A). A possible explanation for this observed effect is a slight increase of binding site volume. Thus, the combined data demonstrate that the amino acid at position 73 plays an important role in altering the substrate preference of a BCADHT to that of a SADHT, but not *vice versa* (*i.e. Pu*DHT is not sensitive toward mutations in that position).

Influence of alignment position 138 on Mg^{2+} coordination

Apart from the FeS cluster, members of the ilvD/EDD superfamily also require a Mg^{2+} ion (or other divalent cations) in the active site for their catalytic function. The majority of the residues coordinating this ion are highly conserved and include D180 and E524, as well as a carboxylated lysine (KCX181) and two water molecules (Figure 1C). Some modest variability is observed with the ligand in alignment position 138, where an aspartate is present in BCADHTs but a glutamate in SA- and PADHTs (Figure 3E). Our docking studies reveal that sugar acid substrates replace the water molecules and coordinate the Mg^{2+} ion with one carboxyl oxygen and the C2–OH, thereby completing the octahedral coordination sphere of the metal ion

Table 2. Relative activity of *Pu*DHT, *Ft*DHAD and *Ss*DHAD variants against substrates with varying size. The highest activity observed in the wild-type enzyme was set as 100%. The sequence numbers are the alignment positions.[a]

	D-glycerate	L-threonate	D-xylonate	D-gluconate	DHIV
*Pu*DHT – WT	0.65 ± 0.06	11.90 ± 2.04	41.38 ± 2.20	100.00 ± 2.45	n.a.
E138D	n.a.	0.71 ± 0.06	1.23 ± 0.03	4.23 ± 0.42	n.a.
E138A	n.d.	n.d.	n.d.	0.05 ± 0.00	n.d.
T224E	n.a.	n.a.	n.a.	n.a.	n.a.
G261F	n.a.	n.a.	n.a.	n.a.	n.a.
G261Y	n.a.	n.a.	n.a.	n.a.	n.a.
T554G, A555T (SGGT)	n.a.	n.a.	n.a.	n.a.	n.a.
H653F (C-terminus)	4.02 ± 0.22	4.31 ± 0.93	5.56 ± 0.06	0.50 ± 0.02	n.a.
*Ft*DHAD – WT	7.44 ± 0.56	18.04 ± 1.30	2.81 ± 0.07	0.02 ± 0.00	100.00 ± 4.91
P73G	1.76 ± 0.1	4.68 ± 1.72	1.55 ± 0.14	0.22 ± 0.01	64.20 ± 7.59
D138E	0.10 ± 0.00	2.04 ± 0.13	0.15 ± 0.00	n.a.	16.55 ± 1.15
D138A	n.d.	n.d.	n.d.	n.d.	0.04 ± 0.00
V224E	0.02 ± 0.01	0.14 ± 0.01	0.03 ± 0.01	n.a.	0.60 ± 0.08
Y261G	0.03 ± 0.00	0.32 ± 0.00	0.11 ± 0.00	n.a.	1.03 ± 0.04
G554T, T555A (SGTA)	n.a.	0.08 ± 0.00	n.a.	n.a.	n.a.
*Ss*DHAD – WT	1.12 ± 0.00	1.58 ± 0.09	100.00 ± 1.54	35.5 ± 1.78	34.73 ± 1.16
D138E	0.92 ± 0.00	1.28 ± 0.02	11.53 ± 1.50	16.76 ± 2.30	22.54 ± 3.34
D138A	n.d.	n.d.	0.23 ± 0.01	n.d.	n.d.
A554T, T555A (SGTA)	n.a.	n.a.	n.a.	n.a.	n.a.

[a] n.d. = not determined; n.a. = relative activity < 0.01%.

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

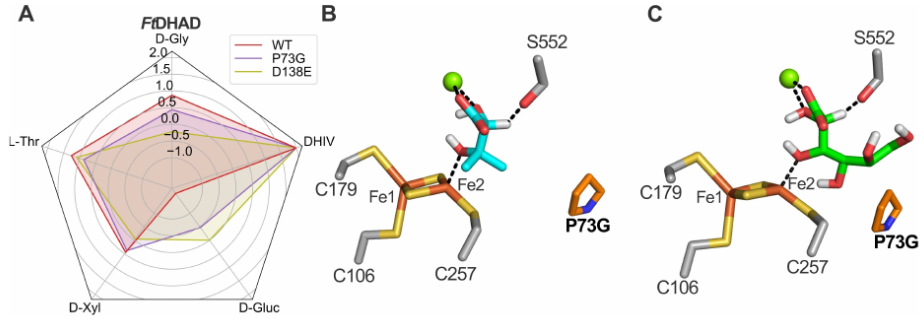


Figure 4. Effect of site-directed mutations on the substrate profile of *Ft*DHAD. Relative activities of *Ft*DHAD variants shown in a radar plot in logarithmic scale, with the highest observed activity set to 100% (A). Exact values are provided in Table 2. Docking of DHIV (B) and D-gluconate (C) in the active site of *Ft*DHAD. P73 is shown as orange sticks. Alignment numbers are used for the residue numbering.

(Figure 3D). To evaluate the effect of the non-conserved residue in position 138 on the catalytic properties of the dehydratases, we performed the following mutations: E138D and E138A in *Pu*DHT and D138E and D138A in *Ft*DHAD and *Ss*DHAD (Table 2). In all variants containing the alanine substitution a virtually complete loss of activity was observed for all substrates, likely to be due to impaired Mg^{2+} binding. The E138D mutation in *Pu*DHT had a similar effect, possibly because the shorter aspartate side chain is not able to interact effectively with the Mg^{2+}. Less dramatic is the effect of the D138E substitution in *Ft*DHAD and *Ss*DHAD, but in general a reduction of the activity was observed for each substrate tested. The *in silico* simulations could not provide a conclusive explanation for this effect, but the longer side chain of glutamate may lead to a shift of the Mg^{2+} ion away from the [2Fe–2S] cluster, and thereby prevent optimal substrate orientation. Importantly, while the residue in position 138 clearly plays an essential role in Mg^{2+} binding, none of the mutations in this position had a significant effect on the substrate specificity of the dehydratases.

The effect of a negative charge in position 224 in the active site

The residues in alignment position 224 are either a hydrophobic branched chain amino acid in BCADHTs, a hydrophilic threonine in SADHTs or a negatively charged glutamate in the PADHT *Ss*DHAD (Figure 3). The promiscuity of *Ss*DHAD is beneficial for the cell-free bioproduction of isobutanol (Scheme 1), a trait we aimed to establish in SADHTs and BCADHTs. We hypothesized that the negative charge at position 224 plays an important role in the promiscuity of PADHTs. Accordingly, T224E and V224E variants of *PuDHT* and *Ft*DHAD, respectively, were generated. This mutation, however, completely abolished the activity of *Pu*DHT toward all substrates tested (Table 2), and a similar effect was observed for *Ft*DHAD, although the mutant retained modest (<1%) activity for most of the substrates. Thus, substituting an amino acid at this position in *Pu*DHT and *Ft*DHAD into a negatively charged glutamate is not sufficient to introduce the promiscuity of *Ss*DHAD into SADHTs or BCADHTs.

Binding site volume (alignment position 261)

Another marked difference between the three classes of DHTs is observed in position 261. While glycine is present in the SADHTs, this small side chain is replaced by the much bulkier tyrosine and phenylalanine in the BCADHTs and PADHTs, respectively. An analysis of the active sites of the homology models for *Pu*DHT, *Ft*DHAD and *Ss*DHAD (Figure 3) reveals that these residues point inside the active site of these enzymes. Consequently, the SADHTs have a considerably larger active site volume than the other DHTs. Hence, in order to test if the active site volume plays an important role in their substrate selection, G261F and G261Y mutations were introduced in *Pu*DHT and the Y261G in *Ft*DHAD (we did not perform the corresponding mutation in *Ss*DHAD, *i.e.* F261G, as our primary interest was in broadening the substrate scope of *Pu*DHT and *Ft*DHAD to a level similar to that of *Ss*DHAD). The substitutions Y261G in *Ft*DHAD resulted in drastic loss of activity, while the substitutions G261F and G261Y in *Pu*DHT both completely abolished the activity toward all substrates. Thus, changing the binding volume of *Ft*DHAD and *Pu*DHT by single substitution is not enough to alter their substrate preferences.

The role of the SGXX motif on the catalytic properties (alignment positions 552–555)

The crystal structure of the SADHT *Cc*XylDHT highlighted the crucial role of the highly conserved serine residue in position 552 in the reaction mechanism of ilvD/EDD dehydratases (see above).[13] It has been proposed that the deprotonated side chain of Ser552 initiates the catalytic cycle by abstracting a proton from the C2 atom of the substrate. Furthermore, structural information from another SADHT, *Rl*ArDHT, also

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

ascribes an important role for the threonine at position 554 in the mechanism, stabilizing the catalytic serine during the reaction.[14] In BCADHTs and PADHTs a glycine or alanine, respectively, is located in this position. However, in these enzymes a threonine is present in position 555 (SGGT and SGAT, respectively), whereas an alanine is present in SADHTs instead (SGTA). In order to assess the importance of the residues in these positions for the mechanism of these ilvD/EDD dehydratases, a series of mutants were generated (Tables 2 and S3). For *Pu*DHT, the threonine in position 554 is indeed crucial for the activity toward its preferred substrate, D-gluconate, with (>0.5% remaining upon substitution with glycine or valine). Interestingly, the alanine in position 555 also plays an important role; upon its replacement by another threonine only ~0.7% of the activity remains. Furthermore, replacing the wild-type Thr554-Ala555 sequence in *Pu*DHT with the Gly554-Thr555 sequence present in BCADHTs leads to a completely inactive double mutant (irrespective of the substrate used). A similar trend was observed for *Ft*DHAD, where the conversion of its native SGGT motif to the SGTA motif present in SADHTs led to the virtually complete inactivation of the enzyme (again irrespective of the substrates used; Table 2). The threonine in this motif is also important for the function of *Ft*DHAD, but contrasting the observations made for *Pu*DHT, its replacement by a non-polar amino acid does not completely abolish the catalytic function of the enzyme (the SGGA variant retains over 18% of the activity). However, in both enzymes it is essential that a non-polar side chain is located next to this threonine - replacing it with another threonine is highly detrimental to the activities of these enzymes (Table S3). The same observations were also made for corresponding mutations in the PADHT *Ss*DHAD. Hence, while functionally essential, residues of the SGxx motif do not play an important role in determining the substrate preference of these enzymes. This interpretation is supported by *in silico* mutation studies, suggesting that a potential interaction between the substrate and Thr554 in *Pu*DHT may also exist in the G554T–T555A double mutant of *Ft*DHAD (*i.e.* a variant with the SGTA motif present in DHT). Since the residues of the SGxx motif are located on a rather flexible loop, we performed MD simulations with wild-type enzymes and variant forms with inverted motifs to probe if the mutations alter the loop dynamics. However, no significant difference was observed (data not shown). It is thus likely that long-range interactions, possibly mediated via an extensive hydrogen bond network, may play an important role in the mechanism. An in-depth analysis of such effects is beyond the scope of the current study.

Altering the substrate specificity of DHADs toward D-glycerate

The hotspots identified in this study play an important role in the function of the ilvD/EDD dehydratases, and in some cases the substrate preference could be broadened through targeted mutations (*e.g.* the P73G mutation which widens the substrate preference of *Ft*DHAD to accept D-gluconate; Table 2). However, none of the mutations in these hotspots altered the substrate preferences of the SA- and PADHTs. Furthermore, none of the variants displayed increased activity toward D-glycerate, an important substrate in biomanufacturing processes (Scheme 1). The crystal structures of *Cc*XylDHT and *Rl*ArDHT and the homology model of *Pu*DHT reveal that the conserved C-terminal histidine residue of the SADHTs extends toward the active site, located at the dimer interface. In contrast, in the available crystal structures of the BCADHTs *Ath*DHAD and *Mt*DHAD, as well as the homology models of *Ft*DHAD and *Ss*DHAD, the C-terminal residue is not conserved and is not in the vicinity of the active site (Figures S1 and S4). Docking simulations were thus performed with the homology model of *Pu*DHT and various sugar acid substrates to probe if the C-terminal histidine contributes to catalysis in SADHTs. We hypothesized that this residue may play a role in positioning sugar acid substrates in the active site, a hypothesis that is based on the observation that *Pu*DHT is more reactive toward larger sugar acids (Table 1). Indeed, the docking simulations demonstrate that both D-gluconate and D-xylonate are able to form a hydrogen bond with the C-terminal histidine via their C5–OH groups; L-threonate and D-glycerate are too short to form a similar H-bond (Figure S5). These findings were further supported by MD simulations with enzyme-D-gluconate and enzyme-L-threonate complexes. In these simulations, both C5–OH and C6–OH of D-gluconate alternately form a hydrogen bond with the terminal histidine residue of *Pu*DHT (Figures 5 and S6), an interaction that is conserved throughout the entire MD simulation. In contrast, MD simulations with the docked enzyme-L-threonate complex confirmed that no interactions between this substrate and the C-terminal histidine is formed. In order to substantiate these computational predictions, two saturation libraries for *Pu*DHT were generated. One library was designed to target the terminal histidine residue, and the other added an amino acid prior to this terminal histidine, thus extending the C-terminal end by one amino acid. The two libraries were screened using different sugar acids and DHIV as substrates. The extension of the C-terminal end showed a highly detrimental effect, completely inactivating the enzyme regardless of which substrate was used in the assays (Figure S7). Saturating the terminal histidine to other amino acids demonstrated a similar effect for the reaction with D-gluconate. However, one variant (H653F) showed an improved activity toward D-glycerate. While wild-type *Pu*DHT is minimally active toward D-glycerate (less than 1% when compared to D-gluconate), the variant is ~six-fold more reactive toward this substrate, and the ratio of D-gluconate/D-glycerate activity has improved >1000-fold in comparison to the wild-type enzyme (Table 2 and Figure 5A). Hence, the C-terminal end plays indeed an important role in the mechanism of SADHTs. Extending the length of the C-terminal end did not promote interactions with the smaller substrates, possibly because of steric clashes. However, replacing the native histidine by a slightly larger phenylalanine altered the size of the active site cavity sufficiently to exclude the large D-gluconate substrate and favor binding of smaller sugar acids.

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

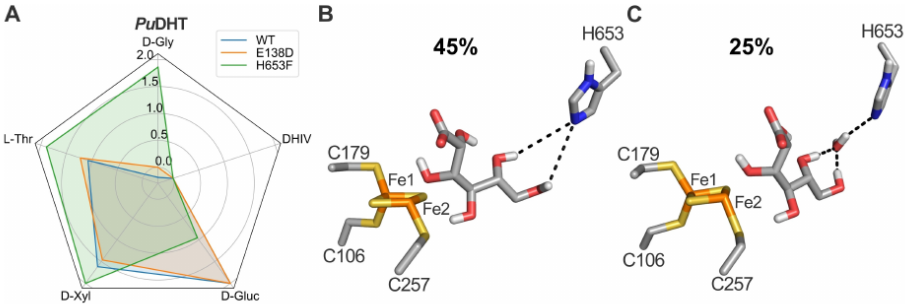


Figure 5. Effect of site-directed mutations on the substrate profile of *Pu*DHT. Relative activities of *Pu*DHT variants shown in a radar plot in logarithmic scale with the highest observed activity set to 100% (A). Exact values are provided in Table 2. Two preferred binding orientations for D-gluconate in the active site of *Pu*DHT were modelled (B, C), where the percentage indicates the frequency of the MD simulations in which the respective binding mode was observed. Alignment numbers (Figure S1) are used for the residue numbering.

Stereoselective substrate specificity of *Pu*DHT and *Ft*DHAD

We also probed the stereoselective substrate specificity of SADHTs and BCADHTs. SADHTs only accept sugar acid substrates that have the (*R*) and (*S*) configuration at positions C2 and C3, respectively.[8] Similarly, spinach DHAD (*So*DHAD), a BCADHT, has a preference for substrates with *(R)* configuration at position C2.[24] Here, we examined the stereoselective substrate specificity of *Pu*DHT and *Ft*DHAD using several C5 sugar acids with different configurations in positions C2 and C3, namely D-ribonate (*R*,*R*), D-arabinonate (*S*,*R*), D-xylonate (*R*,*S*) and D-lyxonate (*S*,*S*). *Pu*DHT, similar to other SADHTs, is only active toward D-xylonate, while the BCADHT *Ft*DHAD is equally active toward D-ribonate and D-xylonate (Figure 6). Molecular docking simulations using the homology models of *Pu*DHT and *Ft*DHAD demonstrate the *(R)* conformation at position C2 is essential as only in this stereoisomer both the OH group at C2 and a carboxyl oxygen at C1 are able to coordinate the catalytically essential Mg^{2+} ion. The docking simulations also indicate that D-ribonate can be accommodated in the active site of *Ft*DHAD, maintaining its orientation and distance to Ser552, which initiates the catalytic cycle by abstracting a proton from the substrate (see above and Figure S8). However, docking simulations were not able to rationalize the stricter stereochemical requirement of SADHTs at position C3.

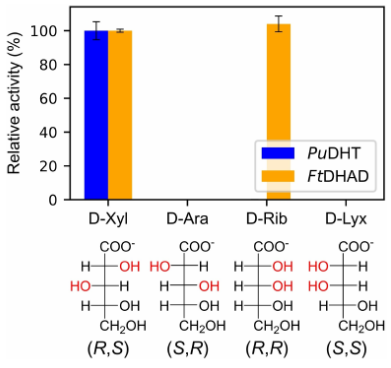


Figure 6. The effect of the stereochemistry at positions C2 and C3 on the substrate acceptance of SADHTs (represented by *Pu*DHT) and BCADHTs (represented by *Ft*DHAD). Exact values are provided in Table S4.

Conclusions

Dihydroxyacid dehydratases, in particular those that contain a [2Fe–2S] cluster in their active sites, encompass enzymes with broad substrate spectra. Three distinct classes (SADHTs, BCADHTs and PADHTs) can be discerned based on their phylogenetic relationship (Figure 2) and substrate profiles (Table 1). We built homology models of three representative enzymes (*Pu*DHT, *Ft*DHAD and *Ss*DHAD) from each class. Together with *in silico* simulations and an analysis of sequence conservation, we were able to predict several hotspots that may play an important role in the function of these enzymes. Site-directed mutations within these hotspots demonstrated that all of them affect the substrate selectivity and/or enzyme activity (Tables 2 and S3; Figures 4 and 5). We probed the catalytic roles of several of these hotspots in more detail. For example, the amino acid at alignment position 138 plays a role in coordinating the catalytically essential Mg^{2+} ion across all of these enzymes. Furthermore, their substrate promiscuity (or lack thereof) may be governed by three amino acids located at alignment positions 73, 224 and 261. Substituting the native proline at position 73 by a glycine (P73G) improved the substrate promiscuity of *Ft*DHAD; its P73G variant has >10-fold improved activity toward D-gluconate, while the activity toward DHIV (the biological substrate of the wild-type enzyme) remains high. We used structural and sequence information to engineer dehydratases with enhanced preference for the non-natural substrate D-glycerate, which is an intermediate in the bio-transformation of D-glucose and glycerol to various chemicals

ChemBioChem
Research Article
doi.org/10.1002/cbic.202200088
Chemistry Europe
European Chemical Societies Publishing

(Scheme 1). The C-terminal histidine was identified as relevant hotspot, as it stabilizes and orients larger sugar acid substrates in SADHTs. Replacing this histidine in *Pu*DHT by a bulkier phenylalanine indeed shifted the substrate preference of this enzyme toward shorter sugar acids, improving the activity toward D-glycerate ~six-fold. In summary, the combination of *in silico* and mutagenesis studies with representatives from the three classes of the ilvD/EDD superfamily provides a "roadmap" for the engineering of optimized dehydratases for biotransformation that are of great interest to the chemical industry.

Experimental Section

Enzyme production and site-directed mutagenesis: All dehydratases in this study, except the ones from *S. solfataricus* (*Ss*DHAD) were produced using *E. coli* BL21 (DE3) in Terrific Broth (TB) and purified as described previously.[9] *Ss*DHAD and its variants were produced using *E. coli* BL21 (DE3) in Autoinduction media and purified as well as activated as described previously.[4] Site-directed or saturation mutagenesis experiments were performed using either QuikChange, two-step PCR, or overlap extension PCR. More details are described in the Supporting Information.

Enzyme activity measurements: Kinetic parameters of the enzymes were measured using HPLC as described previously.[9] Assays with wild type and variant forms of *Pu*DHT were performed at 30 °C, while wild type and variant forms of *Ft*DHAD and *Ss*DHAD were assayed at 50 °C. More details are described in the Supporting Information.

Molecular modelling: Homology models were produced based on the templates with the highest sequence identity. The homology models were generated with Modeller and the resulting model with the best DOPE score was selected for further use. Molecular Mechanics parameters were generated for the [2Fe–2S] cluster and the deprotonated serine. Follow-up MD simulations were conducted with AMBER. Molecular docking simulations were performed with AutoDock, and *in silico* mutations were performed with PyMOL using Dunbrack's rotamer library. More details are described in the Supporting Information.

Supporting information: Supporting experimental and computational methodology, multiple sequence alignment and structural representations of the homology models and a conversion table between sequence and alignment numbers. Activity landscape of the saturation library of alignment position 73 and C-terminal residue, structural properties of the binding sites and substrate binding in DHAD (variants). Docked poses of D-xylonate and D-ribonate in DHAD. Activity data for produced DHAD (variants) and the primers and auxiliary enzymes used in this study.

Accession codes of enzymes discussed in this study (UniProtKB); *Pu*DHT: A0A4R3LQ44, *Ft*DHAD: A0A1I2J0Y3, *Ss*DHAD: Q97UB2, *Rl*ArDT: B5ZZ34, *Mt*DHAD: A0A0E8UWV6, *At*DHAD: Q9LIR4, *Cc*XylDHT: Q9A9Z2.

Acknowledgements

O. M. and I. A. thank the Deutsche Forschungsgemeinschaft for financial support via SFB 1035, project A10 and CIPSM. S. S. and V. S. thank German Federal Ministry for Education and Research (BMBF) through HotSysAPP project (Grant No. 031L0078F) and BINOM project (Grant No. 031B1055). We thank Franziska Totzeck, Sarah Fink, Silvia Bergt and Sophia Zhou for useful insights regarding the modelling and mutation hotspot prediction process. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords: biocatalysts · bioinformatics · dehydratases · enzyme catalysis · structure-activity relationships

[1] M. R. Connor, J. C. Liao, *Curr. Opin. Biotechnol.* **2009**, *20*, 307–315.
[2] S. Atsumi, T. Hanai, J. C. Liao, *Nature* **2008**, *451*, 86–90.
[3] J. K. Guterl, D. Garbe, J. Carsten, F. Steffler, B. Sommer, S. Reiße, A. Philipp, M. Haack, B. Rühmann, A. Koltermann, et al., *ChemSusChem* **2012**, *5*, 2165–2172.
[4] J. M. Carsten, A. Schmidt, V. Sieber, *J. Biotechnol.* **2015**, *211*, 31–41.
[5] T. J. Gmelch, J. M. Sperl, V. Sieber, *Sci. Rep.* **2019**, *9*, 11754.
[6] Z. Li, J. Yan, J. Sun, P. Xu, C. Ma, C. Gao, *Commun. Chem.* **2018**, *1*, 1–7.
[7] C. Gao, Z. Li, L. Zhang, C. Wang, K. Li, C. Ma, P. Xu, *Green Chem.* **2015**, *17*, 804–807.
[8] M. Andberg, N. Aro-kärkkäinen, P. Carlson, M. Oja, S. Bozonnet, M. Toivari, N. Hakulinen, M. O. Donohue, *Appl. Microbiol. Biotechnol.* **2016**, *100*, 7549–7563.
[9] S. Sutiono, M. Teshima, B. Beer, G. Schenk, V. Sieber, *ACS Catal.* **2020**, *10*, 3110–3118.
[10] S. E. Egan, R. Fliege, S. Tong, A. Shibata, R. E. Wolf, T. Conway, *J. Bacteriol.* **1992**, *174*, 4638–4646.
[11] Y. Yan, Q. Liu, X. Zang, S. Yuan, U. Bat-Erdene, C. Nguyen, J. Gan, J. Zhou, S. E. Jacobsen, Y. Tang, *Nature* **2018**, *559*, 415–418.
[12] G. Bashiri, T. L. Grove, S. S. Hegde, T. Lagautriere, G. J. Gerfen, S. C. Almo, C. J. Squire, J. S. Blanchard, E. N. Baker, *J. Biol. Chem.* **2019**, *294*, 13158–13170.
[13] M. M. Rahman, M. Andberg, A. Koivula, J. Rouvinen, N. Hakulinen, *Sci. Rep.* **2018**, *8*, 23–25.
[14] M. M. Rahman, M. Andberg, S. K. Thangaraj, T. Parkkinen, M. Penttila, A. Koivula, J. Rouvinen, N. Hakulinen, *ACS Chem. Biol.* **2017**, *12*, 1919–1927.
[15] D. H. Flint, M. H. Emptage, M. G. Finnegan, W. Fu, M. K. Johnson, *J. Biol. Chem.* **1993**, *268*, 14732–14742.
[16] J. M. Sperl, J. M. Carsten, J. K. Guterl, P. Lommes, V. Sieber, *ACS Catal.* **2016**, *6*, 6329–6334.
[17] H. Gao, T. Azam, S. Randeniya, J. Couturier, N. Rouhier, M. K. Johnson, *J. Biol. Chem.* **2018**, *293*, 4422–4433.
[18] A. S. Dahms, *Biochem. Biophys. Res. Commun.* **1974**, *60*, 1433–1439.
[19] R. Weimberg, *J. Biol. Chem.* **1961**, *236*, 629–635.
[20] K. Katoh, D. M. Standley, *Mol. Biol. Evol.* **2013**, *30*, 772–780.
[21] I. Letunic, P. Bork, *Nucleic Acids Res.* **2019**, *47*, W256–W259.
[22] P. Benkert, M. Biasini, T. Schwede, *Bioinformatics* **2011**, *27*, 343–350.
[23] P. Zhang, B. S. MacTavish, G. Yang, M. Chen, J. Roh, K. R. Newsome, S. D. Bruner, Y. Ding, *ACS Chem. Biol.* **2020**, *15*, 2281–2288.
[24] M. C. Pirrung, C. P. Holmes, D. M. Horowitz, D. S. Nunn, *J. Am. Chem. Soc.* **1991**, *113*, 1020–1025.

