Extending the Reach of Computational Approaches to Model Enzyme Catalysis

BEAT ANTON AMREIN
Abstract


Recent years have seen tremendous developments in methods for computational modeling of (bio-)molecular systems. Ever larger reactive systems are being studied with high accuracy approaches, and high-level QM/MM calculations are being routinely performed. However, applying high-accuracy methods to large biological systems is computationally expensive and becomes problematic when conformational sampling is needed. To address this challenge, classical force field based approaches such as free energy perturbation (FEP) and empirical valence bond calculations (EVB) have been employed in this work. Specifically:

1. Force-field independent metal parameters have been developed for a range of alkaline earth and transition metal ions, which successfully reproduce experimental solvation free energies, metal-oxygen distances, and coordination numbers. These are valuable for the computational study of biological systems.

2. Experimental studies have shown that the epoxide hydrolase from *Solanum tuberosum* (StEH1) is not only an enantioselective enzyme, but for smaller substrates, displays enantioconvergent behavior. For StEH1, two detailed studies, involving combined experimental and computational efforts have been performed: We first used *trans*-stilbene oxide to establish the basic reaction mechanism of this enzyme. Importantly, a highly conserved and earlier ignored histidine was identified to be important for catalysis. Following from this, EVB and experiment have been used to investigate the enantioconvergence of the StEH1-catalyzed hydrolysis of styrene oxide. This combined approach involved wildtype StEH1 and an engineered enzyme variant, and established a molecular understanding of enantioconvergent behavior of StEH1.

3. A novel framework was developed for the Computer-Aided Directed Evolution of Enzymes (CADEE), in order to be able to quickly prepare, simulate, and analyze hundreds of enzyme variants. CADEE’s easy applicability is demonstrated in the form of an educational example.

In conclusion, classical approaches are a computationally economical means to achieve extensive conformational sampling. Using the EVB approach has enabled me to obtain a molecular understanding of complex enzymatic systems. I have also increased the reach of the EVB approach, through the implementation of CADEE, which enables efficient and highly parallel *in silico* testing of hundreds-to-thousands of individual enzyme variants.

Keywords: epoxide hydrolase, enantioselectivity, regioselectivity, enantioconvergence, biocatalysis, empirical valence bond, computational directed evolution

Beat Anton Amrein, Department of Cell and Molecular Biology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

© Beat Anton Amrein 2017

ISSN 1651-6214
ISBN 978-91-554-9816-0
urn:nbn:se:uu:diva-314686 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-314686)
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the publishers.
Additional Publications


Reprints were made with permission from the publishers.
Contents

1 Introduction ........................................................................................................................................ 9
  1.1 Chemistry and Catalysis ........................................................................................................... 9
  1.2 Biochemistry, Biocatalysis and Enzymes ................................................................................ 10
  1.3 Regio- and Enantioselectivity ................................................................................................ 11
  1.4 Solanum Tuberosum Epoxide Hydrolase 1 ........................................................................... 13
  1.5 Molecular Modeling Approaches ............................................................................................ 16
  1.6 Outline ...................................................................................................................................... 19

2 Methods and Theoretical Background ........................................................................................... 20
  2.1 Enzyme Kinetics ..................................................................................................................... 20
  2.2 Sequence Analysis .................................................................................................................. 23
    2.2.1 Sequence Space ................................................................................................................ 23
  2.3 Structural Analysis and Tools ................................................................................................ 24
    2.3.1 Obtaining Protein Structures ......................................................................................... 24
    2.3.2 Limitations of High-Resolution Structures .................................................................... 25
    2.3.3 Structures of Protein Variants ....................................................................................... 25
  2.4 Transition State Theory ........................................................................................................... 26
  2.5 Molecular Simulations ............................................................................................................ 27
    2.5.1 Molecular Mechanical Simulations ............................................................................... 28
    2.5.2 Molecular Dynamics Simulations ............................................................................... 28
    2.5.3 Quantum Mechanical Simulations ............................................................................... 29
    2.5.4 Multiscale Modeling ..................................................................................................... 30
    2.5.5 Free Energy Perturbation .............................................................................................. 30
    2.5.6 Empirical Valence Bond Approach ............................................................................... 31

3 Methodology Advances .................................................................................................................. 35
  3.1 Paper I ...................................................................................................................................... 35
  3.2 Paper IV ................................................................................................................................... 38
    3.2.1 Mutatagensis ................................................................................................................... 39
    3.2.2 Efficient & Parallel Execution ....................................................................................... 40
    3.2.3 Automatic Analysis ....................................................................................................... 40
    3.2.4 Suggested Workflow ..................................................................................................... 40

4 EVB Applications ........................................................................................................................... 43
  4.1 Paper II .................................................................................................................................... 43
    4.1.1 Protonation of H104 and H300 ................................................................................... 43
    4.1.2 Conservation of E35 and H104 .................................................................................... 45
4.1.3 Oxyanion Holes ................................................................. 45
4.1.4 E35Q, H300N and Double Mutant ...................................... 45
4.2 Paper III ................................................................................ 46
  4.2.1 Engineered Variant and Crystal Structure ......................... 46
  4.2.2 Binding Modes ............................................................... 47

5 Conclusions and Future Perspective ........................................... 49

6 Populärvetenskaplig Sammanfattning ......................................... 51

7 Acknowledgments .................................................................... 53

8 References .............................................................................. 55
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Ångström, $10^{-10}$ m</td>
</tr>
<tr>
<td>DB</td>
<td>Data Bank</td>
</tr>
<tr>
<td>DFT</td>
<td>Density Functional Theory</td>
</tr>
<tr>
<td>ee</td>
<td>Enantiomeric Excess</td>
</tr>
<tr>
<td>EVB</td>
<td>Empirical Valence Bond</td>
</tr>
<tr>
<td>fs</td>
<td>Femtosecond, $10^{-15}$ s</td>
</tr>
<tr>
<td>$k_{cat}$</td>
<td>Catalytic Rate Under Optimum (Saturating) Conditions, $s^{-1}$</td>
</tr>
<tr>
<td>$K_M$</td>
<td>Michaelis-Menten Constant</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>MM</td>
<td>Molecular Mechanics</td>
</tr>
<tr>
<td>MO–QM</td>
<td>Molecular Orbital Quantum Mechanics</td>
</tr>
<tr>
<td>ns</td>
<td>Nanosecond, $10^{-9}$ s</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>ps</td>
<td>Picosecond, $10^{-12}$ s</td>
</tr>
<tr>
<td>QM</td>
<td>Quantum Mechanics</td>
</tr>
<tr>
<td>QM/MM</td>
<td>Quantum Mechanics / Molecular Mechanics</td>
</tr>
<tr>
<td>StEH1</td>
<td><em>Solanum tuberosum</em> Epoxide Hydrolase I</td>
</tr>
<tr>
<td>VB</td>
<td>Valence Bond</td>
</tr>
</tbody>
</table>
Computer Programs Used in this Work

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bash</td>
<td>Unix Shell</td>
<td><a href="https://www.gnu.org/software/bash">https://www.gnu.org/software/bash</a></td>
</tr>
<tr>
<td>CADEE</td>
<td>Computer-Aided Directed Evolution of Enzymes</td>
<td><a href="https://github.com/kamerlinlab/caede">https://github.com/kamerlinlab/caede</a></td>
</tr>
<tr>
<td>Chimera</td>
<td>Molecular Modeling System</td>
<td><a href="https://www.egl.ucsf.edu/chimera">https://www.egl.ucsf.edu/chimera</a></td>
</tr>
<tr>
<td>ClustalΩ</td>
<td>Multiple Sequence Alignment</td>
<td><a href="https://www.ebi.ac.uk/Tools/msa/clustalo">https://www.ebi.ac.uk/Tools/msa/clustalo</a></td>
</tr>
<tr>
<td>GROMACS</td>
<td>Molecular Dynamics Simulation Package</td>
<td><a href="http://www.gromacs.org">http://www.gromacs.org</a></td>
</tr>
<tr>
<td>Git</td>
<td>Version Control System</td>
<td><a href="https://git-scm.com">https://git-scm.com</a></td>
</tr>
<tr>
<td>Maestro</td>
<td>Graphical Interface for Molecular Modelling by Schrödinger, LLC</td>
<td><a href="https://www.Schr%C3%B6dinger.com/maestro">https://www.Schrödinger.com/maestro</a></td>
</tr>
<tr>
<td>MOLARIS</td>
<td>Molecular Dynamics &amp; EVB</td>
<td><a href="http://laeuto.usc.edu/software.html">http://laeuto.usc.edu/software.html</a></td>
</tr>
<tr>
<td>PyMOL</td>
<td>Molecular Graphics System</td>
<td><a href="https://www.pymol.org">https://www.pymol.org</a></td>
</tr>
<tr>
<td>Python</td>
<td>Scripting Language</td>
<td><a href="https://www.python.org">https://www.python.org</a></td>
</tr>
<tr>
<td>Q</td>
<td>Molecular Dynamics Package</td>
<td><a href="http://xray.bmc.uu.se/~aqwww/q">http://xray.bmc.uu.se/~aqwww/q</a></td>
</tr>
<tr>
<td>SCWRL4</td>
<td>Protein Side Chain Prediction</td>
<td><a href="http://dunbrack.fccc.edu/scwrl4">http://dunbrack.fccc.edu/scwrl4</a></td>
</tr>
<tr>
<td>VMD</td>
<td>Visual Molecular Dynamics</td>
<td><a href="http://www.ks.uiuc.edu/Research/vmd">http://www.ks.uiuc.edu/Research/vmd</a></td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Chemistry and Catalysis

The world is full of chemical reactions, and the products of chemical reactions are everywhere around us in the form of clothing made from synthetic fibers like nylon, the PET bottles that beverages are stored in, the fuel that is burned both to propel cars and trucks or to caramelize sugar on top of delicious treats. Mankind has used chemical reactions for thousands of years. Prehistoric humans used chemical reactions that ranged from simple combustion for heat and cooking [1, 2] to fermentation to create cheese [3, 4] or to craft alcoholic beverages [5, 6].

At the beginning of the 20th century, a fundamental catalyst was discovered by Fritz Haber that was later brought to industrial scale by Carl Bosch: the Haber-Bosch process used for the industrial fixation of nitrogen [7–9]. Today’s industrial reactions that involve heterogeneous catalysts are used on very large scales (see e.g. [10, 11]).

While industrial scale reactions like the Haber-Bosch process are very important reactions, a trend towards green chemistry is observed and synthetic pathways for more atom efficiency are sought after, pathways involving less toxic intermediates, fewer byproducts and avoiding toxic solvents. In this context, biocatalysts have found to have great potential for industrial usage [12], see also Table 1.1. In general, catalysts lower the activation barrier for a reaction, compared to the same reaction in a different environment, in other words catalysts are increasing the rate at which a reaction occurs, often by many orders of magnitude [13].

Understanding how a catalyst functions can point toward how it can be improved. The reaction mechanism of a reaction can potentially lead to an understanding of how the reaction can best be improved. Consequently, great experimental efforts are undertaken to identify reaction mechanisms [10, 11, 14–16].

