Computational Enzyme Evolution and Design: Studies of Protein Dynamics and Substrate Binding

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Dušan Petrović
aus Smederevo, Serbien

Jülich, Oktober 2017

aus dem Institute of Complex Systems: Strukturbiochemie
des Forschungszentrum Jülich

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:
[1] Jun.-Prof. Dr. Birgit Strodel
[2] Prof. Dr. Christel Marian
[3] Prof. Dr. Johannes Kästner

Tag der mündlichen Prüfung: 14.02.2018

Statement of authorship

I, Dušan Petrović, hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my investigations. I have fully acknowledged and referenced the ideas and work of others, whether published or unpublished, in my thesis. My thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for a degree at this or any other university. Where the results are produced in collaboration with others, I have mentioned my contributions.

D. P.

Abstract

Enzymes are biomacromolecules able to catalyze chemical reactions. Because of their popularity as catalysts in many industries, enzymology became an advanced field of science. Computational enzymology is an emerging field that provides further insights for the structural and mechanistic characterization of enzymes, and it became a common accessory to the experimental biochemical work. In this thesis, computational enzymology was applied to three different problems.

In the first application, glucose oxidase (GOx) was studied to show how protein dynamics can lead to the formation of both catalytically competent and incompetent enzyme structures. The catalytic histidine residue of GOx is flexible, and molecular dynamics (MD) simulations revealed that this histidine could occupy two possible states of which one is catalytically competent, and the other is not. For quantifying the flexibility of this histidine over an evolutionary trajectory, two types of enhanced sampling MD simulations were performed: Hamiltonian replica exchange (HREX) and umbrella sampling MD. The results of these simulations allowed to explain the catalytic efficiencies of the different GOx mutants. Since MD simulations depend on empirical sets of parameters, called force fields, commonly used force fields were tested for how reliably they represent the different side-chain conformations. For some force fields, including the one used for GOx, a very good agreement with experimental results were obtained, which corroborates the conclusions drawn for GOx.

Enzyme selectivity was the focus of the second project. Selectivity, with respect to the chemical reaction that is catalyzed, is an important property of enzymes, in particular for their applications in organic synthesis. Molecular simulations can be used to shift the selectivity of enzymes, as demonstrated here in the case of cytochrome P450 BM3. The shifts of chemo- and regioselectivity of the oxidation of the 14-membered macrocycle β-cembrenediol were engineered into the V78A/F87A variant of this enzyme, by applying point mutations. The mutations were suggested based on HREX-MD simulations and the enzyme-substrate binding free energies decomposed on a per-residue basis.

Substrate binding, which is a major step in the catalytic cycle, was investigated in detail in the third project of this thesis. While many enzymes bind their substrates in buried cavities, some esterases degrading natural polymers (e.g., cellulose, hemicellulose, chitin) developed additional carbohydrate binding modules (CBMs) enabling them to bind to the surface of a substrate. CBMs were shown to be promiscuous, as they can also bind poly(ethylene terephthalate) (PET) and could potentially be fused with PETase enzymes for plastic degradation. MD simulations were able to explain the differences in the PET binding propensities of different CBM peptides. In addition, suggestions were made for mutations to the best-binding CBM for further enhancing their PET affinity.

The three projects illustrate how MD simulations can be used for providing a rationale for enzyme evolution and enzyme design purposes. In particular, enhanced sampling simulations such as HREX-MD and umbrella sampling MD provide a thorough understanding of the conformational ensembles of enzymes or enzyme-substrate complexes, and the free energy profiles connecting important enzyme states. Given the computational efficiency and ease of use of HREX-MD, this approach is the method of choice for a comprehensive and fast scanning of the energy landscape of the enzyme dynamics or substrate binding as was demonstrated in this work.

Zusammenfassung

Enzyme sind Biomakromoleküle, welche in der Lage sind, chemische Reaktionen zu katalysieren. Aufgrund ihrer Beliebtheit als Katalysatoren in vielen Industriezweigen wurde die Enzymologie zu einem wichtigen Feld der Wissenschaft. Die computergestützte Enzymologie ist ein aufstrebender Fachbereich, der weitreichende Erkenntnisse für die strukturelle und mechanistische Charakterisierung von Enzymen liefert und zusehends Einsatz im Bereich des Enzymdesigns findet. In dieser Arbeit wurde die computergestützte Enzymologie auf drei verschiedene Fragestellungen angewendet.

In der ersten Fragestellung wurde Glucoseoxidase (GOx) mit dem Ziel untersucht zu erklären, wie die Proteindynamik zur Bildung von katalytisch kompetenten als auch inkompetenten Enzymstrukturen führen kann. Molekulardynamik-(MD-)Simulationen haben offengelegt, dass der katalytische Histidinrest von GOx flexibel ist und zwei mögliche Zustände annehmen kann, von denen einer für die Katalyse geeignet ist und der andere nicht. Zur Quantifizierung der Flexibilität dieses Histidinrests über eine evolutionäre GOx-Trajektorie wurden zwei Arten von erweitertem Sampling durchgeführt: Hamiltonian Replica Exchange-(HREX-) und Umbrella Sampling-MD-Simulationen. Die Ergebnisse dieser Simulationen erlaubten es, die katalytische Effizienz der verschiedenen GOx-Mutanten zu erklären. Da MD-Simulationen von empirischen Parametern, sogenannten Kraftfeldern, abhängen, wurden die gebräuchlichsten Kraftfelder im Hinblick ihrer Fähigkeit getestet, wie zuverlässig sie Seitenkettenpopulationen repräsentieren können. Für einige Kraftfelder, einschließlich des für GOx verwendeten Kraftfeldes, wurde eine sehr gute Übereinstimmung mit experimentellen Daten erzielt. Dieses Ergebnis untermauert die für GOx gezogenen Schlussfolgerungen hinsichtlich der Seitenkettendynamik des Histidinrests im aktiven Zentrum.

Die Enzymselektivität stand im Mittelpunkt des zweiten Projektes. Die Selektivität in Bezug auf die chemische Reaktion, welche sie katalysieren, ist eine wichtige Eigenschaft von Enzymen, insbesondere für deren Anwendung in der organischen Synthese. Mittels MD-Simulationen ist es möglich, die Selektivität von Enzymen zu modifizieren, wie in dieser Arbeit im Falle von Cytochrom P450 BM3 gezeigt wurde. Die Veränderung der Chemo- und Regionselektivät der Oxidation des 14-gliedrigen Makrozyklus β-Cembrenediol wurde durch Punktmutationen

der Enzymvariante V78A/F87A realisiert. Die Mutationen wurden auf der Grundlage von HREX-MD-Simulationen und der Enzym-Substrat-Bindungsenergien, welche in ihre Anteile je Aminosäure zerlegt wurden, vorgeschlagen.

Die Substratbindung, welche einen wichtigen Schritt im Katalysezyklus darstellt, wurde im dritten Projekt dieser Arbeit untersucht. Während viele Enzyme ihre Substrate in Vertiefungen im Innern des Proteins binden, entwickelten einige Esterasen, die natürliche Polymere abbauen (z.B. Cellulose, Hemicellulose, Chitin), zusätzliche Kohlenhydrat-Bindungsmodule (CBMs), die es ihnen ermöglichen an die Oberfläche des Substrates zu binden. Einige dieser CBM-Peptide können auch Poly(ethylenterephthalat) (PET) binden, was die attraktive Möglichkeit eröffnet, diese CBMs mit PETase-Enzymen für den Abbau von Plastik zu kombinieren. Mithilfe von MD-Simulationen konnten die unterschiedlichen PET-Bindungseigenschaften verschiedener CBM-Peptide erklärt werden. Darüber hinaus wurden Vorschläge für Mutationen unterbreitet, um die Affinität für PET des am besten bindenden CBM-Peptids weiter zu verbessern.

Diese drei Projekte veranschaulichen, wie MD-Simulationen es ermöglichen, Prinzipien der Enzymevolution aufzudecken und wertvolle Beiträge für das Design von Enzymen zu liefern. Insbesondere Simulationen mit verbessertem Sampling wie beispielsweise HREX-MD und Umbrella Sampling-MD tragen zu einem besseren Verständnis der Konformationsenesmble von Enzym-Substrat-Komplexen bei und erlauben die Bestimmung von Energieprofilen zwischen wichtigen Zuständen während der Katalyse. Angesichts der rechnerischen Effizienz und der Benutzerfreundlichkeit von HREX-MD ist dieser Ansatz eine vielversprechende Methode in der computergestützten Enzymologie, wie in dieser Arbeit gezeigt wurde.

Publications

Publications that are part of the thesis:

I D. Petrović, D. Frank, S.C.L. Kamerlin, K. Hoffmann, B. Strodel: Shuffling active site substate populations affects catalytic activity: The case of glucose oxidase. *ACS Catal.* **2017**, *7*: 6188–6197.

II D. Petrović, X. Wang, B. Strodel: How accurately do force fields represent protein side chain ensembles? *Proteins: Struct., Funct., Bioinf.* **2018**, *doi: 10.1002/prot.25525.*

III P. Le-Huu, D. Petrović, B. Strodel, V.B. Urlacher: One-pot, two-step hydroxylation of the macrocyclic diterpenoid β-cembrenediol catalyzed by P450 BM3 mutants. *ChemCatChem* **2016**, *8*: 3755–3761.

IV D. Petrović, A. Bokel, M. Allan, V.B. Urlacher, B. Strodel: Simulation-guided design of cytochrome P450 for chemo- and regioselective macrocyclic oxidation. *J. Chem. Inf. Model.* **2018**, *58*: 848–858.