Manuscript received: February 9, 2022
Revised manuscript received: March 2, 2022
Accepted manuscript online: March 9, 2022
Version of record online: March 23, 2022

ChemBioChem

Supporting Information

Structure-Guided Modulation of the Catalytic Properties of [2Fe–2S]-Dependent Dehydratases

Okke Melse[+], Samuel Sutiono[+], Magdalena Haslbeck, Gerhard Schenk, Iris Antes[#], and Volker Sieber*

Contents

1. Methods

Molecular Biology

Generation of mutant and production of dehydratases variants

The native sequence (WT) of *Ss*DHAD, *Ft*DHAD, and *Pu*DHT were cloned to pET28 as described in previous studies. All mutants were generated using either standard QuikChange protocol, or two-step PCR (PCR without annealing temperature), or overlap extension PCR. The list of primers is presented in **Table S7**. The plasmid containing mutant genes were sent for sequencing to confirm the mutation. The correct plasmid was then used to transform *E. coli* BL21 (DE3). Expression of all variants of *Ft*DHAD and *Pu*DHT was done as previously described. All the variants contain hexahistidine tag in the N-terminus, thus were purified using Ni NTA column and the final buffer was exchanged to 50 mM HEPES pH 8 as described previously.[1] For *Ss*DHAD WT and its respective variants, the expression was performed in autoinduction media at 25 °C as described previously.[2] Due to the instability of *Ss*DHAD in the presence of imidazole, the enzymes were heat purified in the presence of β-mercaptoethanol as described previously.[2] The final buffer was exchanged to 50 mM HEPES pH 8.

Kinetic characterization

All sugar acids were either purchased or synthesized as described previously.[1] Dihydroxy isovalerate (DHIV) was synthesized from pyruvate using two step enzymatic approach (acetolactate synthase from *Bacilus subtilis* (*Bs*ALS), ketolacid reductoisomerase from *Meiothermus ruber* (*Mt*KARI) and formate dehydrogenase from *Candida boidinii* (*Cb*FDH) for cofactor recycling). The yield and concentration of DHIV is estimated by HPLC.

All enzyme activities were measured using HPLC as described previously.[1] In short, all sugar acids and 2-keto-3-deoxy sugar acids were separated using an anion exchange column (Metrosep A Supp 16-250, Metrohm, Germany). DHIV and 2-ketoisovalerate (KIV) were separated using an ion-exclusion column (Rezex ROA-Organic Acid H+(8%), Phenomenex, Germany). Both HPLC systems were monitored using UV detector.

Final concentration of each substrate used was 25 mM. The activity was calculated by following the formation of corresponding product over time up to 24 h. *Ss*DHAD and *Ft*DHAD activities were determined at 50 °C, while *Pu*DHT was measured at 30 °C. The lowest measurable activity is 0.034 mU/mg, recorded at the lowest concentration of the standard (0.1 mM) by HPLC after a 24 h reaction with 2.5 mg/ml enzyme (maximum enzyme used).

Screening of dehydratase libraries

Expression of the dehydratases libraries (*Ft*_P73X, *Pu*_I73X, *Pu*_H653X, *Pu*DHT_N652_H653insX) were performed in 96-Deep Well Plates (DWP). In short, *E. coli* BL21 (DE3) harbouring plasmids expressing respective libraries were grown on LB-agar containing kanamycin (50 µg/ml). Single colony was inoculated in each well of 96-DWP containing autoinduction media with kanamycin (100 µg/ml). The culture was grown into saturation at 25 °C for 36 h by shaking the DWP at 1000

rpm. The culture was harvested by transferring 200 µl of cell culture to 96 U-bottom plate and centrifuging the plate at 4000 xg for 15 min. After the supernatant was decanted, the cell pellet was stored at -80 °C until further use.

The cell was lysed by transferring 200 µl of lysis buffer (50 mM HEPES pH 8, 5 mM $MgCl_2$, and lysozyme 1 mg/ml) and incubating the plate at 1000 rpm, 40 °C for 1 h. The cell debris was pelleted by centrifugation at 4000 xg for 15 min. Activity toward D-glycerate was measured using coupled assay using lactate dehydrogenase (LDH) for porcine heart (purchased from Sigma) and NADH. Activity toward L-threonate and DHIV was measured using coupled assay using the thermostable variant of *Lactococcus lactis* ketoacid decarboxylase (7M.D) and alcohol dehydrogenase from *Bacillus stearothermophilus* (*Bst*ADH) and NADH. Activity toward D-gluconate was measured by coupled assay using aldolase from *Picrophilus torridus* (*Pt*KdgA) and LDH and NADH. The oxidation of NADH was monitored at 340 nm. All coupling enzymes were used in excess. Wild type (WT) enzyme of *Ft*DHAD was used as control for *Ft*_P73X, while *Pu*DHT WT was used for (*Pu*_I73X, *Pu*_H653X, *Pu*DHT_N652_H653insX). List of auxiliary enzymes are presented in **Table S8**.

Computational Biology

Template identification and homology modelling

To identify potential templates for the homology modelling, a protein BLAST[3] search was performed against the PDB database on the sequences of the target enzymes (*Ft*DHAD, *Pu*DHT and *Ss*DHAD). This resulted in four possible templates (**Table S5**). A multiple sequence alignment (MSA) was performed with Clustal Omega[4] using these four template DHADs and the target DHADs, and visualized using ESPript.[5] From the templates within the same DHAD class (as described in the main text) as the target DHAD, the two structures with the highest sequence identity to the target DHAD were selected to serve as template during the homology modeling. Only the chains showing the active loop conformation, *i.e.* the loop containing the catalytic serine were used. For *At*DHAD, this active loop conformation was modelled inside the structure based on the structure of *Mt*DHAD using Modeller release 9.18[6] before the generation of the homology models. Since we generate homology models of the dimer, template structures with only one monomer were duplicated and manually converted into a dimer structure. The homology models were generated using Modeller; five different models were generated for each target DHAD, and the top-ranked model based on the DOPE score was selected. The binding sites of these homology models were further refined by converting the Mg^{2+} coordinating lysine to a carboxylated lysine (KCX), after which the sidechain was placed such that the Mg^{2+} ion was coordinated in an octahedral geometry. Furthermore, conserved water molecules were manually placed inside the binding site. The protonation state of the protein residues was predicted with the PROPKA3.0 software package[7,8] at pH 7.5 containing the catalytic serine in its deprotonated form, followed by a visual check. A QMEAN analysis was performed to get a quality estimation of the generated homology models.[9] Both QMEAN4 and QMEAN6 were calculated, where the first is a scoring function consisting of a combination of structural aspects, and the latter additionally contains

terms describing the agreement of the structure with sequence-based predictions of secondary structure elements and solvent-accessibility. The QMEAN calculations were performed via the Swiss-Model webserver: https://swissmodel.expasy.org/qmean/

System preparation, parameterization and simulation setup

There are no straightforward parameters for Molecular Mechanics-based simulations of the [2Fe-2S] cluster. However, because of its role in substrate binding and catalysis, proper sampling of this cluster is important. Therefore, we parameterized a bonded model of the [2Fe-2S] cluster in DHADs. We applied a similar strategy as described by Carvalho *et al*.[10], but with Fe2 having a vacant coordination site. The parameterization was performed on the *Mt*DHAD structure (PDB: 6ovt), as this structure is applied as template for two of the target DHADs, and is resolved with a high resolution.[11] The two cysteine residues coordinating Fe1 were treated equivalently during the parameterization, as well as the two sulfur ions in the [2Fe-2S] cluster. The geometry of the 'small model' generated by MCPB.py[12], consisting of the [2Fe-2S] and sidechains of the coordinating cysteine residues, was optimized at the B3LYP/6-31G(d) level of theory using Gaussian09 (revision E.01)[13], after which the bonded terms were parameterized using the Seminario[14] method on the optimized model. Charges were fitted applying the RESP procedure[15] based on a Merz-Singh-Kohlmann population analysis[16,17] on a model consisting of the [2Fe-2S] cluster and capped coordinating cysteine residues. The ChgModB scheme was applied, meaning that all charges of the backbone heavy atoms were restrained to the values from the ff14SB[18] force field. An additional improper dihedral term was added (force constant: 10 kcal·mol^{-1}, periodicity: 2, phase: 180 degrees) to the [2Fe-2S] cluster to ensure an in-plane conformation of this cluster. The final parameters are provided in **Table S6**. The nonbonded parameters of the Mg^{2+} ion were set to the values of the IOD set from Li *et al*.[19]. The substrates were parameterized as follows: Van der Waals parameters of the substrate atoms were taken from the General Amber Force Field[20] (GAFF), and the charges were retrieved from a RESP fit based on a HF/6-31G(d) geometry optimized structure of the substrates. The force field parameters for the deprotonated serine residue were retrieved via the R.E.D. server.[15,21–23] Force field parameters for the deprotonated serine which were not in the ff14SB force field were taken from GAFF.

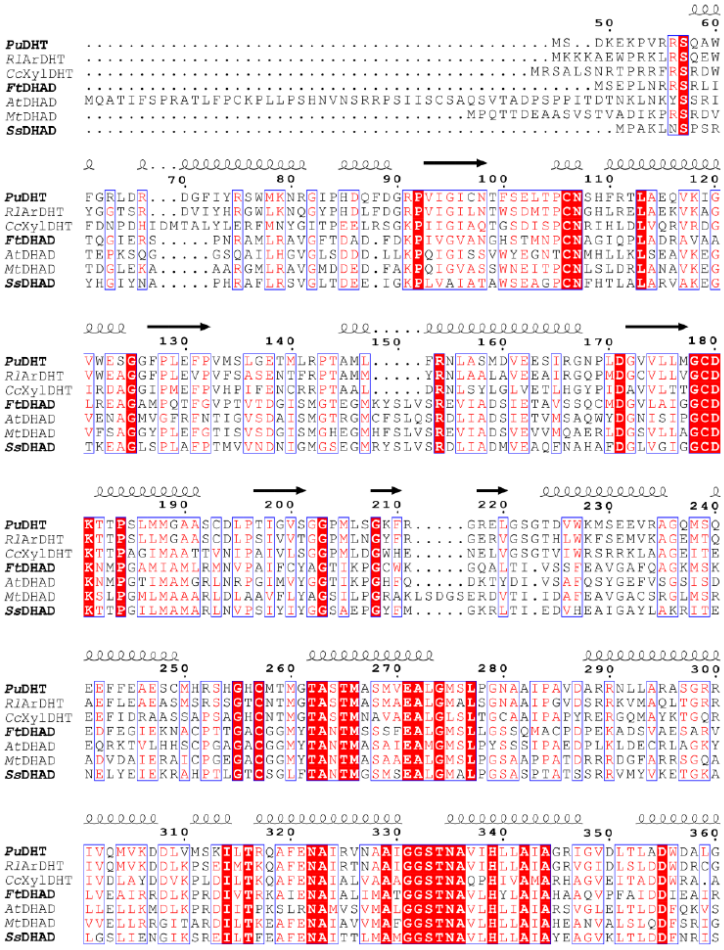
The protonated structures (see above) were solvated in an octahedral box consisting of TIP3P[24] water applying a buffer region of 12 Å, and counterions were added to neutralize the system using the LeaP module of the Amber18/AmberTools18 software package[25]. A distance restraint between the Mg^{2+} ion and Fe2 from the [2Fe-2S] cluster with force constant 100 kcal·mol^{-1}·Å^{-1} was applied to retain a proper binding site geometry during all simulations. Energy minimization was performed with the XMIN method in Sander from the Amber18/AmberTools18 software package applying a 20.0 kcal·mol·Å^{-2} positional restraint to the protein atoms, gradually bringing the density from 0.8 g·cm^{-3} to 1.0 g·cm^{-3} with a step size of 0.02 g·cm^{-3} by adjusting the box size. At target density, a final minimization step was performed without positional restraints. The system was subsequently heated with the following procedure: 10 ps from 0-5 K, 5-10 K, and 10-20 K respectively, in a NVT ensemble, applying the Langevin thermostat[26] with a collision

frequency of 7.0 ps^{-1}, and a 3.0 $kcal \cdot mol^{-1} \cdot Å^{-2}$ positional restraints on all atoms. Subsequently, the system was further heated applying the positional restraints solely on the substrate and backbone heavy atoms in the following sequence: 50 ps 20-50 K, 100 ps 50-100 K, 100 ps 100-200 K, followed by a 200 ps equilibration at 200 K without positional restraints. Afterwards, the temperature was increased to the target temperature (300 K for *Pu*DHT and *Mt*DHAD, 323 K for *Ft*DHAD and *Ss*DHAD) in 400 ps. Finally, the system was equilibrated for 500 ps in a NPT ensemble at the target temperature, applying the Berendsen barostat[27] with a relaxation time of 1 ps and compressibility of $44 \cdot 10^{-6}$ bar^{-1} to keep the pressure constant at 1 bar. Furthermore, periodic boundary conditions were applied, and the SHAKE algorithm[28] was used on all bonds involving hydrogen atoms. A nonbonded cut-off of 12 Å was used, the particle mesh Ewald method[29] was applied for long-range electrostatics, and an integration step of 1 fs was used. After the heatup, a 100-200 ns production Molecular Dynamics (MD) simulation was performed for each system using the pmemd.CUDA MD engine from Amber18, saving the velocities and coordinates of all atoms every 10 ps.

Molecular docking and in silico mutations

Docking simulations were performed with AutoDock 4.2.6[30]. A 60x60x60 points docking grid with 0.375 Å spacing, centered on the middle between the Mg^{2+} ion and Fe2 from the [2Fe-2S] cluster, was generated with AutoGrid. For each docking simulation, 150 poses were generated with a population size of 150, and a maximum of $2.5 \cdot 10^6$ energy evaluations with AutoDock. In the top-ranked poses obtained using the AutoDock scoring function, the substrates were mostly bound in a catalytically non-competent orientation; these poses were therefore assumed to be incorrect (see detailed description in main text). Since there are no experimental structures of a dehydratase with a bound substrate, the scoring function could not be refined to improve the performance to rank the docked poses. Additionally, the presence of the highly polarized [2Fe-2S] cluster and Mg^{2+} ion is also likely to impair the performance of the scoring function in these systems. Therefore, we selected suitable docking poses based on their ability to promote the dehydratation reaction as described in the main text. In practice, the following distances were measured for every pose, and summed up to serve as docking score (in case of the substrate carboxyl oxygen, the distance to the closest oxygen was measured): the distance between a carboxyl oxygen and the hydrogen (NH) from the backbone of Thr262, the distance between a substrate carboxyl oxygen and the OH of the side chain of Thr262, the distance between a substrate carboxyl oxygen and the Mg^{2+} ion, the distance between C2-OH and the Mg^{2+} ion, the distance between C2-H and the oxygen of Ser552 and the distance between C3-OH and Fe2. Visualization and *in silico* mutagenesis was performed with PyMol[31] using Dunbrack's backbone-dependent rotamer library[32].

2. Multiple Sequence Alignment



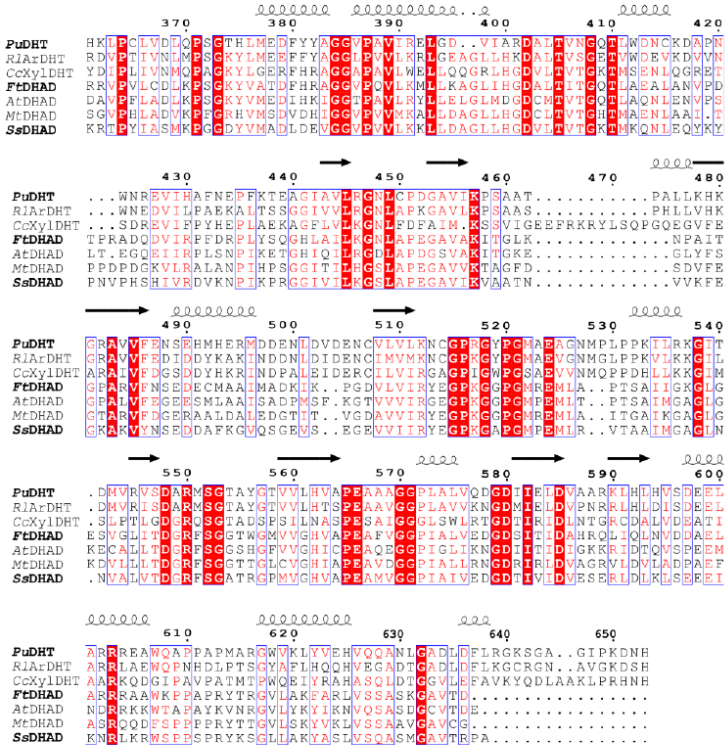


Figure S1. Full Multiple Sequence Alignment (MSA) of target- and template DHAD sequences. Enzymes in bold are the enzymes studied here. The numbers above the alignment indicate the alignment numbers, which are used throughout the manuscript. Helices and arrows above the MSA indicate the location of alpha-helices and beta-sheets, respectively, as obtained in the X-Ray structure of *Mt*DHAD (PDB-ID: 6ovt). Residues with red background indicate fully conserved residues at the respective alignment position, and residues in red font indicate a certain degree of conservation within the studied sequences. Blue boxes indicate cross-group conservation.

Table S1. Conversion between alignment numbers and protein sequence numbering

		Alignment position				
UNIPROT ID	**PDB-ID**	**73 (chain B)**	**224**	**261**	**552**	**138**
A0A4R3LQ44	*Pu*DHT	I24 (602)	T165	G202	S476	E89
A0A1I2J0Y3	*Ft*DHAD	P20 (582)	V164	Y201	S475	D84
Q97UB2	*Ss*DHAD	P18 (576)	E162	F199	S472	D82
B5ZZ34	5j84 -*Rl*ArDHT	Y26	T167	G204	S480	E91
A0A0E8UWV6	6ovt - *Mt*DHAD	A32	I181	Y218	S491	D96
Q9LIR4	5ze4 - *At*DHAD	G29 (604)	V173	Y210	S484	D93
Q9A9Z2	5oyn - *Cc*XylDHT	L26	T168	G205	S490	E92

3. Supplementary Figures

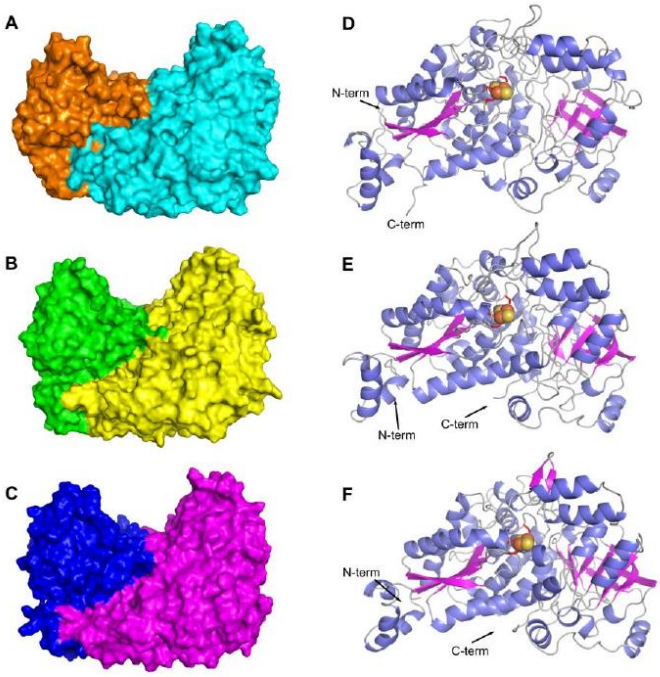


Figure S2. Structural representation of the homology models for *Pu*DHT (A,D), *Ft*DHAD (B,E) and *Ss*DHAD (C,F). The dimer is shown as surface representation (A-C) and a monomer structure is shown in cartoon representation (D-F) with the [2Fe-2S] cluster shown as spheres and its coordinating serine residues as red sticks.

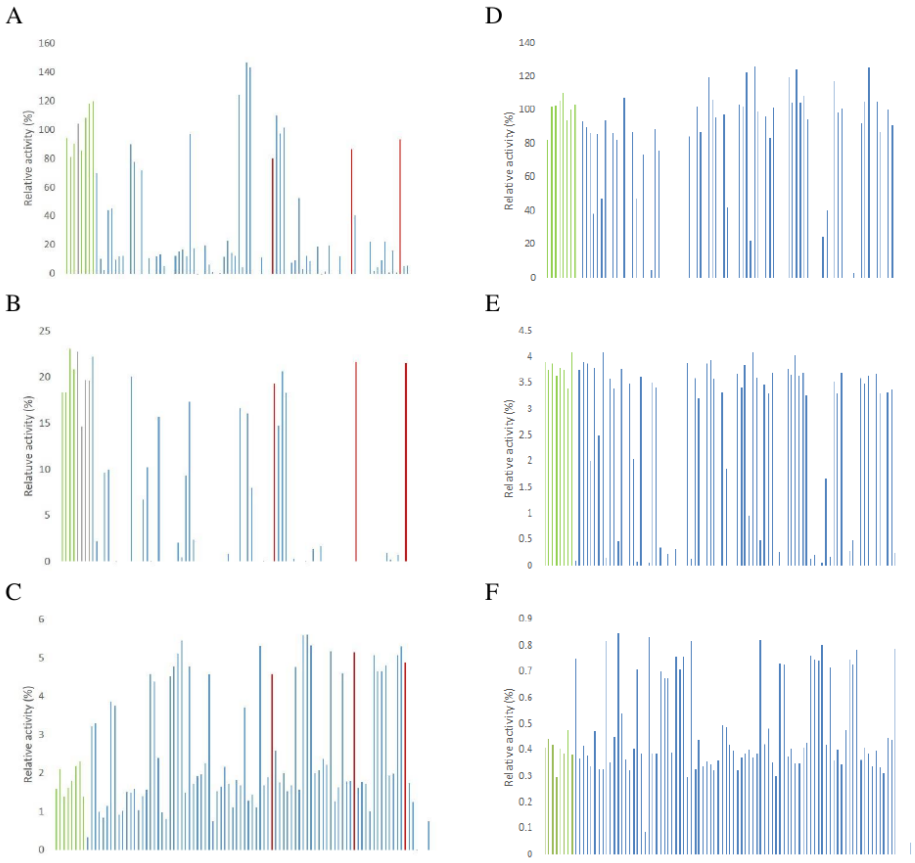


Figure S3. Activity landscape of P73X libraries for *Ft*DHAD toward DHIV (A), L-threonate (B), and D-gluconate (C) and I73X of *Pu*DHT toward D-gluconate (D), L-threonate (E), and DHIV (F). Relative activity is based on the average wild type activity (WT) of FtDHAD toward DHIV (A, B, C) or *Pu*DHT in D-gluconate (D, E, F). Green bars are the WT control of each respective enzyme. Negative controls are media only and are situated on the right. Red bars are *Ft*DHAD variants, which have P73G substitution.

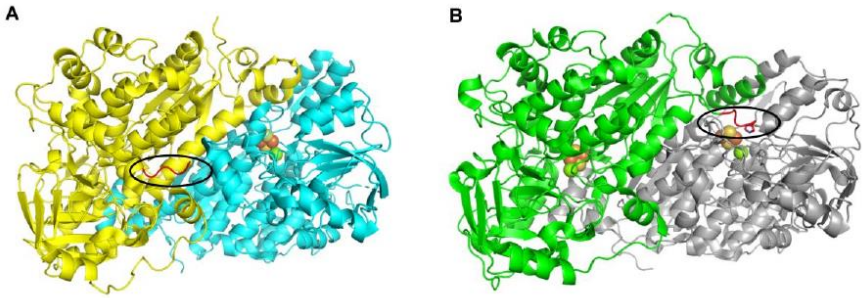


Figure S4. Location of C-terminus in BCADHTs (A) and SADHTs (B). The dimers are shown in a cartoon representation, FeS and Mg^{2+} are shown as spheres, and the C-terminus is highlighted in red. The X-Ray structure of *Mt*DHAD (PDB: 6ovt) and *Rl*ArDHT (PDB: 5j84) were used as representative for BCADHTs and SADHTs respectively.

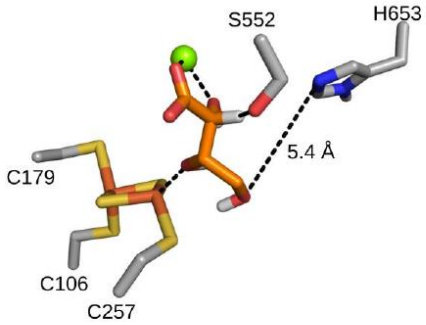


Figure S5. Docked L-threonate in the *Pu*DHT-WT, illustrating that there is no H-bond possible between the substrate and C-terminal histidine.

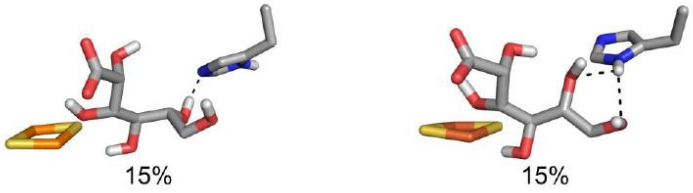


Figure S6. Remainder of observed interactions between D-gluconate and H653 during a 200 ns MD simulation of *Pu*DHT-WT. The most-sampled conformations are shown in the main text. The percentages below the binding states indicate the occurrence of the respective state during the MD simulation.

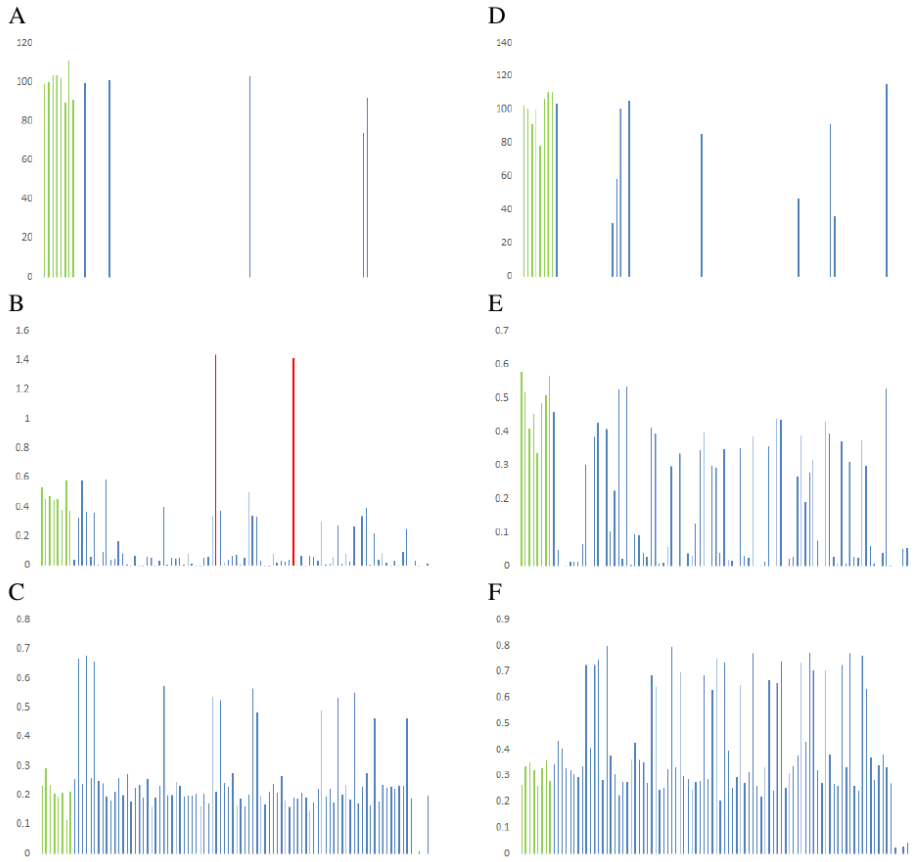


Figure S7. Activity landscape of *Pu*DHT_H653X toward D-gluconate (A), D-glycerate (B), and DHIV (C) and *Pu*DHT_N652_H653insX toward D-gluconate (D), D-glycerate (E), and DHIV (F). Relative activity is based on the average WT activity of *Pu*DHT in D-gluconate. Green bars are the WT control of each respective enzyme. Negative controls are media only and are situated on the right. Red bar are *Pu*DHT variants, which have H653F substitution.