Since the end of the 19th century, significant contributions to the field of reaction kinetics and structure determination have been made by S. Arrhenius; V. Henri; Michaelis and Menten; Eyring, Evans and Polani; R. Franklin; Watson and Crick; Kendrew and Perutz. Modern analytical methods also contribute to experimental results, which can help to explain reaction mechanisms. But even with modern analytical methodologies, it can prove to be extremely difficult to understand a reaction experimentally for multiple reasons. Data may be very difficult to obtain experimentally: For example, structural information may be very hard to obtain or it may be unavailable (e.g., membrane proteins
[17] or intrinsically unstructured proteins [18]). Also, the experimental tools needed to obtain information may not exist (e.g., as when a substrate has many possible but experimentally indistinguishable conformations [19]. As we will see later in this thesis, thanks to the computing power of hardware and software that is available today (and which is undergoing continuous development), it is possible to simulate reactions in silico to provide additional insight [20, 21]. For some specialized problems, computational approaches have even shown potential for the design of new catalysts [22, 23]).

1.2 Biochemistry, Biocatalysis and Enzymes

Biological processes adapt to their native environments, namely the temperature, pressure, and solvent of the biological cell. Proteins that are biocatalytically active (i.e., enzymes) that can handle reactions proficiently under the conditions in living cells, even at the diffusion limit, have evolved [24–26]. At the same time, although highly specific and efficient enzymes are needed for some reactions, biological organisms need to adapt new functions (e.g., to respond to new molecules in their environments). As early as 1976, Jensen suggested that catalytic promiscuity (i.e., the ability of some enzymes to turnover multiple substrates) is important to evolve new functions [27]. Today, this natural process is mimicked by protein engineers and facilitates a better understanding of the evolution of new functions [28]. In this context, various computational tools have been developed to predict enzyme promiscuity and also to reconstruct ancestral proteins [29, 30]. (For more details refer to our perspective on the catalytic promiscuity of the alkaline phosphatase superfamily [31]).

Biocatalysts are catalysts that work in aqueous solutions, usually in the cell cytoplasm, with high concentrations of protein and other biological molecules present [32, 33]. For many reactions it is therefore crucial for an organism to have very selective catalysts. This, together with their high efficiency and the ability to function without precious metals, makes biocatalysts potent contributors to greener chemistry. Also, enzymes can be very selective in terms of chemo-, regio- and enantioselectivity. This can allow for greater atom-efficiency in chemistry because cleaning a product mixture both wastes molecules and is tedious. This, together with high efficiency, allows some enzymes to be used in industrial applications (see Table 1.1), and the number of enzymes applicable to industrial processes is increasing [34, 35].

Important factors of this development are that enzymes work under mild conditions and that they can be very selective. The application of enzymes can, however, be limited by the properties of their environment, namely physiological conditions with low substrate and enzyme loads. Enzymes are hence increasingly tuned toward process requirements [43, 44], but many challenging problems remain unsolved [44]. Various computational approaches have been put forward to address those challenges [23, 45–48].
<table>
<thead>
<tr>
<th>Starting Materials</th>
<th>Catalyst</th>
<th>Products</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$-Casein</td>
<td>Chymosin</td>
<td>para-$\kappa$-Casein &amp; $\kappa$-Macropeptide</td>
<td>Cheese Production [36]</td>
</tr>
<tr>
<td>(rac)-2-Methyl-valeraldehyde</td>
<td>Ketoreductase</td>
<td>(R)-2-Methyl pentanol</td>
<td>Pharma &amp; Liquid Crystals [37]</td>
</tr>
<tr>
<td>Ketones</td>
<td>Transaminase</td>
<td>Chiral Amines</td>
<td>Sitagliptin Manufacture [12]</td>
</tr>
<tr>
<td>Fats, Triaglycerol</td>
<td>Lipase</td>
<td>Free free fatty acids and glycerol</td>
<td>Food, Diverse [38]</td>
</tr>
<tr>
<td>Cellulose</td>
<td>Cellulase</td>
<td>Diverse</td>
<td>Biofuel, Textiles [39]</td>
</tr>
<tr>
<td>Protein</td>
<td>Protease</td>
<td>Peptides</td>
<td>Food Industry [40]</td>
</tr>
<tr>
<td>Starch</td>
<td>Amylase</td>
<td>Maltose</td>
<td>Food Industry [41]</td>
</tr>
<tr>
<td>Niacin</td>
<td>Achromobacter xylosoxidans LK1</td>
<td>6-Hydroxy-nicotinic acid</td>
<td>Insecticides (precursor) [42]</td>
</tr>
</tbody>
</table>

Table 1.1. Listed are various biocatalytically important reactions and their corresponding enzymes. Biocatalytic processes are used in industry for a wide range of reactions ranging from chemical synthesis to biofuels. Please note that most references are reviews of the class of enzymes.

In this work, I focus on the potato enzyme *Solanum tuberosum* epoxide hydrolase 1 (StEH1), which is a member of a protein family of $\alpha/\beta$ hydrolase fold enzymes. A range of epoxide hydrolases have been identified, which are converting epoxides to vicinal diols [49–52]. An example of an epoxide is shown in Figure 1.1. Epoxide hydrolases are capable of enantioselective and even enantioconvergent synthesis of a variety of epoxides [53, 54]. The product of these catalysts are vicinal diols, which are used for the synthesis of fine chemicals [49, 53, 55].

### 1.3 Regio- and Enantioselectivity

Selectivity in chemical reactions is central to the design of new chemical pathways and to the improvement of existing reactions. Selectivity is distinguished in chemo-, regio- and enantioselectivity. Chemo-selectivity usually refers to a reaction that occurs only at a specific functional group of atoms and bonds
of a molecule. Regio-selectivity means that if two equally functional groups are available, determining which of the two will react depends on the substitution at the functional group. Finally, enantioselectivity refers to a reaction that involves chiral molecules of which one enantiomer is formed preferably over the other one. Using the Cahn-Ingold-Prelog nomenclature (see e.g. [56]) enantioselectivity, is the preferred formation of one enantiomer, in such a way that the ratio between the \((R)\)- and \((S)\)-enantiomer is not 1:1. Enantioselectivity requires usually a chiral catalyst or cofactor, to induce chirality. An example of a chiral molecule is displayed in Figure 1.1. Importantly, biocatalysts such as enzymes are built from chiral building blocks (the natural amino acids) and provide a chiral scaffold. (The word chiral is derived from the Greek word χεῖρ (kheir) for hand.)

![Figure 1.1](image.png)

*Figure 1.1.* The two stereoisomers of styrene oxide. The \((R)\)-enantiomer is on the left, and the \((S)\)-enantiomer is on the right. Stereoisomery, or chirality, also referred to as enantiosomery is a property of molecules that consist of the same bonds and atoms, but that are not superimposable, they are mirror images of each other.

Enantiopure molecules display optical rotation of light if polarized light is directed through them. In many chiral compounds, a single atom is the origin of the property, and these atoms are referred to as “chiral.” Sometimes this is indicated with an asterisk (*; see e.g. Figures 1.1 and 1.2). Under nonenantioselective reaction conditions, chiral centers are usually formed as an equal mixture of \((R)\)- and \((S)\)-products. If a catalyst is able to shift the equilibrium between \((R)\)- and \((S)\)-products, the catalyst is said to be enantioselective. It is important to note that all chemical properties of two enantiomers are equal, with the exception of optical activity. This means that it is impossible to separate a racemic mixture by conventional means (e.g., via distillation). The compounds must either first be converted to diastereoisomers (i.e., by adding an additional stereocenter so that the stereoisomers are no longer mirror images), and then separated and converted back to the original compounds, or chiral chromatographic methods must be employed. In many cases, specific enantiomers must be isolated (e.g., in pharmaceutical applications), and it is crucial
to avoid racemization whenever possible to avoid having inactive, “wasted” molecules and difficult purification steps.

![Enantiomeric molecules:](image)

Figure 1.2. Enantiomeric molecules: (A) The (R)- and (S)-enantiomers of limonene are responsible for the odor of oranges and lemons, respectively. (B) The (R)- and (S)-enantiomers of carvone are responsible for the odor of caraway and mint, respectively. (C) Ibuprofen is a potent, chiral, anti-inflammatory drug that is sold as a racemic mixture of (R)- and (S)-ibuprofen. Thanks to an isomerase that mammals possess, the less active (R)-enantiomer is converted into the active (S)-enantiomer \textit{in vivo} [57]. (D) Thalidomide was once used as a cure for morning sickness [58]. Unfortunately, the compound is teratogenic and has severe effects on the fetus leading to specific birth defects if taken during pregnancy (see \textit{e.g.}, [59]). It was suggested that primarily only one enantiomer is responsible for teratogenic effects [60], but this is subject to debate [61]. Additionally, the chiral center is unstable in protonated media and racemization occurs \textit{in vivo}. Thalidomide is no longer used to treat morning sickness but it is currently used to cure leprosy [62] and to treat cancer [63].

1.4 Solanum Tuberosum Epoxide Hydrolase 1

Epoxide hydrolases catalyze the conversion of epoxides to vicinal diols. Depending on the organism, the biological functions of this enzyme class cover a wide range from detoxification over catabolism to cellular signaling [64]. Some epoxide hydrolases cover a wide range of substrates and, interestingly, enantioselectivity was also observed [54]. A candidate of particular interest for studies is the potato \textit{Solanum tuberosum} epoxide hydrolase 1 (StEH1) [65, 66]. The wild-type protein and its active site are shown in Figure 1.3. In earlier studies this enzyme was found to yield optically pure products for small substrates [67].

Engineered protein variants, together with earlier experiments and related epoxide hydrolases, have been used to gain insight into the reaction mechanism of the enzyme. Experimental evidence supports a three-step reaction mechanism after substrate binding (Michaelis Complex): The initial attack of D105 opens the epoxide at either epoxide carbon and is then followed by a general base-activated conserved active-site water molecule, which is believed to be
Figure 1.3. The potato (*Solanum tuberosum*) epoxide hydrolase 1 (StEH1), PDB-ID: 2CJP. (A) The wild-type StEH1 displayed as surface view with part of the active site visible in the center of the image. (B) The substrate-free active site of wild-type StEH1 featuring the catalytically relevant residues: For the first step, these are the nucleophile D105, the oxyanion-hole tyrosines Y154 and Y235, and histidine H300 activated by H300. Finally, the tetrahedral intermediate is converted to product and then released to close the catalytic cycle (see Figure 1.4).

The mechanistic cycle is unable to explain the origin of the regio- and enantioselectivity of StEH1. This is exactly where computational modeling unfolds its potential. A variety of substrates that were subjected to kinetic measurements with different protein variants and crystal structures of some protein variants allowed us to build a computational model that could reproduce the experimental data. With the help of this computational model, we were able to establish the mechanism in even more detail and identified a previously neglected catalytically important residue – H104 – in the active site, see Paper II. We have then, thanks to the established details of the reaction mechanism, obtained insight to detailed regio- and enantioselectivity, see Paper III.