Publications beyond the scope of this thesis:

Research articles

v M. Schulte, <u>D. Petrović</u>, P. Neudecker, R. Hartmann, J. Pietruszka, S. Willbold, D. Willbold, V. Panwalkar: Conformational sampling of the intrinsically disordered C-terminal tail of DERA is important for enzyme catalysis. *ACS Catal.* **2018**, *8*: 3971–3984.

vi B. Uluca, T. Viennet, <u>D. Petrović</u>, H. Shaykhalishahi, F. Weirich, A. Gönülalan, B. Strodel, M. Etzkorn, W. Hoyer, H. Heise: DNP-enhanced MAS NMR: A tool to snapshot conformational ensembles of α-synuclein in different states. *Biophys. J.* **2018**, *114*: 1614–1162.

vii Y.S. Kulkarni, Q. Liao, <u>D. Petrović</u>, D.M. Krüger, B. Strodel, T.L. Amyes, J. Richard, S.C.L. Kamerlin: Enzyme architecture: Modeling the operation of a hydrophobic clamp in catalysis by triosephosphate isomerase. *J. Am. Chem. Soc.* **2017**, *139*: 10514–10525.

viii O. Weiergräber, <u>D. Petrović</u>, A. Kislat, M. Pattky, J. Fabig, R. Batra-Safferling, K. Hänel, C. Huhn, B. Strodel, B. Homey, D. Willbold: Structure and dynamics of human chemokine CCL16–Implications for biological activity. *(submitted)*

Books and book chapters

ix M. Zlatović, <u>D. Petrović</u>: Molecular modelling basics. University of Belgrade, Faculty of Chemistry: Belgrade, Serbia, **2016**. ISBN: 978-86-7220-078-2. *(original book title (Serb): Osnovi molekulskog modelovanja.)*

x M. Schwarten, O. Weiergräber, <u>D. Petrović</u>, B. Strodel, D. Willbold: Structural studies of autophagy-related proteins. *(submitted)*

Acknowledgments

Doing a Ph.D. is like running a marathon. Or, at least, it looks like that to me. Oh, wait. I never ran. But still... Many people were cheering for me on the way to the goal, and they deserve my appreciations:

My supervisor, **Birgit**, for helping me develop and realize my ideas, and supporting my work in general. It has been a great pleasure working with you.

The Strodel group: **Philipp**, **Martin**, **Marianne**, **Qinghua**, **Oliver**, **Xue**, **Michael**, and **Bogdan**, who took care of the pleasant atmosphere and many funny moments that made Jülich a nicer place.

Lynn, for hosting me for several months in Uppsala, and to her group members, especially **Paul**, **Miha**, and **Yashraj**, from whom I learned the basics of EVB.

My students: **Veronika**, **Matthew**, **Alexander**, and **Olga**, for the hard work they put into their projects and for all the good ideas they contributed.

My experimental collaborators: **Priska**, **Joanna**, and **Ansgar**, who were willing to bear with my theories and to test the hypotheses I proposed. Thank you. The simulation work is really no fun without experiments.

The IHRS BioSoft grad school colleagues, and especially **Claudia**, for helping me shape my science for the broader audience.

Last but not least, I own my deepest gratitude to my family who was ever supportive during my education.

D. P.

Contents

List of Figures

List of Tables

List of Acronyms

ADH alcohol dehydrogenase. 21

CBM carbohydrate-binding module. 2, 39–41, 43, 46

cI compound I. 32, 35, 36

CoA coenzyme A. 6

CYP cytochrome P450. 2, 31, 32, 35, 46

DFT density-functional theory. 10

DHAP dihydroxyacetone phosphate. 5

FAD flavin adenine dinucleotide. 6, 21

GAP D-glyceraldehyde phosphate. 5

GB3 third immunoglobulin G-binding domain of protein G. 2, 27, 28, 46

GIAO gauge-independent atomic orbital. 33

GOx glucose oxidase. 2, 21–27, 45, 46

HREX Hamiltonian replica exchange. 2, 20, 25, 26, 35, 45–47

LJ Lennard-Jones. 13, 15, 20

MD molecular dynamics. 1, 2, 10, 11, 13, 15–20, 23–28, 35, 40–47

MM molecular mechanics. 9, 13, 36

NAD nicotinamide adenine dinucleotide. 6, 21

NAG *N*-acetylglucosamine. 4

NAM *N*-acetylmuramic acid. 4

NMR nuclear magnetic resonance. 1, 2, 27, 32, 33, 40, 46

PBC periodic boundary conditions. 41

PDB protein data bank. 44

PET poly(ethylene terephthalate). 2, 39–44, 46

QM quantum mechanics. 2, 27, 32, 44, 46

QM/MM quantum mechanics/molecular mechanics. 10, 24, 47

RDC residual dipolar coupling. 27

REX replica exchange. 2, 16, 19

RMSD root-mean-square deviation. 40, 41, 43

SA simulated annealing. 16

TREX temperature replica exchange. 19

US umbrella sampling. 1, 2, 16–19, 26, 45–47

WHAM weighted histogram analysis method. 18

WT wild-type. 22, 25–27

Sometimes science is more art than science, Morty.
A lotta people don't get that.
— Rick and Morty (S1:E6)

1 Introduction

The present thesis on ***Computational enzyme evolution and design*** reports on the new achievements in the field of *in silico* enzymology. Herein, the use of several enzymes was necessary due to their specific characteristics relating to various methodological problems. The thesis presents workflows and benchmarks developed and applied to the enzyme evolution and design studies. All studies were tightly related to experiments: either explaining experimental observations, using them as benchmark data sets, or even predicting new properties that would further be confirmed experimentally. The link between experiments and simulations is crucial for shedding light on a problem from different sides, enabling a better understanding of the problem, and providing more creative solutions.

Chapter 2 introduces enzymes as an important class of proteins and discusses their general classification. This chapter further describes enzyme architecture and prerequisites for catalysis. As some enzymes studied in this thesis require the help from special small molecules, called *cofactors*, to catalyze a reaction, the structure and chemistry of common cofactors are described. Finally, the important physicochemical properties of enzymes, including their *kinetics* and *thermodynamics*, are briefly discussed.

Chapter 3 describes the theory and main applications of the commonly used methods in *in silico* enzymology. Broadly defined, these methods could be divided into two big groups: those that rely on *classical mechanics* and those involving *quantum mechanics*. Undoubtedly, the most influential classical method is *molecular dynamics* (MD), which is based on empirically derived sets of parameters (i.e., *force fields*) to model enzyme structures and their time-dependent behavior. In many cases, *unbiased* MD simulations suffer from the timescale issue. Namely, many processes of interest in enzymology are on time scales longer than currently achievable by MD; for example, *nuclear magnetic resonance* (NMR) observables often report on processes on the μs to s time scales, while MD usually samples ns–μs processes. A handful of ms-long MD simulations have been reported in the literature; however, they typically require specialized hardware solutions (i.e., the ANTON computer at D.E. Shaw Research, NY, USA). Several *enhanced sampling* techniques were developed to overcome the timescale problem. The most commonly used techniques are *umbrella sampling* (US-MD),

metadynamics and *replica exchange* (REX) MD. This thesis reports on the *in silico* enzymology applications of a popular replica exchange flavor—*Hamiltonian replica exchange* molecular dynamics (HREX-MD).

Chapter 4 deals with the enzyme evolution and dynamics, particularly at the level of the amino acid side chains. **Paper I** describes the implications of the side chain dynamics on the catalytic properties of glucose oxidase (GOx). Because unbiased MD simulations were not able to probe the side chain dynamics due to the timescale issue, the enhanced sampling methods (HREX-MD and US-MD) were used to investigate the reaction profiles and energetics of the catalytic histidine's rotation. To confirm that the AMBER force field was a good choice for calculating relative side chain populations in GOx, **Paper II** presents a force fields benchmark of side chain dynamics in two small proteins: ubiquitin and the third immunoglobulin G-binding domain of protein G (GB3), for which extensive experimental NMR data are available.

Chapter 5 concerns the enzyme design for more chemo- and regioselective cytochromes P450 (CYPs). **Paper III** describes the reactivity of some P450 BM3 mutants that can convert β-cembrenediol to its oxidized products. As their stereochemistries could not be determined based on NMR methods alone, quantum mechanics (QM) calculations were used to predict spectra of all possible products to identify the most probable stereomers. Knowing the absolute stereochemistries of the products was crucial for **Paper IV**, which reports on the redesign of P450 BM3 using the computational enzyme design workflow based on HREX-MD simulations. The protocol consists of identifying all possible substrate binding modes in a spacious active site and imposing interactions that would stabilize or destabilize certain binding modes, leading to the shift in selectivity as demonstrated by experiments.

Chapter 6 describes the protein binding of an unusual substrate. Unlike most cases where a small substrate can penetrate into the active site of an enzyme, poly(ethylene terephthalate) (PET) esterases need to bind to a polymer surface for the chemical reaction to occur. A similar problem arises in cellulases, which typically contain a carbohydrate-binding module (CBM) that binds to the substrate and anchors the enzyme long enough to perform the bond cleavage. This chapter deals with the selection and dynamics of several CBMs that were thought to have the potential to bind to PET. The development of an ordered PET model enabled the study of peptide binding to a PET surface using MD simulations, and the analysis of the MD simulations suggested that the binding is dominated by a mixture of π-stacking and H-bond interactions. These insights help in the identification and design of better CBMs.

2 Principles of enzymology

Enzymes are biomacromolecules able to catalyze chemical reactions (e.g., breaking a disaccharide into monosaccharides shown in Figure 2.1). From the chemical nature perspective, enzymes are almost exclusively proteins, while few catalytic RNAs are known as ribozymes. In nature, enzymes catalyze numerous reactions in metabolic pathways. Therefore, understanding how enzymes function is a necessity for the fundamental understanding of nature, but also for medical applications (e.g., producing lactase enzyme to help lactose-intolerant patients or inhibiting the HIV-1 protease as drug target). The commercial applications of enzymes go far beyond the pharmaceutical industry, and many enzymes (e.g., proteases, amylases, and lipases) are used in households on a daily basis, for example in washing powders or contact lens cleaning solutions.
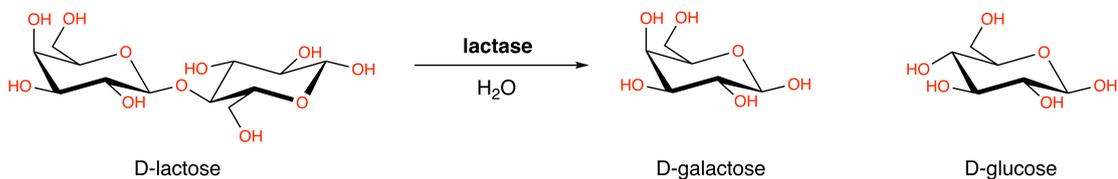
Figure 2.1: The substrate for enzyme lactase is lactose, a disaccharide found in milk. Upon enzymatic reaction, lactose is split into two products: galactose and glucose.