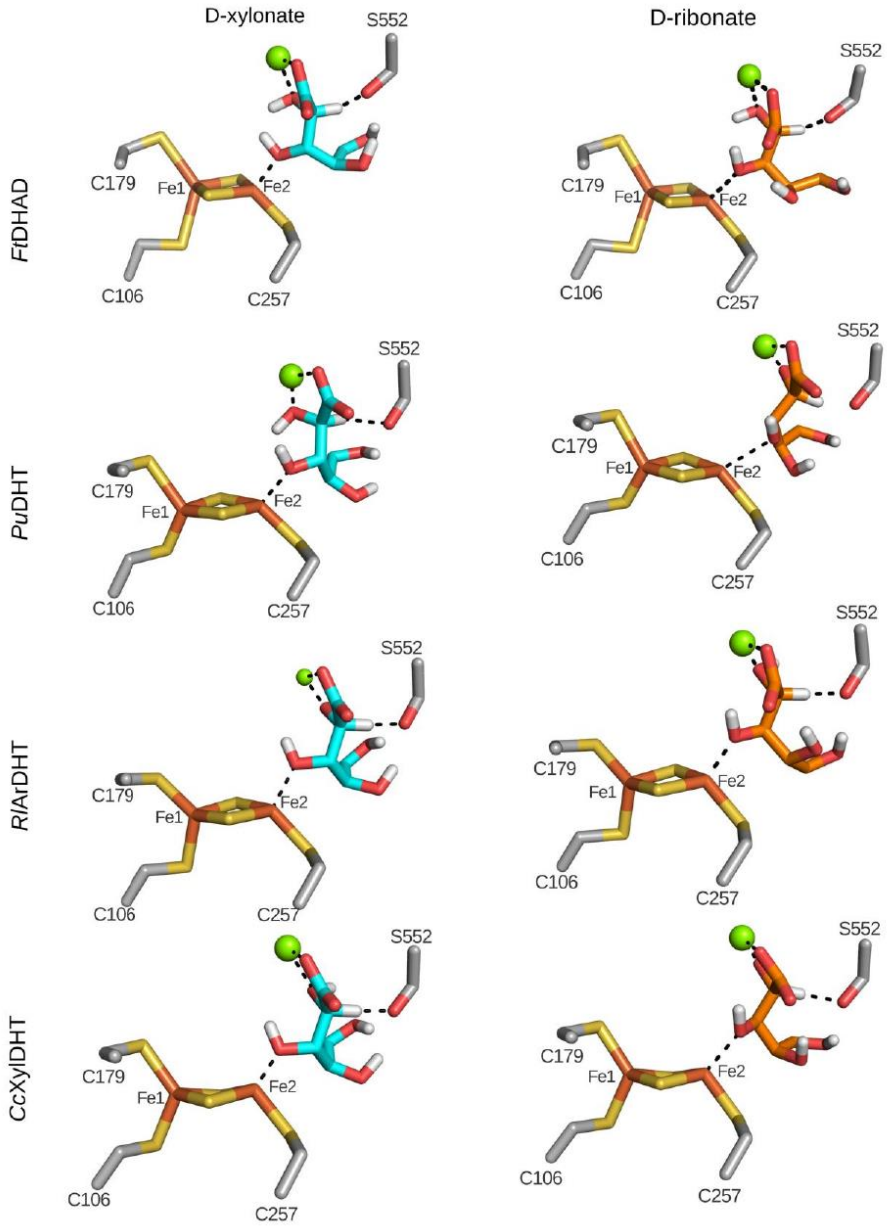


Figure S8. Docked poses of D-xylonate (left) and D-ribonate (right) in DHADs.

4. Supplementary Tables

Table S2. Quality estimation of homology models.

Target	Templates applied	QMEAN4	QMEAN6
*Pu*DHT	*Rl*ArDHT & *Cc*XylDHT	-0.62	-1.05
*Ft*DHAD	*At*DHAD & *Mt*DHAD	-1.05	-0.88
*Ss*DHAD	*At*DHAD & *Mt*DHAD	-1.34	-0.93

Table S3. Relative activity of alignment position 554/555 variants. The activity is measured relative to the wild-type. Alignment numbers are used for residue numbering.

	Rel. activity (%)
PuDHT (D-Gluconate)	
*Pu*DHT-WT (*SGTA*)	100.0 ± 3.5
*Pu*DHT-T554V (*SGVA*)	0.0 ± 0.0
*Pu*DHT-T554G (*SGGA*)	0.34 ± 0.0
*Pu*DHT-A555T (*SGTT*)	0.71 ± 0.0
*Pu*DHT-T554G, A555T (*SGGT*)	0.0 ± 0.0
FtDHAD (DHIV)	
*Ft*DHAD-WT (*SGGT*)	100.0 ± 4.9
*Ft*DHAD-G554T (*SGTT*)	0.0 ± 0.0
*Ft*DHAD-T555V (*SGGV*)	1.6 ± 0.0
*Ft*DHAD-T555A (*SGGA*)	18.4 ± 1.3
*Ft*DHAD-G554T, T555A (*SGTA*)	0.0 ± 0.0
SsDHAD (D-Xylonate)	
*Ss*DHAD-WT (*SGAT*)	100.0 ± 1.5
*Ss*DHAD-A554T (*SGTT*)	0.3 ± 0.1
*Ss*DHAD-T555V (*SGAV*)	3.1 ± 0.0
*Ss*DHAD-T555A (*SGAA*)	2.2 ± 0.0
*Ss*DHAD-A554T, T555A (*SGTA*)	0.0 ± 0.0

Table S4. Relative activity of *Pu*DHT and *Ft*DHAD in D-xylonate with different configurations of C2-OH and C3-OH.

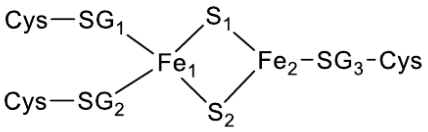
	*Pu*DHT - WT	*Ft*DHAD - WT
D-Xyl (*R*,*S*)	100.00 ± 5.31	100.00 ± 0.88
D-Ara (*S*,*R*)	0.00 ± 0.00	0.00 ± 0.00
D-Rib (*R*,*R*)	0.09 ± 0.06	104.00 ± 4.66
D-Lyx (*S*,*S*)	0.02 ± 0.00	0.00 ± 0.00

Table S5. Sequence identity between target- and template dehydratases

	*At*DHAD (PDB: 5ze4)	*Mt*DHAD (PDB: 6ovt)	*Cc*XylDHT (PDB: 5oyn)	*Rl*ArDHT (PDB: 5j84)
*Pu*DHT	34	37	**39**	**64**
*Ft*DHAD	**49**	**51**	31	36
*Ss*DHAD	44	**51**	32	38

Table S6. Nonbonded parameters of [2Fe-2S]

Atom	Mass	LJ Radius	LJ Depth	Charge
SG1†	32.06	2.0000	0.2500	-0.3970
SG2†	32.06	2.0000	0.2500	-0.3970
SG3	32.06	2.0000	0.2500	-0.4624
S1‡	32.06	2.0000	0.2500	-0.4098
S2‡	32.06	2.0000	0.2500	-0.4098
Fe1	55.8500	1.3860	0.0136	+0.4077
Fe2	55.8500	1.3860	0.0136	+0.4991



†,‡ Both Cys1 and Cys2, and the [2Fe-2S] sulfur ions were enforced to be equal during the parameterization, both for the fitting of the bonded parameters and the RESP partial charge fitting.

5. List of primers and auxiliary enzymes used

Table S7. List of primers used in this study. The amino acid number refers to the actual position of each enzyme

Primers	Sequence (5´→ 3´)
FtDHAD_P20X.F	CGCTCANNSAATCGCGCGATGTTGCGCG
FtDHAD_P20X.R	GCGATTSNNTGAGCGCTCGATGCCTTGGGTG
FtDHAD_D84E/A.F	GTCACCGMAGGCATCTCCATGGGCACCGAGGG
FtDHAD_D84E/A.R	AGATGCCTKCGGTGACGGTCGGCACACCGAAGG
FtDHAD_V164E.F	ACCATCGAATCGTCCTTCGAGGCCGTCGGTGCG
FtDHAD_V164E.R	AGGACGATTCGATGGTCAGCGCCTGACCCTTCC
FtDHAD_Y201G.F	GGCATGGGTACCGCGAATACGATGTCCTCGTCG
FtDHAD_Y201G.R	TTCGCGGTACCCATGCCGCCGCAGGCGCCGG
FtDHAD_T478V/A.F	GGCGGAGYCTGGGGTATGGTCGTGGGACACGTC
FtDHAD_T478V/A.R	TACCCCAGRCTCCGCCGGAGAAACGGCCATCGG
FtDHAD_GT477/8TT/A.F	TCCGGCACCRCCTGGGGTATGGTCGTGGGACACG
FtDHAD_GT477/8TT/A.R	ACCCCAGGYGGTGCCGGAGAAACGGCCATCGGTG
SsDHAD_T475A/V.F	GGTGCAGYCCGTGGTCCGATGGTTGGTCATG
SsDHAD_T475A/V.R	ACCACGGRCTGCACCGCTAAAACGACCATCGG
SsDHAD_AT474/5TT/A.F	TAGCGGTACCRCCCGTGGTCCGATGGTTGGTCATG
SsDHAD_AT474/5TT/A.R	CCACGGGYGGTACCGCTAAAACGACCATCGGTAAC
PuDHT_I24X.F	GGTTTCNNSTACCGTTCGTGGATGAAGAATCGAGG
PuDHT_I24X.R	AACGGTASNNGAAACCGTCGCGGTCAAGCC
PuDHT_E89D/A.F	CTGGGCGMTACGATGTTGCGGCCTACGGCCATG
PuDHT_E89D/A.R	ACATCGTAKCGCCCAGCGACATGACCGGAAACTCG
PuDHT_T165E.F	TTCAGGCGAAGACGTCTGGAAGATGTCTGAAGAAG
PuDHT_T165E.R	AGACGTCTTCGCCTGAACCCAGTTCGCGGCC
PuDHT_G202F/Y.F	ACGATGTWTACGGCATCCACGATGGCCAGCATG
PuDHT_G202F/Y.R	ATGCCGTAWACATCGTCATGCAATGCCCATGC
PuDHT_T478G/V.F	AGCGGCGKTGCGTATGGCACGGTGGTTCTGC
PuDHT_T478G/V.R	ATACGCAMCGCCGCTCATGCGGGCGTCGGAC
PuDHT_A479T.F	GGCACCACCTATGGCACGGTGGTTCTGCATGTG
PuDHT_A479T.R	TGCCATAGGTGGTGCCGCTCATGCGGGCGTC
PuDHT_TA478/9GT.F	AGCGGCGGCACCTATGGCACGGTGGTTCTGCATG
PuDHT_TA478/9GT.R	TGCCATAGGTGCCGCCGCTCATGCGGGCGTC
PuDHT_H575X.F	GACAATNNSTAAGAATTCGAGCTCCGTCGACAAGC
PuDHT_H575X.R	AATTCTTASNNATTGTCTTTGGGTATGCCCGCGCC
PuDHT_Ins575X.F	GACAATNNSCACTAAGAATTCGAGCTCCGTCGAC
PuDHT_Ins575X.R	CTTAGTGSNNATTGTCTTTGGGTATGCCCGCGCCG
SsDHAD_D82E/A.F	TTGTGAATGMAAATATTGGCATGGGTAGCGAAGGTATGCGTTATAG
SsDHAD_D82E/A.R	AATATTTKCATTCACAACCATGGTCGGAAATGCCAGCGGAGACAG
SsDHAD_D82E.F	TTGTGAATGAGAATATTGGCATGGGTAGCGAAGGTATGCGTTATAG
SsDHAD_D82E.R	AATATTCTCATTCAAACCATGGTCGGAAATGCCAGCGGAGACAG

15

SsDHAD_T475V.F	GGTGCAGTCCGTGGTCCGATGGTTGGTCATG
SsDHAD_T475V.R	ACCACGGACTGCACCGCTAAAACGACCATCGG
SsDHAD_AT474/5TA.F	TAGCGGTACCGCCCGTGGTCCGATGGTTGGTATG
SsDHAD_AT474/5TA.R	CCACGGGCGGTACCGCTAAAACGACCATCGGTAAC

Table S8. List of auxiliary enzymes used in this study

Enzyme	Microorganisms	NCBI Ref. Seq.*	Substitution	Reference
*Bs*ALS	*Bacillus subtilis*	WP_003244057.1	Wild type	[33]
*Mr*KARI	*Meiothermus ruber*	WP_013014163.1	T84S	[34]
*Cb*FDH	*Candida boidinii*	O13437.1	Wild type	[35]
*Pt*KdgA	*Picrophilus torridus*	WP_048059513.1	Wild type	[36]
*Bst*ADH	*Bacillus stearothermophilus*	KFL15473.1	Wild type	[33]
*Ll*KdcA (7M.D)	*Lactococcus lactis*	PDB: 2vbf	See publication	[37]

* The NCBI reference sequences/PDB numbers presented correspond to the wild type sequences

6. References

[1] S. Sutiono, M. Teshima, B. Beer, G. Schenk, V. Sieber, *ACS Catal.* **2020**, *10*, 3110–3118.

[2] J. M. Carsten, A. Schmidt, V. Sieber, *J. Biotechnol.* **2015**, *211*, 31–41.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403–410.

[4] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., *Mol. Syst. Biol.* **2011**, *7*, 539.

[5] X. Robert, P. Gouet, *Nucleic Acids Res.* **2014**, *42*, 320–324.

[6] A. Šali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779–815.

[7] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, J. H. Jensen, *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

[8] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, J. H. Jensen, *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

[9] P. Benkert, M. Biasini, T. Schwede, *Bioinformatics* **2011**, *27*, 343–350.

[10] A. T. P. Carvalho, Teixeira Ana F. S., M. J. Ramos, *J. Comput. Chem.* **2013**, *34*, 1540–1548.

[11] G. Bashiri, T. L. Grove, S. S. Hegde, T. Lagautriere, G. J. Gerfen, S. C. Almo, C. J. Squire, J. S. Blanchard, E. N. Baker, *J. Biol. Chem.* **2019**, *294*, 13158–13170.

[12] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

[13] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, et al., *Gaussian 09, Revision E.01, Gaussian, Inc., Wallingford CT*, **2009**.

[14] J. M. Seminario, *Int. J. Quantum Chem.* **1996**, *60*, 1271–1277.

[15] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[16] U. C. Singh, P. A. Kollman, *J. Comput. Chem.* **1984**, *5*, 129–145.

[17] B. H. Besler, K. M. Merz Jr., P. A. Kollman, *J. Comput. Chem.* **1990**, *11*, 431–439.

[18] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[19] P. Li, B. P. Roberts, D. K. Chakravorty, K. M. Merz, *J. Chem. Theory Comput.* **2013**, *9*, 2733–2748.

[20] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[21] E. Vanquelef, S. Simon, G. Marquant, E. Garcia, G. Klimerak, J. C. Delepine, P. Cieplak, F.-Y.

Dupradeau, *Nucleic Acids Res.* **2011**, *39*, W511–W517.

[22] F.-Y. Dupradeau, A. Pigache, T. Zaffran, C. Savineau, R. Lelong, N. Grivel, D. Lelong, W. Rosanski, P. Cieplak, *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821–7839.

[23] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, et al., *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[24] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926–935.

[25] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, et al., *AMBER 18*, University Of California, San Francisco, **2018**.

[26] R. J. Loncharich, B. R. Brooks, R. W. Pastor, *Biopolymers* **1992**, *32*, 523–535.

[27] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[28] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327–341.

[29] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[30] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, A. J. Olson, *J. Comput. Chem.* **2009**, *30*, 2785–2791.

[31] Schrodinger LLC, **n.d.**

[32] M. V. Shapovalov, R. L. Dunbrack Jr., *Structure* **2011**, *19*, 844–858.

[33] J. K. Guterl, D. Garbe, J. Carsten, F. Steffler, B. Sommer, S. Reiße, A. Philipp, M. Haack, B. Rühmann, A. Koltermann, et al., *ChemSusChem* **2012**, *5*, 2165–2172.

[34] S. Reiße, D. Garbe, T. Brück, *Biochimie* **2015**, *108*, 76–84.

[35] H. Slusarczyk, S. Felber, M. R. Kula, M. Pohl, *Eur. J. Biochem.* **2000**, *267*, 1280–1289.

[36] T. J. Gmelch, J. M. Sperl, V. Sieber, *Sci. Rep.* **2019**, *9*, 11754.

[37] S. Sutiono, J. Carsten, V. Sieber, *ChemSusChem* **2018**, *11*, 3335–3344.

This page was left blank intentionally.

3.5 Thiazoline-Specific Amidohydrolase PurAH is the Gatekeeper of Bottromycin Biosynthesis

In this study, the metallo-dependent amidohydrolase from *Streptomyces purpureus* (PurAH) was characterized, and its role in the biosynthesis of bottromycin A2, a ribosomally synthesized and post-translationally modified peptide (RiPP), was elucidated. Bottromycins are natural products with antimicrobial activity, especially against Methicillin-resistant *Staphylococcus aureus* (MRSA), and other dangerous human pathogens, and its biosynthesis is getting increasing attention.[277-279] The biosynthesis starts with BotA, a precursor peptide which is tailored by numerous enzymes, including several YcaO enzymes (*i.e.* bacterial enzymes involved in thiazole-containing antibiotics).[280] The reaction of interest in this study is the macroamidine formation by the YcaO enzyme BotCD together with BotAH. Since recombinant expression of BotAH led to insoluble protein, a homologous enzyme PurAH (72% sequence identity) was expressed and studied instead. In a previous publication, co-authors Laura Franz and Jesko Koehnke already showed that PurCD (a homologue of BotCD) can catalyze both macroamidine formation and its reopening, thus PurCD catalyzes the reaction in both directions.[278] In this study, PurAH was shown to be highly selective to the macroamidine-containing bottromycin precursor, and cleaves the peptide C-terminal of the thaizoline. Moreover, the results showed that BotCD cannot reopen the macroamidine ring of the cleaved peptide. Thus, the concerted efforts of macroamidine formation by PurCD, followed by peptide cleavage by PurAH, lead to the irreversible formation of the macroamidine-containing bottromycin precursor *in vitro*. In other words, PurAH acts as the 'gatekeeper' during the *in vitro* biosynthesis of bottromycin, by preventing reopening of the macroamidine ring. Furthermore, an X-Ray structure of PurAH in the apo-form was determined at 1.73 Å resolution. The activity of PurAH on some mutated peptides was measured, and site-directed mutagenesis of PurAH was performed in order to study the binding of the bottromycin precursor to PurAH.

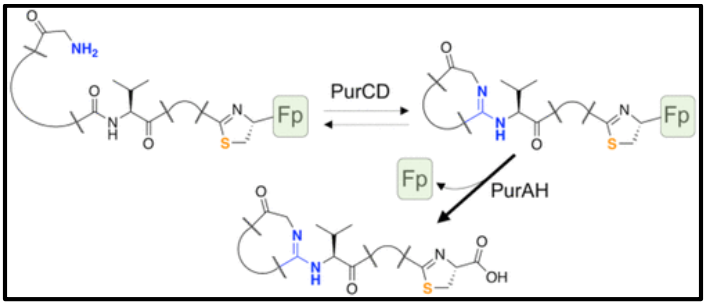
Jesko Koehnke designed the study, and Asfandyar Sikandar and Laura Franz performed and analyzed the *in vitro* experiments. Okke Melse performed molecular docking studies with DynaDock to support the site-directed mutagenesis, and also conducted bioinformatics analysis and binding site volume measurements in PurAH and structural homologues. All authors were involved in writing of the manuscript.

Thiazoline-Specific Amidohydrolase PurAH is the Gatekeeper of Bottromycin Biosynthesis

Asfandyar Sikandar, Laura Franz, Okke Melse, Iris Antes, Jesko Koehnke

Journal of the American Chemical Society

2019, **141**(25), 9748-9752
https://doi.org/10.1021/jacs.8b12231



This work is not embedded, but cited as work relevant for this dissertation.

3.6 Broad Spectrum Antibiotic-Degrading Metallo-β-Lactamases are Phylogenetically Diverse

Antibiotic resistance is a major public health problem, reported as a current crisis by the World Health Organization.[281] β-lactamases, which can be divided in serine-β-lactamases (class A, C, and D) and metallo-β-lactamases (MBLs, class B), play an important role in the antibiotic resistance by bacteria.[214, 215] Since MBLs are not inhibited by clinically available antagonists, they are responsible for a large number of the antibiotic resistance problems.[218] Most MBLs contain two Zn^{2+} ions in their active site, with B3-type MBLs typically carrying the active site motif HHH/DHH for the α- and β-site, respectively. Two variations have been observed with the active site motifs **Q**HH/DHH and H**R**H/D**QK** instead (variations shown in bold).[222, 223] In this study, the above-mentioned B3-Q and B3-RQK MBLs (named by their active site motif variations), among others, were studied in more detail.[221] It was found that B3-RQK is sensitive to the serine-β-lactamase inhibitor clavulanic acid, which was never observed for any MBL so far. Moreover, no density for Zn^{2+} ions was observed in the determined X-Ray structures, suggesting reduced Zn^{2+} affinity. In X-Ray structures of B3-RQK variants, where either the α-site, β-site, or both sites were "back-mutated" to the original motif found in B3 MBLs (HHH/DHH), Zn^{2+} binding was observed again in parallel with increased activity and resistance against clavulanic acid. Thus, molecular docking with subsequent QM/MM geometry refinement simulations were performed, both in bimetallic and monometallic MBL, by which a binding mode for clavulanic acid in B3-RQK MBLs could be determined. The QM/MM calculations further resulted in a suggestion for the inhibition mechanism of clavulanic acid, namely via Zn^{2+} displacement from the low affinity β-site. The *in silico* studies suggested K263 to play an essential role in clavulanic acid binding, which is in agreement with experimental data. Therefore, the results in this study suggest that modifying clavulanic acid to target H263 may increase the therapeutic range of clavulanic acid, and perhaps lead to drugs acting as MBL inhibitors.

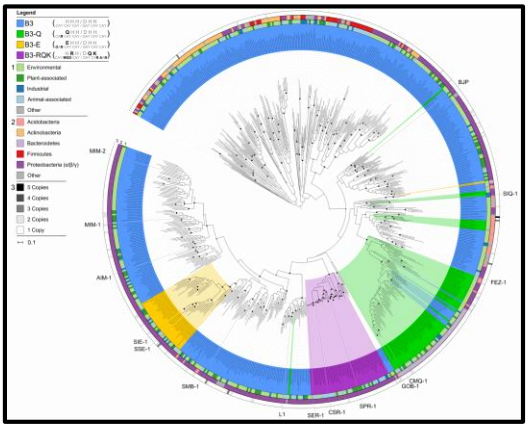
Marcelo Monteiro Pedroso, David W. Waite, Nataša Mitić, Ross P. McGeary, and Gerhard Schenk devised the study, while David W. Waite and Philip Hugenholtz performed the phylogenetic analysis. Marcelo Monteiro Pedroso, Nataša Mitić, Liam Wilson, and Luke W. Guddat performed the experimental work. Okke Melse and Iris Antes designed the *in silico* analysis, and Okke Melse performed and analyzed the molecular docking and QM/MM simulations, and drafted the first hypotheses to structurally rationalize the observed differences in Zn^{2+} and inhibitor binding. All authors were involved in writing the final version of the manuscript.

Broad Spectrum Antibiotic-Degrading Metallo-β-Lactamases are Phylogenetically Diverse

Marcelo Monteiro Pedroso, David W. Waite, Okke Melse, Liam Wilson, Nataša Mitić, Ross P. McGeary, Iris Antes, Luke W. Guddat, Philip Hugenholtz, Gerhard Schenk

Protein & Cell

2020, **11**(8), 613-617
https://doi.org/10.1007/s13238-020-00736-4



This work is not embedded, but cited as work relevant for this dissertation.

This page was left blank intentionally.

4 DISCUSSION

4.1 COMPUTER-AIDED IDENTIFICATION OF POTENTIAL BIOCATALYSTS AND LIGAND BINDING SITES

4.1.1 Target and biocatalyst identification

Biomolecular simulations have shown to make a useful contribution in numerous scientific areas, including enzyme engineering and structure-based drug discovery.[2, 16, 147, 282] However, in order to perform simulations which are able to guide enzyme engineering or drug discovery projects, pre-knowledge about the target structure (ideally with known structure) and the binding site is essential. This information is often not available, and experimental approaches to retrieve this knowledge, such as target fishing and high-throughput screening, are time-consuming tasks.[283] Bioinformatics tools can support in this process, for example via public biological databases, which collect and combine large amounts of functional information about known proteins, and stores them in a searchable manner. Besides for data collection, bioinformatics tools can also be used to predict target proteins, including enzymes. Ligand similarity search algorithms are often applied for this purpose, for example in SciFinder[284] and PubChem[285]. In these tools, a large number of databases are searched for a structurally similar ligand or substrate compared to the particular ligand of interest. The general idea is that if a target is known for an identified similar ligand, this target may also bind to the ligand of interest. Other bioinformatics tools generally apply profile analysis (*e.g.* interaction fingerprints), network-based searches, shape-based screening or virtual screening.[283, 286-288] Most of these methods are however developed for drug (off-)target identification, and not to predict potential biocatalysts, *i.e.* an enzyme which can not only bind a certain substrate, but also catalyze a target reaction.

4.1.2 EnzymeMatch: challenges, features, and perspectives

In order to fill this gap, this dissertation describes the development of EnzymeMatch: a python-based algorithm which applies a structure-based interaction point matching approach combined with a database search. In short, a query containing an optimal interaction point network at the protein side is automatically predicted based on the target substrate (*.mol2* or *.sdf* file formats supported). Subsequently, a database search is performed within the BioLiP database, which contains information about binding sites of nearly all structures in the Protein Data Bank, in order to efficiently identify enzymes containing the required interaction types, and thus potentially able to catalyze the target reaction.[131]

Finally, an interaction pattern matching is performed in all enzyme candidates applying a graph-theoretical approach and triangle matching, which allows for efficient, and accurate predictions.