As the *trans*-stilbene oxide (TSO) substrate, styrene oxide (SO) can be opened at both epoxide carbons (C1 and C2). In contrast to TSO, however, the product is not always the meso-diol, but it is experimentally distinguishable. It was found in earlier work that the wild-type enzyme is enantioconvergent and converts both *(R)*-SO and *(S)*-SO preferably to the *(R)*-product (see also Figure 1.5) [53, 69].

StEH1 was investigated using a combined experimental and computational approach and different enzyme variants have been tested with all stereoisomers of the substrate. In Paper II and Paper III EVB calculations have been performed on StEH1 with different mutant structures and two substrates, *trans*-stilbene oxide (TSO, Paper II) and styrene oxide (SO, Paper III), see also Fig-
Figure 1.4. The generalized reaction mechanism of the catalytic cycle catalyzed by StEH1. The computational work focused on the first two reactions steps; the nucleophilic epoxide ring-opening and the formation of the tetrahedral intermediate. The third step was not modeled because the breakdown of the tetrahedral intermediate is not expected to be rate-limiting [68].
Figure 1.5. The SO substrate as illustration of how a chiral catalyst’s enantioconver-
gence is explained: if, as in the case for the wild-type variant of StEH1, (S)-SO is
opened on C1, the stereo center is inverted and the (R)-diol forms. If at the same
time a catalyst preferably opens the (R)-SO at C2, again the (R)-diol will be the product. As
a consequence, only (R)-diol is formed and enantioconvergent behavior is observed.

Figure 1.6. The substrates investigated in this work are trans-stilbene oxide (TSO)
and styrene oxide (SO). Because TSO is a symmetric substrate, its hydrolysis product
is the superimposable meso-diol and so its regio-selectivity cannot be determined ex-
perimentally. Styrene oxide is not symmetric, and ratios between (R)- and (S)-diols
can be determined experimentally.

1.5 Molecular Modeling Approaches
In modern society, computation is used to gain additional insight into sys-
tems of interest. In fact, early computational chemistry was used by humans
long before computers became available, and some of those computational ap-
proaches are used today on modern computers with ever more complex sys-
tems. Early approaches in computational chemistry did rely on human computa-
tional power, as exemplified by the acknowledgments in Robert Mulliken’s
work on electronic population analysis (which is still in used to calculate Mul-
likén charges): “The writer is indebted to Mr. W. Jaunzemis for calculations
whose results are embodied in the tables...” [70]
In the last 60 years, tremendous exponential growth of computing power
was achieved. Since nearly 30 years, this growth has been achieved by adding
more and more processor-cores and so code has to be optimized to run on multiple processors, see Figure 1.8).

Two major computational modeling approaches are in use today. Approaches based on molecular mechanics (MM) rely on force fields to simulate a molecule with Newtonian limitations and scale, depending on the implementation details between $O(n \log(n))$ to $O(n^2)$, see also Figure 1.7. While the $O(n \cdot \log(n))$ algorithm involves a computationally expensive reverse Fourier transformation, it scales much better with larger problem sizes.

![Big-Ω Notation](image)

*Figure 1.7.* It is important to estimate the computational costs of algorithms. The big $Ω$ notation describes the worst-case performance of an algorithm, when the argument $n$ tends to infinity, and it provides a worst-case estimate of how many operations an algorithm will need to arrive at a result given a problem of size $n$. For example, $O(n)$ means that the computational cost will increase linearly with the problem size, and $O(n^2)$ that it will increase to the square, as the problem size is increased. The big $Ω$ notation estimates only how many iterations the algorithm needs. In practice, a higher-order algorithm may be faster for a range of problem sizes if the costs per iteration are lower.

One limitation of classical force fields is that they are unable to describe excited states and thus usually limited to nonreactive simulations (*i.e.*, no bonds can be formed or broken). To describe excited states numerical approximations to solve the Schrödinger equation are employed. These approaches are referred as quantum mechanical (QM) approaches and scale, depending on the implementation details, usually between $O(n^3)$ (DFT) to $O(n^4)$ (CCSD(T)), $O(n^5)$ (CCSDT) [71, 72].

Many of the QM approaches used today are hence limited to system sizes ranging from between a few dozens to about 1,000 atoms, depending on the available resources and approach used. Efforts are ongoing to reduce the order of quantum mechanical approaches required and to achieve linear scal-
Figure 1.8. The fastest supercomputers over the course of the past 50 years. Notably, a smartphone processor could outrun the fastest computers from 1980 and an office computer, especially if equipped with a modern graphic processing unit (GPU), can compete with supercomputers from the 1990s. Another important point is that the speed gain with single-core silicon chips has slowed around 1980 and this has been compensated for by the introduction of parallel computing with multiple processor cores. The ’Triolith-Cluster’ was the fastest supercomputer (or cluster) used in this work.

...
To sum up, modeling reactions in large systems such as proteins is inefficient when done with QM-only methods. While it can be done, it is often very expensive, sufficient sampling is hard to achieve, and results may not be accurate. One alternative is to use reactive force fields, such as ReaxFF, where the force field is based on bond orders (not explicit bonds) and can undergo reactions for classes of parameterized reactions [84].

Ideally, one would use a method that scales as MM-based approaches do but with the accuracy of a QM-based approach. One possible solution to this problem is multi-scale modeling, which was used already in the 1970s. In multi-scale modeling, a combination of different levels of accuracy for different parts of a system is used. For example, two levels of theory could be applied, QM for the reactive part of the system and MM for the rest of the system. The QM part is often simulated based on molecular orbital QM approaches. It is, however, also possible to use a valence bond based approach and to couple it with MM. In this thesis, the main workhorse is the empirical valence bond approach as implemented in Q5, which will be described in the methods section.

1.6 Outline

The main motivation of this thesis is to gain a deeper understanding of the reaction mechanism of the potato enzyme *Solanum tuberosum* epoxide hydrolase 1 (StEH1). It also involved the associated method development of CADEE to automate *in silico* enzyme evolution. In the next chapter, I will discuss the methods used and the technical background. In the method development chapter, I present work on the octahedral dummy model, which I initially planned to use in other metal-containing enzymes such as ADH-A. I will then present CADEE: Computer-Aided Directed Evolution of Enzymes, a framework to simplify preparation and parallel computation of *in silico* generated protein variants. Finally, I will discuss the epoxide hydrolase, StEH1 which was investigated using a combination of kinetic measurements, crystallographic investigations, and computational work.
2. Methods and Theoretical Background

2.1 Enzyme Kinetics

Reaction kinetics can be determined by measuring the concentration of reactant or product over time. This is straightforward, particularly when the backward reaction is very slow or when the product is removed from the reaction and the backward reaction can be ignored. A relationship between temperature and reaction-rate was suggested by Arrhenius and is known as the Arrhenius equation. Later, the Eyring-Polanyi equation was introduced to describe reaction kinetics. Of particular use for enzymatic reactions are the contributions of Michaelis and Menten [85], who developed a well-known simplified model that is used to characterize enzyme kinetics (see also Figure 2.1).

\[
\begin{align*}
[S] & \gg K_M \implies V = V_{\text{max}} = k_{\text{cat}}[E_0] \\
[S] & \approx K_M \implies V = \frac{V_{\text{max}}}{2} \\
[S] & \ll K_M \implies V = V_0 = \frac{V_{\text{max}}[S]}{K_M + [S]}
\end{align*}
\]

*Figure 2.1.* If kinetic measurements are performed on enzymes, typical data obtained using the Michaelis-Menten kinetics are: the Michaelis-Menten constant \(K_M\) and \(k_{\text{cat}}\), a measure of the rate of catalytic product-formation under optimum conditions. They can be obtained by plotting the substrate concentration \([S]\) against the reaction velocity \(V\). This plot illustrates specific cases for low and high concentration of substrate, and for \(K_M = [S]\).

The Michaelis-Menten constant \(K_M\), is the concentration of substrate needed for the enzyme to perform at half of its maximum reaction rate. As measure of the catalytic product-formation under optimum conditions \(k_{\text{cat}}\) is used. It is representing the number of product molecules formed per second (s\(^{-1}\)) and in complex systems with several steps, \(k_{\text{cat}}\) is the rate constant of the rate-limiting step. A measure for catalytic efficiency is \(k_{\text{cat}}/K_M\) which takes both substrate binding and conversion to product into account.
Assuming a simple model system, an enzymatic reaction can be viewed as an initial equilibrium of enzyme and substrate (E+S) that encounter and form an enzyme-substrate complex (ES), react, and finally separate as enzyme and product (E+P) (see Scheme 1 and Figure 2.2). The backward reaction of the rate-limiting step ($k_2$) is here assumed to be negligible, which is naturally not necessarily the case.

\[ \Delta G^0 \]

\[ \Delta G_{\text{uncat}}^\ddagger \]

\[ \Delta G_{\text{cat}}^\ddagger \]

\[ k_{-1} \]

\[ k_2 \]

\[ E+P \]

\[ E+S \]

\[ ES \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]

\[ \text{Energy} \]

\[ \text{Reaction Coordinate} \]
To define $K_M$, we separate the concentrations and rates:

$$(k_{-1} + k_2)/k_1 = [E][S]/[ES]$$

(2.4)

And define $K_M$ as:

$$K_M = (k_{-1} + k_2)/k_1$$

(2.5)

Then we substitute $K_M$ back into eq. 2.4 to arrive at:

$$K_M = [E][S]/[ES]$$

(2.6)

Substituting 2.2 into eq. 2.6 gives:

$$K_M = ([E]_{tot} - [ES]) [S]/[ES]$$

(2.7)

Next, we isolate $[ES]$:

$$[ES]([S] + K_M) = [E]_{tot}[S]$$

(2.8)

$$[ES] = \frac{[E]_{tot}[S]}{K_M + [S]}$$

(2.9)

Substituting 2.9 into eq. 2.1 gives:

$$v_0 = k_2 \frac{[E]_{tot}[S]}{K_M + [S]}$$

(2.10)

The maximum velocity of the enzymatic reaction is given by:

$$V_{max} = k_2 [E]_{tot}$$

(2.11)

Replacing $k_2$ with 2.11 yields the Michaelis-Menten equation:

$$v_0 = \frac{V_{max}[S]}{K_M + [S]}$$

(2.12)

Where the initial rate of the reaction ($v_0$) is set in relation to the substrate concentration $[S]$, which allows the determination of maximum rate $V_{max}$ and the Michaelis-Menten constant $K_M$. Additionally, by using more sophisticated experimental approaches it is possible to determine the barriers of intermediate (EI) steps (see Scheme 2).

$$\begin{align*}
E + S & \xrightleftharpoons[k_{-1}]{k_1} ES & \xrightleftharpoons[k_{-2}]{k_2} EI & \xrightarrow[k_3]{k_3} E + P
\end{align*}$$

(Scheme 2)

This can be achieved by using pre-steady-state kinetic measurements, if the intermediate is experimentally observable. The enzyme reaction is often measured using stopped-flow procedures: substrate is exposed to an enzyme, and
the intermediate is observed over time. It is possible, depending on the reaction setup, to observe different steps of a reaction. Depending on the mechanism of the reaction investigated, a model for the observed rate can be constructed. For a case like depicted in Scheme 2 assuming $[E+S]$ and $[ES]$ are in equilibrium, for example:

$$k_{obs} = \frac{k_2[S]}{K_S + [S]} + k_{-2} + k_3$$

(2.13)

Here, $K_S$ represents the dissociation constant ($k_{-1}/k_1$). The smaller the dissociation constant, the higher the probability that product will be formed after the Michaelis complex has formed. This data can then be used, often in conjunction with other experimental observables (i.e. with a 3D structure of a particular enzyme) to articulate and validate a reaction mechanism and to obtain information on how an enzyme can be improved.