In the early 20th century, the chemical nature of enzymes was unknown. Eduard Buchner, who worked with yeast extracts, was able to enzymatically break down sucrose into glucose and fructose outside of the living cells.[1,2] For "*his biochemical researches and his discovery of cell-free fermentation,*" Buchner received the 1907 Nobel Prize in Chemistry.[3] However, it took more than twenty years before James Sumner crystallized urease and showed that this enzyme consists of only protein.[4] Further work on showing the protein nature of enzymes was accomplished by John Northrop through his research on digestive enzymes (pepsin, trypsin, and chymotrypsin).[5] The 1946 Nobel Prize in Chemistry was awarded to Sumner for "*his discovery that enzymes can be crystallized*" and to Northrop and Stanley for "*their preparation of enzymes and virus proteins in a pure form*".[6]

Sumner's success in crystallizing proteins led to the possibility of solving their 3D structures using X-ray crystallography. While myoglobin and hemoglobin were the first high-resolution crystal structures solved (in 1960),[7,8] the first solved structure of an enzyme was released in 1965 for the hen-egg-white lysozyme (Figure 2.2).[9] The ability to crystallize enzymes and to solve their structures at the atomistic level was the origin of structural enzymology and the study of the structure–function relationships.[10] Only several years after the determination of the lysozyme crystal structure, Arieh Warshel and Michael Levitt developed theoretical methods to study reaction mechanisms.[11]
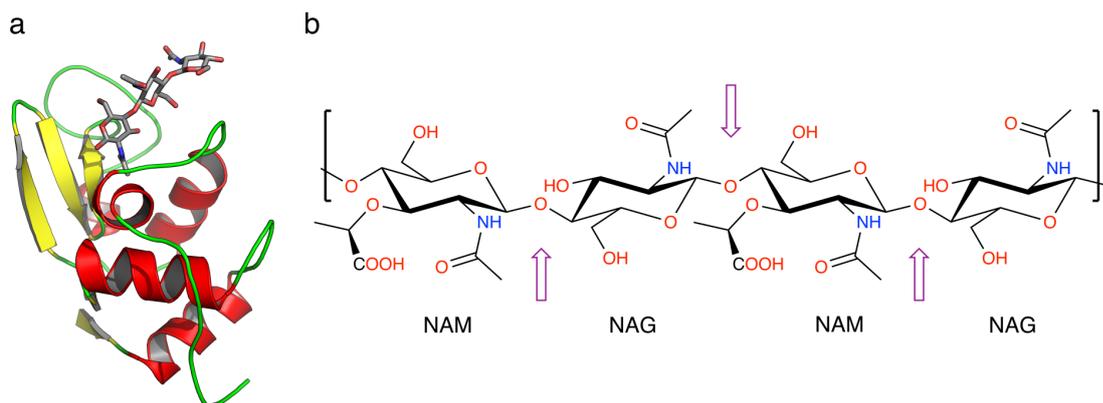
Figure 2.2: Lysozyme, often found in cell walls of gram-positive bacteria, catalyzes the hydrolysis of $\beta(1{\rightarrow}4)$-linkages in peptidoglycans. (a) Crystal structure of hen-egg-white lysozyme with an inhibitor bound. (b) The points of lysozyme cleavage are indicated by the arrows on a glycan consisting of *N*-acetylmuramic acid (NAM) and *N*-acetylglucosamine (NAG).

According to the International Union of Biochemistry and Molecular Biology, enzymes can be classified into six main groups based on the chemistry they catalyze:

1. *Oxidoreductases* catalyze oxidation and reduction reactions, and some members of this class are discussed further in Chapters 4 and 5.

2. *Transferases* speed up the transfer of a functional group from a donor to an acceptor. The notable members of this class are transaminases (exchange of the amine and the keto groups) and kinases (transfer the phosphate group—phosphorylation).

3. *Hydrolases* catalyze the hydrolysis reaction. This class contains, among others, esterases (cleave esters into alcohols and acids) that are further discussed in Chapter 6, proteases (cleave peptide bonds), and phosphatases (cleave phosphoric acid monoesters).

4. *Lyases* also catalyze elimination of chemical groups from a substrate; however, they use mechanisms which do not depend on oxidation or hydrolysis. Some representatives of this class are decarboxylases (remove a carboxyl group) and dehydratases (remove a water molecule from a substrate to create a double C=C bond).

5. *Isomerases* convert one isomer of a substrate to another by facilitating intramolecular rearrangements. One of the most studied members of this class is triosephosphate isomerase, which catalyzes reversible isomerization between dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde phosphate (GAP).

6. *Ligases* catalyze the formation of new covalent bonds. DNA ligase is essential for joining DNA strands (the formation of a phosphodiester bond), while acetone carboxylase catalyzes the formation of a new C–C bond between acetone and CO_2.

2.1 Enzyme architecture

While the enzyme size can vary drastically, from small monomeric proteins to large polymeric machines, only several residues directly participate in the chemical step of catalysis. These residues, often between two and four, are annotated as the *catalytic site*. Other residues directly surrounding the substrate are referred to as the *binding site*. The *active site* is a common umbrella term for the catalytic and the binding sites. The protein scaffold around the active site has several important functions in enzymes. These include, for example, providing occluded and hydrophobic pocket, maintaining the proper geometrical orientation of a substrate and the catalytic site, or providing allosteric effects to tune activity.[12–17] The enzyme architecture of triosephosphate isomerase is shown in Figure 2.3.
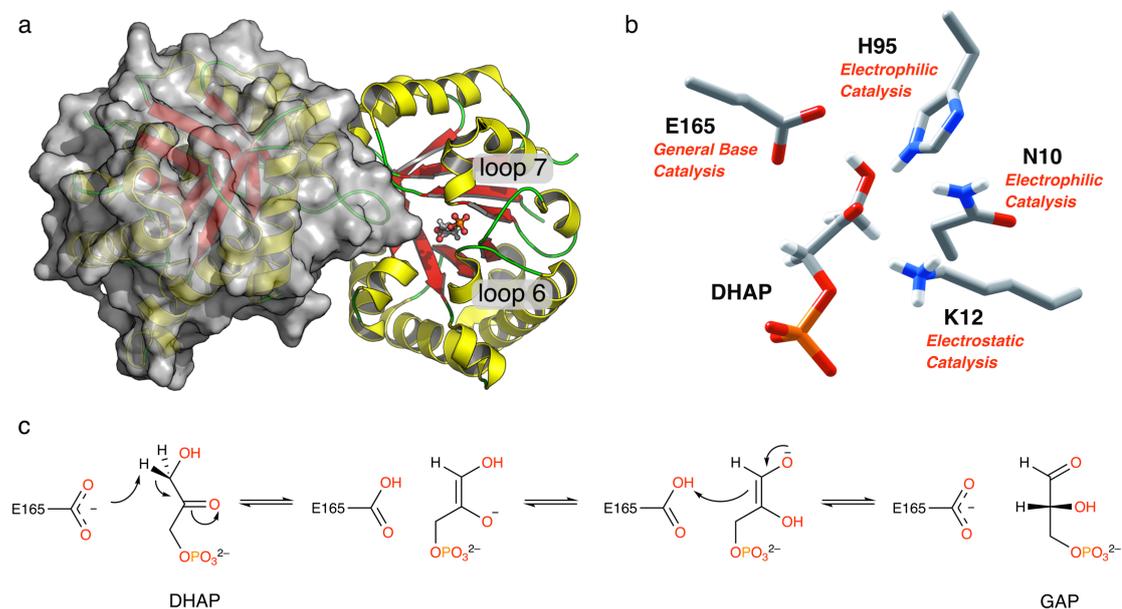
Figure 2.3: Triosephosphate isomerase. (a) Quarternary structure with bound DHAP and highlighted loops 6 and 7 that participate in the substrate clamping and enhance catalysis. (b) The catalytic site includes Glu165 as the general base catalyst and Asn10, Lys12, and His95 as the electrophilic/electrostatic catalysts. (c) Reaction mechanism. *Reprinted with permission from ref. 18. Copyright 2017 American Chemical Society.*

Some enzymes harness the power of small organic molecules or metal ions to perform catalysis; such species are commonly named *cofactors*. Cofactors bind to the active site and actively participate in catalysis as intermediates during electron or functional group transfers. Besides transition metals (e.g., Mn^{2+}, Fe^{3+}, Co^{2+}, Cu^{2+}, Zn^{2+}), Mg^{2+} and Ca^{2+} are also important cofactors in metabolic pathways. Organic molecules acting as cofactors are called *coenzymes*, and if they are tightly but noncovalently bound to the protein, they are referred to as the *prosthetic groups*. Three common coenzymes are shown in Figure 2.4: flavin adenine dinucleotide (FAD), nicotinamide adenine dinucleotide (NAD), and coenzyme A (CoA). While FAD and NAD are involved in electron transfer, CoA transfers acyl group in the metabolism of fatty acids.
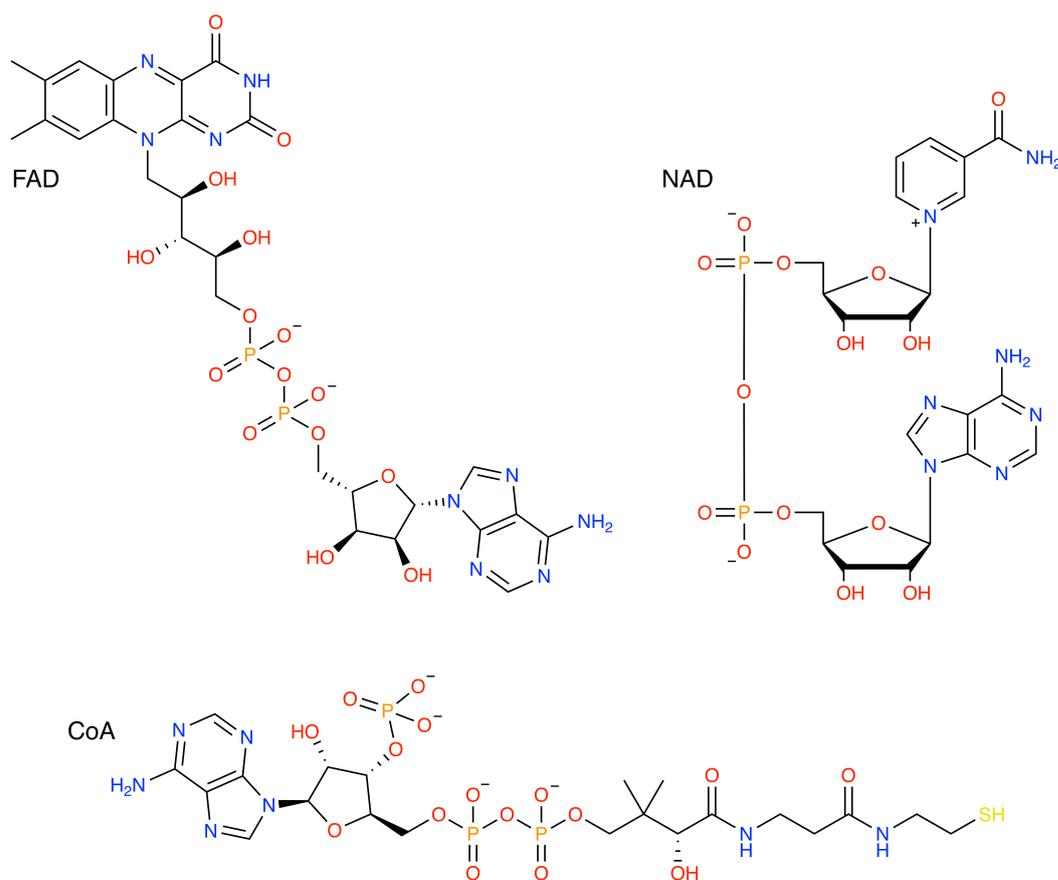
Figure 2.4: Chemical structures of three coenzymes commonly found in nature: FAD, NAD, and CoA.