Based on discussions with experimentalists, several conditions were defined which EnzymeMatch should fulfill (see Chapter 1.7.1). In short: (I) the algorithm should be able to efficiently screen a large number of enzymes, (II) the program should be easy to install and use by non-expert users, (III) the user should be able to fine-tune the search- and matching procedure without having to modify the code, and (IV) the output should be comprehensible, and contain relevant information for experimentalists to further assess, express and engineer the predicted enzymes. In order to address the first condition, EnzymeMatch matches enzymes and the target substrate applying graph theory instead of extensive conformational sampling and docking techniques. In this way, only vectors connecting interaction points need to be compared, which can be implemented in a much more efficient manner than traditional conformational sampling and docking approaches. The second condition (*i.e.* easy to install and usable by non-expert users) was fulfilled by writing this algorithm as a python project, with as little dependencies as possible. Python code has the advantage that it is easy to run on multiple operating platforms, easy to share, *e.g.* via GitHub or other platforms, and rather easy to read for new developers who want to further develop the algorithm. The third condition (*i.e.* user-modifiable settings) was met by using input files in a human-readable format, in which all settings can be defined. Furthermore, the algorithm carefully checks the validity of (the combination of) applied keywords and their settings during the input file parsing. Finally, the last condition (*i.e.* clear, informative, and extendable output) was fulfilled by saving information retrieved from the BioLiP database for each matched entry, which include information such as binding site residues, EC numbers, known ligands, etc. Moreover, the residues that were predicted to interact with the ligand are provided in the output as well. A small additional external python script was written, which matches all predicted enzymes with the Protein Data Bank, to further provide information about the enzyme annotation, origin organism and expression system.

Protein- and ligand flexibility is often not considered in bioinformatics tools aiming to support in target finding, while protein- and ligand flexibility is known to strongly influence ligand binding (see Chapter 1.5.2 for an extensive discussion on this topic in the context of molecular docking simulations). This lack of describing protein- and ligand flexibility is often caused by the strong simplification applied in these algorithms, *e.g.* in profile- or network analysis approaches, and due to the lack of structural information used in these approaches.[286, 288] However, due to the large amount and ever-growing availability of protein structures in the Protein Data Bank (**Figure 1**), and potentially on-demand structure

prediction by AlphaFold (see Chapter 1.3.2), structure-based approaches show promising potential. In EnzymeMatch, ligand flexibility is automatically considered during the Automatic Query Design (AQD) step. Here, a user-defined number of substrate conformations are automatically generated, which are all used to predict the optimal geometry of interaction points at the protein side. Moreover, the generated substrate conformations also provide information about flexible and rigid regions of the substrate, which is also considered during the interaction pattern matching phase. Smaller deviations from the optimal interaction geometry are allowed when the respective interaction point interacts with a rigid region of the substrate, and more deviation is allowed for interaction points interacting with more flexible regions of the substrate (**Figure 9**). This is implemented by using individual offset values for each vector in the interaction pattern, as described in Chapter 3.1. Protein flexibility can also be considered during the interaction point matching in EnzymeMatch, which is performed by the use of rotamer libraries (see Chapter 1.3.1), where all possible combinations of rotamers for all flexible binding site residues are predicted. However, this has the consequence that much more interaction point patterns of the respective enzyme need to be matched to the input interaction point pattern (*i.e.* the optimal interaction point pattern predicted by AQD). This significantly increases computation time, and is thus only reasonable when a small database is screened. Besides consideration of protein- and ligand flexibility, several additional features were implemented in EnzymeMatch. For example, a mode which automatically downloads the required protein structures from the Protein Data Bank was added, as well as a scoring function, which provides a measure on how the interaction patterns matched (note that this requires evaluation of all possible interaction patterns in a binding site in order to find the best-fit, and therefore negatively affects computation time).

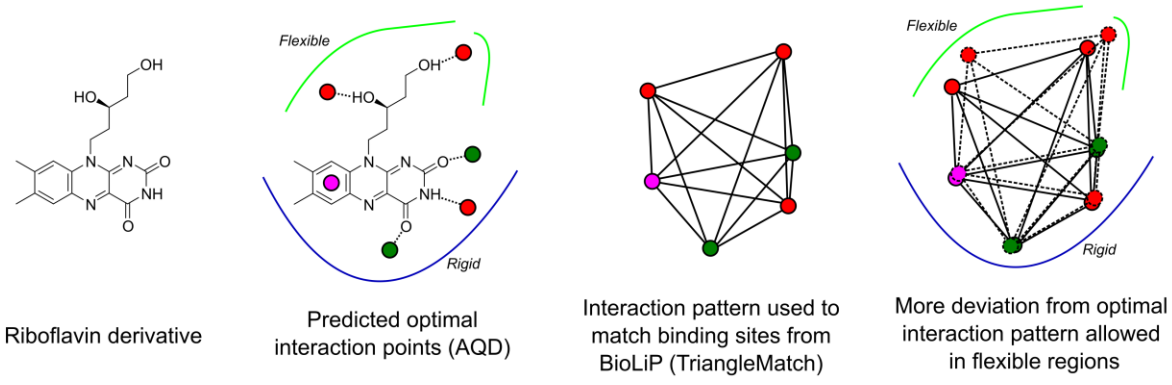


Figure 9. Illustration of the interaction point matching algorithm in EnzymeMatch including the consideration of ligand flexibility. Colored nodes indicate optimal positions of different types of interaction points at the protein side, and the edges represent the Euclidean distances between them, which are matched during the TriangleMatch phase. Dotted spheres and edges indicate a hypothetically matched binding site from the BioLiP database.

Performance evaluation of EnzymeMatch is not straightforward because of the expected high number of enzymes which may be active on the target substrate, but not identified as such, making *e.g.* Receiver Operator Plots hard to interpret. Note that identifying these so far unidentified enzymes, which are able to catalyze a certain reaction on the target substrate, is exactly the aim of EnzymeMatch. Therefore, an ideal evaluation analysis should include experimental activity measurements on the top-predicted enzymes. However, this is a very time-consuming and labor intensive process, as not all enzymes may be easily expressed in the lab, and one does generally not know if the expressed enzyme is properly folded. Therefore, we evaluated EnzymeMatch by performing docking simulations in the top-predicted enzymes, which was performed for a range of target substrates of varying size and rigidity. For all evaluated systems, top-ranked enzymes could be confirmed to be good candidates by molecular docking simulations. For example, in an EnzymeMatch run searching for enzymes binding riboflavin, a lumazine synthase from *Aquifex aeolicus* was matched (fit: 0.713 Å), which after some literature search appeared to be the second last enzyme in the biosynthesis pathway of RBF.[289] Molecular docking studies further illustrated that riboflavin fits perfectly in the binding site of the lumazine synthase. Furthermore, EnzymeMatch predicted the synthetic glucocorticoid dexamethasone to be catalyzed by human aldo-keto reductase, an enzyme known to bind testosterone, which is structurally highly similar to dexamethasone. Finally, for S-hexylglutathione, numerous glutathione-binding enzymes were identified (*e.g.* glutathione S-transferase), which is structurally identical to S-hexylglutathione, except the hexylgroup connected to the thiol group of glutathione. These examples show the potential of EnzymeMatch, but the performance remains limited to the enzymes which are present in the Protein Data Bank.

4.1.3 Protein-ligand interaction points in search queries

The idea of representing protein binding sites as a set of protein-ligand interaction points is not completely new, neither is the application of these interaction points to search through the Protein Data Bank. For example, Relibase and Relibase+, nowadays part of CSD-CrossMiner from the Cambridge Crystallographic Data Centre, compiled a database containing protein-ligand interactions and structural information, such as distance, angle and interaction partners. All information was retrieved from the Protein Data Bank. This information can be used to analyze preferred interaction patterns of certain chemical groups, including information about preferred interaction partners and interaction geometry.[290] More recently, another web application was published: GeoMine, as part of Protein*Plus*, which is a web-based portal with its focus on protein-ligand interaction developed by the Computational Molecular Design Group from the University of Hamburg.[291] GeoMine starts with loading a PDB structure, which immediately highlights the first main difference with EnzymeMatch, as EnzymeMatch does not need a protein-ligand complex as

starting point. In GeoMine, one can define certain atoms, distances and interactions in a graphical user interface, which can be added to the query. The presence of this query is subsequently searched in the Protein Data Bank. Thus, in principle, a binding site similarity search is performed taking the co-crystallized ligand as part of the binding site, meaning that GeoMine can be used to find proteins with a similar binding site rather than a binding site for new substrates, which is required to identify potential new biocatalysts. In contrast, EnzymeMatch does not require a protein structure (which is often not known when searching for a potential biocatalyst), as the optimal interaction points at the protein side are automatically predicted by the AQD module based on the target substrate. Moreover, GeoMine only matches interactions, thus a protein is only matched when a similar protein-ligand interaction was observed in the X-Ray structure of the protein to what has been observed in the input structure. This strongly limits the amount of possible predictions, and limits the predictions to proteins which were co-crystallized with a ligand. EnzymeMatch does not have this limitation, since a geometric interaction pattern is matched instead, which is generated from interaction points at the protein side. In other words, EnzymeMatch predicts if a protein-ligand interaction could be formed, rather than matching on known protein-ligand interactions, making the amount of potential targets (here: enzymes) much larger.

Altogether, EnzymeMatch can efficiently and rigorously predict enzymes which can potentially bind a target substrate and catalyze a required chemical reaction. Due to the careful design choices, it is easy to use, also for non-expert users. This, together with numerous additional features described above, makes EnzymeMatch a potentially valuable tool to screen for potential biocatalysts.

4.1.4 Consideration of protein- and ligand flexibility during binding site identification

Dependent on the application, an identified target is not sufficient. For example, in structure-based drug design, as well as in protein design or -engineering studies, the next challenge is the definition of (a) binding site(s). When an experimental structure is available in complex with a ligand, the definition of the binding site can often be retrieved from the structure. However, many structures are resolved in their apo-form (*i.e.* without presence of ligand or cofactor), making a prediction of the binding site essential. Since experimental elucidation of ligand binding sites is challenging, numerous computational models have been developed, which are extensively described in Chapter 1.4. Moreover, large conformational changes can occur upon ligand binding, as described by the induced fit- and conformational selection models (**Figure 2**), which can result in opening or closing events of the binding site and/or entrance tunnel, or lead to conformational changes affecting substrate specificity. The majority of available binding site identification methods neglect protein- and ligand flexibility despite recognition that protein-ligand

interactions can be strongly affected by the protein dynamics.[136, 155, 156, 169] Cavity-mapping algorithms such as POVME and mkgridXf can work on an ensemble of protein structures retrieved from a molecular dynamics simulation of the apo structure, but this does not account for induced fit effects upon ligand binding.[125, 126] Furthermore, most of the available methods were evaluated using protein structures resolved in the ligand-bound state, meaning that the binding site was already adapted on the presence of the ligand. However, in actual applications of these binding site identification methods, the algorithms are applied on an apo structure, where one cannot expect that the binding site is present in the conformation required to accommodate ligand binding.

Thus, DynaBiS was developed in this dissertation to improve the binding site identification for large and highly flexible ligands, such as peptides, in protein apo structures, as described in detail in Chapter 3.2. The performance of DynaBiS was evaluated using a diverse evaluation set consisting of 8 small ligand and 7 peptide protein-ligand complexes. The apo structures were used for 11 of the systems as well, to evaluate the performance of DynaBiS in a more real-life example. [252] The evaluation set consisted of binding sites for ligands ranging from small organic ligands to large, and highly flexible peptides. We showed that in the top-5 predicted binding sites, the correct binding site was present for all but one of the evaluation systems, and for 19 out of 26 binding sites, the correct binding site was predicted correctly as top-ranked binding site. Thereby DynaBiS strongly outperformed the AutoDock blind docking method, as well as the commonly used algorithm AutoSite.[134, 137] We further showed that the sampling algorithm in DynaBiS was able to sample all known binding sites, as all known binding sites were identified as potential binding sites and further evaluated in the pocket sampling step.

Because DynaBiS uses both the protein and ligand structure, the predicted binding sites are ligand-specific. This means that only binding sites which might be able to accommodate binding of the target ligand are considered. The majority of other available binding site identification methods including AutoLigand, AutoSite, POVME, and Q-SiteFinder, predict binding sites independent of a target ligand.[120, 125, 133, 134] These methods thus also consider binding sites which are too small (*e.g.* small ligand or metal ion binding sites), or do not contain the required physico-chemical properties to bind the target ligand. Besides DynaBiS, Pep-SiteFinder is one of the few ligand-specific binding site identification methods. Pep-SiteFinder is however limited to the identification of peptide binding sites, as the algorithm consists of a blind docking approach with the ATTRACT force field and pre-generated peptide conformations.[139] Moreover, Pep-SiteFinder applies a rigid docking strategy, in contrast to the fully flexible treatment of both the ligand and protein by DynaBiS.

The major limitation of DynaBiS lies in the scoring function, as the pepscore is used to rank the sampled binding sites. The pepscore was developed for peptide docking within the DynaDock algorithm, and trained on a set consisting of 15 peptides by Iris Antes.[169] Thus, the performance of DynaBiS could be further improved by the design of a scoring function specifically designed for the identification of binding sites rather than docked poses, and trained on a set consisting of both peptides and small ligands. Nevertheless, the strong performance of DynaBiS illustrates that explicit consideration of protein-ligand flexibility can strongly improve binding site identification, especially for binding sites for large and flexible ligands.

4.1.5 Fragment-based molecular docking combined with QM/MM simulations

As soon as the binding site is known, *e.g.* identified by DynaBiS, molecular docking simulations can be applied to predict the binding mode of a ligand in the binding site. These molecular docking algorithms perform rather well for small ligands, but docking of peptides, or other large and flexible ligands, remains challenging.[292, 293] For example, Chapter 3.5 describes the characterization of the amidohydrolase from *Streptomyces purpureus* (PurAH), which led to the identification of this enzyme as the "gatekeeper" of bottromycin synthesis.[229] These conclusions were supported by an in-depth analysis of the binding site including binding site volume calculations, as well as structural comparison to other homologues. Further, *in silico* analysis and molecular dynamics simulations also led to the suggestion of several interesting mutation positions to probe the promiscuity of PurAH, which were further analyzed *in vitro* by site-directed mutagenesis. Furthermore, the X-Ray structure of PurAH was resolved in this study, however no electron density was observed for enzyme-bound substrate (PDB-ID: 6i5s). However, it would be of interest to know how the bottromycin precursor exactly binds in the PurAH binding site, as this can guide engineering studies leading to a more efficient bottromycin biocatalysis. Initial molecular docking simulations applying AutoDock and FlexX, which both rely on the rigid protein approximation, did not result in any poses which were biochemically feasible (*i.e.* in which the ligand is positioned such that the hydrolysis reaction could occur). Since an apo-structure of the protein is used during docking (there is no complex structure available for this or a related protein, as is common for RiPPs enzymes), these failed docking simulations indicate that an induced fit of the binding site upon ligand binding may be important. This necessitates the flexible treatment of the active site during the docking process. Side-chain flexibility, as in AutoDock (see Chapter 1.5.2) might not suffice in the case of PurAH due to the large binding site (**Figure 5B**). Moreover, due to the lack of ligand-bound structures, it is unknown (I) which and how many side chains are flexible and (II) if and to what extend backbone movement is involved in binding site adaptation. Extensive previous studies demonstrated that in such cases molecular dynamics-based

molecular docking methods as implemented in DynaDock are needed to obtain realistic bound protein-ligand conformations.[169, 185, 294] Therefore, DynaDock may be the most suitable algorithm for the treatment of binding site flexibility in the docking pipeline in PurAH.

As described above, docking of ligands with many rotatable bonds, such as peptides, remains challenging due to their large degrees of freedom. This was recognized in 1996 by Rarey *et al*.[150] and led to the development of the FlexX incremental construction algorithm. In this method, the ligand is fragmented into small, rigid substructures prior to docking, and the docking procedure consists of the incremental reconstruction of the ligand from these fragments under consideration of all possible conformational angles between these fragments. Nowadays, fragment-based docking is a common way to dock large peptides. For example, Liao *et al*.[295] showed that the docking performance of a test case consisting of 17 peptides could be improved significantly by first docking the individual fragments (peptide was fragmented in two halves in this study) with AutoDock Vina. After placement, the fragments were combined and refined by molecular dynamics simulations. However, in that strategy, protein flexibility is only included in the final simulation, as the docking of the fragments is performed using a rigid protein conformation.

A molecular docking study of the bottromycin precursor in PurAH thus requires a combination of DynaDock and a fragment-based docking strategy. To illustrate if this suggested docking protocol could indeed be suitable for PurAH, this docking protocol was evaluated for this discussion. The substrate was truncated with a methylamine after four follower peptide residues, which was subsequently subdivided in four fragments (**Figure 10A**). Since the cleavage site of the substrate is known, a pharmacophore constraint (see Chapter 1.5.4) enforcing the cleavage site to be located near the catalytic hydroxide ion was applied during docking of the first fragment, *i.e.* the thiazoline moiety. The pharmacophore constraint was defined as follows: (I) The distance between the geometric centers of the ligand and the binding site (defined as the geometric center of the two Zn^{2+} ions) was enforced to be less than 0.35 nm, and (II) all vacant coordination sites of both Zn^{2+} ions needed to be filled after base placement. Since the binding mode of the base fragment strongly affects the placement of the remaining fragments, a QM/MM geometry optimization was performed on this docked base fragment. This refinement optimizes the placement of the base fragment, and ensures a correct coordination geometry of the Zn^{2+} ions and the catalytic hydroxide ion in the active site. Subsequently, the two adjacent fragments were docked, followed by the macrocycle moiety. Finally, a 150 ns molecular dynamics simulation was performed, which after a short equilibration phase showed stable substrate binding (**Figure 10D**). The final pose suggested here is

the representative pose after clustering the converged section of the molecular dynamics simulation. Besides the native substrate, the substrate analogue which was used in the *in vitro* measurements was also docked applying the same protocol, which resulted in a highly similar docked pose (**Figure 10B**), indicating that the experimental measurements in our PurAH study will most likely be valid for the native substrate as well.

This docked pose fills the vacant coordination positions of both Zn^{2+} ions: Zn_1^{2+} is coordinated by D7 of the substrate leading to a tetrahedral coordination geometry (H210, H229 and OH^-), and an interaction with the thiazoline residue fulfills the octahedral coordination geometry for Zn_2^{2+} (H94, H96, D348, Kcx183, OH^-; **Figure 10A,B**). This observation is in agreement with the substrate mutation study described in Chapter 3.5, as PurAH did not process a substrate carrying a valine instead of aspartate at position 7, probably due to the lacking interaction with Zn_1^{2+}. A substrate with an alanine at this position was however still partially processed.[229] To understand this observation, a short molecular dynamics simulation was performed with this modified substrate, which showed a slight reorientation of the substrate such that the backbone carbonyl oxygen of F6 the takes over the coordination of Zn_1^{2+} (data not shown), which could potentially explain the observed activity for this substrate. Furthermore, a stable pi-stacking interaction between F6 of the substrate and W118 was predicted, an interaction which is assumed to play an important role in substrate binding. Finally, the macrocycle moiety of the substrate, which we showed to be essential for substrate recognition by PurAH, binds in a hydrophobic patch in the binding site. This docked pose also fits the performed mutations in PurAH rather well, as substrates carrying the mutations P2A, V4L, F6W or D7N were all determined to be PurAH substrates, and full conversion was observed. Moreover, the Y185F mutation in PurAH showed reduced activity to the natural substrate due to impaired hydrogen-bonding with D8, and the D348N variant was fully inactive, either due to loss of the catalytic base, or impaired Zn^{2+} binding.[229] During the DynaDock OPMD refinement simulations, as well as the subsequent molecular dynamics simulations, opening of a hydrophobic sub-pocket was observed. The opening/closing of this sub-pocket is controlled by R279, and binds M9 of the substrate during the docking procedure (**Figure 10E**). This observation is in agreement with the poor side-chain electron density observed for R279 in the X-Ray structure of PurAH, indicating side-chain flexibility. Moreover, *in vitro* studies with mutated substrates agree with the observation of this sub-pocket, as a substrate carrying the M9A mutation was barely processed, while a substrate carrying the M9F mutation, which narrowly fits into the sub-pocket, was partially processed. The required opening of the sub-pocket may explain the failed initial docking simulations, and thereby further illustrate the importance of docking algorithms allowing for a fully flexible treatment of both the protein and the ligand.

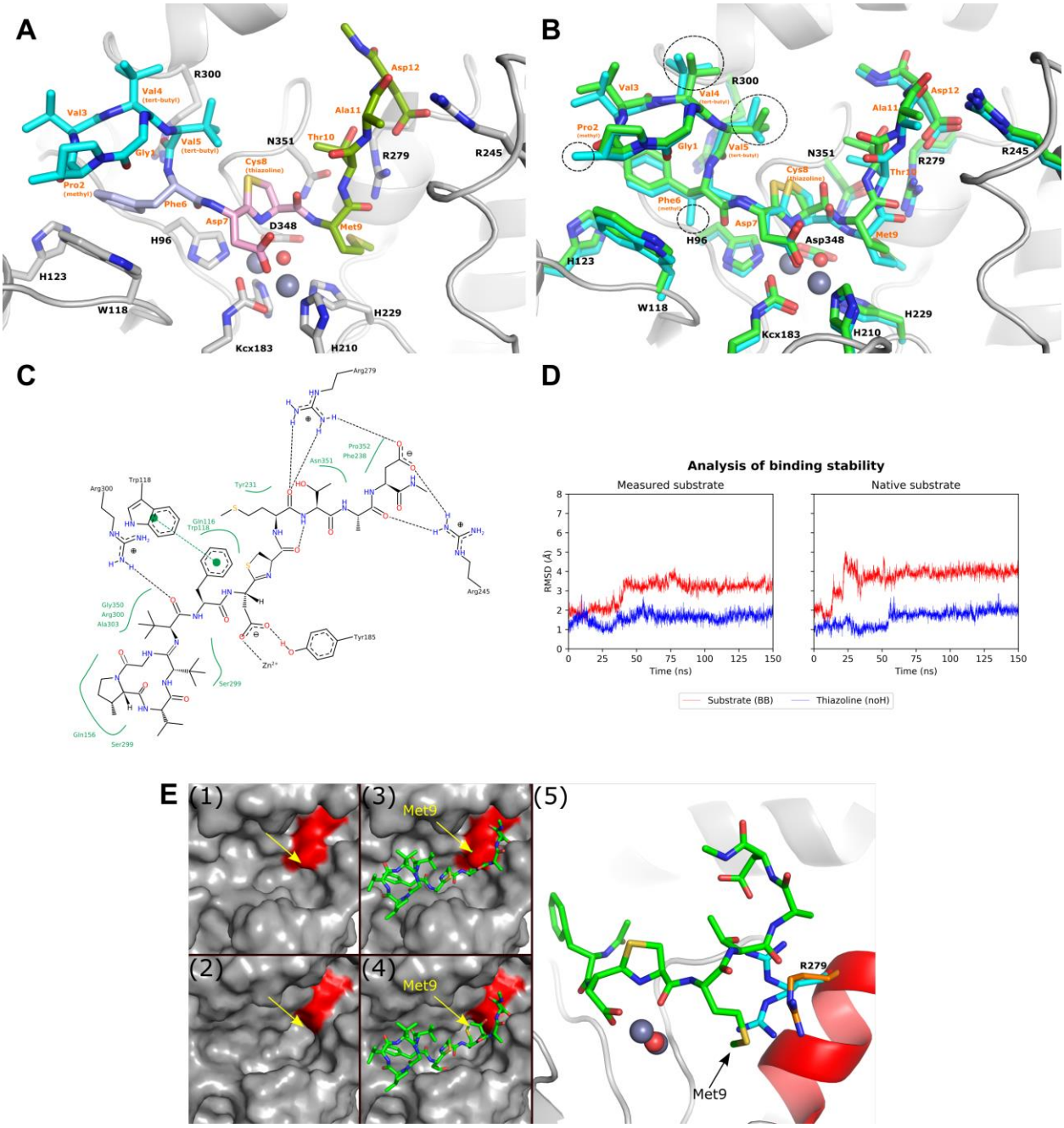


Figure 10. The results of the docking study of (A) a truncated native substrate (with the substrate colored according to the applied fragmentation), and (B) a truncated substrate analogue as used in the *in vitro* experiments, with the native and measured substrate in cyan and green sticks, respectively. The substrate residues are labeled by their three-letter code, the enzyme residues by their one-letter code (Kcx represents a carboxylated lysine residue), and differences between the native and measured substrate are highlighted in black dashed circles. The substrate-enzyme interactions (C) are shown in a PoseView scheme, and (D) the RMSD course of 150 ns molecular dynamics simulations for the measured and native substrate are plotted, where the RMSD of the substrate backbone atoms is shown in red, and the heavy-atom RMSD of the thiazoline residue (*i.e.* the cleavage site) in blue. The multi-panel section (E) shows the suggested opening of the arginine-controlled pocket. A surface representation of the binding site with closed (panel 1 & 3) and open (panel 2 & 4) sub-pocket, with (panel 1 & 2) and without (panel 3 & 4) docked substrate. The arrows indicate the sub-pocket occupied by Met9, and the region controlled by R279 is colored red. The two alternate conformations of R279 derived from the electron density are shown in cyan (panel 5), and the corresponding conformation after refinement is shown in orange.

Enzymes belonging to the YcaO superfamily (*i.e.* enzymes involved in the biosynthesis of thiazole-containing antibiotics), as well as other enzymes involved in the biosynthesis of RiPPs with antimicrobial activity, are still actively studied and show potential to become useful biocatalysts due to their unique structural features.[229, 232, 278, 279] Therefore, knowledge about the binding mode of the respective substrates in these enzymes is highly valuable, and can guide protein engineering studies to improve enzyme activity and/or its substrate scope. The detailed knowledge about the binding mode of the bottromycin precursor in PurAH presented here can thus guide the development of biocatalysts for bottromycin derivatives with antibiotic activity, and other RiPPs.