2.2 Sequence Analysis

The sequence of a protein is the 1-dimensional construction plan of the same protein. It is therefore possible to compare proteins, by comparing the sequences of related proteins to each other. Clearly, comparing two proteins can be easy if their sequences are very closely related (e.g., if only a single point mutation is the difference). While two identical sequences can be compared manually, it becomes very tedious to compare two or more sequences that may have insertions and deletions or low sequence similarity (which is also referred to as sequence identity). This is emphasized by the number of known protein sequences: Today, more than 70,000,000 unique sequences are available on www.uniprot.com (as of Jan 27, 2017; more than 550,000 manually annotated and 73,000,000 automatically annotated sequences are accessible online). Tools were developed to manage these protein sequences, to organize and align sequences to each other, and to compare sequences for similarity. Among the best known approaches are the Basic Local Alignment Search Tool (BLAST) [86] and for multiple sequence alignments Clustal $\Omega$ [87]. Various ways to display the similarity of sequences exist and WebLogo [88] is a well-known alternative to display raw sequences.

2.2.1 Sequence Space

It is hard to imagine the sequence space of proteins and enzymes because of its exponential growth (20$^N$, with $N$ representing the sequence length). A brute force approach, (i.e. trying all possible combinations), on a sequence space of just 60 amino-acids, would yield 20$^{60} \approx 10^{78}$ unique sequences, a number comparable to the estimated number of atoms in the known universe (10$^{79}$) [89]. Hence, it is unfeasible to try out the total sequence space of a protein. Instead,
directed evolution experiments are used, with the inherent risk of getting stuck in local minima [90, 91]. Here, computational tools and structural information are best used in combination with directed evolution approaches. Because the sequence space grows exponentially, mutagenesis of more than one residues of a protein becomes very large quickly. For example, consider a case in which a protein is engineered and two or more positions are mutated simultaneously. While a single-point mutation to all 20 natural amino acids would produce 20 variants, the number of variants in a double or triple point-saturation will grow exponentially (by a factor of 20) and lead to 400 (dual-point saturation, 20 x 20) or 8,000 (triple saturation, 20 x 20 x 20) enzyme variants. The plaintext of all the sequences generated when simultaneously permuting 8 positions from a 300 amino-acid protein would need nearly 8 TB of storage alone. This growth can be slightly reduced (it remains exponential) if a subset of amino acids is used as has been suggested and applied experimentally [92]. The example above would scale with 12 amino acids, to produce 12, 144, 1,728 mutant sequences or 130 GB of text.

2.3 Structural Analysis and Tools

2.3.1 Obtaining Protein Structures

An enormously large number of conformations are theoretically accessible to enzymes. Nevertheless, they often fold into one biologically active conformation, with a few notable exceptions such as intrinsically unstructured proteins [18]. To predict the fold of a protein using only sequence information remains an unsolved challenge. Even with very powerful computing resources like Folding@Home [93] or Anton 2 [94], it remains a problem to fold proteins correctly in silico [95, 96]. This means, that without experimentally obtained 3D structures of proteins, it is very difficult to obtain insight in the 3-dimensional organization of proteins.

Therefore, whenever possible, a 3D structure is obtained experimentally, usually by using X-ray diffraction [97]: The protein is expressed, purified, and crystallized, and then X-ray diffraction data is collected. If all previous steps were successful and the crystal did diffract, a density map can be obtained. This in turn allows 3D coordinates to be computed (e.g., using Coot [98]).

The 3D data is stored in various databases. One well-known example is the Protein Data Bank (www.rcsb.org, [99]), where 125,000 structures are available (as of January 5, 2017). It is, however, not trivial to express and crystallize a protein, and even when it is possible, the crystals may not diffract or may be of poor quality. As an alternative to crystallization, the protein may be expressed in an isotope-enriched medium and subjected to nuclear magnetic resonance (NMR) spectroscopy, provided it is sufficiently soluble to yield a good signal. Other emerging technologies such as single-molecule X-ray laser diffraction
[100] and cryo-electron microscopy (cryo-EM) are being developed. Recently a 1.8 Å cryo-EM structure was published [101].

2.3.2 Limitations of High-Resolution Structures

X-ray structures usually lack hydrogen atoms, and so hydrogen is usually added during post-processing (e.g., before starting a molecular dynamic simulation). The reason for this is that the electron density around hydrogen atoms is low and often polarized toward heavy atoms, and, as the lightest element, hydrogen, is mobile even at low temperatures and the electron density is spread out. Additionally, it is important to note that crystal structures are often obtained at a pH that allows the protein to crystallize, which is not necessarily the same as the physiological pH of the host organism. To address the issues that arise from protonation, several tools are available, including MolProbity [102], which can be used to find the most likely protonation states and even side chain flips. Side chain flips occur, when the density of a crystal is not allowing to clearly assign the orientation of side chains like histidine, arginine, or glutamine. These cases can be found and resolved with hydrogen bond network analysis. Additionally, tools like PROPKA can be employed to estimate $pK_a$ [103].

2.3.3 Structures of Protein Variants

Proteins are subject to both natural evolution and human engineering. For example, random point mutations can happen to an organism if it is exposed to ionizing radiation. Adding a radiation source and hence accelerating the natural mutation rate is used to improve crop [104]. With the sophisticated biotechnological tools available today, it is also possible to edit the DNA of an organism and directly alter it. For example, point mutations may be introduced to alter the function of an enzyme.

Experimentally, this can be used to test whether a residue assumed to be crucial for catalysis is indeed crucial (e.g., by mutating an aspartate that is assumed to be the nucleophile, to an alanine). If this enzyme is later determined to be inactive, chances are that this specific aspartate side chain is indeed crucial for catalysis, especially if the structure of the enzyme did not change. In such a scenario, a single mutation rendered the enzyme inactive, while leaving it structurally intact. In other cases, a mutation may introduce a significant change in the structure of a protein, as for example when StEH1 was modified and H300 was mutated to asparagine [19]. In many other cases, the protein retains its 3D structure, as in case of the Y149F-mutant [105]. For such substitutions, a mutant structure may be obtained by “mutating” the corresponding side chain in silico.

To introduce or change protein side chains, a number of tools have become available. For example, the Ramachandran plot gives an estimate if the angles
of the protein backbone are in their energetically preferred region (see fig. 2.3). Also, thanks to the tremendous number of solved X-ray structures that are now available, Dunbrack et al. [108, 109] and others, have created rotamer libraries of typical residue orientations, and computational modeling and visualization tools such as Chimera [110] or PyMOL [111] offer built-in mutagenesis tools. These tools usually provide an interface to rotamer libraries and allow users to choose their preferred side chain orientation. Some scoring function is usually provided as well, making the tools even more user friendly (e.g., PyMOL’s “Mutagenesis Wizard”).

A number of programs have been written to automate the selection of the best rotamer. A well-known program is SCWRL4 [112]. It is able to construct protein side chain orientations, given just the backbone trace. Importantly for this work, SCWRL4 also solves a combinatorial problem when using libraries: if a residue is too large and neighboring residues need to be moved, or if two neighboring residues clash, it quickly become difficult to find the best solution. Thanks to a heuristic, SCWRL4 always obtains a solution for such cases. Even, if many deeply nested and interdependent combinations of side chain orientations exist.

2.4 Transition State Theory

Early advances towards the fundamental understanding of chemical reactions in the gas phase date back to 1889, when Svante Arrhenius introduced his famous equation:

\[ k = A e^{-\frac{E_a}{RT}} \]  

(2.14)
Where, \( k \) = rate constant, \( A \) = pre-exponential factor, \( E_a \) = activation energy, \( R \) = gas constant, \( T \) = Temperature in degrees Kelvin.

To rationalize this finding, we assume a reaction similar to the one depicted in Scheme 3.

\[
\begin{align*}
E + S & \underset{\text{ES}^\dagger}{\longrightarrow} \rightarrow E + P \\
\text{Scheme 3}
\end{align*}
\]

Equation (2.14) can then be rewritten:

\[
ln(k) = ln(A) - \frac{E_a}{R} \left(\frac{1}{T}\right)
\]

(2.15)

To make use of this relationship, the temperature is plotted against the rate constant \((1/T)\) vs. \(\ln(k)\), so a linear equation can be fitted and both the pre-exponential factor and \(E_a/R\) can be extracted from \(y\)-intercept and slope, respectively.

The Eyring equation (eq. 2.16) was found in 1935 independently by Henry Eyring [15] and Evans & Polanyi [113]. The exponent features, in contrast to eq. 2.14, the Gibbs activation energy and replaces the pre-exponential factor with \(k_B T/h\) (if \(\kappa = 1\) (see eq. 2.17)).

\[
k = \frac{k_B T}{h} e^{-\frac{\Delta G^\dagger}{kT}}
\]

(2.16)

With the same symbols as in the Arrhenius equation and additionally \(k_B\) = Boltzmann constant, \(h\) = Planck’s constant and \(\Delta G^\dagger\) = Gibbs activation energy.

Transition state theory (TST) allows us to convert experimentally observed kinetic energies to activation energies [114]. TST suggests, that reactions proceed through a maximum, crossing from reactant to product state. This maximum is referred to as a transition state. It is the highest point in energy between reactant and product. Thanks to the work by Eyring it is possible to calculate the activation energy of a reaction from a macroscopically observed reaction rate:

\[
k = \kappa \frac{k_B T}{h} e^{-\frac{\Delta G^\dagger}{kT}}
\]

(2.17)

With the same symbols as the Arrhenius equation and additionally:

\(k_B\) = Boltzmann constant, \(h\) = Planck’s constant, \(\Delta G^\dagger\) = Gibbs activation energy and \(\kappa\) = Transmission coefficient, usually assumed to be 1.

### 2.5 Molecular Simulations

Two fundamentally different ways to perform molecular simulations exist. One is based on MM and, the other on QM. MM methods rely on a set of
parameters for all bonded and non-bonded terms used in a system to produce the potential energy of the system:

\[ U(R) = \sum U_{bonded} + \sum U_{nonbonded} \]

The major advantage of using MM to model a molecule is that it is fast, even with many particles. The disadvantages of MM-based approaches are that they are bound to Newtonian mechanics, unable to model excited states and usually bound to harmonic potentials for bonded terms. As a result, classical MM approaches cannot support reactions directly, \textit{i.e.} no formation or breaking of bonds. To form and break bonds a more detailed description is needed, and the Schrödinger equation is solved approximately:

\[ H\Psi = E\Psi \]

With \( H \) being the Hamiltonian operator, describing the total energy of the system. \( \Psi \) represents the wavefunction and \( E \) the energy values of \( \Psi \). Two main approaches are used to solve it: those based on valence bonds (VB) and those that are based on molecular orbitals (MO).