2.2 Enzyme kinetics and thermodynamics

According to the *transition state theory*, transformation of reactants into products occurs via an unstable transition state characterized by the maximum on a free energy surface (Figure 2.5a). Enzymes increase reaction rate by lowering the activation free energy, as demonstrated in Figure 2.5b.
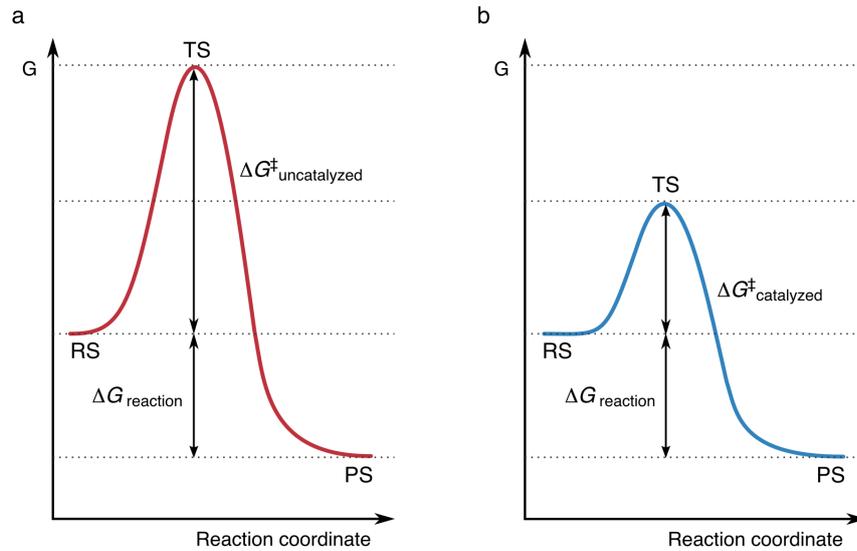
Figure 2.5: Reaction progress in (a) uncatalyzed and (b) catalyzed reaction. The RS, TS, and PS labels denote the reaction, transition, and product states, respectively. The activation ($\Delta G^{\ddagger}$) and Gibbs free energy of reaction ($\Delta G_{\text{reaction}}$) are indicated on the plots.

To explain the enzymatic rate enhancement, Leonor Michaelis and Maud Leonora Menten proposed a kinetic model which describes reaction rate dependence on the concentrations of substrate and enzyme (the so-called Michaelis-Menten kinetics). In a reaction:

$$\text{E} + \text{S} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} \text{E·S} \xrightarrow{k_{\text{cat}}} \text{E} + \text{P}$$
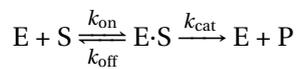
E, S, and P are respectively the enzyme, substrate, and product, while k_{on} and k_{off} are the association and dissociation rate constants for the enzyme–substrate complex, respectively, and k_{cat} is the rate constant for the product formation. The rate of catalysis, v, can be defined as the amount of the product formed in a unit of time:

$$v = \frac{d[\text{P}]}{dt} \tag{2.1}$$

where [P] is product concentration. At some initial substrate concentration $[\text{S}]_0$, v is proportional to the total enzyme concentration $[\text{E}]_0$. For a given $[\text{E}]_0$, v increases linearly with [S] for low $[\text{S}]_0$ values, and it asymptotically approaches the maximal velocity (V_{max}) at high $[\text{S}]_0$ values:

$$v = V_{\text{max}} \frac{[\text{S}]}{K_{\text{M}} + [\text{S}]} \tag{2.2}$$

and

$$K_{\mathrm{M}} = \frac{k_{\mathrm{off}} + k_{\mathrm{cat}}}{k_{\mathrm{on}}} = \frac{[\mathrm{E}]\,[\mathrm{S}]}{[\mathrm{E}\cdot\mathrm{S}]} \tag{2.3}$$

where K_{M} is the Michaelis constant, i.e., the concentration of a substrate at $V_{\mathrm{max}}/2$. The Michaelis-Menten kinetics is graphically shown in Figure 2.6.
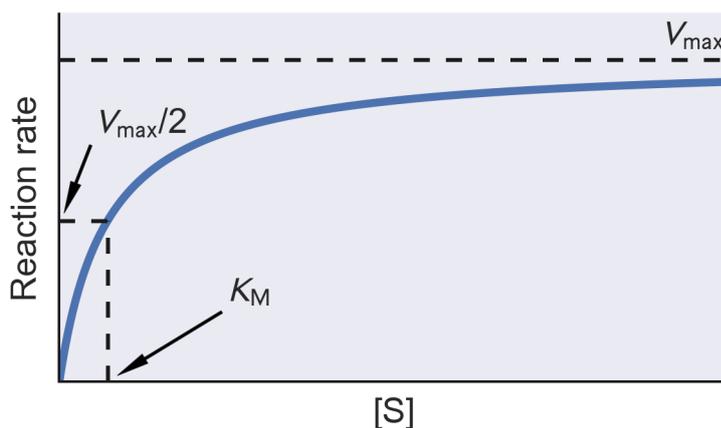
Figure 2.6: The Michaelis-Menten kinetics describes the dependence of the reaction rate for an enzyme catalyzed reaction on the substrate concentration [S].

From the chemical reaction of the enzyme catalyzed transformation of S to P, the reaction rate of the product release is equivalent to the turnover number k_{cat}, i.e., the number of the catalytic cycles in a unit of time:

$$k_{\mathrm{cat}} = \frac{V_{\mathrm{max}}}{[\mathrm{E}]_0} \tag{2.4}$$

Together with K_{M} and k_{cat}, the catalytic efficiency, $k_{\mathrm{cat}}/K_{\mathrm{M}}$, represents the effective rate constant:

$$\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}} = \frac{k_{\mathrm{on}}\,k_{\mathrm{cat}}}{k_{\mathrm{off}} + k_{\mathrm{cat}}} \tag{2.5}$$

Moreover, the catalytic efficiency is a useful parameter for comparison of different enzymes (e.g., mutants) catalyzing the same reaction. The $k_{\mathrm{cat}}/K_{\mathrm{M}}$ value is maximal when $k_{\mathrm{cat}} \gg k_{\mathrm{on}}$, in which case the reaction is diffusion limited.

3 Computational enzymology

The ancient Greek philosophy provided many ideas on how nature functions. In his work *On the Heavens* (Περὶ οὐρανοῦ), Aristotle, in 350 BC, described cosmology and the principles of motion of the *heavenly bodies*. He claimed that terrestrial bodies move to their *natural place*, and introduced the terms *natural motion* and *forced motion*. Although later shown to be mistaken in some of his theories, Aristotle's philosophy was an important step for the advancement of the scientific thought.

It was not until the late 17th century, when Isaac Newton published, in *Philosophiæ Naturalis Principia Mathematica*, the three laws that became the basis of *classical mechanics*—Newton's laws of motion. In addition to the law of inertia and the law of action and reaction, Newton's second law indicates that the acceleration, *a*, of an object is equal to the force, *F*, acting upon it divided by the mass, *m*, of this object:

$$a = \frac{F}{m} \tag{3.1}$$

Newton's second law is the foundation of the computational classical or molecular mechanics (MM) method to simulate the motion of atomic nuclei. Initially, this approach was used for finding the natural or *equilibrium* positions of atoms in small molecules (i.e., geometry optimization). Together with the advances in the computer hardware and software, the primary application of classical mechanics shifted toward the investigation of the time dependent motion (i.e., dynamics) of macromolecular systems involving tens or hundreds of thousands of atoms.

The discovery of *quantum mechanics* pioneered the physical sciences of the 20th century, providing a new view on the world by describing the motion of the subatomic particles—electrons. The electronic structure calculations, based on the *Schrödinger equation*, proved to be especially important for chemistry. In their 1927 paper, Walter Heitler and Fritz London described, for the first time, the quantum concept of a chemical bond in the dihydrogen

molecule.[19] Further advances included the development of the *ab initio* quantum mechanics by Douglas Hartree and Vladimir Aleksandrovich Fock, which was a cornerstone for contemporary quantum chemistry.[20, 21] Walter Kohn and Lu Jeu Sham built on the Hartree-Fock method to develop the modern-day *density-functional theory* (DFT).[22, 23] The 1998 Nobel Prize in Chemistry was awarded to Kohn "*for his development of the density-functional theory*" and to Pople "*for his development of computational methods in quantum chemistry.*"[24]

At the current state of the art in computational chemistry, a rule of thumb is to use quantum mechanics for small organic and inorganic molecules where the electronic effects are dominant, and molecular mechanics in cases where the electronic effects can be approximated. For systems approaching the speed of light, the relativistic mechanics and quantum field theory should be used instead of the classical and quantum mechanics, respectively.