4.2 BIOMOLECULAR SIMULATIONS OF METALLOPROTEINS

4.2.1 Classical simulations

Biomolecular force field-based description of interactions between a metal ion and its biochemical (*e.g.* protein) environment is highly challenging, because the complex nature of a metal ion needs to be described with (non-bonded) force field parameters of only a single atom. Therefore, numerous versatile metal ion models were presented over the last years. The first models consist of standard LJ 12-6 parameters, *i.e.* the well depth and R_{min} values. Such a description is however often too limited to accurately describe all properties of a metal ion.[275] Zhang *et al*.[296] did though manage to simulate accurate solvation free energies and several structural properties with a traditional LJ 12-6 model, but needed a rather high value for the well depth (ε): they applied a well depth of 295.5289 kcal·mol^{-1}, while the well depths for the non-bonded Zn^{2+} models from Li *et al*. ranged from $7.16 \cdot 10^{-4}$ to $1.49 \cdot 10^{-2}$ kcal·mol^{-1}. The benchmarking study (see Chapter 3.3) showed that the LJ 12-6 non-bonded models often fail to describe the interaction between Zn^{2+} and non-charged ligating atoms, such as the imidazole nitrogen in histidine, generally leading to distorted coordination geometries. While the 12-6-4 LJ-type model improved the description to soft-bases, we showed that this model can only simulate octahedral coordination geometries, and over-estimate the interaction with charged ligating atoms. Dummy-atom models showed the best performance, but can only simulate the coordination geometry they were designed for. While this works perfectly for an octahedral coordination geometry, a tetrahedral geometry remains challenging to simulate. Thus, the development of a tetrahedral dummy atom model specifically parameterized for tetrahedral Zn^{2+} in a protein environment would be highly useful. The parameterization of a tetrahedral Zn^{2+} model is however not so straightforward. While octahedral dummy atom models are parameterized in bulk water, where the non-bonded force field parameters can be tuned to reproduce

the hydration free energy and ion-oxygen distance, tetrahedral Zn^{2+} does not have a counterpart in water. Thus, another parameterization strategy needs to be developed to design a tetrahedral Zn^{2+} model. In this dissertation, an attempt to parameterize such a tetrahedral model by reproducing the interaction energy profile calculated at the QM level (MP2/cc-pVTZ level of theory; **Figure 11**) is described (Chapter 3.3). The Lennard-Jones parameters of a tetrahedral Zn^{2+} dummy-atom model were fitted applying a weighted least-squared fitting approach, and both a LJ 12-6 and a 12-6-4 LJ-type model was designed. These models clearly reproduced the potential energy surface (PES) better in comparison to the already published models (**Figure 11**). However, these models did not result in a stable tetrahedral Zn^{2+} coordination in molecular dynamics simulations of CAII and VIM-2, indicating that solely fitting a dummy-atom model to reproduce a Zn^{2+}-water interaction energy scan is not suitable to design a tetrahedral Zn^{2+} model. The above strategy was also performed applying different charge distributions over the dummy atoms and the Zn^{2+} ion, but this did neither result in a better fit of the PES, nor to more accurate simulations. Moreover, this approach was further extended to additionally fit interaction energy profiles between a Zn^{2+} ion and amino acid analogues of histidine, cysteine, aspartate, and glutamate. A proper fit of these interaction energy patterns appeared to be challenging, as visualized here for a PES scan between Zn^{2+} and a cysteine analogue (**Figure 11B**), thus this approach is probably not a suitable fitting procedure to develop a tetrahedral dummy atom model for Zn^{2+}.

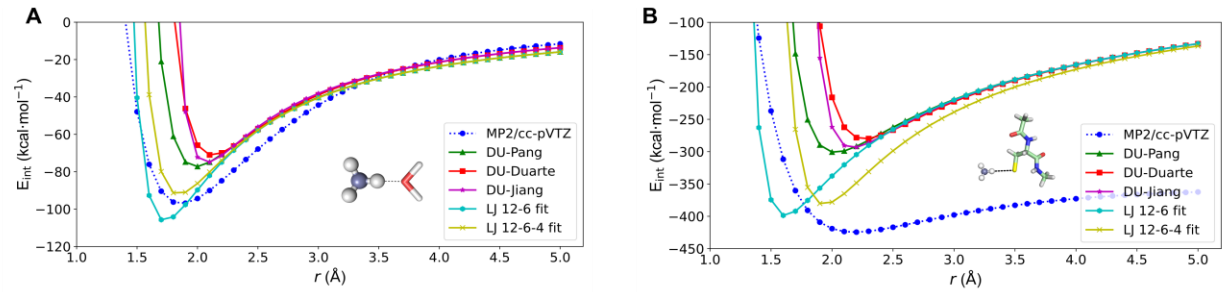


Figure 11. Potential energy surfaces between Zn^{2+} and (A) TIP3P water and (B) a cysteine analogue. The obtained surfaces for published parameters are shown (depicted as 'DU + first author'; note that DU-Duarte and DU-Jiang are octahedral dummy atom models) as well as the surfaces obtained from the newly fitted tetrahedral models in this dissertation. The reference surface measured at the MP2/cc-pVTZ level of theory is shown as a blue dotted line. The classical energy represents the full interaction energy, *i.e.* containing both the Lennard-Jones and Coulomb potentials.

The benchmarking study presented in this dissertation showed promising performance of a rather recent non-bonded model by Macchiagodena *et al.*[276, 297] In this model, new Zn^{2+} parameters were presented together with additional new Zn^{2+}-coordinating residues (histidine, cysteine, aspartate and glutamate) for the AMBER force field. These residues contain modified LJ-parameters and an altered charge distribution to improve the description with a Zn^{2+} ion. While showing excellent performance in this benchmarking

study regarding the description of Zn^{2+}-protein interactions (*e.g.* stable coordination and correct Zn^{2+}-ligating atom distances), it was observed that interactions with non-protein residues, such as ligands, are not properly described by this model. This led to unstable coordination of the Zn^{2+} ions in the ligand binding sites during the molecular dynamics simulations of CAII and VIM-2. Interestingly, replacing the Zn^{2+} parameters by the compromise model of Li *et al*.[275] led to more stable simulations and a correct tetrahedral coordination geometry of Zn^{2+} in CAII. These results indicate that there is potential to further improve this model to overcome its current limitations, which may lead to a Zn^{2+} model able to correctly simulate Zn^{2+} in a tetrahedral coordination geometry in proteins.

4.2.2 Hybrid QM/MM simulations: metallo-β-lactamases

The above-described limitations of a classical description of a metal site are especially problematic when a change of coordination geometry needs to be sampled. The dummy-atom and bonded models all enforce/support a certain coordination geometry by design, but also the non-bonded models clearly prefer a certain geometry (see Chapter 3.3). In the case of the studies performed in metallo-β-lactamase CSR-1 (*Cronobacter sakasakii*), the expected coordination geometries for both Zn^{2+} ions were unknown, since the binding mode of the inhibitor clavulanic acid was unclear (see Chapter 3.6). Therefore, QM/MM-geometry optimization was performed on a number of docked substrates to further refine the binding mode, as well as the coordination by the Zn^{2+} ion. The choice of QM-potential is not so obvious due to the large number of available QM-potentials, and the large difference in accuracy and required computing time between them. Most approaches aiming to improve the scoring capability of the docked substrates rely on semi-empirical (SE) methods due to their reduced computational costs (see Chapter 1.5.4 for a more in-depth discussion of this topic). However, because of the limited accuracy of the SE methods, a DFT functional was applied instead in this study to ensure accurate description of the Zn^{2+} ion and its surroundings.

The computations in this study elucidated the binding mode of the inhibitor clavulanic acid to CSR-1, as well as the interaction between clavulanic acid and K263, thereby suggesting an inhibition mechanism of clavulanic acid by displacing the Zn^{2+} ion in the β-site. The predicted binding mode is highly similar to what has been observed in the X-Ray structure of L1 MBL from *Stenotrophomonas maltophilia*, another B3-MBL, in complex with the hydrolysis product of moxalactam (**Figure 12**; PDB-ID: 2aio).[298] Only one other X-Ray structure of a B3-MBL in complex with a ligand is known, *i.e.* MIM-1 MBL from *Novosphingobium pentaromativorans* (PDB-ID: 6auf), which contains citric acid in the active site. However, no supporting literature has been published yet describing this structure. Both these enzymes contain the common

HHH/DHH motif for the α- and β-site, respectively. In L1, both Zn^{2+} ions are coordinated in an octahedral coordination geometry. The coordination geometry in the X-Ray structure of MIM-1 is less clear: the Zn^{2+} ion in the α-site is either present in a distorted tetrahedral, or distorted octahedral geometry with a vacant coordination site, dependent on the interpretation, while the Zn^{2+} ion in the β-site was found in a distorted octahedral geometry (**Figure 12**). Comparing these geometries to the predicted binding mode of clavulanic acid in the triple mutant of CSR-1 (*i.e.* with back-mutated HHH/DHH motif), both Zn^{2+} ions were observed in a perfect tetrahedral coordination geometry, which is the most common coordination geometry for Zn^{2+} in proteins according to a study by Laitaoja *et al*.[226] More interesting is the docked pose in the wild-type of CSR-1 (HRH/DQK motif), since inhibition by clavulanic acid was observed in this enzyme.[221] Here, the Zn^{2+} ion in the α-site was shown to be tetrahedral (**Figure 12**), which is in line with the expectations due to the missing H118. The β-site is expected to be empty because of highly unstable Zn^{2+} binding, which is in agreement with the missing density in the X-Ray structure. The hydrolysis product of clavulanic acid was also docked because one cannot rule out that the observed inhibition is caused by hydrolyzed clavulanic acid. In this pose, a trigonal bipyramidal coordination was observed for the Zn^{2+} ion in the α-site: the nitrogen in the hydrolyzed product was able to additionally coordinate the Zn^{2+} ion, which was looking for additional ligating partners due to the lacking H118 (**Figure 12**). Thus, comparing the predicted binding mode of clavulanic acid in CSR-1 and the coordination geometries of the Zn^{2+} ions between CSR-1 and other available X-Ray structures of B3-type MBLs indicate that the resulting Zn^{2+} coordination is in line with the expectations due to a missing ligating residue in CSR-1. Moreover, the predicted interactions between the Zn^{2+} ions and the ligands were similar to what has been observed in the X-Ray structures of homologues. All by all, the observed similarities with homologous structures further support the accuracy of the predicted protein-ligand complex, as well as the suggested inhibition mechanism. Therefore, it is likely that the conclusions draw in Chapter 3.6 can be transferred to other MBLs in the B3-RQK family as well.

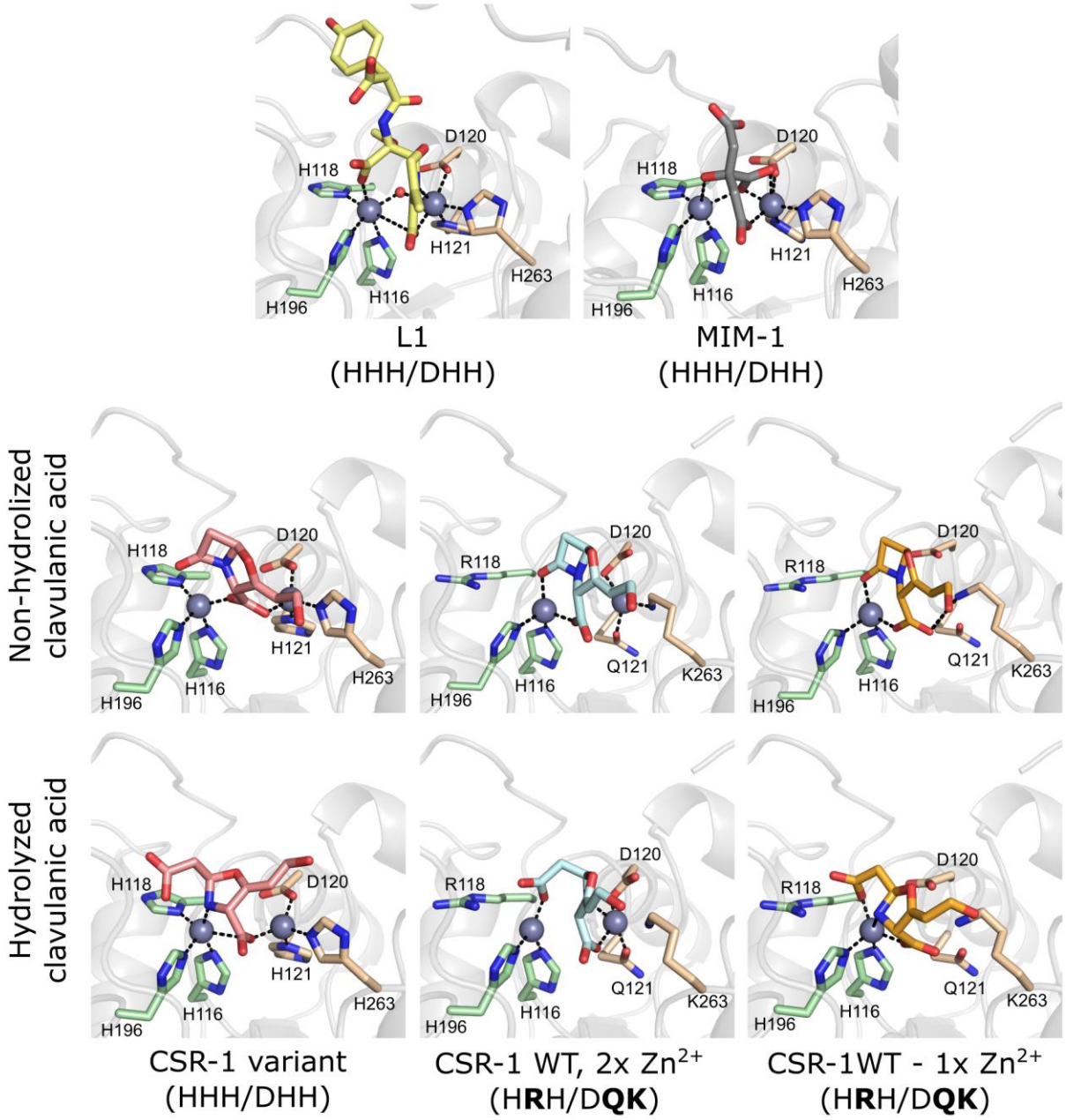


Figure 12. Comparison of binding modes observed in B3-type metallo-β-lactamases (MBLs). The ligating residues in the α- and β-site are shown as green and orange sticks, respectively. The active site motif is shown in brackets after the annotation, and variations to the most common motif are indicated in bold. Residues are numbered according to the BBL numbering scheme for class B β-lactamases. L1 MBL from *Stenotrophomonas maltophilia* (PDB-ID: 2aio) was co-crystallized with hydrolyzed moxalactam, and MIM-1 MBL from *Novosphingobium pentaromativorans* (PDB-ID: 6auf) was co-crystallized with citric acid in the active site. The binding modes of clavulanic acid in CSR-1 WT (PDB-ID: 6dq2) and a triple mutant of CSR-1 (PDB-ID: 6dr8) were obtained from docking simulations followed by QM/MM simulations, as described in this dissertation (Chapter 3.6).

4.3 RATIONAL ENZYME ENGINEERING: DIHYDROXY-ACID DEHYDRATASES

A rational engineering approach of dihydroxy-acid dehydratases (DHADs) is described in this dissertation as well (Chapter 3.4). Based on the multiple sequence alignments and superposition of structural models, alignment numbers were defined for all residues, allowing comparison of residues present at the same position in the three-dimensional structure, rather than a position in the sequence. Moreover, homology models of several DHADs belonging to different DHAD clusters were generated. Subsequent computational analysis of sequence- and structural data resulted in the identification of interesting mutation hotspots. Via subsequent *in vitro* site-directed and saturation mutagenesis, the role of these residues was elucidated, and engineered variants with improved activity and/or altered substrate scope were produced. This information can be used to further study enzymes belonging to the ilvD/EDD superfamily and guide engineering studies to optimize properties of dehydratases. For example, we engineered the recently discovered *Pu*DHT, and observed >6-fold improved activity for D-glycerate by introducing the H575F mutation, leading to a specific activity of 6.18 U/mg. This represents a >20-fold higher activity compared to the recently engineered *Ss*DHAD variant (I161M/Y145S/G205K) by Wang *et al*.[299], who applied a semi-rational engineering approach including iterative saturation mutagenesis, illustrating the relevance of these results. Compared to the wild-type *Ss*DHAD enzyme, which is generally applied in the biosynthesis of isobutanol or L-lactate from glucose or glycerol, this *Pu*DHT variant shows >200-fold improved activity.[8, 299, 300] Moreover, the activity of *Pu*DHT was measured at 30˚C vs 70˚C for *Ss*DHAD, which is much closer to typical temperatures at which enzymatic cascades are conducted.

In this study, a new classification scheme was proposed for [2Fe-2S]-dependent dehydratases belonging to the ilvD/EDD superfamily based on their substrate preference, in order to be able to clearly discriminate between them. Briefly, the dehydratases catalyzing predominantly DHIV are classified as branched chain acid dehydratases (BCADHTs), which includes dehydratases homologous to the DHAD from *Fontinomas thermophilia* (*Ft*DHAD), while dehydratases that are most active to sugar acids are called sugar acid dehydratases (SADHTs). The latter contain the DHAD from *Paralcaligenes ureilytius* (*Pu*DHT), the D-xylonate dehydratase from *Caulobacter crescentus* (*Cc*DHT), and the L-arabinonate dehydratase from *Rhizobium leguminosarum* (*Rl*ArDHT), among others. Finally, there are also dehydratases characterized with a less clear substrate preference, such as the DHAD from *Saccharolobus solfataricus* (*Ss*DHAD), which we therefore classified in promiscuous acid dehydratases (PADHTs). However, there are also dehydratases characterized which bind a [4Fe-4S] cluster instead, such as the DHAD from *Escherichia coli* (*Ec*DHAD) and the 6-phosphogluconate dehydratases (6-PGDHTs) from *Zymomonas mobilis* (*Zm*6PGDHT).[301, 302]

Moreover, Bayaraa *et al.[303]* recently characterized the DHAD from *Campylobacter jejuni* (*Cj*DHAD) and *Staphylococcus aureus* (*Sa*DHAD), and suggest that these enzymes bind a [4Fe-4S] cluster, but this could not be confirmed due to the lack of structural models. A phylogenetic tree generated from characterized dehydratases belonging to these classes (**Figure 13A**) shows the evolutionary relationship between them. The classification described above based on the substrate preference of dehydratases is also reflected in its phylogenetics, as enzymes belonging to these classes cluster together in their own clades. This phylogenetic study also resolves that BCADHTs and PADHTs are rather well connected, with sequence identities ranging from 31-79 % for the BCADHTs and PADHTs included in this dissertation (**Figure 13B**). Moreover, while the dihydroxy acid dehydratases (BCADHT, PADHT, and [4Fe-4S]-DHADs) share somewhat similarity between them, with sequence identities >28 %, the sugar acid dehydratases (SADHTs, as well as 6-PGDHT) are more distant from the other classes (seq. identity 17—26 % to the DHADs), illustrating an earlier separation during evolution. Interestingly the dehydratases which require a [4Fe-4S] cluster for their activity ([4Fe-4s]-DHADs and 6-PGDHTs) do not share a high sequence similarity to each-other, but are more similar to other [2Fe-2S]-dependent dehydratases, which may indicate that a small number of mutations might be sufficient to change the preference of a dehydratase from a [2Fe-2S] to a [4Fe-4S] cluster.

Several conserved residues and motifs can be observed in ilvD/EDD enzymes, such as the fully-conserved CDK motif (**Figure 13C**), which contains an FeS-coordinating cysteine, as well as an aspartate and (carboxylated) lysine residue, which both are involved in the coordination of a Mg^{2+} ion. Furthermore, the other Mg^{2+}-coordinating residues are fully conserved, with position 89 (*Pu*DHT numbering) either an aspartate or glutamate, as well as the catalytically essential serine (position 476 for *Pu*DHT) in a SGXX motif, with X being either T, S, or A, which was one of the motifs studied in this dissertation (Chapter 3.4; **Appendix 1**). While the FeS-coordinating cysteine residues C125 (CDK motif) and C198 (both *Pu*DHT numbering) are fully conserved within the ilvD/EDD superfamily, C57 in the PCN motif is only conserved within the [2Fe-2S]-containing dehydratases (*i.e.* SADHT, BCADHT, and PADHT). In this alignment position, a PGH and SAH motif is observed for [4Fe-4S]-DHADs and 6-PGDHTs, respectively (**Figure 13C**). Due to the nearby histidine (PGH and SAH), it has been suggested by numerous people including Bashiri *et al*.[212] that this histidine may coordinate the FeS instead.[212, 304] Rahman *et al*.[210] however suggested C112 (*So*6PGDHT numbering) within the CDG motif as a putative FeS-coordinating residue for 6-PGDHTs, but this cysteine is not present in the [4Fe-4S]-DHADs.

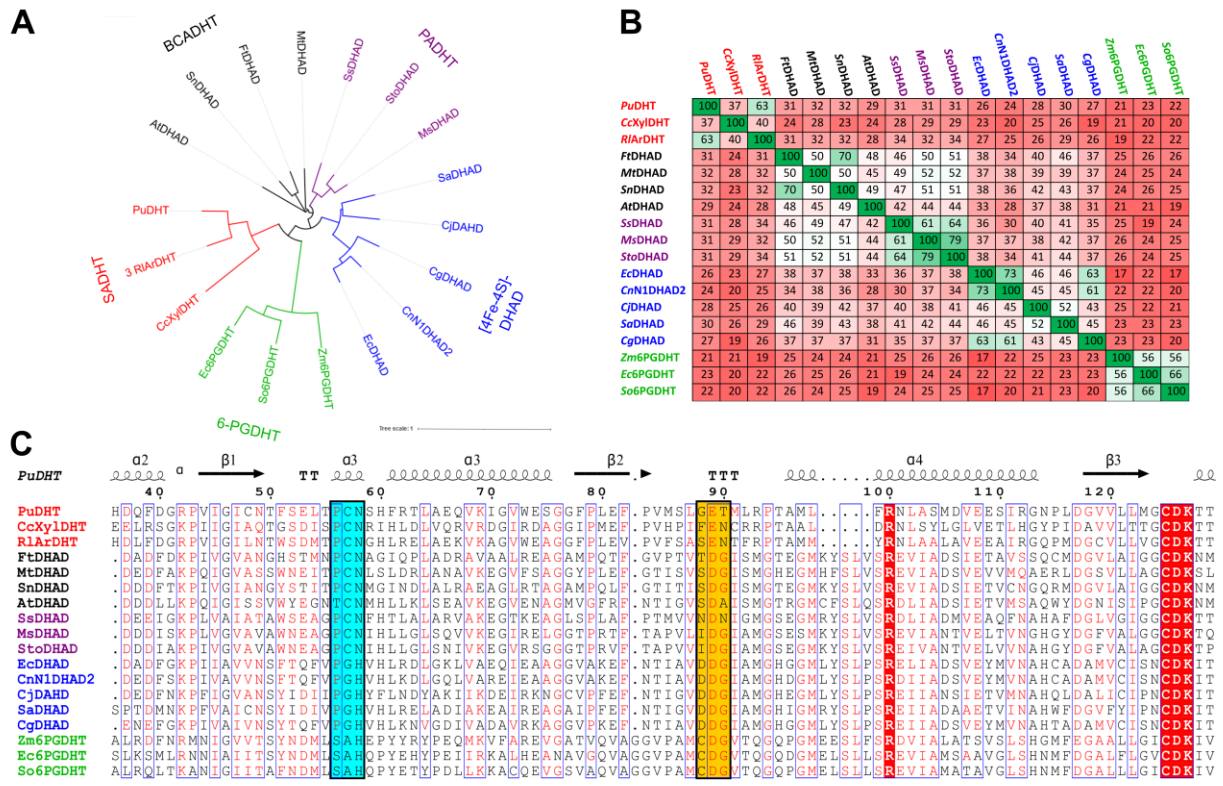


Figure 13. Sequence analysis of a selection of dehydratases belonging to the ilvD/EDD superfamily. The (A) phylogenetic tree and (B) sequence identity matrix are shown, illustrating the (evolutionary) relationship between these sequences. In (C) the multiple sequence alignment, conserved residues are depicted in red, with motifs containing the investigated FeS coordinating residues highlighted in cyan and yellow boxes. All dehydratases are colored based on their classification: 6-phosphogluconate dehydratases (6-PGDHTs; green), [4Fe-4S]-containing DHADs (blue), promiscuous acid dehydratases (PADHTs; magenta), sugar acid dehydratases (SADHTs; red) and branched chain acid dehydratases (BCADHTs; black).

By constructing a structural model of *Cj*DHAD followed by modelling of a [4Fe-4S] cluster in the active site, it can be observed that the suggested histidine (PGH and SAH motif) aligning the PCN motif is most likely too remote to coordinate the FeS cluster (**Figure 14**). This was also observed in AlphaFold structures of other [4Fe-4S]-DHADs and 6-PGDHTs (data not shown). However, this structural analysis indicates that D77 (*Cj*DHAD numbering) may be a good candidate to coordinate the third iron of the [4Fe-4S] cluster (**Figure 14**). While being less common than cysteine residues, aspartate is able to coordinate FeS clusters as well, as shown by Yonemoto *et al*. in a [NiFe]-hydrogenase.[305] Alternatively, D78 could also coordinate the [4Fe-4S] cluster, but since this residue is involved in Mg^{2+} binding, coordination by D77 is more likely. Moreover, this D77 is conserved in a DDG motif in [4Fe-4S]-DHADs, and aligned to a CDG motif in 6-PGDHTs, *i.e.* the cysteine suggested by Rahman *et al*. There is an X-Ray structure available for a 6-PGDHT from *Shewanella oneidensis* (*So*6PGDHT; PDB-ID: 2gp4), however, the model does not fit well in the observed electron density. Moreover, no FeS-coordinating cysteine residues could be found, partially due to the many unresolved loops around the active site. Rahman *et al*.[210] re-refined this structure

which led to a better fit to the observed electron density, where C112 (part of the CDG motif) and C154 are located in the active site, which could potentially coordinate the FeS cluster. However, the third cysteine was present in an unresolved loop. Therefore, AlphaFold structures for 6-PGDHTs, including *So*6PGDHT were generated for this discussion. These structures indeed suggest the cysteine within the CDG motif to coordinate the [4Fe-4S] in 6-PGDHTs (**Figure 14**). The final coordination position for the last iron is probably vacant, similarly to what has been observed for [2Fe-2S] dehydratases, as this iron may coordinate the substrate, and abstract the C3'OH proton to initiate the reaction.[40, 209]

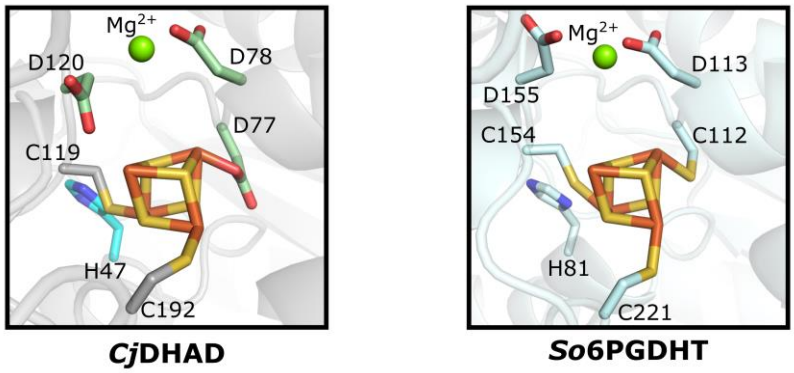


Figure 14. Structural models of the DHAD from *Campylobacter jejuni* (*Cj*DHAD) generated by AlphaFold, followed by manual modeling of a [4Fe-4S] cluster in the active site. The [4Fe-4S] cluster and important binding site residues, as well as His47, are shown in sticks, while the Mg^{2+} ion is represented as a green sphere.