### 2.5.1 Molecular Mechanical Simulations

The force field based approaches, rely on parametrization of all interactions of molecules in the form of:

\[
U(R) = \sum_{bonds} k_{i}^{bond}(r_{i} - r_{0})^{2} + \sum_{angles} k_{i}^{angle}(\theta_{i} - \theta_{0})^{2} + \sum_{dihedrals} k_{i}^{dihedral}[1 + \cos(\eta_{i}\phi_{i} + \delta_{i})] + \sum_{i \neq j} 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i \neq j} \frac{q_{i}q_{j}}{r_{ij}} \]

(2.18)

where the first three terms describe bond, angle, and dihedral, respectively, and the fourth term describes the nonbonded interactions, usually split into van der Waals and Coulomb terms. For a given compound, all parameters have to be known \((r_{i}, k_{i}^{bond}, \theta_{0}, k_{i}^{angle}, \eta_{i}, \phi_{i}, \delta_{i}, k_{i}^{dihedral}, \sigma_{ij}, \varepsilon_{ij}, q_{i}, q_{j})\) and consistently defined, for all atoms that are part of the force field simulation.

### 2.5.2 Molecular Dynamics Simulations

Approaches based on molecular dynamics (MD) aim to simulate a molecule over time, between femto-seconds (fs) to microseconds (\( \mu \)s), and with special
hardware (ANTON) milliseconds (ms) [94], to obtain information of the conformational landscape of a molecule. Often, molecules have many degrees of freedom, making it very hard or impossible to find the global minimum energy conformation and simulations can get trapped in local minima. Additionally, large molecules have many energetically similar configurations that are all accessible with temperatures around 300K. It is therefore of central importance to sample a molecule’s energetic landscape long enough and with different initial conditions, to obtain sufficient sampling and collect an understanding for the states a molecule visits over time. This approach is routinely used to perform simulations of protein in solvent and is implemented in many simulation packages (e.g. GROMACS [115, 116], AMBER [117], CHARMM [118] and Q [119]). Depending on the simulated system, code, and hardware used, 1 second compute time translates to $10^{-14}$ (single core [120]) to $10^{-9}$ seconds (ANTON 2 [94]).

2.5.3 Quantum Mechanical Simulations

Quantum mechanical approaches (QM) simulate a molecule entirely, including electrons. Since electrons have much less mass than the nuclei they are almost instantaneously adapting to the movement of the nuclei and this can be used to separate the nuclei from the electrons. This fundamental approximation is called Born-Oppenheimer approximation.

The goal of the QM approaches is to calculate the wave function of a molecule, as can be done by solving the Schrödinger equation. The Schrödinger equation has however only an analytical solution for the simplest atoms (e.g. hydrogen atom) and it needs to be solved numerically for other atoms and molecules. Using the Born-Oppenheimer approximation, the Schrödinger equation can be solved for the electrons separately. The Hartree-Fock method approximates further, by calculating the electron interaction with the meanfield of the other electrons, thus neglecting electron-electron correlations. To include electron correlation QM methods based on Møller–Plesset perturbation theory (e.g. MP2), configuration interaction (CI) or coupled cluster (CC) can be employed. These methods are, however, computationally expensive ($N^4$ to $N^8$), growing with the number of electrons in the system. Hence they are impractical for large systems. Density function theory (DFT) is an approach that can be used for larger systems, which does not aim to approximate the wavefunction, but instead the electron density. A shortfall of DFT is the description of exchange-correlation and different functionals should be tested for a system to avoid non-physical DFT artefacts.
2.5.4 Multiscale Modeling

In 2013 the Nobel prize in chemistry was awarded to Karplus, Lewit, and Warshel for the development of multi-scale modeling. The idea the Laureates came up with is the combination of different levels of theory, or complexity, for different parts of a molecule or a system of molecules, like a solvated protein. Modern QM/MM approaches build on this idea and use a force field for a large part of the system, and density functional theory (DFT) or another QM approach for the reactive part of the system.

In the empirical valence bond (EVB) approach, which was extensively used in this work, the reactive part of a system is modeled with valence bond. Additionally, in this work spherical reaction boundaries are used, (i.e. surface-constrained all-atom solvent (SCAAS) [119, 121]), see Figure 2.4. With EVB, the reactive atoms are part of the force field, and the bonds formed and/or broken are modelled with Morse potentials instead of harmonic potentials. This approximation of reactivity has both an impact on the accuracy of the active bonds, but also on the compute time needed to simulate one timestep. Provided that the intrinsic energy needed for bond-formation and/or bond-breaking does not change significantly for the same reaction in vacuum, water or protein, the EVB allows for significantly more sampling of the environment. In turn the energetic contributions from the environment (e.g. solvent or protein) can be sampled longer and this can help to obtain converging energetics.

There are other ways to add reactivity to a force field. For example, ReaxFF is an approach that can enable a force field based approach to bond breaking and bond forming processes. Considerable parametrization of the reactive force field is needed to describe different classes of systems [84, 122].

2.5.5 Free Energy Perturbation

In molecular dynamics, it is not possible to sample high energy states in a physically meaningful way. It is, however, possible to describe two states A and B with a force field. If the two states are close and sampled enough (i.e., there is enough overlap), it is possible to calculate the transition from one state to another: The procedure known as free energy perturbation (FEP) can be used, based on the work of Robert Zwanzig from 1954 [123]:

$$\Delta G = -RT \ln \left \langle \exp \left \{ -\frac{U_B - U_A}{k_B T} \right \} \right \rangle_A$$  \hspace{1cm} (2.19)

This only works if the overlap between the two states A and B is large enough (i.e., the difference between state A and B is small). If not enough sampling is available due to low overlap, intermediate potentials can be created, usually with a linear combination of states A and B and a mapping parameter $\theta_m$:
Figure 2.4. A: Schematic illustration of the spherical boundaries used in this work. The protein is displayed in grey cartoon. Three different layers of atoms are distinguished: In the center, depicted as bond and sticks, the reacting atoms. In a second layer the atoms are unrestrained and may, in contrast to the third layer, move freely. The third layer restrains the mobility of atoms and the everything outside the third layer is heavily restrained and practically immobilized. B: The two valence bond states ($\Psi_1$ and $\Psi_2$) for the first step of nucleophilic epoxide hydrolysis. C: The free energy surfaces of the two valence bond states. With $\lambda$ representing the reorganization energy needed to change from one to the other surface.

$$U_{\text{eff}}(\theta_m) = \varepsilon_m = (1 - \theta_m)U_A + \theta_m U_B \quad (0 \leq \theta_m \leq 1) \quad (2.20)$$

The mapping parameter $\theta_m$ is then incrementally changed from 0 to 1, so that the overlap between two intermediate states is large enough to converge. All that is left is the combination, as described below (see eq. 2.28) of all the intermediate steps, which results in a change from state A to B [124–126].

2.5.6 Empirical Valence Bond Approach

The EVB approach adds reactivity to force field based simulations. As its name implies, empirical fitting of a set of parameters to match a reference system is needed. Once the parameters are fitted (e.g., to the reaction of interest in water), they are fixed and do not change for the same reaction in other solvents. The exact same parameters as obtained by the reference reaction are then applied to the actual system of interest (e.g., the same reaction in protein) [127–129].
The valence bond (VB) theory and molecular orbital (MO) theory both describe phenomena on the molecular level. In the VB approach, the molecule is handled as a linear combination of valence bond states. Unlike MO theory, which describes the wave functions of delocalized electrons by combining all possible orbitals looking for the probability of finding an electron at given positions in space, the VB accounts for the probability of finding a molecule at a given state [130].

Because the EVB approach uses inexpensive Morse-potential (eq. 2.21) terms to model reactivity, (i.e. bond forming and -breaking) it is well suited to account for the energetic contributions from the environment (i.e., vacuum, solvent, or protein) and the computation of the corresponding energetics is much faster than with MO-based QM/MM calculations. This is because it is not necessary to compute the wave function for every time step.

\[ M = D_e \exp\{-2a(r - r_0)\} - 2e\exp\{-a(r - r_0)\} \]  
(2.21)

Building from Marcus Theory, Warshel & Weiss introduced the EVB approach in 1980. The fundamental idea was to determine the solvation effects of the environment on the system of interest. [127]

The diabatic states \( (H_{11} \text{ and } H_{22}) \) are described as in eq. 2.24 for the \( i^{th} \) state, with the first term \( \alpha_{gas}^i \) representing the gas-phase energy. The easiest example of an EVB calculation is the calculation in the gas phase where the diagonal elements \( H_{ii} = \varepsilon_i = \alpha_{gas}^i + U_{\text{intra}}(R, Q) \), and where the second term \( U_{\text{intra}}(R, Q) \) accounts for the intramolecular interactions within the solute.

The off-diagonal term in eq. 2.22 may be implemented as a description of exponential coupling functions, and together with \( \alpha_{gas}^i \) empirically fitted to a reference energy:

\[ H_{ij} = A e^{-\mu(r-r_0)} \]  
(2.22)

With eqs. 2.24 and 2.22, the Hamiltonian matrix can be constructed as:

\[ H = \begin{pmatrix} H_{11} & H_{21} \\ H_{12} & H_{22} \end{pmatrix} \]  
(2.23)

The case of a gas phase calculation can be extended to account for the solvent or protein effects, by extending \( H_{ii} \) with elements from the solvent and solvent interaction:

\[ H_{ii} = \varepsilon_i = \alpha_{gas}^i + U_{\text{intra}}^i(R, Q)+U_{\text{inter}}^i(R, Q, r, q)+U_{\text{solvent}}^i(r, q) \]  
(2.24)

In eq. 2.24, \( R \) and \( Q \) are representing coordinates and charges of the reacting atoms (solute) while \( r \) and \( q \) are the coordinates and charges of the surrounding atoms (solvent), usually in vacuum, solvent or protein. The term \( U_{\text{intra}}^i(R, Q) \) represents the intramolecular interactions of the solute as already
mentioned above and $U_{\text{inter}}^i(R, Q, r, q)$ is the interaction between solute and the surrounding solvent. Finally, $U_{\text{solvent}}^i(r, q)$ represents the solvent energies.

Now, the adiabatic ground-state energy and its corresponding eigenvector are obtained from the lowest eigenvalue by solving the secular equation:

$$HC = E_g C$$  \hspace{1cm} (2.25)

For the two-state EVB system, the solution to this eigenvalue problem is:

$$E_g = \frac{1}{2}[(\varepsilon_1 + \varepsilon_2) - \sqrt{(\varepsilon_1 - \varepsilon_2)^2 + 4H_{12}^2}]$$  \hspace{1cm} (2.26)

The solution of 2.26 (i.e., the lowest eigenvalue) is the ground state energy of the system. The activation free-energy $\Delta G^\Delta$, can then be obtained by changing the system from one state to the other. A “mapping potential” in the form of:

$$\varepsilon_m = (1 - \theta_m)\varepsilon_1 + \theta_m\varepsilon_2 \quad (0 \leq \theta_m \leq 1)$$  \hspace{1cm} (2.27)

can be used to change the system gradually from one state to the other, where $\theta_m$ is changed incrementally from 0 to 1. Then, the free energy $\Delta G_m$ of changing $\theta_m$ from 0 to 1 can be obtained by free-energy perturbation/umbrella sampling (FEP/US) [128].