Multiscale models that describe the behavior of a system at a range of levels (from electronic, over atomistic and coarse-grained, to mesoscale) are beneficial for addressing multiple phenomena simultaneously. In some cases, the electronic effects can play a dominant role in macromolecular behavior; e.g., enzymes are proteins that can effectively form and cleave chemical bonds. As chemical bonds are a quantum phenomenon, it is not possible to model chemical reactions with pure classical mechanics. Hybrid *quantum mechanics/molecular mechanics* (QM/MM) approaches were developed for such cases. In QM/MM, the active site of an enzyme is treated at the quantum level, enabling the proper description of the chemical bonding, while the rest of the protein is treated classically, to allow conformational sampling at a much lower computational cost.[25] "*For the development of multiscale models for complex chemical systems,*" Martin Karplus, Michael Levitt, and Arieh Warshel were awarded the 2013 Nobel Prize in Chemistry.[26]

3.1 Molecular dynamics

As already introduced, MD is an invaluable tool for studying the structure and dynamics of macromolecules such as proteins. While the equations of motion that MD relies on are already developed, the outcome of a simulation considerably depends on the used set of parameters to describe the atoms and their interactions. These parameters, together with the mathematical models of the interaction potentials, are called force fields. Many biomolecular force fields are available, often with similar mathematical representations but unique sets of parameters.

The current MD simulations are typically performed on the ns- to μs-timescales. However, many functionally important motions in proteins (e.g., loop dynamics, allosteric changes, and protein folding) happen on the μs- to s-timescales.[27] Since the computer hardware limits performing such long simulations, several different enhanced sampling methods were developed to overcome this limitation. In the enhanced sampling methods, one biases the system in a certain way to speed up the motion of interest.[28]

3.1.1 Equations of motion

To simulate the time dependent behavior of a system, MD relies on Newton's second law (Equation 3.1). Given the initial set (at time $t = t_0$) of (1) *atomic coordinates* (obtained, for example, from a crystal structure), (2) *velocities* (usually assigned under the Maxwell-Boltzmann distribution at a given temperature), and (3) the *potential energy function*, one can estimate the new coordinates and velocities of the system after a short time interval, Δt, assuming constant interactions for very short times (e.g., 1–2 fs). The MD trajectory is constructed by allowing the system to move many times with the same short time step.

Equation 3.1 can be written in the differential form where the force (i.e., the negative of the energy gradient) is:

$$-\frac{\mathrm{d}U}{\mathrm{d}\mathbf{r}_i} = m\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} \tag{3.2}$$

Here, U represents the potential energy of a system with certain conformation $\mathbf{r}_i$, where $\mathbf{r}_i$ contains all atomic coordinates of the system. After some time, Δt, the new positions of the atomic coordinates, $\mathbf{r}_{i+1}$, are given by a Taylor expansion:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_i(\Delta t) + \frac{1}{2}\mathbf{a}_i(\Delta t)^2 \tag{3.3}$$

where the velocities, $\mathbf{v}_i$, and accelerations, $\mathbf{a}_i$, are the first and second derivative of the positions with respect to time, respectively. The positions $\mathbf{r}_{i-1}$ at time $t - \Delta t$ are:

$$\mathbf{r}_{i-1} = \mathbf{r}_i - \mathbf{v}_i(\Delta t) + \frac{1}{2}\mathbf{a}_i(\Delta t)^2 \tag{3.4}$$

The positions of a system at any time can then be estimated from the addition of Equations 3.3 and 3.4:

$$\mathbf{r}_{i+1} = 2\mathbf{r}_i - \mathbf{r}_{i-1} + \mathbf{a}_i(\Delta t)^2 \tag{3.5}$$

where the acceleration at each time step is calculated from the force (Equations 3.1 and 3.2). The *Verlet* integration algorithm (Equation 3.5),[29] where the velocities are not considered explicitly, has a problem with running simulations at a constant temperature. Namely, the temperature T, at time t, is calculated from the average kinetic energy at that time, and

if it does not correspond to the desired temperature, velocities need to be scaled. In the NVT (constant number of particles, volume, and temperature) and NpT (constant number of particles, pressure, and temperature) ensembles, where the temperature is controlled by a thermostat (i.e., coupled to a heat bath), the energy is added or removed from the system gradually in time, with a coupling parameter τ. For example, for Berendsen thermostat,[30] the velocity scale factor is given as:

$$\sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{desired}}{T_{actual}} - 1\right)}$$

Similarly, the constant pressure p can be maintained with a barostat (i.e., coupled to a pressure bath). In the Berendsen barostat,[30] this is achieved by scaling the coordinates of the system with:

$$\sqrt[3]{1 + \kappa\frac{\Delta t}{\tau}\left(p_{actual} - p_{desired}\right)}$$

where κ is the compressibility of the system.

The velocity problem of the Verlet algorithm can be resolved by the *leap frog* algorithm.[31] By estimating Equations 3.3 and 3.4 at half time steps, $i + \frac{1}{2}$, the addition equivalent to Equation 3.5 gives:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_{i+\frac{1}{2}}(\Delta t) \tag{3.6}$$

The velocity is then:

$$\mathbf{v}_{i+\frac{1}{2}} = \mathbf{v}_{i-\frac{1}{2}} + \mathbf{a}_i(\Delta t) \tag{3.7}$$

which indicates that the positions and velocities are not updated at the same time, but with half a time step difference.

3.1.2 Force field

In MM, and subsequently in MD, the force field energy is calculated as a sum of several bonded (stretching, bending, and torsion) and nonbonded (Lennard-Jones (LJ) and electrostatic) components (Figure 3.1):

$$U_{FF} = U_{stretch} + U_{bend} + U_{torsion} + U_{LJ} + U_{electrostatic} \tag{3.8}$$

Bonded interactions

Nonbonded interactions
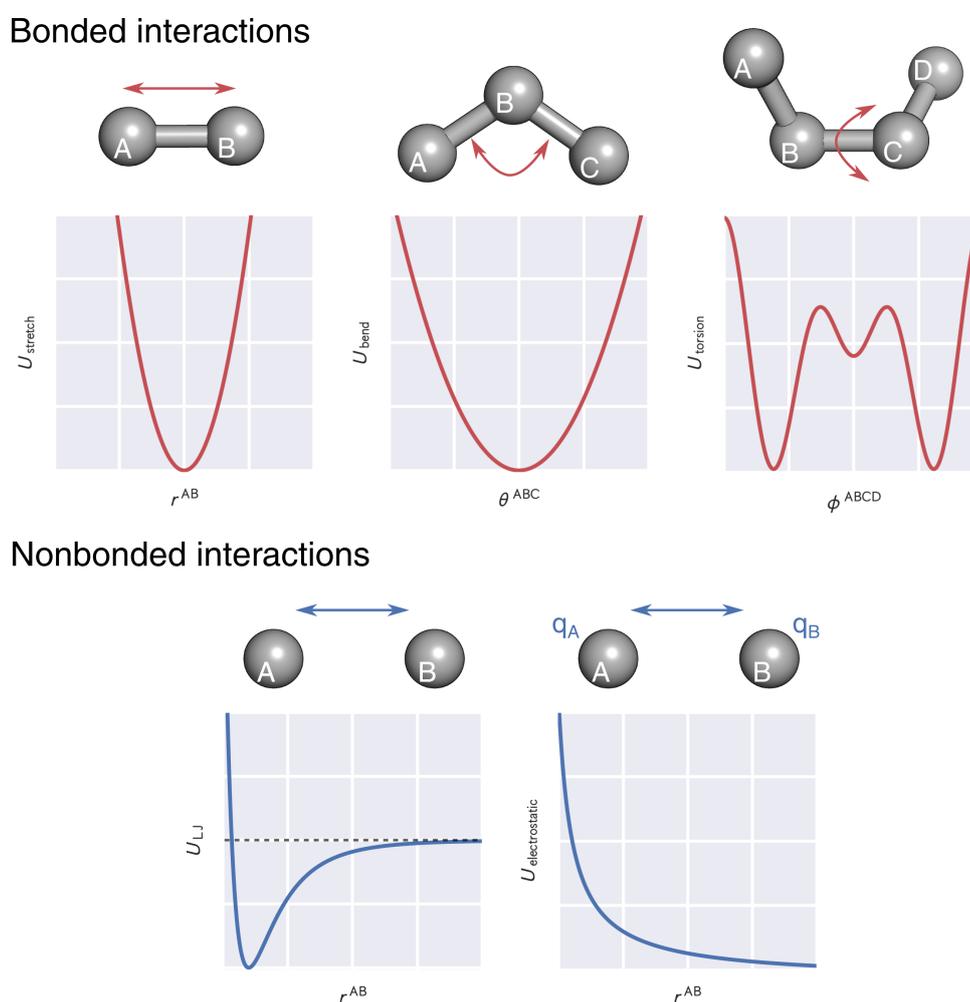
Figure 3.1: The bonded interactions model the chemical bonds, angles, and dihedral angles, while the nonbonded interactions describe the Lennard-Jones and electrostatic interactions.

Bonded interactions

The basic form of the stretch energy can be written in the form of a harmonic oscillator:

$$U_{stretch}\left(r^{AB}\right) = \frac{1}{2}k_r\left(r^{AB} - r_0\right)^2 \tag{3.9}$$

where r^{AB} is the distance between atoms A and B, r_0 is the equilibrium bond length, and k_r is the stretching force constant. The quadratic potential is simple, yet it is sufficient to describe the molecular geometry at the state of equilibrium. When the correct limiting behavior is necessary (e.g., in bond breaking and bond forming), the harmonic oscillator is unable to describe these events involving bond stretching to infinity, and then the Morse potential should be used.

The bending energy describes the change in the angle between three consecutive atoms that are forming two bonds. Like the previous term, U_{bend} is typically represented by a harmonic function:

$$U_{bend}\left(\theta^{ABC}\right) = \frac{1}{2}k_\theta\left(\theta^{ABC} - \theta_0\right)^2 \tag{3.10}$$

where θ^{ABC} and θ_0 are the angle between atoms A, B, and C and the equilibrium value of the angle, respectively, and k_θ is the bending force constant.