In summary, based on sequence and structural analysis, this author suggests that besides the two fully conserved cysteine residues, the aspartate or cysteine in a respective DDG ([4Fe-4S]-DHADs) or CDG motif (6-PGDHTs) coordinates the [4Fe-4S] cluster. This hypothesis requires further experimental elucidation by site-directed mutagenesis and EPR studies, which could finally provide an explanation for the so far unclear, but often discussed FeS coordination in [4Fe-4S]-dependent dehydratases.

This page was left blank intentionally.

5 CONCLUSIONS AND OUTLOOK

In this dissertation, a complete enzyme engineering pathway has been followed and partially designed, with special emphasis on metalloproteins as target. The first part of this dissertation describes the development of the bioinformatics algorithm *EnzymeMatch*, which can support the search of enzymes catalyzing a target reaction on a user-specified substrate, and thus potentially fulfilling the role as biocatalyst. So far, existing algorithms require the structure of an enzyme as input, and search for proteins with a similar binding site/active site geometry. EnzymeMatch however only requires the target substrate as input, as it automatically designs an optimal binding site around it, and searches for enzymes with a matching binding site in the Protein Data Bank. The implementation of a graph theoretical method in EnzymeMatch allows for a highly efficient screening of all Protein Data Bank entries, as well as a straightforward ranking of the identified matched enzymes.

Alternatively, if the target enzyme or protein is already known, one may still not know the exact location of the binding site, which is essential information for (semi-)rational protein engineering studies. Since experimental elucidation of the binding site is challenging, several computational approaches have been developed to take over this task. However, the majority of these methods rely on the rigid-protein approximation, an approximation which is known to be often invalid, especially when only information about apo-structures is available. The second part of this dissertation describes the development and evaluation of *DynaBiS*, which explicitly includes both protein- and ligand flexibility. DynaBiS applies soft-core potentials and allows protein-ligand overlap to identify binding sites for large and flexible ligands, such as peptides. DynaBiS was shown to outperform the commonly applied binding site identification algorithms AutoSite and AutoDock, in particular for peptide binding sites. Moreover, the sampling algorithm behind DynaBiS was able to simulate the correct binding sites for all 26 evaluation systems, while predicting the correct site for 19 systems as top-ranked prediction. For all but one system, the correct binding site was present in the top five ranked binding site predictions. This strong performance illustrates the importance of a flexible treatment of both the protein and ligand in a computational binding site identification method. To make DynaBiS more broadly applicable, it would be worth implementing the DynaBiS algorithm in the newest DynaCell code (*i.e.* a wider in-house simulation package containing *e.g.* the DynaDock and OPMD algorithms), which is planned to be presented to the public in the near future. Moreover, scoring remains the biggest challenge in both molecular docking and binding site identification. The pepscore was applied as scoring function in DynaBiS, which is a force-field scoring function trained on 15 peptide binding sites during the development of DynaDock by Iris Antes.[169]

Thus, DynaBiS may strongly benefit from a further optimized scoring function, potentially containing additional empirical terms as well. Such a scoring function could also be developed for small molecules, which may positively affect the scoring capability of DynaBiS for small ligand binding sites.

In order to rationally design a protein engineering study, information about the role of binding site residues is essential. Molecular dynamics simulations can provide such information, making them a popular tool in rational protein engineering studies. However, many proteins, especially enzymes, contain a metal ion as cofactor in the binding/active site. Simulations of these metal sites with biomolecular force fields remains challenging, especially if the metal ion is present in a highly flexible site with numerous potential ligating atoms, which is often true in the active sites of enzymes. To this end, the third part of this dissertation describes a benchmarking study of numerous metal ion models to describe Zn^{2+} ions in ligand binding sites. This benchmarking study was performed for Zn^{2+}, as this is a highly versatile, biologically available metal ion, which is often found in catalytically relevant metalloproteins.[225, 226, 306] The performance of these Zn^{2+} models was evaluated on several aspects, including the ability to sample the correct coordination geometry, the preference of certain types of ligating atoms, as well as the overall stability of the simulation. This resulted in large performance differences and allowed suggestions of suitable simulation conditions for varying modelling approaches. This data further indicated that Zn^{2+} ions adopting a tetrahedral coordination are still not accurately described by existing Zn^{2+} models. Since the dummy-atom models performed best for the octahedral site and were able to enforce a coordination geometry while still allowing for ligand exchange, this type of model may be most suitable for a tetrahedral Zn^{2+} model as well. An initial design approach for a new tetrahedral Zn^{2+} model was not successful, illustrating that stable metal ligation is not fully reflected in ion-water interaction energy profiles. However, the model developed by Macchiagodena *et al*.[276, 297], in which new force field residues were developed for Zn^{2+}-coordinating residues was very promising. Therefore, a combination of this approach with a tetrahedral dummy-atom model may be an interesting parameterization procedure to explore.

Finally, the application of rational protein engineering with a strong computational input has been illustrated in several studies described in this dissertation. For example, sequence, structure, and activity relationships within [2Fe-2S]-dependent dehydratases were deduced, which led to the proposal of a new classification scheme reflecting substrate preference and evolutionary relationships. Moreover, mutation hotspots were identified which were further investigated with site-directed and saturation mutagenesis. This resulted in several variants with altered substrate preference and improved activity towards non-

native substrates. These results can guide further dehydratase engineering, ultimately leading to a more efficient production of biofuels and other fine chemicals. The structural information and sequence, structure, and activity relationships can be further extended to [4Fe-4S]-dependent dehydratases and potentially elucidate the so-far unknown coordination of the [4Fe-4S] cluster. For example, in this dissertation, the suggestion is raised that the first aspartate residue in the conserved DDG motif within [4Fe-4S]-dependent dehydratases coordinates the [4Fe-4S] cluster, rather than the previously suggested histidine in PGH/SAH motif. Moreover, in 6-phosphogluconate dehydrogenases, the cysteine in a CDG motif may coordinate the [4Fe-4S] cluster, which is present at the same alignment position as the DDG motif for [4Fe-4S]-dependent dehydratases. Experimental elucidation of this hypothesis, for example via site-directed mutagenesis, would be scientifically interesting. This would resolve the "missing cysteine" problem, which concerns the unknown [4Fe-4S]-coordinating residue, and provide valuable information about the structure of these dehydratases and promising biocatalysts.

The application of biomolecular simulations in metalloproteins has been illustrated in two structure-based drug discovery settings as well, namely in the characterization of an amidohydrolase PurAH, and in the rationalization of observed inhibition in metallo-β-lactamases. Both applications provide useful insights which can support the development and/or biosynthesis of drugs: in the first application a natural product with antibiotic activity, and in the second application a metallo-β-lactamase inhibitor. The development of such drugs are expected to have major impact in biomedicine, and computational chemistry can play a significant supporting role to accomplish these goals.

All by all, this dissertation describes the development of two new algorithms which can support in the identification of potential biocatalysts (EnzymeMatch) or identify so-far unknown binding sites for difficult targets (DynaBiS), and suggests suitable simulation conditions for metalloproteins. The role of computational biology has been illustrated in both an enzyme engineering setting showcased by the development of several [2Fe-2S]-dependent dehydratase variants, as well as in a drug-discovery setting showcased by molecular characterization of PurAH, and the rationalization of metallo-β-lactamase inhibition by clavulanic acid. The results and new algorithms presented here open up new strategies to further advance both rational enzyme engineering and structure-based drug discovery.

This page was left blank intentionally.

6 REFERENCES

1. Pauling, L., *Molecular architecture and biological reactions.* Chem. Eng. News, 1946. **24**(10): p. 1375-1377.
2. Lutz, S., and S.M. Iamurri, *Protein Engineering: Past, Present, and Future*, in *Protein Engineering: Methods and Protocols*, U.T. Bornscheuer and M. Höhne, Editors. 2018, Springer New York: New York, NY. p. 1-12.
3. Michaelis, L., and M.L. Menten, *Die kinetik der invertinwirkung.* Biochem. Z., 1913. **49**(333-369): p. 352.
4. Johnson, K.A., and R.S. Goody, *The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper.* Biochemistry, 2011. **50**(39): p. 8264-8269.
5. Chapman, J., A.E. Ismail, and C.Z. Dinu, *Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks.* Catalysts, 2018. **8**(6): p. 238.
6. Schmid, A., J.S. Dordick, B. Hauer, A. Kiener, M. Wubbolts*, et al.*, *Industrial biocatalysis today and tomorrow.* Nature, 2001. **409**(6817): p. 258-268.
7. Bilal, M., H.M.N. Iqbal, H. Hu, W. Wang, and X. Zhang, *Metabolic engineering and enzyme-mediated processing: A biotechnological venture towards biofuel production – A review.* Renew. Sust. Energ. Rev., 2018. **82**: p. 436-447.
8. Guterl, J.-K., D. Garbe, J. Carsten, F. Steffler, B. Sommer*, et al.*, *Cell-Free Metabolic Engineering: Production of Chemicals by Minimized Reaction Cascades.* ChemSusChem, 2012. **5**(11): p. 2165-2172.
9. Atsumi, S., T. Hanai, and J.C. Liao, *Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels.* Nature, 2008. **451**(7174): p. 86-89.
10. Cipolatti, E.P., M.C. Cerqueira Pinto, R.O. Henriques, J.C.C. da Silva Pinto, A.M. de Castro*, et al.*, *Chapter 5 - Enzymes in Green Chemistry: The State of the Art in Chemical Transformations*, in *Advances in Enzyme Technology*, R.S. Singh, et al., Editors. 2019, Elsevier. p. 137-151.
11. Abdelraheem, E.M.M., H. Busch, U. Hanefeld, and F. Tonin, *Biocatalysis explained: from pharmaceutical to bulk chemical production.* React. Chem. Eng., 2019. **4**(11): p. 1878-1894.
12. Li, S., X. Yang, S. Yang, M. Zhu, and X. Wang, *Technology Prospecting on Enzymes: Application, Marketing and Engineering.* Comput. Struct. Biotechnol. J., 2012. **2**(3): p. e201209017.
13. Wang, S., X. Meng, H. Zhou, Y. Liu, F. Secundo*, et al.*, *Enzyme Stability and Activity in Non-Aqueous Reaction Systems: A Mini Review.* Catalysts, 2016. **6**(2): p. 32.
14. Homaei, A.A., R. Sariri, F. Vianello, and R. Stevanato, *Enzyme immobilization: an update.* J. Chem. Biol., 2013. **6**(4): p. 185-205.
15. Woodley, J.M., *Accelerating the implementation of biocatalysis in industry.* Appl. Microbiol. Biotechnol., 2019. **103**(12): p. 4733-4739.
16. Sharma, A., G. Gupta, T. Ahmad, S. Mansoor, and B. Kaur, *Enzyme Engineering: Current Trends and Future Perspectives.* Food Rev. Int., 2021. **37**(2): p. 121-154.
17. Yoo, Y.J., Y. Feng, Y.H. Kim, and C.F.J. Yagonia, *Engineering Tools for Enzymes*, in *Fundamentals of Enzyme Engineering*. 2017, Springer Netherlands: Dordrecht. p. 87-100.
18. Reetz, M.T., M. Bocola, J.D. Carballeira, D. Zha, and A. Vogel, *Expanding the Range of Substrate Acceptance of Enzymes: Combinatorial Active-Site Saturation Test.* Angew. Chem. Int. Ed., 2005. **44**(27): p. 4192-4196.
19. Reetz, M.T., and J.D. Carballeira, *Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes.* Nat. Protoc., 2007. **2**(4): p. 891-903.
20. Reetz, M., *Making Enzymes Suitable for Organic Chemistry by Rational Protein Design.* ChemBioChem, 2022. **n/a**(n/a): p. e202200049.

21. Qu, G., A. Li, C.G. Acevedo-Rocha, Z. Sun, and M.T. Reetz, *The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes.* Angew. Chem. Int. Ed., 2020. **59**(32): p. 13204-13231.
22. Xu, J., Y. Cen, W. Singh, J. Fan, L. Wu*, et al.*, *Stereodivergent Protein Engineering of a Lipase To Access All Possible Stereoisomers of Chiral Esters with Two Stereocenters.* J. Am. Chem. Soc., 2019. **141**(19): p. 7934-7945.
23. Kuipers, R.K., H.-J. Joosten, W.J.H. van Berkel, N.G.H. Leferink, E. Rooijen*, et al.*, *3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities.* Proteins: Struct. Funct. Bioinform., 2010. **78**(9): p. 2101-2113.
24. Genz, M., O. Melse, S. Schmidt, C. Vickers, M. Dörr*, et al.*, *Engineering the Amine Transaminase from Vibrio fluvialis towards Branched-Chain Substrates.* ChemCatChem, 2016. **8**(20): p. 3199-3202.
25. van den Bergh, T., G. Tamo, A. Nobili, Y. Tao, T. Tan*, et al.*, *CorNet: Assigning function to networks of co-evolving residues by automated literature mining.* PLoS One, 2017. **12**(5): p. e0176427.
26. Da Costa, M., O. Gevaert, S. Van Overtveldt, J. Lange, H.-J. Joosten*, et al.*, *Structure-function relationships in NDP-sugar active SDR enzymes: Fingerprints for functional annotation and enzyme engineering.* Biotechnol. Adv., 2021. **48**: p. 107705.
27. Schymkowitz, J., J. Borg, F. Stricher, R. Nys, F. Rousseau*, et al.*, *The FoldX web server: an online force field.* Nucleic Acids Res., 2005. **33**(suppl_2): p. W382-W388.
28. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations.* J. Mol. Biol., 2002. **320**(2): p. 369-387.
29. Buß, O., J. Rudat, and K. Ochsenreither, *FoldX as Protein Engineering Tool: Better Than Random Based Approaches?* Comput. Struct. Biotechnol. J., 2018. **16**: p. 25-33.
30. Schwarte, A., M. Genz, L. Skalden, A. Nobili, C. Vickers*, et al.*, *NewProt – a protein engineering portal.* Protein Eng. Des. Sel., 2017. **30**(6): p. 441-447.
31. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat*, et al.*, *The Protein Data Bank.* Nucleic Acids Res., 2000. **28**(1): p. 235-242.
32. Korendovych, I.V., *Rational and Semirational Protein Design*, in *Protein Engineering: Methods and Protocols*, U.T. Bornscheuer and M. Höhne, Editors. 2018, Springer New York: New York, NY. p. 15-23.
33. Chen, K.H., S.P. Le, X. Han, J.M. Frias, and J.S. Nowick, *Alanine scan reveals modifiable residues in teixobactin.* Chem. Commun., 2017. **53**(82): p. 11357-11359.
34. Sanchez-Ruiz, J.M., *Protein kinetic stability.* Biophys. Chem., 2010. **148**(1): p. 1-15.
35. Childers, M.C., and V. Daggett, *Insights from molecular dynamics simulations for computational protein design.* Mol. Syst. Des. Eng., 2017. **2**(1): p. 9-33.
36. Gilis, D., and M. Rooman, *PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins.* Protein Eng. Des. Sel., 2000. **13**(12): p. 849-856.
37. Parthiban, V., M.M. Gromiha, and D. Schomburg, *CUPSAT: prediction of protein stability upon point mutations.* Nucleic Acids Res., 2006. **34**(suppl_2): p. W239-W242.
38. Capriotti, E., P. Fariselli, I. Rossi, and R. Casadio, *A three-state prediction of single point mutations on protein stability changes.* BMC Bioinformatics, 2008. **9**(2): p. S6.
39. Höltje, H.D., W. Sippl, D. Rognan, and G. Folkers, *Molecular Modeling. Basic Principles and Applications. 3rd Edition*. 2008, Weinheim: Wiley-VCH Verlag GmbH.
40. Melse, O., S. Sutiono, M. Haslbeck, G. Schenk, I. Antes*, et al.*, *Structure-Guided Modulation of the Catalytic Properties of [2Fe−2S]-Dependent Dehydratases.* ChemBioChem, 2022. **23**(10): p. e202200088.
41. Steiner, K., and H. Schwab, *Recent Advances in Rational Approaches for Enzyme Engineering.* Comput. Struct. Biotechnol. J., 2012. **2**(3): p. e201209010.

42. Grisewood, M.J., N.P. Gifford, R.J. Pantazes, Y. Li, P.C. Cirino*, et al.*, *OptZyme: Computational Enzyme Redesign Using Transition State Analogues.* PLoS One, 2013. **8**(10): p. e75358.
43. Jochens, H., and U.T. Bornscheuer, *Natural Diversity to Guide Focused Directed Evolution.* ChemBioChem, 2010. **11**(13): p. 1861-1866.
44. Röthlisberger, D., O. Khersonsky, A.M. Wollacott, L. Jiang, J. DeChancie*, et al.*, *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-195.
45. Siegel, J.B., A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert*, et al.*, *Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction.* Science, 2010. **329**(5989): p. 309-313.
46. Jiang, L., E.A. Althoff, F.R. Clemente, L. Doyle, D. Röthlisberger*, et al.*, *De Novo Computational Design of Retro-Aldol Enzymes.* Science, 2008. **319**(5868): p. 1387-1391.
47. Pan, X., and T. Kortemme, *Recent advances in de novo protein design: Principles, methods, and applications.* J. Biol. Chem., 2021. **296**: p. 100558.
48. Rouhani, M., F. Khodabakhsh, D. Norouzian, R.A. Cohan, and V. Valizadeh, *Molecular dynamics simulation for rational protein engineering: Present and future prospectus.* J. Mol. Graph. Model., 2018. **84**: p. 43-53.
49. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-590.
50. Karplus, M., *Molecular dynamics of biological macromolecules: A brief history and perspective.* Biopolymers, 2003. **68**(3): p. 350-358.
51. Spiwok, V., *Predictive Power of Biomolecular Simulations*, in *Biomolecular Simulations in Structure-Based Drug Discovery*. 2018. p. 1-26.
52. Huang, P.-S., S.E. Boyken, and D. Baker, *The coming of age of de novo protein design.* Nature, 2016. **537**(7620): p. 320-327.
53. C.I. Branden, J.T., *Introduction to Protein Structure (2nd ed.)*. 1998, New York: Garland Science, Taylor & Francis Group.
54. Martí-Renom, M.A., A.C. Stuart, A. Fiser, R. Sánchez, F. Melo*, et al.*, *Comparative protein structure modeling of genes and genomes.* Annu. Rev. Biophys. Biomol. Struct., 2000. **29**: p. 291-325.
55. Waterhouse, A., M. Bertoni, S. Bienert, G. Studer, G. Tauriello*, et al.*, *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Res., 2018. **46**(W1): p. W296-W303.
56. Schmidt, T., A. Bergner, and T. Schwede, *Modelling three-dimensional protein structures for applications in drug design.* Drug Discov. Today, 2014. **19**(7): p. 890-897.
57. Kundrotas, P.J., Z. Zhu, J. Janin, and I.A. Vakser, *Templates are available to model nearly all complexes of structurally characterized proteins.* Proc. Natl. Acad. Sci. U.S.A., 2012. **109**(24): p. 9438-9441.
58. Kiefer, F., K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, *The SWISS-MODEL Repository and associated resources.* Nucleic Acids Res., 2008. **37**(suppl_1): p. D387-D392.
59. Fiser, A., R.K.G. Do, and A. šAli, *Modeling of loops in protein structures.* Protein Sci., 2000. **9**(9): p. 1753-1773.
60. Song, Y., F. DiMaio, Ray Y.-R. Wang, D. Kim, C. Miles*, et al.*, *High-Resolution Comparative Modeling with RosettaCM.* Structure, 2013. **21**(10): p. 1735-1742.
61. Yang, J., R. Yan, A. Roy, D. Xu, J. Poisson*, et al.*, *The I-TASSER Suite: protein structure and function prediction.* Nat. Methods, 2015. **12**(1): p. 7-8.
62. Webb, B., and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr. Protoc. Bioinform., 2016. **54**(1): p. 5.6.1-5.6.37.
63. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang*, et al.*, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. **25**(17): p. 3389-3402.

64. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res., 2020. **49**(D1): p. D480-D489.
65. Sadowski, M.I., and D.T. Jones, *The sequence–structure relationship and protein function prediction.* Curr. Opin. Struct. Biol., 2009. **19**(3): p. 357-362.
66. Krissinel, E., *On the relationship between sequence and structure similarities in proteomics.* Bioinformatics, 2007. **23**(6): p. 717-723.
67. Alexander, P.A., Y. He, Y. Chen, J. Orban, and P.N. Bryan, *The design and characterization of two proteins with 88% sequence identity but different structure and function.* Proc. Natl. Acad. Sci. U.S.A., 2007. **104**(29): p. 11963-11968.
68. Lesk, A.M., and C. Chothia, *How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins.* J. Mol. Biol., 1980. **136**(3): p. 225-270.
69. Studer, G., G. Tauriello, S. Bienert, M. Biasini, N. Johner*, et al.*, *ProMod3—A versatile homology modelling toolbox.* PLoS Comput. Biol., 2021. **17**(1): p. e1008667.
70. Šali, A., and T.L. Blundell, *Comparative Protein Modelling by Satisfaction of Spatial Restraints.* J. Mol. Biol., 1993. **234**(3): p. 779-815.
71. Li, Y., *Conformational Sampling in Template-free Protein Loop Structure Modeling: An Overview.* Comput. Struct. Biotechnol. J., 2013. **5**(6): p. e201302003.
72. Yajia, Z., K. Hauser, and L. Jingru. *Unbiased, scalable sampling of closed kinematic chains*. in *2013 IEEE International Conference on Robotics and Automation*. 2013.
73. Mandell, D.J., E.A. Coutsias, and T. Kortemme, *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nat. Methods, 2009. **6**(8): p. 551-552.
74. van Vlijmen, H.W.T., and M. Karplus, *PDB-based protein loop prediction: parameters for selection and methods for optimization11Edited by P. E. Wright.* J. Mol. Biol., 1997. **267**(4): p. 975-1001.
75. Huang, X., R. Pearce, and Y. Zhang, *FASPR: an open-source tool for fast and accurate protein side-chain packing.* Bioinformatics, 2020. **36**(12): p. 3758-3765.
76. Colbes, J., R.I. Corona, C. Lezcano, D. Rodríguez, and C.A. Brizuela, *Protein side-chain packing problem: is there still room for improvement?* Brief. Bioinform., 2016. **18**(6): p. 1033-1043.
77. Francis-Lyon, P., and P. Koehl, *Protein side-chain modeling with a protein-dependent optimized rotamer library.* Proteins: Struct. Funct. Bioinform., 2014. **82**(9): p. 2000-2017.
78. Zhang, W., and Y. Duan, *Grow to Fit Molecular Dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement.* Protein Eng. Des. Sel., 2006. **19**(2): p. 55-65.
79. Shapovalov, Maxim V., and Roland L. Dunbrack, Jr., *A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions.* Structure, 2011. **19**(6): p. 844-858.
80. Towse, C.-L., Steven J. Rysavy, Ivan M. Vulovic, and V. Daggett, *New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities.* Structure, 2016. **24**(1): p. 187-199.
81. Lovell, S.C., J.M. Word, J.S. Richardson, and D.C. Richardson, *The penultimate rotamer library.* Proteins: Struct. Funct. Bioinform., 2000. **40**(3): p. 389-408.
82. Miao, Z., and Y. Cao, *Quantifying side-chain conformational variations in protein structure.* Sci. Rep., 2016. **6**(1): p. 37024.
83. Krivov, G.G., M.V. Shapovalov, and R.L. Dunbrack Jr., *Improved prediction of protein side-chain conformations with SCWRL4.* Proteins: Struct. Funct. Bioinform., 2009. **77**(4): p. 778-795.
84. Lu, M., A.D. Dousis, and J. Ma, *OPUS-Rota: a fast and accurate method for side-chain modeling.* Protein Sci., 2008. **17**(9): p. 1576-1585.
85. Xu, G., T. Ma, J. Du, Q. Wang, and J. Ma, *OPUS-Rota2: An Improved Fast and Accurate Side-Chain Modeling Method.* J. Chem. Theory Comput., 2019. **15**(9): p. 5154-5160.

86. Hartmann, C., I. Antes, and T. Lengauer, *IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models.* Protein Sci., 2007. **16**(7): p. 1294-1307.
87. Benkert, P., M. Biasini, and T. Schwede, *Toward the estimation of the absolute quality of individual protein structure models.* Bioinformatics, 2011. **27**(3): p. 343-350.
88. Shen, M.-y., and A. Sali, *Statistical potential for assessment and prediction of protein structures.* Protein Sci., 2006. **15**(11): p. 2507-2524.
89. Benkert, P., S.C.E. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins: Struct. Funct. Bioinform., 2008. **71**(1): p. 261-277.
90. Studer, G., C. Rempfer, A.M. Waterhouse, R. Gumienny, J. Haas*, et al.*, *QMEANDisCo—distance constraints applied on model quality estimation.* Bioinformatics, 2019. **36**(6): p. 1765-1771.
91. Laskowski, R.A., M.W. MacArthur, D.S. Moss, and J.M. Thornton, *PROCHECK: a program to check the stereochemical quality of protein structures.* J. Appl. Crystallogr., 1993. **26**(2): p. 283-291.
92. Park, H., S. Ovchinnikov, D.E. Kim, F. DiMaio, and D. Baker, *Protein homology model refinement by large-scale energy optimization.* Proc. Natl. Acad. Sci. U.S.A., 2018. **115**(12): p. 3054-3059.
93. Raval, A., S. Piana, M.P. Eastwood, R.O. Dror, and D.E. Shaw, *Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.* Proteins: Struct. Funct. Bioinform., 2012. **80**(8): p. 2071-2079.
94. *Protein Structure Prediction Center*. n.d. accessed at 20.12.2021]; Available from: https://predictioncenter.org/.
95. Service, R.F., *'The game has changed.' AI triumphs at protein folding.* Science, 2020. **370**(6521): p. 1144-1145.
96. Kryshtafovych, A., T. Schwede, M. Topf, K. Fidelis, and J. Moult, *Critical assessment of methods of protein structure prediction (CASP)—Round XIII.* Proteins: Struct. Funct. Bioinform., 2019. **87**(12): p. 1011-1020.
97. Kryshtafovych, A., T. Schwede, M. Topf, K. Fidelis, and J. Moult, *Critical assessment of methods of protein structure prediction (CASP)—Round XIV.* Proteins: Struct. Funct. Bioinform., 2021. **89**(12): p. 1607-1617.
98. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov*, et al.*, *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
99. Senior, A.W., R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre*, et al.*, *Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13).* Proteins: Struct. Funct. Bioinform., 2019. **87**(12): p. 1141-1148.
100. Senior, A.W., R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre*, et al.*, *Improved protein structure prediction using potentials from deep learning.* Nature, 2020. **577**(7792): p. 706-710.
101. Skolnick, J., M. Gao, H. Zhou, and S. Singh, *AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function.* J. Chem. Inf. Model., 2021. **61**(10): p. 4827-4831.
102. Rubiera, C.O., *AlphaFold 2 is here: what's behind the structure prediction miracle*. 2021, Oxford Protein Informatics Group.
103. Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov*, et al.*, *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
104. Zhang, J., Q. Wang, B. Barz, Z. He, I. Kosztin*, et al.*, *MUFOLD: A new solution for protein 3D structure prediction.* Proteins: Struct. Funct. Bioinform., 2010. **78**(5): p. 1137-1152.
105. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones*, et al. Attention is all you need*. in *Adv. Neural Inf. Process. Syst.* 2017.