To obtain the full energy surface between different diabatic states, (e.g., from state 1 to state 2), the changes in $\varepsilon_m$ need to be gradual enough to be able to construct the free-energy functional:

$$\Delta G(x') = \Delta G_m - \beta^{-1} \ln \langle \delta(x - x') \rangle \times \exp\{-\beta[E_g(x) - \varepsilon_m(x)]\}$$  \hspace{1cm} (2.28)

As mentioned earlier, with EVB a fundamental approximation is that the off-diagonals, and $\alpha_{\text{gas}}^i$ do not change significantly when changing the environment of the reaction (i.e., when transferring the system from water to protein) [131, 132].

$$H_{12} = \sqrt{(\varepsilon_1 - E_g)(\varepsilon_2 - E_g)}$$  \hspace{1cm} (2.29)

It is possible to estimate the origin of the catalytic effect to the EVB result by approximating the activation free energy with the Hwang–Åqvist–Warshel (HAW) equation:

$$\Delta G^\Delta = W + \frac{(\Delta G^0 + \lambda)^2}{4\lambda} - H_{12}(x) + \frac{H_{12}(R_0)}{(\Delta G^0 + \lambda)} - \Gamma$$  \hspace{1cm} (2.30)

$W$, the “work term” corresponds to the free energy needed to bring the reactant pair to the interaction distance $R_0$. $\Delta G^0$ represents the reaction free
energy and $\lambda$ is the reorganization energy. $H_{12}(x)$ and $H_{12}(R_0)$ describe the average value of $H_{12}$ at the transition and reactant state, respectively, and finally, $\Gamma$ represents the nuclear quantum-mechanical correction [128, 133].

It is possible to approximate the reorganization energy [134] by using the relationship

$$\lambda = \frac{1}{2}(\langle \Delta \varepsilon \rangle_2 - \langle \Delta \varepsilon \rangle_1)$$

(2.31)

where $\langle \Delta \varepsilon \rangle_1$ and $\langle \Delta \varepsilon \rangle_2$ represent the average difference between $\varepsilon_1$ and $\varepsilon_2$. For a schematic overview refer to Figure 2.5.

![Figure 2.5](image-url)

*Figure 2.5. An illustration of the relationship between two diabatic states ($\varepsilon_1$ and $\varepsilon_2$) in water (left) and enzyme (right): Both, the reorganization energy $\lambda$ and the activation barrier are lower in the enzymatic reaction. Published in ref. [120], licensed under CC BY.*

This concludes the methodology chapter and next, I will present articles that have been published in the context of this thesis. The computational modeling of metals is a particularly challenging task and so I will next introduce the non-bonded octahedral dummy model.
3. Methodology Advances

3.1 Paper I

Force Field Independent Metal Parameters Using a Nonbonded Dummy Model

Metals play ubiquitous roles in proteins and enzymes: Nearly 1/3 of the protein structures stored on the PDB contain metal ions [135]. Also, metals are found in all classes of enzymes and for example 44% of all known oxidoreductases contain a metal-center [136]. Clearly accurate metal models are important to model metallo proteins.

Modeling metal ions is challenging because metals are easily ionized and highly reactive. Hence, one would preferably model metals with QM approaches which is increasingly done in QM/MM studies [137–139]. It is however often still problematic to reproduce physical properties accurately with QM methods [140, 141]. Additionally, QM approaches are computationally expensive. This is accented, especially if one wishes to perform free energy calculations, which require significant conformational sampling.

Different strategies to model metal ions with classical simulations have been published previously. In Paper I, we have described them broadly as three classes: the non-bonded soft-sphere model, the covalently bonded approach and the non-bonded dummy model approach. The simple non-bonded soft-sphere models are describing metal-ligand interactions trough electrostatic and van der Waals potentials. While both earth-alkali and alkali metals have been modeled successfully with this approach, the model appeared to be inadequate for modeling multinuclear metal centers [142]. An additional challenge was the simultaneous reproduction of both metal - water distances and free energies of solvation. The second approach using covalently bonded approaches is not allowing ligand exchange and/or interconversion between different coordination geometries [143] and generally suffers from predefined bonds and challenges with a large number of parameters and additionally it needs reparametrization for new systems [135].

The final approach is using a dummy model, originally developed by Åqvist and Warshel for the octahedral dummy model of Mn$^{2+}$ ions [144]. The dummy model approach aims to provide an alternative metal model to deal with shortcomings of the other approaches and reproduce experimental data. For the octahedral dummy model a metal ion consists of six $\delta^+$ charged dummy-particles

35
arranged around a metal center which is in turn is charged $n-6\delta$ (see Figure 3.1). The dummy atoms are kept in octahedral geometry by strong bond- and angle-parameters between the metal center and the dummy-atoms. This approach pre-aligns the charge of the metal in octahedral geometry. Note, that the dummy model can be used for other coordination geometries, as demonstrated for tetrahedral zinc [145].

![Diagram of octahedral dummy model](figure31.png)  

**Figure 3.1.** The octahedral dummy model. A: The dummy model as Lewis structure, with the $\delta^+$ partial charges on dummy atoms and the $n - 6\delta^-$ charge on the metal-center. B: The octahedral dummy model in a protein active site. Reprinted with permission from ref [146]. Copyright 2014 American Chemical Society.

In this work we tested previously available parameters for Mn$^{2+}$, Zn$^{2+}$, Mg$^{2+}$, and Ca$^{2+}$, refined them, and developed new parameters for Ni$^{2+}$, Co$^{2+}$, and Fe$^{2+}$. We parametrized the dummy models to reproduce important experimental properties of metals that had been published, namely the solvation free energy [147], the ligand-metal distances [148], and also the ligand dissociation energies of Mg$^{2+}$ for the glyoxalase I system (see below) [149]. We were able to demonstrate that our dummy-models remain stable both in water and in enzyme, and we showed that the dummies reproduce the above mentioned key experimental properties simultaneously: the free energy of solvation and ion-water distances. Finally, we ran molecular dynamic simulations of the dummies in a protein system to demonstrate general applicability.

We used our parameters to demonstrate their applicability in glyoxalase I. Additionally, my coauthors have employed the octahedral dummy model for simulations of metal ions in different systems [150–153]. In the past three years since Paper I was published, Liao et al. [154] have published a Cu$^{2+}$ model that successfully reproduces the Jahn-Teller distortion, a difficult property to model. Additional octahedral dummy model parameters have been published by Jiang et al. [155, 156]. Very recently a comprehensive review about metal models using classical mechanics has been published [157]. All of which show, that the non-bonded octahedral dummy model is a versatile approach to model metal ions.

To parametrize the metal ions, we used a FEP protocol to obtain the free energy of solvation:
\[ \Delta G_{\text{solv}} = \Delta G_{M^0 \rightarrow M^{n+}}^{FEP} + \Delta G_{\text{Born}} + \Delta G_{\text{cav}} + \Delta G_{\text{std.state}}^{\text{corr}} \]  

Here, the first term is obtained by charging the metal in a FEP procedure:

\[ \Delta G_{M^0 \rightarrow M^{n+}}^{FEP} = -RT \sum_{m=1}^{n-1} \ln \left( e^{-\left( U_{m+1} - U_m / kT \right)} \right) \]  

as explained in section 2.5.5 (with \( U \) as in eq. 2.20). The \( \Delta G_{\text{Born}} \) term corrects for the error introduced by the finite interaction cutoff radius used for the electrostatic interactions, and it can be estimated with:

\[ \Delta G_{\text{Born}} = -166 \frac{Q_i^2}{R_{\text{Born}}} \left( 1 - \frac{1}{\varepsilon(T)} \right) \]  

Where the solute charge is represented by \( Q_i \), the cavity radius of the cavity is \( R_{\text{Born}} \) and the dielectric constant of the medium and the solute are embedded in \( \varepsilon(T) \).

The last two terms of eq. 3.2 cancel out; \( \Delta G_{\text{cav}} \) refers the cost of creating a cavity in the solvent and \( \Delta G_{\text{std.state}}^{\text{corr}} \) is a correction for moving the metal ion from the vacuum to solution.

Having access to improved parameters allows more accurate modeling of new systems. In the next section, CADEE, a tool for \textit{in silico} directed evolution will be introduced.
3.2 Paper IV

CADEE: Computer-Aided Directed Evolution of Enzymes

One goal of my thesis was the development of an automated way to generate and run EVB simulations to allow for in silico directed protein evolution.

In Paper IV we illustrated that rational enzyme design approaches have had significant impact and rational design approaches are still applied today. Rational design is however relying on and limited by the availability of experimental information [158, 159]. Approaches for the directed evolution of enzymes have since revolutionized the development of biocatalysts: This systematic approach employs gene diversification and rapid screening techniques to quickly evolve an existing enzyme design. This enables access to catalytically superior enzymes that would be inaccessible by rational design because the effect of the introduced mutations is sometimes unpredictable [160–162]. Additionally, de novo enzyme designs often need refinement.

With modern experimental approaches throughputs of $10^8$ to $10^{10}$ screened variants per experiment are possible [163]. This large number of constructs can however already be challenged with only 8 amino acids mutated to all 20 variations at each position ($20^8 \approx 2.6 \cdot 10^{10}$, see also Figure 1.7 on page 17). Unfortunately, directed evolution approaches are prone to get trapped in local minima in the sequence space, which can render further evolution impossible [90, 91]. In these cases, a combination of computational and structural approaches can help to rationalize these and to possibly find a way out of a local minimum [90, 91]. In addition, computational approaches that cover a wide range of applications have been published: These include the development of de novo enzyme designs with new properties, hotspot prediction tools with machine-learning, molecular dynamical and quantum mechanical approaches [23, 46, 164, 165]. Also, approaches to perform in silico directed evolution experiments have been put forward in the literature [166]. Most computational enzyme design attempts rely on high-level quantum-mechanical approaches and are computationally expensive. Unfortunately, this problem only accentuates when such approaches should be used to screen thousands of enzyme mutants. To choose computationally cheaper semi-empirical QM/MM approaches would allow for additional sampling and screening, however the inherently limited accuracy of such methods is often problematic [79–81].

Recent studies demonstrate the speed and efficiency of the EVB approach as a tool for computational enzyme design [167–169]. The EVB approach has the distinct advantage of being accurate and fast enough to allow for sufficient sampling. At the same time it, was earlier reliably used for diverse systems [19, 170–174]. Favorably for the computational directed evolution of enzymes, the EVB approach computes relative energetics and hence allows ef-
Figure 3.2. Simulation flowchart: After initialization, CADEE locates the initialization directory and searches for simulation packets (simpacks). They are subsequently distributed on the computational resources and computed with Q. Thanks to the fact that the individual simpacks can be computed independently of each other, CADEE scales *pleasingly parallel* with multiprocessor systems. Published in ref. [120], licensed under CC BY.

sufficient comparison of different enzyme constructs: The EVB is coupled with detailed knowledge of the reaction mechanism and careful parametrization to reference system.