The torsional energy reflects the rotation around a dihedral angle between four consecutive atoms that are forming three bonds. Unlike $U_{stretch}$ and U_{bend}, $U_{torsion}$ has a different form, mainly due to the required periodicity. Namely, when the dihedral is rotated for the full circle, the energy should take the identical value:

$$U_{torsion}\left(\phi^{ABCD}\right) = k_\phi\left[1 + cos\left(n\phi^{ABCD} + \delta\right)\right] \tag{3.11}$$

where ϕ^{ABCD} is the dihedral formed by atoms A, B, C, and D that form three consecutive bonds, δ is the phase, n describes the periodicity of the function, that is, the number of minima (e.g., for $n = 1$ the function is periodic by $360°$, for $n = 2$ the function is periodic by $180°$), and k_ϕ is the torsion force constant. More than one function is often necessary to properly describe the torsion around a dihedral angle, and the energy is then composed of the sum of all the torsional terms.

A special potential—typically denoted *improper dihedral*—is often used to describe the out-

of-plane bending energy, and it can be either of harmonic or periodic type. It is generally used to enforce planarity, for example, of aromatic rings or the peptide bond.

Nonbonded interactions

The LJ energy represents the repulsion and dispersion contributions to the interaction of nonbonded atoms:

$$U_{LJ}\left(r^{AB}\right) = 4\epsilon^{AB}\left[\left(\frac{r_0}{r^{AB}}\right)^{12} - \left(\frac{r_0}{r^{AB}}\right)^6\right] \tag{3.12}$$

where r^{AB} is the distance between two particles, ϵ^{AB} is the depth of the LJ well, and r_0 is the distance at which $U_{LJ} = 0$. The first term (r^{-12}) models repulsion, while the second (r^{-6}) models attraction (i.e., van der Waals interactions). At short nuclear distances, U_{LJ} quickly takes large positive values, becoming extremely repulsive. At large distances, $U_{LJ}\left(r^{AB}\right)$ approaches zero. The $U_{LJ}\left(r^{AB}\right)$ value is modeled as zero after a certain cutoff (typically 9–12 Å) to reduce the computational cost.

The force field assigns partial charges to each atom in the system, which contributes to the electrostatic energy, $U_{electrostatic}$. As this term describes the interactions between the point charges, the most appropriate relation to mathematically describe it is the Coulomb potential:

$$U_{electrostatic}\left(r^{AB}\right) = \frac{1}{4\pi\epsilon_0}\frac{q^A q^B}{r^{AB}} \tag{3.13}$$

where r^{AB} is the distance between particles A and B, q^A and q^B are the charges of the two atoms, and ϵ_0 is the vacuum permittivity.

3.2   Enhanced sampling MD

Unbiased atomistic MD simulations, as described above, are a powerful tool to study enzyme motion. However, MD suffers from the time scale issue. At the current state of the art, unbiased MD simulations can, at best, describe the conformational ensemble for the processes happening at the ns- to low μs time scales. Assigning free energies to different states in the conformational ensemble requires significantly more sampling. The *energy landscape* (i.e., all energy minima and maxima) needs to be crossed multiple times to obtain statistically significant results, and even for ns-long processes, μs-long simulations are required to quantify the landscape properly.[32]

Several alternative approaches were developed to overcome the time scale problem. For example, the atomistic representation could be simplified by coarse graining the system. Such coarse-grained force fields combine several atoms into one bead so that the whole amino acid is represented by only a few beads.[33] The smaller number of particles in the system, together with the possibility to use a larger time step during a coarse-grained MD simulation, allows sampling the conformational space significantly faster. The applied assumptions are, however, connected with information loss, especially for the side chain dynamics, as coarse grained force fields often use only one bead for the whole side chain.

Special methods were developed to efficiently cross high energy barriers and sample the complete conformational landscape with atomistic accuracy. These enhanced sampling algorithms bias a simulation either by lowering the barrier or by physically forcing the system over the barrier. Such methods could crudely be divided into two classes: (I) those that bias a system over a specific reaction coordinate and (II) those that bias the whole system uniformly, without the explicit definition of a reaction coordinate. The group I includes methods such as US-MD[34] and metadynamics.[35] The US-MD constraints the system in a series of states along the reaction coordinate (the so-called US windows), effectively pushing the system over the barriers. Metadynamics, on the other hand, explores a minimum on the energy landscape and adds the "computational sand" to increase the basin energy, which effectively reduces the energy barrier and enables the system to cross over it more easily and to explore other parts of the energy landscape.

Although the group I methods are outstanding in estimating the energies of minima and maxima, they require a previously predefined reaction coordinate, and they bias the system along this single coordinate (or a small number of coordinates) only. Therefore, the results of these methods are heavily dependent on the selection of the reaction coordinate. The group II methods, on the other hand, do not require a specific reaction coordinate, as they bias a whole system (or a part of a system) uniformly. The representatives of this group are simulated annealing (SA)[36] and REX-MD.[37] In SA-MD the system is heated, which provides more kinetic energy to cross the energy barriers, and subsequently cooled down, allowing the system to fall into the closest local minimum. After several rounds of SA-MD, the system should explore all energy minima, and the global minimum can then be identified. In REX-MD, which could be considered as an advanced SA-MD technique, several replicas of the system are simulated simultaneously, e.g., at different temperatures, and allowed to exchange the coordinates from time to time. These exchanges allow crossing the barriers at higher temperatures while running the reference replica at the relevant temperature ensemble (e.g., at 300 K).

3.2.1   Umbrella sampling

At a given temperature T, a system samples the accessible conformations $\mathbf{r}$, with probability, $P(\mathbf{r})$:

$$P(\mathbf{r}) \propto e^{-\frac{U(\mathbf{r})}{k_B T}} \tag{3.14}$$

where k_B is the Boltzmann constant. Reducing the high dimensionality of $\mathbf{r}$ to a single reaction coordinate, ξ, the probability of the system with respect to the reaction coordinate can be expressed as the free energy, $G(\xi)$:

$$G(\xi) = -k_B T \ln P(\xi) \tag{3.15}$$

and where the probability distribution along the specified reaction coordinate, $P(\xi)$, is:

$$P(\xi) \propto \int e^{-\frac{U(\mathbf{r})}{k_B T}} \delta(\xi - \xi(\mathbf{r})) d\mathbf{r} \tag{3.16}$$

As a consequence of the energy barriers, $P(\xi)$ converges slowly and an additional harmonic potential, $V(\xi)$, is added in US-MD simulations to restrain the conformational sampling of the system close to the desired space, ξ_0:

$$V(\xi) = \frac{1}{2} k(\xi - \xi_0)^2 \tag{3.17}$$

The biased probability, $P'(\xi)$, is then:

$$\begin{aligned} P'(\xi) &\propto \int e^{-\frac{U(\mathbf{r}) + V(\xi(\mathbf{r}))}{k_B T}} \delta(\xi - \xi(\mathbf{r})) d\mathbf{r} \\ &\propto e^{-\frac{V(\xi(\mathbf{r}))}{k_B T}} P(\xi) \end{aligned} \tag{3.18}$$

From this, the biased free energy, $G'(\xi)$, can be expressed as:

$$G'(\xi) = -k_B T \ln P'(\xi) = G(\xi) + V(\xi) + c \qquad (3.19)$$

where c is the free energy shift. Thus, the unbiased free-energy $G(\xi)$ can be written as:

$$G(\xi) = G'(\xi) - V(\xi) + c \qquad (3.20)$$

However, to combine different windows, the data needs to be weighted based on the error magnitude from the corresponding histograms. One possibility for doing this is the weighted histogram analysis method (WHAM) that determines the probability distribution and free energy shifts iteratively to self-consistency.[38]

In a typical US-MD simulation, one uses many harmonic potentials to construct different windows and move along the reaction coordinate (i.e., change the ξ_0 value). This procedure is of particular importance for sampling the conformations at the energy barriers, which are often not stable and, therefore, require a high biasing potential. A separate $P'(\xi)$ is estimated for every window, with windows being distributed in such a way that the probabilities of neighboring windows substantially overlap. The results of all windows are then combined. The probability distribution of a mock US-MD simulation along a distance coordinate (e.g., the distance between the substrate and the active site of an enzyme) is shown in Figure 3.2.
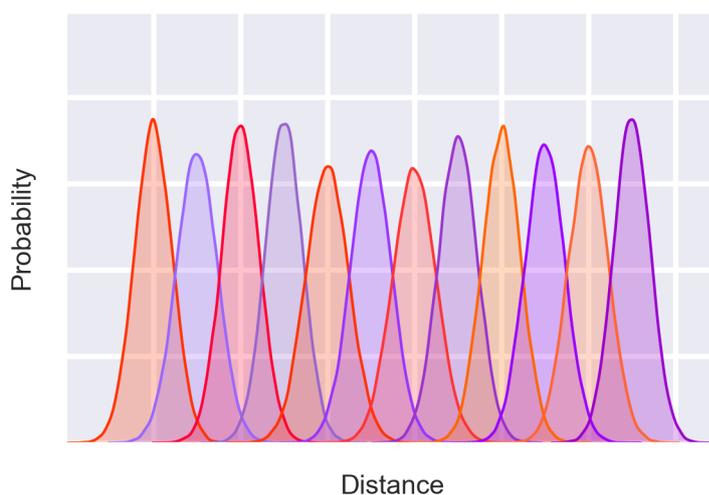
Figure 3.2: Probability distribution from a 12-window US-MD simulation along a reaction coordinate representing the distance between two groups.

3.2.2 Replica exchange

In many cases where the US-MD reaction coordinate is not an intrinsic geometric property, such as an angle or a torsion, finding a proper coordinate is not a trivial task. Instead of biasing the system along with a predefined reaction coordinate ξ, one may accelerate the sampling by changing temperature or the effective energy function, also called Hamiltonian, of the system under study, allowing to cross energy barriers more efficiently. One such method is REX-MD, where N replicas of a system are simulated under different conditions, e.g., temperatures. In temperature REX (TREX) MD simulations, all replicas use the same energy function, U_{FF} (Equation 3.8), but they are coupled to heat baths at different temperatures. At the reference temperature (e.g., 300 K), the system is allowed to sample the proper dynamical properties, while the replicas at the higher temperatures can efficiently cross the energy barriers.