106. Khamsi, M.A., and W.A. Kirk, *An introduction to metric spaces and fixed point theory*. Vol. 53. 2011: John Wiley & Sons.
107. Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg*, et al.*, *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins: Struct. Funct. Bioinform., 2006. **65**(3): p. 712-725.
108. Satya, P.G., *Protein Flexibility: A Challenging Issue of Drug Discovery.* Curr. Chem. Biol., 2018. **12**(1): p. 3-13.
109. Cramer, P., *AlphaFold2 and the future of structural biology.* Nat. Struct. Mol. Biol., 2021. **28**(9): p. 704-705.
110. Higgins, M.K., *Can We AlphaFold Our Way Out of the Next Pandemic?* J. Mol. Biol., 2021. **433**(20): p. 167093.
111. Zhao, J., Y. Cao, and L. Zhang, *Exploring the computational methods for protein-ligand binding site prediction.* Comput. Struct. Biotechnol. J., 2020. **18**: p. 417-426.
112. Armon, A., D. Graur, and N. Ben-Tal, *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* J. Mol. Biol., 2001. **307**(1): p. 447-463.
113. Ding, Y., J. Tang, and F. Guo, *Identification of Protein–Ligand Binding Sites by Sequence Information and Ensemble Classifier.* J. Chem. Inf. Model., 2017. **57**(12): p. 3149-3161.
114. Sotriffer, C., and G. Klebe, *Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design.* Il Farmaco, 2002. **57**(3): p. 243-251.
115. Brady, G.P., and P.F.W. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* J. Comput. Aided Mol. Des., 2000. **14**(4): p. 383-401.
116. Levitt, D.G., and L.J. Banaszak, *POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids.* J. Mol. Graph., 1992. **10**(4): p. 229-234.
117. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J. Mol. Graph. Model., 1997. **15**(6): p. 359-363.
118. Liang, J., C. Woodward, and H. Edelsbrunner, *Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design.* Protein Sci., 1998. **7**(9): p. 1884-1897.
119. Venkatachalam, C.M., X. Jiang, T. Oldfield, and M. Waldman, *LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.* J. Mol. Graph. Model., 2003. **21**(4): p. 289-307.
120. Laurie, A.T.R., and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-1916.
121. Hernandez, M., D. Ghersi, and R. Sanchez, *SITEHOUND-web: a server for ligand binding site identification in protein structures.* Nucleic Acids Res., 2009. **37**(suppl_2): p. W413-W416.
122. Ngan, C.-H., D.R. Hall, B. Zerbe, L.E. Grove, D. Kozakov*, et al.*, *FTSite: high accuracy detection of ligand binding sites on unbound protein structures.* Bioinformatics, 2011. **28**(2): p. 286-287.
123. Lin, Y., S. Yoo, and R. Sanchez, *SiteComp: a server for ligand binding site analysis in protein structures.* Bioinformatics, 2012. **28**(8): p. 1172-1173.
124. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nat. Chem. Biol., 2009. **5**(11): p. 789-796.
125. Wagner, J.R., J. Sørensen, N. Hensley, C. Wong, C. Zhu*, et al.*, *POVME 3.0: Software for Mapping Binding Pocket Flexibility.* J. Chem. Theory Comput., 2017. **13**(9): p. 4584-4592.
126. Monet, D., N. Desdouits, M. Nilges, and A. Blondel, *mkgridXf: Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics.* J. Chem. Inf. Model., 2019. **59**(8): p. 3506-3518.

127. Brylinski, M., and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proc. Natl. Acad. Sci. U.S.A., 2008. **105**(1): p. 129.
128. Roche, D.B., S.J. Tetchner, and L.J. McGuffin, *FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins.* BMC Bioinformatics, 2011. **12**(1): p. 160.
129. Wass, M.N., L.A. Kelley, and M.J.E. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res., 2010. **38**(suppl_2): p. W469-W473.
130. Yang, J., A. Roy, and Y. Zhang, *Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment.* Bioinformatics, 2013. **29**(20): p. 2588-2595.
131. Yang, J., A. Roy, and Y. Zhang, *BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions.* Nucleic Acids Res., 2012. **41**(D1): p. D1096-D1103.
132. Roy, A., and Y. Zhang, *Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement.* Structure, 2012. **20**(6): p. 987-997.
133. Harris, R., A.J. Olson, and D.S. Goodsell, *Automated prediction of ligand-binding sites in proteins.* Proteins: Struct. Funct. Bioinform., 2008. **70**(4): p. 1506-1517.
134. Ravindranath, P.A., and M.F. Sanner, *AutoSite: an automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms.* Bioinformatics, 2016. **32**(20): p. 3142-3149.
135. Morris, G.M., R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew*, et al.*, *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J. Comput. Chem., 2009. **30**(16): p. 2785-2791.
136. Hetényi, C., and D. van der Spoel, *Efficient docking of peptides to proteins without prior knowledge of the binding site.* Protein Sci., 2002. **11**(7): p. 1729-1737.
137. Hetényi, C., and D. van der Spoel, *Blind docking of drug-sized compounds to proteins with up to a thousand residues.* FEBS Lett., 2006. **580**(5): p. 1447-1450.
138. Campbell, S.J., N.D. Gold, R.M. Jackson, and D.R. Westhead, *Ligand binding: functional site location, similarity and docking.* Curr. Opin. Struct. Biol., 2003. **13**(3): p. 389-395.
139. Saladin, A., J. Rey, P. Thévenet, M. Zacharias, G. Moroy*, et al.*, *PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces.* Nucleic Acids Res., 2014. **42**(W1): p. W221-W226.
140. Fiorucci, S., and M. Zacharias, *Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT.* Proteins: Struct. Funct. Bioinform., 2010. **78**(15): p. 3131-3139.
141. Leach, A.R., *Molecular Modelling: Principles and Applications, 2nd Edition*. 2001, Harlow: Pearson Education.
142. Salmaso, V., and S. Moro, *Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview.* Front. Pharmacol., 2018. **9**(923).
143. Fischer, E., *Einfluss der Configuration auf die Wirkung der Enzyme.* Ber. Dtsch. Chem. Ges., 1894. **27**(3): p. 2985-2993.
144. Pagadala, N.S., K. Syed, and J. Tuszynski, *Software for molecular docking: a review.* Biophys. Rev., 2017. **9**(2): p. 91-102.
145. Kuntz, I.D., J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin, *A geometric approach to macromolecule-ligand interactions.* J. Mol. Biol., 1982. **161**(2): p. 269-288.
146. Pinzi, L., and G. Rastelli, *Molecular Docking: Shifting Paradigms in Drug Discovery.* Int. J. Mol. Sci., 2019. **20**(18): p. 4331.
147. Torres, P.H.M., A.C.R. Sodero, P. Jofily, and F.P. Silva-Jr, *Key Topics in Molecular Docking for Drug Design.* Int. J. Mol. Sci., 2019. **20**(18): p. 4574.

148. Welch, W., J. Ruppert, and A.N. Jain, *Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites.* Chem. Biol., 1996. **3**(6): p. 449-462.
149. Leach, A.R., and I.D. Kuntz, *Conformational analysis of flexible ligands in macromolecular receptor sites.* J. Comput. Chem., 1992. **13**(6): p. 730-748.
150. Rarey, M., B. Kramer, T. Lengauer, and G. Klebe, *A Fast Flexible Docking Method using an Incremental Construction Algorithm.* J. Mol. Biol., 1996. **261**(3): p. 470-489.
151. Fan, J., A. Fu, and L. Zhang, *Progress in molecular docking.* Quant. Biol., 2019. **7**(2): p. 83-89.
152. Jones, G., P. Willett, R.C. Glen, A.R. Leach, and R. Taylor, *Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen.* J. Mol. Biol., 1997. **267**(3): p. 727-748.
153. Trott, O., and A.J. Olson, *AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J. Comput. Chem., 2010. **31**(2): p. 455-461.
154. Verdonk, M.L., J.C. Cole, M.J. Hartshorn, C.W. Murray, and R.D. Taylor, *Improved protein–ligand docking using GOLD.* Proteins: Struct. Funct. Bioinform., 2003. **52**(4): p. 609-623.
155. Stank, A., D.B. Kokh, J.C. Fuller, and R.C. Wade, *Protein Binding Pocket Dynamics.* Acc. Chem. Res., 2016. **49**(5): p. 809-815.
156. Plattner, N., and F. Noé, *Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models.* Nat. Commun., 2015. **6**(1): p. 7653.
157. Stefan, H., S.A.O.M. H., H. Bingding, R.F. F., C. Gabriele*, et al.*, *Computational approaches to identifying and characterizing protein binding sites for ligand design.* J. Mol. Recognit., 2010. **23**(2): p. 209-219.
158. Keller, B.G., S. Aleksić, and L. Donati, *Markov State Models in Drug Design*, in *Biomolecular Simulations in Structure-Based Drug Discovery*. 2019. p. 67-86.
159. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc. Natl. Acad. Sci. U.S.A., 1958. **44**(2): p. 98-104.
160. Straub, F., and G. Szabolcsi, *Chapter Remarks on the Dynamic Aspects of Enzyme Structure (in Russian)*, in *Molecular biology, problems and perspectives*. 1964, Nauka: Moscow. p. 182-187.
161. Changeux, J.-P., and S. Edelstein, *Conformational selection or induced fit? 50 years of debate resolved.* F1000 Biol. Rep., 2011. **3**: p. 19-19.
162. Csermely, P., R. Palotai, and R. Nussinov, *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events.* Nat. Preced., 2010.
163. Gianni, S., J. Dogan, and P. Jemth, *Distinguishing induced fit from conformational selection.* Biophys. Chem., 2014. **189**: p. 33-39.
164. Ferrari, A.M., B.Q. Wei, L. Costantino, and B.K. Shoichet, *Soft Docking and Multiple Receptor Conformations in Virtual Screening.* J. Med. Chem., 2004. **47**(21): p. 5076-5084.
165. Totrov, M., and R. Abagyan, *Flexible ligand docking to multiple receptor conformations: a practical alternative.* Curr. Opin. Struct. Biol., 2008. **18**(2): p. 178-184.
166. Österberg, F., G.M. Morris, M.F. Sanner, A.J. Olson, and D.S. Goodsell, *Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock.* Proteins: Struct. Funct. Bioinform., 2002. **46**(1): p. 34-40.
167. Claußen, H., C. Buning, M. Rarey, and T. Lengauer, *FlexE: efficient molecular docking considering protein structure variations11Edited by J. Thornton.* J. Mol. Biol., 2001. **308**(2): p. 377-395.
168. Davis, I.W., and D. Baker, *RosettaLigand Docking with Full Ligand and Receptor Flexibility.* J. Mol. Biol., 2009. **385**(2): p. 381-392.
169. Antes, I., *DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility.* Proteins: Struct. Funct. Bioinform., 2010. **78**(5): p. 1084-1104.
170. Li, J., A. Fu, and L. Zhang, *An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking.* Interdiscip. Sci., 2019. **11**(2): p. 320-328.

171. Wan, S., B. Knapp, D.W. Wright, C.M. Deane, and P.V. Coveney, *Rapid, Precise, and Reproducible Prediction of Peptide–MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment.* J. Chem. Theory Comput., 2015. **11**(7): p. 3346-3356.
172. Wang, E., H. Sun, J. Wang, Z. Wang, H. Liu*, et al.*, *End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design.* Chem. Rev., 2019. **119**(16): p. 9478-9508.
173. Srinivasan, J., T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case, *Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices.* J. Am. Chem. Soc., 1998. **120**(37): p. 9401-9409.
174. Guedes, I.A., F.S.S. Pereira, and L.E. Dardenne, *Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges.* Front. Pharmacol., 2018. **9**(1089).
175. Wang, R., L. Lai, and S. Wang, *Further development and validation of empirical scoring functions for structure-based binding affinity prediction.* J. Comput. Aided Mol. Des., 2002. **16**(1): p. 11-26.
176. Murray, C.W., T.R. Auton, and M.D. Eldridge, *Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model.* J. Comput. Aided Mol. Des., 1998. **12**(5): p. 503-519.
177. Kramer, B., M. Rarey, and T. Lengauer, *Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking.* Proteins: Struct. Funct. Bioinform., 1999. **37**(2): p. 228-241.
178. Qing, X., X.Y. Lee, J. De Raeymaecker, J.R. Tame, K.Y. Zhang*, et al.*, *Pharmacophore modeling: advances, limitations, and current utility in drug discovery.* J. Recept. Ligand Channel Res., 2014. **7**: p. 81-92.
179. Muhammed, M.T., and A.-Y. Esin, *Pharmacophore Modeling in Drug Discovery: Methodology and Current Status.* J. Turk. Chem. Soc. A: Chem., 2021. **8**(3): p. 749-762.
180. Wermuth, C.G., C.R. Ganellin, P. Lindberg, and L.A. Mitscher, *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998).* Pure Appl. Chem., 1998. **70**(5): p. 1129-1143.
181. Alonso, H., A.A. Bliznyuk, and J.E. Gready, *Combining docking and molecular dynamic simulations in drug design.* Med. Res. Rev., 2006. **26**(5): p. 531-568.
182. Liu, K., and H. Kokubo, *Exploring the Stability of Ligand Binding Modes to Proteins by Molecular Dynamics Simulations: A Cross-docking Study.* J. Chem. Inf. Model., 2017. **57**(10): p. 2514-2522.
183. Guterres, H., and W. Im, *Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations.* J. Chem. Inf. Model., 2020. **60**(4): p. 2189-2198.
184. Król, M., R.A.G. Chaleil, A.L. Tournier, and P.A. Bates, *Implicit flexibility in protein docking: Cross-docking and local refinement.* Proteins: Struct. Funct. Bioinform., 2007. **69**(4): p. 750-757.
185. Ugur, I., M. Schatte, A. Marion, M. Glaser, M. Boenitz-Dulat*, et al.*, *Ca2+ binding induced sequential allosteric activation of sortase A: An example for ion-triggered conformational selection.* PLoS One, 2018. **13**(10): p. e0205057.
186. Xu, M., and M.A. Lill, *Induced fit docking, and the use of QM/MM methods in docking.* Drug Discov. Today Technol., 2013. **10**(3): p. e411-e418.
187. Burger, S.K., D.C. Thompson, and P.W. Ayers, *Quantum Mechanics/Molecular Mechanics Strategies for Docking Pose Refinement: Distinguishing between Binders and Decoys in Cytochrome c Peroxidase.* J. Chem. Inf. Model., 2011. **51**(1): p. 93-101.
188. Ahmadi, S., L. Barrios Herrera, M. Chehelamirani, J. Hostaš, S. Jalife*, et al.*, *Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review.* Int. J. Quantum Chem, 2018. **118**(9): p. e25558.
189. Senn, H.M., and W. Thiel, *QM/MM Methods for Biomolecular Systems.* Angew. Chem. Int. Ed., 2009. **48**(7): p. 1198-1229.

190. Aucar, M.G., and C.N. Cavasotto, *Molecular Docking Using Quantum Mechanical-Based Methods*, in *Quantum Mechanics in Drug Discovery*, A. Heifetz, Editor. 2020, Springer US: New York, NY. p. 269-284.
191. Pecina, A., S. Haldar, J. Fanfrlík, R. Meier, J. Řezáč*, et al.*, *SQM/COSMO Scoring Function at the DFTB3-D3H4 Level: Unique Identification of Native Protein–Ligand Poses.* J. Chem. Inf. Model., 2017. **57**(2): p. 127-132.
192. Cavasotto, C.N., and M.G. Aucar, *High-Throughput Docking Using Quantum Mechanical Scoring.* Front. Chem., 2020. **8**(246).
193. Chaskar, P., V. Zoete, and U.F. Röhrig, *Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function.* J. Chem. Inf. Model., 2014. **54**(11): p. 3137-3152.
194. Zhang, D., H. Li, H. Wang, and L. Li, *Docking accuracy enhanced by QM-derived protein charges.* Mol. Phys., 2016. **114**(20): p. 3015-3025.
195. Cho, A.E., V. Guallar, B.J. Berne, and R. Friesner, *Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach.* J. Comput. Chem., 2005. **26**(9): p. 915-931.
196. Chaskar, P., V. Zoete, and U.F. Röhrig, *On-the-Fly QM/MM Docking with Attracting Cavities.* J. Chem. Inf. Model., 2017. **57**(1): p. 73-84.
197. Raha, K., and K.M. Merz, *A Quantum Mechanics-Based Scoring Function: Study of Zinc Ion-Mediated Ligand Binding.* J. Am. Chem. Soc., 2004. **126**(4): p. 1020-1021.
198. Pecina, A., R. Meier, J. Fanfrlík, M. Lepšík, J. Řezáč*, et al.*, *The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein–ligand interactions and implicit COSMO solvation.* Chem. Commun., 2016. **52**(16): p. 3312-3315.
199. Ajani, H., A. Pecina, S.M. Eyrilmez, J. Fanfrlík, S. Haldar*, et al.*, *Superior Performance of the SQM/COSMO Scoring Functions in Native Pose Recognition of Diverse Protein–Ligand Complexes in Cognate Docking.* ACS Omega, 2017. **2**(7): p. 4022-4029.
200. Pecina, A., J. Brynda, L. Vrzal, R. Gnanasekaran, M. Hořejší*, et al.*, *Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II–Inhibitor Complexes.* ChemPhysChem, 2018. **19**(7): p. 873-879.
201. Andberg, M., N. Aro-Kärkkäinen, P. Carlson, M. Oja, S. Bozonnet*, et al.*, *Characterization and mutagenesis of two novel iron–sulphur cluster pentonate dehydratases.* Appl. Microbiol. Biotechnol., 2016. **100**(17): p. 7549-7563.
202. Sentheshanmuganathan, S., and S.R. Elsden, *The mechanism of the formation of tyrosol by Saccharomyces cerevisiae.* Biochemical Journal, 1958. **69**(2): p. 210-218.
203. Connor, M.R., and J.C. Liao, *Microbial production of advanced transportation fuels in non-natural hosts.* Curr. Opin. Biotechnol., 2009. **20**(3): p. 307-315.
204. Gmelch, T.J., J.M. Sperl, and V. Sieber, *Optimization of a reduced enzymatic reaction cascade for the production of L-alanine.* Sci. Rep., 2019. **9**(1): p. 11754.
205. Watanabe, S., N. Shimada, K. Tajima, T. Kodaki, and K. Makino, *Identification and Characterization of l-Arabonate Dehydratase, l-2-Keto-3-deoxyarabonate Dehydratase, and l-Arabinolactonase Involved in an Alternative Pathway of l-Arabinose Metabolism.* J. Biol. Chem., 2006. **281**(44): p. 33521-33536.
206. Jiang, Y., W. Liu, T. Cheng, Y. Cao, R. Zhang*, et al.*, *Characterization of D-xylonate dehydratase YjhG from Escherichia coli.* Bioengineered, 2015. **6**(4): p. 227-232.
207. Meloche, H., and W. Wood, *The mechanism of 6-phosphogluconic dehydrase.* J. Biol. Chem., 1964. **239**(10): p. 3505-3510.
208. Pirrung, M.C., C.P. Holmes, D.M. Horowitz, and D.S. Nunn, *Mechanism and stereochemistry of .alpha.,.beta.-dihydroxyacid dehydratase.* J. Am. Chem. Soc., 1991. **113**(3): p. 1020-1025.

209. Rahman, M.M., M. Andberg, S.K. Thangaraj, T. Parkkinen, M. Penttilä*, et al.*, *The Crystal Structure of a Bacterial l-Arabinonate Dehydratase Contains a [2Fe-2S] Cluster.* ACS Chem. Biol., 2017. **12**(7): p. 1919-1927.
210. Rahman, M.M., M. Andberg, A. Koivula, J. Rouvinen, and N. Hakulinen, *The crystal structure of D-xylonate dehydratase reveals functional features of enzymes from the Ilv/ED dehydratase family.* Sci. Rep., 2018. **8**(1): p. 865.
211. Yan, Y., Q. Liu, X. Zang, S. Yuan, U. Bat-Erdene*, et al.*, *Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action.* Nature, 2018. **559**(7714): p. 415-418.
212. Bashiri, G., T.L. Grove, S.S. Hegde, T. Lagautriere, G.J. Gerfen*, et al.*, *The active site of the Mycobacterium tuberculosis branched-chain amino acid biosynthesis enzyme dihydroxyacid dehydratase contains a 2Fe–2S cluster.* J. Biol. Chem., 2019. **294**(35): p. 13158-13170.
213. Zhang, P., B.S. MacTavish, G. Yang, M. Chen, J. Roh*, et al.*, *Cyanobacterial Dihydroxyacid Dehydratases Are a Promising Growth Inhibition Target.* ACS Chem. Biol., 2020. **15**(8): p. 2281-2288.
214. Bush, K., and P.A. Bradford, *Interplay between β-lactamases and new β-lactamase inhibitors.* Nat. Rev. Microbiol., 2019. **17**(5): p. 295-306.
215. Behzadi, P., H.A. García-Perdomo, T.M. Karpiński, and L. Issakhanian, *Metallo-ß-lactamases: a review.* Mol. Biol. Rep., 2020. **47**(8): p. 6281-6294.
216. Ambler, R.P., J. Baddiley, and E.P. Abraham, *The structure of β-lactamases.* Philos. Trans. R. Soc. Lond. B Biol. Sci., 1980. **289**(1036): p. 321-331.
217. Bebrone, C., *Metallo-β-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily.* Biochem. Pharmacol., 2007. **74**(12): p. 1686-1701.
218. McGeary, R.P., D.T. Tan, and G. Schenk, *Progress toward inhibitors of metallo-β-lactamases.* Future Med. Chem., 2017. **9**(7): p. 673-691.
219. Lee, J.H., M. Takahashi, J.H. Jeon, L.-W. Kang, M. Seki*, et al.*, *Dual activity of PNGM-1 pinpoints the evolutionary origin of subclass B3 metallo-β-lactamases: a molecular and evolutionary study.* Emerg. Microbes Infect., 2019. **8**(1): p. 1688-1700.
220. Hou, C.-F.D., E.K. Phelan, M. Miraula, D.L. Ollis, G. Schenk*, et al.*, *Unusual metallo-β-lactamases may constitute a new subgroup in this family of enzymes* Am. J. Mol. Biol., 2014. **4**(1): p. 11-15.
221. Pedroso, M.M., D.W. Waite, O. Melse, L. Wilson, N. Mitić*, et al.*, *Broad spectrum antibiotic-degrading metallo-β-lactamases are phylogenetically diverse.* Protein & Cell, 2020. **11**(8): p. 613-617.
222. Morán-Barrio, J., M.-N. Lisa, N. Larrieux, S.I. Drusin, A.M. Viale*, et al.*, *Crystal Structure of the Metallo-β-Lactamase GOB in the Periplasmic Dizinc Form Reveals an Unusual Metal Site.* Antimicrob. Agents Chemother., 2016. **60**(10): p. 6013-6022.
223. Vella, P., M. Miraula, E. Phelan, E.W.W. Leung, F. Ely*, et al.*, *Identification and characterization of an unusual metallo-β-lactamase from Serratia proteamaculans.* J. Biol. Inorg. Chem., 2013. **18**(7): p. 855-863.
224. Kovalevsky, A., M. Aggarwal, H. Velazquez, M.J. Cuneo, M.P. Blakeley*, et al.*, *"To Be or Not to Be" Protonated: Atomic Details of Human Carbonic Anhydrase-Clinical Drug Complexes by Neutron Crystallography and Simulation.* Structure, 2018. **26**(3): p. 383-390.e3.
225. Krishnamurthy, V.M., G.K. Kaufman, A.R. Urbach, I. Gitlin, K.L. Gudiksen*, et al.*, *Carbonic Anhydrase as a Model for Biophysical and Physical-Organic Studies of Proteins and Protein–Ligand Binding.* Chem. Rev., 2008. **108**(3): p. 946-1051.
226. Laitaoja, M., J. Valjakka, and J. Jänis, *Zinc Coordination Spheres in Protein Structures.* Inorg. Chem., 2013. **52**(19): p. 10983-10991.
227. Melse, O., and I. Antes, *Assessment of Molecular Mechanics-based Zn2+ Models in Mono- and Bimetallic Ligand Binding Sites.* bioRxiv, 2021: p. 2021.06.28.450184.