To achieve a tool for *in silico* directed evolution, different aspects need to be addressed. First, once the user has defined mutation sites, mutagenesis must be introduced in an automated fashion, without human interaction. Second, the simulations must be run efficiently and in parallel, and the I/O load needs to be controlled to avoid I/O overload. Finally, the data had to be analyzed on the fly, and the results needed to be accessible during the run, see also Figure 3.2. All the while, calculation time must be kept to a minimum. The solution for this is CADEE: Computer-Aided Directed Evolution of Enzymes, is able to solve potential clashed by adapting rotamers of surrounding sidechains.

3.2.1 Mutagenesis
There are various approaches for modeling *in silico* mutagenesis. One well known approach is to use rotamer libraries [108, 109, 175, 176]. Because traditional rotamer libraries are inconvenient to use with an automatic program
such as CADEE, particularly in cases with inevitable clashes, an automatic solution had to be found. We decided to use SCWRL4, an automated tool that is used to predict side chain orientations based on rotamer distributions.

3.2.2 Efficient & Parallel Execution

It is obvious that hundreds or thousands of mutants need to be run in parallel. Since the individual simulations of mutants are independent of one another, this is a pleasingly parallel scaling problem. Every additional mutant structure can be run on an additional core. A Python script that uses the mpi4py module was implemented to coordinate the parallel computation mutant-structures. CADEE reads a directory with input files and then runs each simpack until a time limit is reached or all the simpacks have been computed.

3.2.3 Automatic Analysis

Unlike in wild-type simulations when a simulation is first run and then analyzed, this would have been highly unpractical for usage within the CADEE framework. It was therefore evident, that CADEE would analyze and compress runs on the fly, while running. To add comfort and accessibility, we decided to store the data in an SQLite database and read the data from the database for the user.

3.2.4 Suggested Workflow

CADEE assists the user by preparing and running free energy calculations with in silico generated enzyme variants. It is therefore crucial that a well-parameterized reference system is selected as a starting point. For illustrative purposes we present an example in silico directed evolution of triose phosphate isomerase (TIM), where the rate-limiting step of the proton abstraction reaction was modeled, see Figure 3.3 for the general workflow.

Initially, the user selects sites for point saturation, and this can for example be achieved by first running an alanine scan. Results of this initial mutagenesis round (see Figure 3.4) were then used to prepare an individual saturation of user selected hotspots saturations on a selection of residues (in our example, the three residues (Y164, L93, and T172)). Finally, we selected for each position a subset of amino acids and did a combinatorial saturation, see Figure 3.5. For the full data set, please refer to Paper IV.

Note, that this workflow can be adapted easily to additional data about of a system. If an enzyme is known to have two or three interesting residues (e.g. hotspots), no alanine scan or individual saturation is needed. Instead a combinatorial saturation can be performed directly, either with the complete 20 amino acids or with custom amino acid library.
Figure 3.3. The suggested workflow using CADEE: After the reaction mechanism is established and the rate limiting step thoroughly parameterized, sites have to be determined, for example with an alanine scan. Next the determined sites can be subjected to in silico site-directed evolution. This process can be repeated until the user decides to end the in silico mutagenesis. Published in Paper IV [120] (fig. 8), licensed under CC BY.

In the next chapter work on StEH1, an epoxide hydrolase that has been subject to experimental directed evolution, will be presented.
Figure 3.4. Results of an alanine scan performed around the active site of triose phosphate isomerase (TIM). Indicated in red are mutations that were found to have a particularly large impact on the activation barrier, compared to the wild-type enzyme (far left). Derived from fig. 8 of Paper IV [120], licensed under CC BY.

Figure 3.5. Results of combined point mutagenesis on three residues (Y164, L93, and T172) around the active site of triosephosphate isomerase (TIM). Published in Paper IV [120] (fig. 10), licensed under CC BY.
4. EVB Applications

4.1 Paper II

Expanding the Catalytic Triad in Epoxide Hydrolases and Related Enzymes

Enzyme variants of StEH1 were tested with the substrate TSO to establish and validate the reaction mechanism. Experimental and computational work has been combined. Mutagenesis variants of the protein E35Q, Y154F, Y235F, and H300N have been tested by experiment and simulation, and key findings are as follows. H104 was found to be conserved in similar sequences, and we suggest that it acts as charge-relay. Y154 and Y235 are confirmed as being important for oxyanion stabilization. The crystal structure of H300N has revealed that this enzyme variant lacks the active-site hydrolytic water.

The work focused on the general mechanism, and the protonation of the active site residues, primarily the active-site histidines (H104 and H300). While H300 was previously suggested as general base, it was also argued that it is protonated during the nucleophilic attack [177]. Our methodology helped us to demonstrate in particular that H104 is also likely to be crucial to the reaction mechanism. We found that H104 is highly conserved and accompanied by glutamic acid or aspartic acid. In the case of StEH1, it is glutamate E35. We have shown that both the H300N and E35Q mutants have some residual activity and the double mutant E35Q/H300N was found to be inactive in the hydrolytic half-reaction, which suggests that E35 may act as a backup base.

In general, any epoxide can be opened at both of its carbons. In case of StEH1 with TSO substrate, the resulting diols cannot be distinguished experimentally. However, computationally this is possible and the energetic profiles of all four variants (two enantiomers and two carbons) are depicted in Figure 4.1.

4.1.1 Protonation of H104 and H300

We found that an additional histidine H104, adjacent to the suggested general base H300, was important for the energetics, in particular the free energies. While earlier studies on epoxide hydrolases have either neglected H104 entirely [82] or have kept it neutral [177], we found that the experimental energetics were only reproducible computationally, when H104 was protonated.
Figure 4.1. Free-energy profiles for wild-type StEH1 catalyzed hydrolysis of (A) (R,R)-TSO and (B) (S,S)-TSO. In color the calculated (C1 attack, blue; C2 attack, red) and in grey the experimentally derived energies. Reprinted with permission from ref [19]. Copyright 2015 American Chemical Society.

Earlier studies predicted the pK_a with PROPKA suggesting that the residue equivalent of H300 in the crystal structure was protonated [177]. We reproduced this for StEH1 with PROPKA, but we additionally determined the pK_a at different stages of the reaction and we found that the pK_a of H300 dropped significantly upon formation of the Michaelis complex. Importantly, this drop in the pK_a was not the case for H104. Encouraged by these findings, we did a sequence alignment with manually curated α/β-hydrolases (using the ESTHER database [178]) and we found that H104 is highly conserved. This strongly suggests, that H104 is protonated during catalysis and H300 is neutral (see also part A of Figure 4.2).

Figure 4.2. Structures in comparison. A: Snapshot of an MD simulation with the active site with key residues (yellow) of the wild-type enzyme with the covalent intermediate TSO (purple). B: Superimposed crystal structures of important residues of wild-type (yellow) and H300N mutant (blue). Reprinted with permission from ref [19]. Copyright 2015 American Chemical Society.
4.1.2 Conservation of E35 and H104

As described in the previous section, we found that H104 was highly conserved in epoxide hydrolases and \( \alpha/\beta \)-hydrolase sequences. We also found that if the residue equivalent of H104 was a histidine, the equivalent of E35 was either a glutamate or an aspartate. This, together with PROPKA \( pK_a \) predictions, we arrived at the conclusion that E35 and H104 both need to be ionized.

4.1.3 Oxyanion Holes

In earlier work on epoxide hydrolases it was suggested that two oxyanion holes play key roles in the catalytic mechanism. The oxyanion hole formed by the two tyrosines (Y154/Y235) was believed to be important for the first step of stabilizing the earlier epoxide-oxygen, and the oxyanion hole formed by F33 and W106 was suggested to be important for the breakdown of the intermediate and for stabilizing the oxyanion formed.

4.1.4 E35Q, H300N and Double Mutant

It was suggested that H300 is the general base required for activity. Experimental mutation of H300 to H300N revealed, however, that the enzyme was able to retain some activity. We suggested that E35 could be the backup base and we have shown experimentally, that the E35Q/H300N double-mutant was inactive in the hydrolytic step. One important finding with the H300N mutant was that the crystal structure of it revealed a rearrangement of the active-site in the H300N mutant and, importantly, the water consumed in the catalytic cycle was no longer observed in this enzyme variant (see also Figure 4.2).

The general reaction mechanism was established with this part of the work. In the next section, we investigated the enantioconvergent behavior of StEH1 with SO substrate.
4.2 Paper III
Conformational Diversity and Enantioconvergence in Potato Epoxide Hydrolase I

Earlier experiments found that the wild-type StEH1 was displaying enantioconvergence with the substrate styrene oxide (SO) to yield primarily (R)-1-phenylethanol. After the reaction mechanism was established with the TSO substrate (see Paper II), we set out to investigate the enantioconvergence of StEH1 and the substrate SO was selected to investigate enantioconvergence.

It was demonstrated earlier, that StEH1 is able to do enantioconvergent catalysis of SO and in this article we combined experimental, computational and crystallographic work to understand and re-engineer the enzyme to shift the activity of the enzyme. The energy profile is shown in fig. 4.3. As can be seen from our energy calculations, we can explain stereo-selectivity. We found that the additional coordination modes the substrate has in the active site has a major impact on the energetic profile compared to TSO.

![Energy Profiles](image)

*Figure 4.3.* Free energy profiles for the hydrolysis of wild-type StEH1 and R-C1B1 mutant (see section 4.2.1) in the lowest energy binding mode for each enantiomer. Reprinted with permission from ref. [174] - Published by The Royal Society of Chemistry.

4.2.1 Engineered Variant and Crystal Structure

As noted earlier, the wild-type enzyme was enantioconvergent the SO-substrate, namely converting both (S)- and (R)- enantiomer to (R)-diol. This motivated experimental research to invert the selectivity, see Figure 4.4).
A mutant enzyme was engineered to invert the enantioselectivity of the enzyme. This enzyme variant, the quadruple mutant R-C1B1, has four mutations W106L, L109Y, V141K, and I155V. As can be seen from Figure 4.4, the residues are located close to the active site.

![Figure 4.4](image)

**Figure 4.4.** Comparison of the wild-type enzyme and the R-C1B1 variant. The backbone, the catalytically important residues, and the mutated residues are displayed in yellow and are from the wild-type crystal structure (PDB-ID 2CJP). Displayed in purple are the superimposed, substituted residues of R-C1B1 from the crystal structure (PDB-ID 4UFN).

### 4.2.2 Binding Modes
Unlike the TSO substrate, SO does not fill the active site of the enzyme. Therefore, the enzyme can bind the substrate in different modes (see Figure 4.5). Our calculations have shown, that the W106 mutation was very important for the inversion of enantioselectivity and at the same time other mutations have made the simulation of the enzyme more difficult to model, for example the V141K mutation.
Figure 4.5. The (R)-styrene oxide substrate showing the two additional binding modes, the asymmetric SO has over TSO (i.e. TSO has one orientation, because for TSO A=B). Note that W106 was subjected to mutagenesis in R-C1B1 and is leucine.
5. Conclusions and Future Perspective

In short, this work is an example for how classical approaches allow for economical extensive conformational sampling. Using the EVB approach has enabled me to obtain a molecular understanding of complex enzymatic systems. I have also increased the reach of the EVB approach, through the implementation of CADEE, which enables efficient and highly parallel in silico testing of hundreds-to-thousands of individual enzyme variants. Methods based on the EVB have been combined with experimental work. Understanding of the reaction mechanism, the regio-, and the enantioselectivity of StEH1 have been achieved.