In TREX-MD, neighboring replicas are allowed to exchange their coordinates at specified time intervals. The exchanges between the replicas are accepted based on the Metropolis-Hastings criterion,[39, 40] with the probability, α, of:

$$
\begin{aligned}
\alpha &= \min\left(1, \frac{e^{-\left(\frac{U(\mathbf{r}_i)}{k_B T_j} + \frac{U(\mathbf{r}_j)}{k_B T_i}\right)}}{e^{-\left(\frac{U(\mathbf{r}_i)}{k_B T_i} + \frac{U(\mathbf{r}_j)}{k_B T_j}\right)}}\right) = \\
&= \min\left(1, e^{\left[U(\mathbf{r}_i) - U(\mathbf{r}_j)\right]\left[\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right]}\right)
\end{aligned}
\tag{3.21}
$$

A TREX-MD simulation is usually set up in such a way that α reaches the optimal value of 0.3–0.7, ensuring that the sampling is sufficiently enhanced. To achieve this α value, the potential energy distributions of two replicas need to overlap significantly. TREX-MD is very convenient for simulations employing an implicit solvent, which typically have a smaller number of particles and, hence, a limited number of degrees of freedom. As the system size increases, more temperature replicas are required to achieve the optimal exchange rates. That is why the presence of explicit water makes TREX-MD simulations inefficient.

Based on a convenient estimator provided by van der Spoel,[41] for a protein made of ~1,000 atoms (~2,000 degrees of freedom), 8 replicas would be necessary to reach $\alpha = 0.3$ in the temperature range of 300–400 K. Solvating the protein with 4,000 rigid water molecules increases the number of degrees of freedom to ~26,000. Under the same conditions, this TREX-MD simulation would require 40 replicas. Having an even larger solvation box with, e.g., 10,000 water molecules, leads to ~62,000 degrees of freedom and would need more than 60 replicas.

Although TREX-MD quickly becomes inefficient with the addition of water molecules, the protein behavior is modeled more reliably with explicit than with implicit solvent.[42] For this reason, other REX flavors were developed to solve the inefficiency problem of explicit solvent MD simulations. One of them is Hamiltonian replica exchange MD.

Hamiltonian Replica Exchange

The replica exchange solute tempering (or HREX-MD)[43] simulations do not use increasing temperatures to enhance sampling. The temperature is rather kept the same for all replicas. Instead of the temperature, HREX-MD scales the Hamiltonian, or U_{FF} (Equation 3.8), of the system for the different replicas. This allows the system to be split into two regions: hot ($\mathscr{H}$) and cold ($\mathscr{C}$), where the Hamiltonian is scaled for the $\mathscr{H}$ region only. The $\mathscr{H}$ region could be, for example, the protein, while the solvent would be the $\mathscr{C}$ region (Figure 3.3). The $\mathscr{H}$ region could also be only a part of a protein, e.g., a relevant residue or a single loop in protein. Such a variation is refereed to as *partial tempering*.
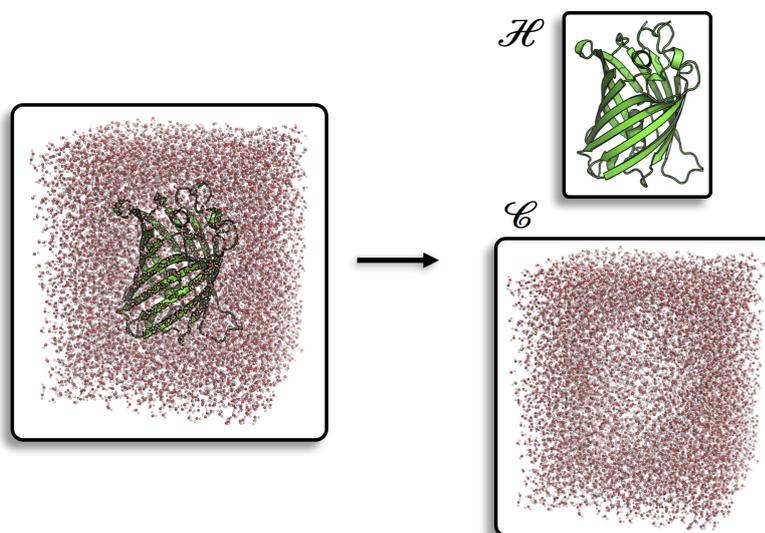
Figure 3.3: A solvated protein system can be split into two regions under the HREX-MD formalism: the hot ($\mathscr{H}$), consisting of the protein, and the cold ($\mathscr{C}$) one, formed by the water molecules.

The Hamiltonian of the $\mathscr{H}$ region depends on the scaling parameter λ, which takes the value of 1 for the reference replica, and values < 1 (in principle down to 0) for the increasing effective temperatures. However, λ rarely takes the zero value (corresponding to infinite temperature) as this would require a large number of replicas. Most often, λ reaches 0.7–0.5 for the minimum value, representing effective temperatures of 450–600 K, respectively.

In the Hamiltonian of the $\mathscr{H}$ region, the atom charges (q^A and q^B) are scaled by $\sqrt{\lambda}$, the LJ energy parameter (ϵ^{AB}) is scaled by λ, and the proper dihedral potential, $U_{torsion}(\phi^{ABCD})$, is also scaled by λ. In a case of partial tempering, if either of atoms A and D, but not both, is in the $\mathscr{H}$ region, the dihedral potential is scaled by $\sqrt{\lambda}$. Scaling Hamiltonians allows for the effective temperature T/λ in the $\mathscr{H}$ region, $T/\sqrt{\lambda}$ for the interface between the two regions, and T for the $\mathscr{C}$ region.[44]

4 Side chain dynamics

Oxidoreductases are a big class of enzymes that catalyze electron transfer from a reductant to an oxidant. They commonly require cofactors to carry out the reaction.[45] Notable members of this group are alcohol dehydrogenase, laccase, cytochrome P450 monooxygenase, and glucose oxidase. The alcohol dehydrogenase (ADH) enzymes catalyze the reduction of alcohols to aldehydes and ketones, and these enzymes are present in many species: in bacteria,[46] yeasts,[47] plants,[48] and animals.[49] Mammalian ADH enzymes use the power of NAD coenzyme and zinc to oxidize otherwise toxic alcohols present in food, or consumed via alcoholic beverages. Laccase catalyzes the copper-dependent one-electron oxidation of phenols.[50] In the food industry, laccase can remove polyphenolic compounds from beer and juice, improving their quality.[51] Together with glucose oxidase (anode), laccase (cathode) is used for the construction of enzymatic biofuel cells.[52] Cytochromes P450 are a widely present group of hemoproteins that catalyze the monooxygenation reaction.[53] More on this superfamily will be discussed in Chapter 5.

Glucose oxidase is an enzyme commonly found in fungi and insects. It catalyzes the oxidation of β-D-glucose to δ-glucono-D-lactone.[54] Similar to other oxidoreductases, GOx needs a noncovalently bound cofactor, FAD, to complete the reaction. Flavoprotein oxidases are an exquisite group of enzymes that use molecular oxygen as electron acceptor.[55] GOx from *Aspergillus niger* is among the most studied members of this family, and many properties of this enzyme served to resolve general features of the whole family.[56]

The enzyme operates by the ping-pong bi-bi mechanism, consisting of reductive and oxidative half-reactions. In the reductive half-reaction, two hydrogen atoms are transferred concertedly from the C1 (anomeric) atom of glucose to the enzyme (its His516 residue and FAD).[57] The oxidative half-reaction happens via two single electron transfer steps involving molecular oxygen, requiring a positive charge in the active site for oxygen binding.[58] The His516 residue of *A. niger* GOx is well conserved among the family.[56] In addition to His516, another semi-conserved residue in the family corresponds to the His559 of *A. niger* GOx, that either remains His or changes to Asn.[56] The general GOx reaction mechanism is shown in Figure 4.1.
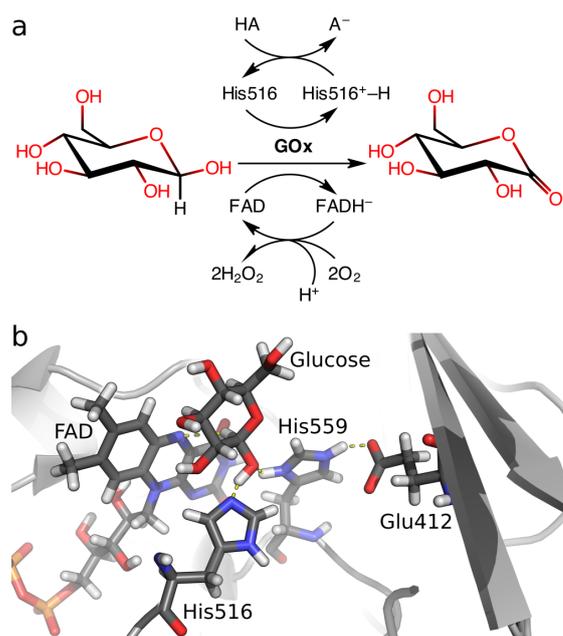
Figure 4.1: (a) Glucose oxidase catalytic mechanism. (b) The active site of *A. niger* GOx. *Reprinted with permission from ref. 59. Copyright 2017 American Chemical Society.*

GOx is an important industrial catalyst, with a multi-billion dollar market in the medical and pharmaceutical industries (i.e., for the construction of glucose biosensors).[60] Other GOx applications are in the food and beverage industry for preservation and low-alcohol-content wine production,[54] in the textile industry for bleaching,[54] for enzymatic biofuel cells that can power implantable devices such as cardiac pacemakers,[61] or for the construction of logic circuits.[62] Many studies have been conducted trying to identify mutations that could enhance some of the key GOx properties, such as the catalytic activity,[63, 64] oxidative[65] and thermostability,[63, 64] oxygen independence,[66] and electron transfer mediation.[67] However, the reason why many of these mutations were beneficial remained mostly unexplained. To bridge this gap, an extensive study of some of these mutations and their effect on the structure and dynamics of GOx was performed within the present work.[59]

4.1   Paper I: Active site preorganization

As a very evolved enzyme, GOx from *A. niger* already possesses excellent catalytic properties. Nonetheless, a recent study identified several mutations that can further increase its catalytic efficiency (k_{cat}/K_M) compared to the wild-type (**WT**) enzyme.[64] At first, three mutations were introduced to form the parent enzyme, **P**: T30V, I94V, and A162T. In the further evolution rounds, two mutants, named **A2** and **F9**, showed superior properties. The **A2** mutant has two additional mutations: R537K and M556V. Further variants of **P**, carrying either of the two additional mutations separately, were created *in silico* to study the effects of the latter two mutations, i.e., **Pk** with the R537K and **Pv** with the M556V mutation. The summary of

the computationally investigated variants, including their kinetic properties,[64] is given in Table 4.1.