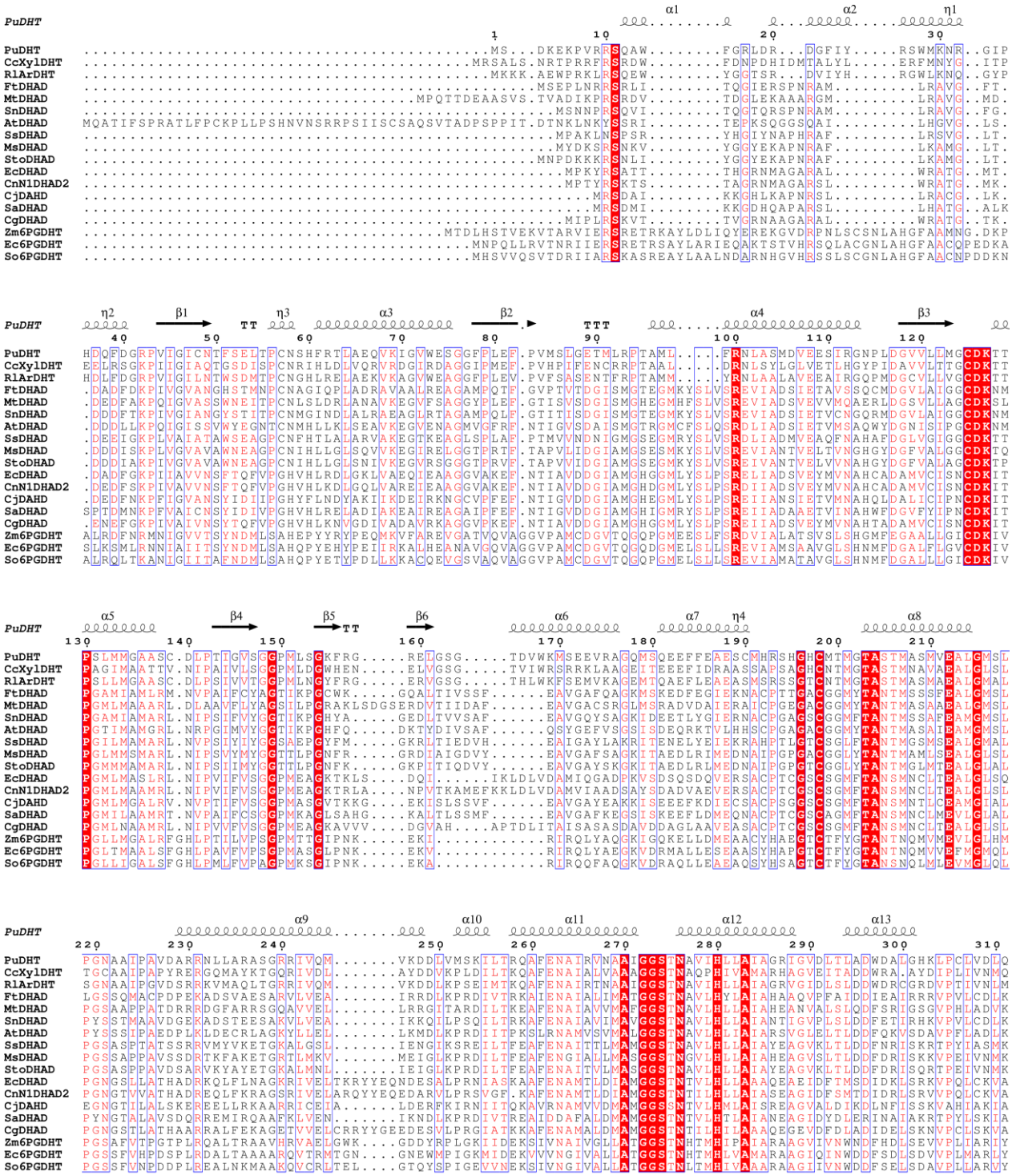
228. Liang, Z., Y. Xue, G. Behravan, B.H. Jonsson, and S. Lindskog, *Importance of the conserved active-site residues Try7, Glu106 and Thr199 for the catalytic function of human carbonic anhydrase II.* Eur. J. Biochem., 1993. **211**(3): p. 821-827.
229. Sikandar, A., L. Franz, O. Melse, I. Antes, and J. Koehnke, *Thiazoline-Specific Amidohydrolase PurAH Is the Gatekeeper of Bottromycin Biosynthesis.* J. Am. Chem. Soc., 2019. **141**(25): p. 9748-9752.
230. Waisvisz, J.M., M.G. van der Hoeven, J. van Peppen, and W.C.M. Zwennis, *Bottromycin. I. A New Sulfur-containing Antibiotic.* J. Am. Chem. Soc., 1957. **79**(16): p. 4520-4521.
231. Yamada, T., M. Yagita, Y. Kobayashi, G. Sennari, H. Shimamura*, et al.*, *Synthesis and Evaluation of Antibacterial Activity of Bottromycins.* J. Org. Chem., 2018. **83**(13): p. 7135-7149.
232. Eyles, T.H., N.M. Vior, and A.W. Truman, *Rapid and Robust Yeast-Mediated Pathway Refactoring Generates Multiple New Bottromycin-Related Metabolites.* ACS Synth. Biol., 2018. **7**(5): p. 1211-1218.
233. Cramer, C.J., *Essentials of computational chemistry: theories and models, Second Edition*. 2004: Wiley.
234. Jing, Z., C. Liu, S.Y. Cheng, R. Qi, B.D. Walker*, et al.*, *Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications.* Annu. Rev. Biophys., 2019. **48**(1): p. 371-394.
235. Kony, D., W. Damm, S. Stoll, and W.F. Van Gunsteren, *An improved OPLS–AA force field for carbohydrates.* J. Comput. Chem., 2002. **23**(15): p. 1416-1429.
236. Oostenbrink, C., A. Villa, A.E. Mark, and W.F. Van Gunsteren, *A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6.* J. Comput. Chem., 2004. **25**(13): p. 1656-1676.
237. Zhu, X., P.E.M. Lopes, and A.D. MacKerell Jr, *Recent developments and applications of the CHARMM force fields.* Wiley Interdiscip. Rev. Comput. Mol. Sci., 2012. **2**(1): p. 167-185.
238. Maier, J.A., C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser*, et al.*, *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J. Chem. Theory Comput., 2015. **11**(8): p. 3696-3713.
239. Li, P., and K.M. Merz, *Metal Ion Modeling Using Classical Mechanics.* Chem. Rev., 2017. **117**(3): p. 1564-1686.
240. Li, P., and K.M. Merz, *Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions.* J. Chem. Theory Comput., 2014. **10**(1): p. 289-297.
241. Li, P., L.F. Song, and K.M. Merz, *Parameterization of Highly Charged Metal Ions Using the 12-6-4 LJ-Type Nonbonded Model in Explicit Water.* J. Phys. Chem. B, 2015. **119**(3): p. 883-895.
242. Åqvist, J., and A. Warshel, *Free energy relationships in metalloenzyme-catalyzed reactions. Calculations of the effects of metal ion substitutions in staphylococcal nuclease.* J. Am. Chem. Soc., 1990. **112**(8): p. 2860-2868.
243. Zuo, Z., and J. Liu, *Assessing the Performance of the Nonbonded Mg2+ Models in a Two-Metal-Dependent Ribonuclease.* J. Chem. Inf. Model., 2019. **59**(1): p. 399-408.
244. Liao, Q., S.C.L. Kamerlin, and B. Strodel, *Development and Application of a Nonbonded Cu2+ Model That Includes the Jahn–Teller Effect.* J. Phys. Chem. Lett., 2015. **6**(13): p. 2657-2662.
245. Jiang, Y., H. Zhang, and T. Tan, *Rational Design of Methodology-Independent Metal Parameters Using a Nonbonded Dummy Model.* J. Chem. Theory Comput., 2016. **12**(7): p. 3250-3260.
246. Duarte, F., P. Bauer, A. Barrozo, B.A. Amrein, M. Purg*, et al.*, *Force Field Independent Metal Parameters Using a Nonbonded Dummy Model.* J. Phys. Chem. B, 2014. **118**(16): p. 4351-4362.
247. Berendsen, H.J., J.v. Postma, W.F. van Gunsteren, A. DiNola, and J. Haak, *Molecular dynamics with coupling to an external bath.* J. Chem. Phys., 1984. **81**(8): p. 3684-3690.
248. Ryckaert, J.-P., G. Ciccotti, and H.J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* J. Comput. Phys., 1977. **23**(3): p. 327-341.

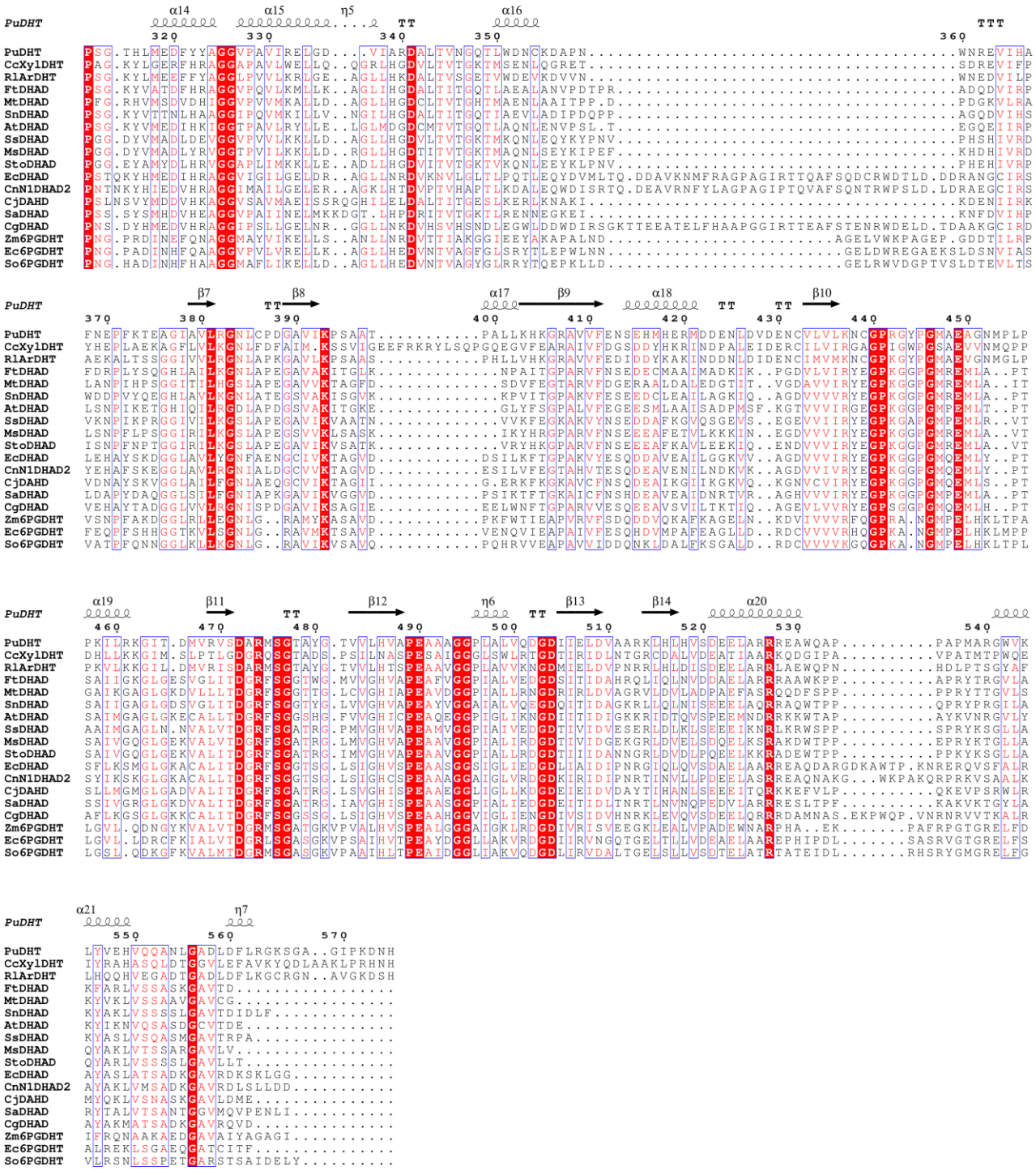
249. Li, P., and K.M. Merz, *MCPB.py: A Python Based Metal Center Parameter Builder.* J. Chem. Inf. Model., 2016. **56**(4): p. 599-604.
250. Yu, Z., P. Li, and K.M. Merz, *Extended Zinc AMBER Force Field (EZAFF).* J. Chem. Theory Comput., 2018. **14**(1): p. 242-254.
251. Seminario, J.M., *Calculation of intramolecular force fields from second-derivative tensors.* Int. J. Quantum Chem, 1996. **60**(7): p. 1271-1277.
252. Melse, O., S. Hecht, and I. Antes, *DynaBiS: A hierarchical sampling algorithm to identify flexible binding sites for large ligands and peptides.* Proteins: Struct. Funct. Bioinform., 2022. **90**(1): p. 18-32.
253. Taylor, R.D., P.J. Jewsbury, and J.W. Essex, *FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function.* J. Comput. Chem., 2003. **24**(13): p. 1637-1656.
254. Cao, L., and U. Ryde, *On the Difference Between Additive and Subtractive QM/MM Calculations.* Front. Chem., 2018. **6**(89).
255. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool.* J. Mol. Biol., 1990. **215**(3): p. 403-410.
256. Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus*, et al.*, *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol. Syst. Biol., 2011. **7**(1): p. 539.
257. Loncharich, R.J., B.R. Brooks, and R.W. Pastor, *Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide.* Biopolymers, 1992. **32**(5): p. 523-535.
258. Olsson, M.H.M., C.R. Søndergaard, M. Rostkowski, and J.H. Jensen, *PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions.* J. Chem. Theory Comput., 2011. **7**(2): p. 525-537.
259. Søndergaard, C.R., M.H.M. Olsson, M. Rostkowski, and J.H. Jensen, *Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values.* J. Chem. Theory Comput., 2011. **7**(7): p. 2284-2295.
260. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein, *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**(2): p. 926-935.
261. Wang, J., R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case, *Development and testing of a general amber force field.* J. Comput. Chem., 2004. **25**(9): p. 1157-74.
262. Frisch, M.J., G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb*, et al.*, *Gaussian 09, Revision E.01, Gaussian, Inc., Wallingford CT*. 2009.
263. Marion, A., M. Groll, D.H. Scharf, K. Scherlach, M. Glaser*, et al.*, *Gliotoxin Biosynthesis: Structure, Mechanism, and Metal Promiscuity of Carboxypeptidase GliJ.* ACS Chem. Biol., 2017. **12**(7): p. 1874-1882.
264. Essmann, U., L. Perera, M.L. Berkowitz, T. Darden, H. Lee*, et al.*, *A smooth particle mesh Ewald method.* J. Chem. Phys., 1995. **103**(19): p. 8577-8593.
265. Tao, J., J.P. Perdew, V.N. Staroverov, and G.E. Scuseria, *Climbing the Density Functional Ladder: Nonempirical Meta--Generalized Gradient Approximation Designed for Molecules and Solids.* Phys. Rev. Lett., 2003. **91**(14): p. 146401.
266. Eichkorn, K., O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs, *Auxiliary basis sets to approximate Coulomb potentials.* Chem. Phys. Lett., 1995. **240**(4): p. 283-290.
267. Eichkorn, K., F. Weigend, O. Treutler, and R. Ahlrichs, *Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials.* Theor. Chem. Acc., 1997. **97**(1): p. 119-124.
268. Grimme, S., J. Antony, S. Ehrlich, and H. Krieg, *A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu.* J. Chem. Phys., 2010. **132**(15): p. 154104.

269. Schäfer, A., H. Horn, and R. Ahlrichs, *Fully optimized contracted Gaussian basis sets for atoms Li to Kr.* J. Chem. Phys., 1992. **97**(4): p. 2571-2577.
270. Weigend, F., and R. Ahlrichs, *Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy.* Phys. Chem. Chem. Phys., 2005. **7**(18): p. 3297-3305.
271. Sherwood, P., A.H. de Vries, M.F. Guest, G. Schreckenbach, C.R.A. Catlow*, et al.*, *QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis.* J. Mol. Struct. Theochem, 2003. **632**(1): p. 1-28.
272. Ahlrichs, R., M. Bär, M. Häser, H. Horn, and C. Kölmel, *Electronic structure calculations on workstation computers: The program system turbomole.* Chem. Phys. Lett., 1989. **162**(3): p. 165-169.
273. Kästner, J., J.M. Carr, T.W. Keal, W. Thiel, A. Wander*, et al.*, *DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations.* J. Phys. Chem. A, 2009. **113**(43): p. 11856-11865.
274. Mader, S.L., A. Bräuer, M. Groll, and V.R.I. Kaila, *Catalytic mechanism and molecular engineering of quinolone biosynthesis in dioxygenase AsqJ.* Nat. Commun., 2018. **9**(1): p. 1168.
275. Li, P., B.P. Roberts, D.K. Chakravorty, and K.M. Merz, *Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent.* J. Chem. Theory Comput., 2013. **9**(6): p. 2733-2748.
276. Macchiagodena, M., M. Pagliai, C. Andreini, A. Rosato, and P. Procacci, *Upgrading and Validation of the AMBER Force Field for Histidine and Cysteine Zinc(II)-Binding Residues in Sites with Four Protein Ligands.* J. Chem. Inf. Model., 2019. **59**(9): p. 3803-3816.
277. Shimamura, H., H. Gouda, K. Nagai, T. Hirose, M. Ichioka*, et al.*, *Structure Determination and Total Synthesis of Bottromycin A2: A Potent Antibiotic against MRSA and VRE.* Angew. Chem. Int. Ed., 2009. **48**(5): p. 914-917.
278. Franz, L., S. Adam, J. Santos-Aberturas, A.W. Truman, and J. Koehnke, *Macroamidine Formation in Bottromycins Is Catalyzed by a Divergent YcaO Enzyme.* J. Am. Chem. Soc., 2017. **139**(50): p. 18158-18161.
279. Schwalen, C.J., G.A. Hudson, S. Kosol, N. Mahanta, G.L. Challis*, et al.*, *In Vitro Biosynthetic Studies of Bottromycin Expand the Enzymatic Capabilities of the YcaO Superfamily.* J. Am. Chem. Soc., 2017. **139**(50): p. 18154-18157.
280. Crone, W.J.K., N.M. Vior, J. Santos-Aberturas, L.G. Schmitz, F.J. Leeper*, et al.*, *Dissecting Bottromycin Biosynthesis Using Comparative Untargeted Metabolomics.* Angew. Chem. Int. Ed., 2016. **55**(33): p. 9639-9643.
281. Talebi Bezmin Abadi, A., A.A. Rizvanov, T. Haertlé, and N.L. Blatt, *World Health Organization Report: Current Crisis of Antibiotic Resistance.* BioNanoScience, 2019. **9**(4): p. 778-788.
282. Lin, X., X. Li, and X. Lin, *A Review on Applications of Computational Methods in Drug Screening and Design.* Molecules, 2020. **25**(6): p. 1375.
283. Wang, S., T.B. Sim, Y.-S. Kim, and Y.-T. Chang, *Tools for target identification and validation.* Curr. Opin. Chem. Biol., 2004. **8**(4): p. 371-377.
284. Gabrielson, S.W., *SciFinder.* J. Med. Libr. Assoc., 2018. **106**(4): p. 588-590.
285. Kim, S., P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu*, et al.*, *PubChem Substance and Compound databases.* Nucleic Acids Res., 2015. **44**(D1): p. D1202-D1213.
286. Katsila, T., G.A. Spyroulias, G.P. Patrinos, and M.-T. Matsoukas, *Computational approaches in target identification and drug discovery.* Comput. Struct. Biotechnol. J., 2016. **14**: p. 177-184.
287. Schenone, M., V. Dančík, B.K. Wagner, and P.A. Clemons, *Target identification and mechanism of action in chemical biology and drug discovery.* Nat. Chem. Biol., 2013. **9**(4): p. 232-240.
288. Jenkins, J.L., A. Bender, and J.W. Davies, *In silico target fishing: Predicting biological targets from chemical structure.* Drug Discov. Today Technol., 2006. **3**(4): p. 413-421.

289. Zhang, X., W. Meining, M. Cushman, I. Haase, M. Fischer*, et al.*, *A Structure-based Model of the Reaction Catalyzed by Lumazine Synthase from Aquifex aeolicus.* J. Mol. Biol., 2003. **328**(1): p. 167-182.
290. Hendlich, M., A. Bergner, J. Günther, and G. Klebe, *Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions††We dedicate this paper to Professor J. D. Dunitz.* J. Mol. Biol., 2003. **326**(2): p. 607-620.
291. Diedrich, K., J. Graef, K. Schöning-Stierand, and M. Rarey, *GeoMine: interactive pattern mining of protein–ligand interfaces in the Protein Data Bank.* Bioinformatics, 2020. **37**(3): p. 424-425.
292. London, N., B. Raveh, and O. Schueler-Furman, *Druggable protein–protein interactions – from hot spots to hot segments.* Curr. Opin. Chem. Biol., 2013. **17**(6): p. 952-959.
293. Ciemny, M., M. Kurcinski, K. Kamel, A. Kolinski, N. Alam*, et al.*, *Protein–peptide docking: opportunities and challenges.* Drug Discov. Today, 2018. **23**(8): p. 1530-1537.
294. Schneider, M., M. Rosam, M. Glaser, A. Patronov, H. Shah*, et al.*, *BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP.* Proteins: Struct. Funct. Bioinform., 2016. **84**(10): p. 1390-1407.
295. Liao, J.-m., Y.-T. Wang, and C.-l.S. Lin, *A fragment-based docking simulation for investigating peptide–protein bindings.* Phys. Chem. Chem. Phys., 2017. **19**(16): p. 10436-10442.
296. Zhang, Y., Y. Jiang, J. Peng, and H. Zhang, *Rational Design of Nonbonded Point Charge Models for Divalent Metal Cations with Lennard-Jones 12-6 Potential.* J. Chem. Inf. Model., 2021. **61**(8): p. 4031-4044.
297. Macchiagodena, M., M. Pagliai, C. Andreini, A. Rosato, and P. Procacci, *Upgraded AMBER Force Field for Zinc-Binding Residues and Ligands for Predicting Structural Properties and Binding Affinities in Zinc-Proteins.* ACS Omega, 2020. **5**(25): p. 15301-15310.
298. Spencer, J., J. Read, R.B. Sessions, S. Howell, G.M. Blackburn*, et al.*, *Antibiotic Recognition by Binuclear Metallo-β-Lactamases Revealed by X-ray Crystallography.* J. Am. Chem. Soc., 2005. **127**(41): p. 14439-14444.
299. Wang, J., G. Qu, L. Xie, C. Gao, Y. Jiang*, et al.*, *Engineering of a thermophilic dihydroxy-acid dehydratase toward glycerate dehydration for in vitro biosystems.* Appl. Microbiol. Biotechnol., 2022.
300. Li, Z., J. Yan, J. Sun, P. Xu, C. Ma*, et al.*, *Production of value-added chemicals from glycerol using in vitro enzymatic cascades.* Commun. Chem., 2018. **1**(1): p. 71.
301. Flint, D.H., M.H. Emptage, M.G. Finnegan, W. Fu, and M.K. Johnson, *The role and properties of the iron-sulfur cluster in Escherichia coli dihydroxy-acid dehydratase.* J. Biol. Chem., 1993. **268**(20): p. 14732-14742.
302. Rodriguez, M., A.G. Wedd, and R.K. Scopes, *6-phosphogluconate dehydratase from Zymomonas mobilis: an iron-sulfur-manganese enzyme.* Biochem. Mol. Biol. Int., 1996. **38**(4): p. 783-9.
303. Bayaraa, T., J. Gaete, S. Sutiono, J. Kurz, T. Lonhienne*, et al.*, *Dihydroxy-acid dehydratases from pathogenic bacteria: emerging drug targets to combat antibiotic resistance.* Chem. Eur. J. **n/a**(n/a).
304. Sutiono, S., *Development of decarboxylases and dehydratases as valuable biocatalysts for the production of fine chemicals*. 2021, Technical University of Munich: Munich.
305. Yonemoto, I.T., B.R. Clarkson, H.O. Smith, and P.D. Weyman, *A broad survey reveals substitution tolerance of residues ligating FeS clusters in [NiFe] hydrogenase.* BMC Biochem., 2014. **15**(1): p. 10.
306. Vahrenkamp, H., *Why does nature use zinc-a personal view.* Dalton Transactions, 2007(42): p. 4751-4759.

7 APPENDICES





Appendix 1. Multiple Sequence Alignment of dehydratases belonging to the ilvD/EDD superfamily. Conserved residues are indicated in red, and secondary structure elements (from the *Pu*DHT structure) are indicated above the alignment.

Protein	Organism	UniProtKB	PDB-ID
Dehydratases			
*Pu*DHT	Paralcaligenes ureilyticus	A0A4R3LQ44	n.a.
*Cc*XylDHT	*Caulobacter crescentus*	Q9A9Z2	5oyn
*Rl*ArDHT	*Rhizobium leguminosarum*	B5ZZ34	5j83, 5j84, 5j85
*Ft*DHAD	*Fontimonas thermophila*	A0A1I2J0Y3	n.a.
*Mt*DHAD	*Mycobacterium tuberculosis*	P9WKJ5	6ovt
*Sn*DHAD	*Synechocystis sp.*	P74689	6nte
*At*DHAD	*Arabidopsis thaliana*	Q9LIR4	5ze4, 5ym0
*Ss*DHAD	*Sulfolobus solfataricus*	Q97UB2	n.a.
*Ms*DHAD	*Metallosphaera sedula*	A4YEN4	n.a.
*Sto*DHAD	*Sulfurisphaera tokodaii*	Q96YK0	n.a.
*Ec*DHAD	*Escherichia coli*	P05791	n.a.
*Cn*N1DHAD2	*Cupriavidus necator*	F8GPL1	n.a.
*Cj*DHAD	*Campylobacter jejuni*	A8FJH6	n.a.
*Sa*DHAD	*Staphylococcus aureus*	P65156	n.a.
*Cg*DHAD	*Corynebacterium glutamicum*	Q8NQZ9	n.a.
*Zm*6PGDHT	*Zymomonas mobilis*	P21909	n.a.
*Ec*6PGDHT	*Escherichia coli*	P0ADF6	n.a.
*So*6PGDHT	*Shewanella oneidensis*	Q8EEA0	2gp4[a]
Metallo-β-lactamases			
CSR-1 MBL	Cronobacter sakazakii	NCBI: WP_007898024.1	6dq2
L1 MBL	*Stenotrophomonas maltophilia*	P52700	2aio
MIM-1 MBL	*Novosphingobium pentaromativorans*	G6EHN2	6auf
VIM-2	*Pseudomonas aeruginosa*	Q9K2N0	6hf5
Others			
PurAH (Bottromycin amidohydrolase)	*Streptomyces purpureus*	A0A5S8WF49	6i5s
CAII (Carbonic Anhydrase II)	*Homo sapiens*	P00918	5nxg

Appendix 2. Overview of proteins described in this dissertation.
[a]: Structure only partially resolved
n.a. = not available

8 ABBREVIATIONS

6PGDHT: 6-Phosphogluconate Dehydratase
AQD: Automatic Query Design
BCADHT: Branched-Chain Acid Dehydratase
CA: Carbonic Anhydrase
CASP: Critical Assessment of Structure Prediction
CAST: Combinatorial Active-Site Saturation Test
COSMO: Conductor like Screening Model
DFT: Density Functional Theory
DHAD: Dihydroxy-acid dehydratase
DHIV: (R)-2,3-dihydroxyisovalerate
DHT: (sugar-acid) dehydratase
DOPE: Discrete Optimized Protein Energy
EC: Enzyme Commission
ED: Entner-Doudoroff (pathway)
FRISM: Focused Rational Iterative Site-specific Mutagenesis
GA: Generic Algorithm
GAFF: General Amber Force Field
HF: Hartree-Fock
ISM: Iterative Saturation Mutagenesis
KIV: 2-ketoisovalerate
LJ: Lennard Jones
MBL: Metallo-β-Lactamase
MC/SA: Monte Carlo/Simulated Annealing
MM: Molecular Mechanics
MM-(PB/GB)SA: Molecular Mechanics – (Poisson Boltzmann/Generalized Born) Surface Area
MSA: Multiple Sequence Alignment
ONIOM: Our own N-layered Integrated Molecular Orbital + Molecular Mechanics
OPMD: Optimized Potential Molecular Dynamics
PADHT: Promiscuous Acid Dehydratase
PASS: Putative Active Site with Spheres
PCR: Polymerase Chain Reaction
PDB: Protein Data Bank
pdf: Probability Density Function
PES: Potential Energy Surface
PSCP: Protein Side-Chain Packing
RiPP: Ribosomally synthesized and post-translationally modified peptide
SASA: Solvent Accessible Surface Area
QM: Quantum Mechanics
QM/MM: Quantum Mechanics/Molecular Mechanics
QMEAN: Qualitative Model Energy Analysis
QMEANDisCo: Qualitative Model Energy Analysis with Distance Constraints
SADHT: Sugar Acid Dehydratase
SE: Semi-Empirical
VIM-2: Verona Integron-encoded Metallo-β-lactamase 2

9 LIST OF TABLES

10 LIST OF FIGURES