New Metal Parameters
In this thesis new metal parameters were developed and demonstrated, expanding the capabilities of classical molecular modeling. These parameters have found wide applications and have been used also by co-authors [150–152]. The parameters have additionally been expanded to cover the Jahn-Teller effect by Liao et al. [154].

New Mechanism
The usage of EVB to model reactions allowed to obtain sufficient sampling in StEH1 and to establish the reaction mechanism for this enzyme, in close collaboration with experimental coworkers. With EVB we have been able to uncover the important role the second active site histidine for catalysis. PROPKA has been used to estimate the $pK_a$s of the involved residues and supports our claim. Additionally, we have found that related enzymes of the $\alpha / \beta$ hydrolase family have a very high conservation of this active-site histidine. This expands the relevance of our mechanistic findings from StEH1 to an entire family of enzymes. Modeling the hydrolysis of TSO allowed us to explore mechanism with reduced complexity because less binding modes needed to be considered. Additionally, experimental kinetic data was available for the wild-type enzyme and different enzyme variants for this substrate.

New Understanding of Enantioconvergence
To investigate enantioselectivity and enantioconvergence, we have modeled the enzyme with the smaller styrene oxide substrate. Earlier experimental work
has shown that StEH1 was able to perform enantioconvergent catalysis with this substrate. Additionally, an experimental structure of a variant of the same enzyme has been solved, that displayed inverted enantioconvergence. This was hence an ideal model system and we have employed EVB to investigate the reaction mechanism of this enzyme. Our calculations have shown, that the W106 mutation was very important for the inversion of enantioselectivity and at the same time other mutations have made the simulation of the enzyme more difficult to model, for example the V141K mutation.

New Software
To expand the reach of computational modeling, CADEE a framework for the Computer-Aided Directed Evolution of Enzymes has been developed. With CADEE a tool is now available to the scientific community, that grants easy access towards in silico evolution of enzymes. For now CADEE is implemented such, that the user supplies CADEE with a reference enzyme. Application of CADEE on systems that have been investigated with EVB earlier could be a test field to do explorations with in silico directed evolution. CADEE is, however, not limited to in silico evolution and can alternatively be used to conveniently collect a large amount of sampling for a single specific enzyme variant.

Personal Perspective
To enable a wider scope of CADEE, I suggest to focus on shortening the average simulation times (machine learning, artificial intelligence). This could allow for triple point mutations ($20^3$, 8000 variants) and possibly even quadruple point saturation on a reduced set of amino acids (e.g. $12^4$, 20736 variants). For this, additional computational resources would be needed and CADEE will have to prepare inputs on the fly. Additionally, significantly less data must be retained.

Future Perspective
This work has shown modern applications of methods based on classical molecular modeling. The computational modeling of enzymes remains challenging and with modern hard- and software the possibilities are extended.
6. Populärvetenskaplig Sammanfattning

Modellering av Enzymkatalys och Nya Möjligheter för Datorberäkningar

I min avhandling har jag använt datorer för att studera och förklara kemiska fenomen på molekylär nivå. De senaste åren har datorer lett till en våldsam teknisk utveckling såväl i privatlivet som inom vetenskapen, vilken jag dragit nytta av inom min forskning.

Jag har använt en metod för att utnyttja de tillgängliga beräkningsresurserna på det mest effektiva sättet. Detta är möjligt främst då vår metod förenklar det studerade systemet.

"Så enkelt som möjligt, men inte enklare!" - Albert Einstein

Vi använder en klassisk beskrivning av de atomära systemen. Vi simulerar molekyler bestående av atomer och bindingar som kulor och fjädrar, vilken i motsats till kvantmekaniska beskrivningar innebär att elektronkonfigurationerna inte beräknas exakt. Detta behövs för att beräkna molekyler och reaktioner så noggrant som möjligt. Kvantmekaniska modeller är dock mycket beräkningsintensiva och därför kan vissa egenskaper hos stora molekyler inte beräknas tillräckligt noggrant eller med tillräckligt många upprepningar (för biologiska system måste olika startkonfigurationer provas och systemet måste ges en viss tid att komma jämvikt). Med den förenklade klassiska beskrivningen har jag forskat på atomer, molekyler och reaktioner. Jag har publicerat fyra arbeten inom följande tre områden:

För det andra har jag forskat på ett enzym från potatis. Enzym är biokatalysatörer som, likt en katalysator i en bil, gör en kemisk reaktion snabbare. Detta potatisenzym kan omvandla så kallade epoxider mycket selektivt. Epoxider är molekyler som har en tre-ring av två kolatomer och en syreatom. Epoxider kan användas som råmaterial till olika finkemikalier och läkemedel. I min första artikel om StEH1, Solanum tuberosum epoxidhydrolas, har jag undersökt i detalj hur enzymet genomför reaktionen. Vi har funnit och publicerat reaktionsmekanismen, och dessutom funnit att det är mycket sannolikt att en viss histidin aminosyra i enzymet är joniserad, samt att detta histidin också finns på motsvarande plats i närbesläktade enzymer. I tidigare datorstudier av liknande system har detta histidin ignorerats eller modellerats felaktigt. Genom mitt nästa steg, och tredje vetenskapliga arbete, har vi kunnat förstå en mycket intressant egenskap hos StEH1 bättre. Denna egenskap kallas enantiokonvergens. Enkelt förklarat finns det molekyler som är spegelbilder av varandra. I fall man genomför en reaktion och får två spegelvända produkter men endast är intresserad av den ena, så förlorar man en stor del av produkten! Därför använder man ofta så kallade enantioselektiva katalysatörer vilka främst omvandlar substratet till den ena produkten (och inte dess spegelbild). StEH1, vårt potatisenzym, är en sådan. Men mycket mer spännande för oss var att StEH1 kan omvandla två spegelvända molekyler till en enda produkt, därav benämningen enantiokonvergens. Experiment med StEH1 har visat att selektiviteten till och med kan vändas. Vi har därför undersökt enantioselektiviteten hos StEH1. I vår modell har vi kunnat återskapa fenomenet och bättre förklara varför StEH1 är enantiokonvergent.


Sammanfattningsvis har jag med datorers hjälp studerat enzymer och genom CADEE utvidgat möjligheterna att förstå komplexa enzymatiska system.
7. Acknowledgments

I want to thank all the people who have been there for me. Without help from all of you, this work would have been impossible to achieve.

First, I would like to thank my supervisor Lynne Kamerlin. Without your help, your solid decisions and both your drive and trust, this work would have hardly been possible. Your determined mindset and your to-the-point decisions were crucial. I am also very grateful for the scientific exchange with foreign groups and collaborators; attending conferences, visiting collaborators and all the visitors that you have brought to the lab.

Second, I would like to thank my co-supervisor Mikael Widersten. It was important to me to have a strong link to the experimental facts and for seeing the positive side of results.

Third, I would like to thank my co-workers Fernanda, Paul, Åsa, Miha, Alex, Fabian, Irek, Yashraj, Agata, Doreen, Johan and Sherry. Without your help, my thesis would be awfully thin.

I thank the European Research Council (ERC) for funding this work.

I also thank the Swedish National Infrastructure for Computing, SNIC for the computation time granted on UPPMAX, HPC2N and NSC. Special thanks to Mats Kronenberg (NSC) for the scaling tests performed on Triolith.

I would like to thank David van der Spoel for helpful feedback at different stages during my PhD. And I would like to thank Peter Kasson for fruitful discussions regarding statistics and AI.

I would like to thank ICM, BMC, Uppsala and Sweden for welcoming me and allowing me to conduct this work. I enjoyed the time in this friendly and safe place, and I am fascinated by the wilderness on the doorstep. I thank all the people I have met here and I am grateful, that I was allowed to spend this time of my life in this unusual country.

I thank my colleagues at the Kamerlin Lab:

Fernanda: For being my “scientific older sister” and you became a good friend. Thank you for teaching me how you tackle challenges. Alex: You have
been a “fantastisch” trail-blazer! Thank you especially for settling me in. Paul: Thank you for working with me on StEH1. Ich thank you for our discussions that helped to improve our work. Miha: Thank you for developing Qscripts and allowing me to interface them =). And thanks for all the fruit! Klaudia & Irek: Thanks for the board games at your place, and thanks for sharing Battlestar Galactica, a great game! Yashraj: Thank you for all the yummi, exotic sweets :). And thank you for borrowing TIM. Ania: Thanks for asking tough questions. Dennis: Thank you for your inspiring laser focus. Fabian: I thank you for our vivid Discussions and for our your tips about writing. Quinghua: Thank you for being so friendly and helpful. David: Thank you for the nice parties at your place, stay charming! Malin: Thank you for sharing chocolate, fruit and stuff. Arina: Thanks for being an awesome CADEE-beta tester and for asking tough questions! I would my office colleagues for the friendly and help and “production environment”; :). Thanks to all the visiting students for bringing new ideas, new music and new jokes =).

The POTATIS Team: Åsa, Paul, Sherry and Arina: Solanum tuberosum. Thank you for sharing knowledge, asking interesting questions and sharing.

The Q Team: Irek, Paul, Miha, Alex and Mauricio: Thanks for your contributions to the code.

I thank for feedback from Fabian and Paul on this work. Also, I would like to thank Quinghua and Dennis for feedback on parts of this work. I thank Arvid for help with translating the sammanfattning.

The following people have contributed to this work in a more indirect way: I am thankful for the morning fikas with Zaza and for planning and accountability. Thanks Brotherhood, for helping with the constant evolving and constructive work we did together. You have been a challenge and we will keep this challenging! For organizing social events, I would like to thank especially Ozzi and Petar. I also want to thank all of you, who have discussed different topics providing strong and diverse alternative perspectives and reasoning, helping me to ‘see the world with different eyes’. I would like to thank people from in- and outside BMC and especially Showgy, Lucia, Umut, Zaza, Johanna, Benedikt and Max. I thank my flatmates: Daniele: Thanks for being much fun! Axel & Albin: Thanks for bringing Sweden :)! I thank all my friends Switzerland for their friendship and support. Ein spezielles Dankeschön an Matthias, guet hemmer diskutiert. Livia: Merci für deine Unterstützung. Ich möchte mich auch bei Max Bollinger und Thomas Häfeli für Ihr fordern und fördern bedanken. Es grosses Dankeschön auch an Kunigunde, Toni und Seppi für eure Unterstützung. Schliesslich bedanke ich mich für die Hilfe und Support meiner Eltern, Hans & Edith und meinen Geschwistern Peter, Roman, Helen, Marianne, Erika und Alexander.

TACK ALLA!
8. References


[41] M. J. Van Der Maarel, B. Van Der Veen, J. C. Uitdehaag, H. Leemhuis, and L. Dijkstra. Properties and Applications of Starch-Converting Enzymes of


978-1-4822-9786-7. Google-Books-ID: BdEqBgAAQBAJ.


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)