Table 4.1: Summary of the simulated GOx mutants and their catalytic properties. The kinetic parameters k_{cat}, K_M, and k_{cat}/K_M are given in the units of s^{-1}, mM, and $mM^{-1}s^{-1}$, respectively.

GOx	k_{cat}	K_M	k_{cat}/K_M	Mutations						
WT	189.4	28.3	6.7	—	—	—	—	—	—	—
P	291.8	15.0	19.5	T30V	—	I94V	—	A162T	—	—
Pk	n.d.	n.d.	n.d.	T30V	—	I94V	—	A162T	R537K	—
Pv	n.d.	n.d.	n.d.	T30V	—	I94V	—	A162T	—	M556V
A2	498.3	18.5	26.9	T30V	—	I94V	—	A162T	R537K	M556V
F9	345.2	19.8	17.5	T30V	R37K	I94V	V106I	A162T	—	M556V

The active site's His516 was previously reported to have a flexible side chain.[68, 69] Further exploration of this observation with both X-ray crystallography and MD simulations showed that the higher catalytic activity of **A2** GOx is a consequence of the conformational lock of His516, restraining it in a conformation preorganized for the catalysis to occur. The active site preorganization was previously shown to have a significant effect on many enzymes.[12, 18, 70–73]

4.1.1 Histidine rotamer library

The chemical structure of an amino acid can be divided into two main components, (1) the *backbone* and (2) the *side chains*. While the backbone atoms are identical for all amino acids, the side chain chemistry contributes to their unique properties. Some side chains can be as simple as a hydrogen atom or a methyl group in Gly and Ala, respectively, or as complicated as the indole ring in Trp. The side chain dynamics is described by the corresponding χ dihedral angles. All amino acids, except Gly and Ala, have at least one χ angle: the χ_1 is typically defined by the N–C_α–C_β–C_γ atoms (the exceptions are Cys and Thr that have sulfur or oxygen at the γ-position). The number of χ angles depends on the nature of amino acid, where Arg is the most complex one with five side chain dihedral angles. On the other side, His residue has only two dihedrals (Figure 4.2a). The χ_2 angle of His is defined by the C_α–C_β–C_γ–N_δ atoms.

The analysis of the protein crystal structures database revealed that most side chains cluster around certain stable geometries. Three such clusters are shown in Figure 4.2b for the His residue along the χ_1 angle: two gauche (g^+ and g^-) and one trans (t) conformer. A discrete residue conformation along all relevant χ angles is known as a rotamer, and all such conformations are compiled in rotamer libraries. Along the χ_2 angle of His, six possible clusters exist. Therefore, combining both χ_1 and χ_2 angles gives rise to eighteen conceivable rotamers. Their relative probabilities from the Dynameomics rotamer library[74] are shown in Figure 4.2c.
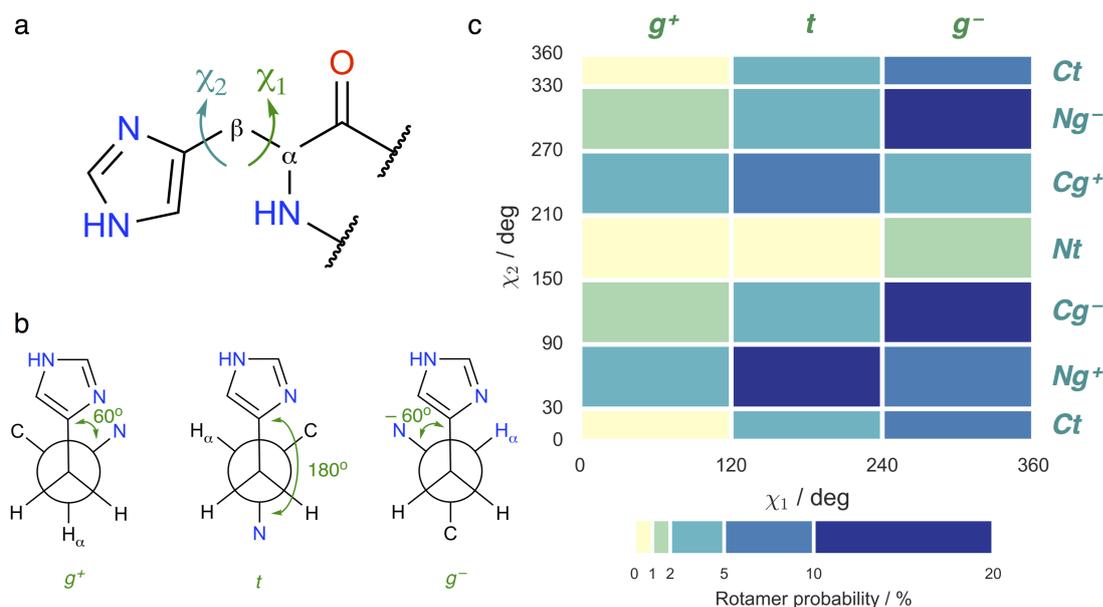
Figure 4.2: (a) The side chain dynamics of a histidine residue can be described by two χ angles. (b) His has three rotamers along the χ_1 angle. (c) His rotamer probabilities from the Dynameomics rotamer library.

4.1.2 His516 rotamers sampled in GOx

Seven available GOx crystal structures revealed that only two His516 rotamers are accessible to the enzyme: the (g^-, Nt) and (g^-, Ng^+) geometries.[59] An overview of the χ angles observed in different GOx crystal structures is presented in Table 4.2. The structures 1GAL,[68] 1CF3,[69] 3QVP,[75] and 3QVR[75] are crystals of the wild-type GOx from *A. niger*, while 5NIT and 5NIW are its mutants.[59] The 1GPE structure originates from GOx from *Penicillium amagasakiense*.[69]

Table 4.2: Overview of the sampled His516 χ_1 and χ_2 dihedral angles among the solved GOx crystal structures.

PDB ID	1GAL	1CF3	3QVP	3QVR	5NIT	5NIW	1GPE
χ_1 / deg	257	254	291	295	293	288	284
χ_2 / deg	225	194	185	195	197	199	64

The (g^-, Nt) rotamer was used to study the mechanism of the GOx reaction with QM/MM,[76] and molecular docking finds a glucose binding mode suitable for this reaction (Figure 4.3a). Furthermore, the same rotamer was observed in a related enzyme, glucose dehydrogenase, that was crystallized with the lactone product (PDB ID: 4YNU).[77] Therefore, this rotamer will be designated as the *catalytic conformation* henceforth. The MD simulations performed in this work suggest that the His516 flipping to the (g^-, Ng^+) rotamer involves an H-bond stabilization by His559. In this geometry, however, the proton transfer from glucose to the enzyme is unfavorable (Figure 4.3b). Hence, this rotamer is called the *noncatalytic conformation*.
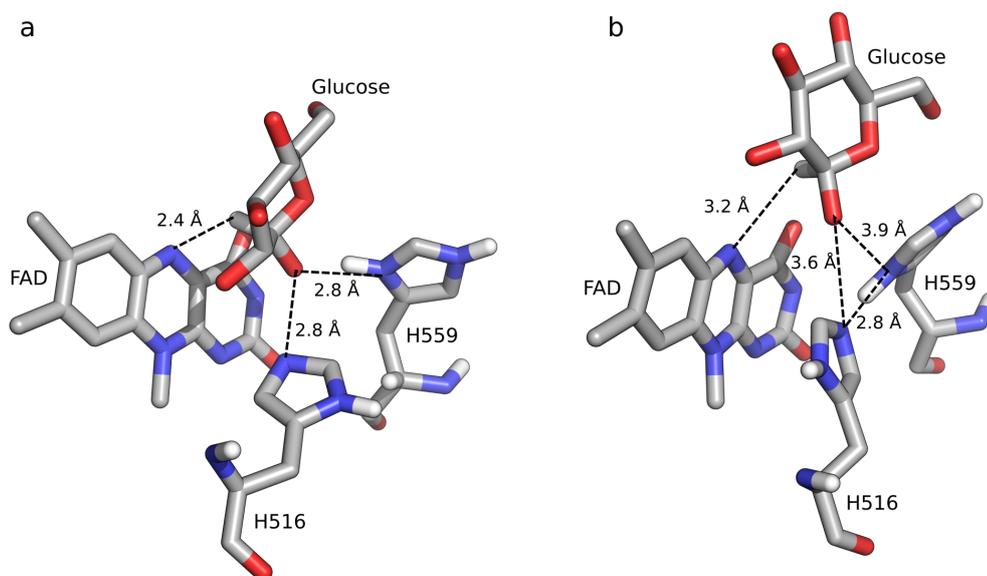
Figure 4.3: Representative structures of the (a) catalytic and (b) noncatalytic conformations of His516 in *A. niger* GOx obtained by MD simulations. *Reprinted with permission from ref. 59. Copyright 2017 American Chemical Society.*

4.1.3 His516 conformational ensembles from MD simulations

The unbiased MD simulations revealed a higher flexibility of the His516 residue in the **WT** GOx compared to the mutants.[59] However, it was not possible to quantify this motion from several 100 ns long simulations as the His ring flipping is a rare event. Therefore, HREX-MD simulations were performed for the laboratory evolutionary of GOx to study the effects of mutations on the His516 side chain dynamics. The χ_1 angle remained conserved to the g^- geometry in all variants, while χ_2 switched between the *Nt* and Ng^+ conformations. In the **WT** enzyme, the two χ_2 conformations were similar in energy, but the noncatalytic (g^-, Ng^+) geometry was somewhat more probable that the catalytic (g^-, *Nt*) rotamer (Figure 4.4a). The introduction of the initial three mutations in the parent mutant had only a small effect on His516, making the catalytic conformation only slightly more probable (Figure 4.4b). The two additional mutations in **A2** GOx make this variant a much more proficient catalyst by almost entirely removing the steric possibility for the occurrence of the noncatalytic conformation (Figure 4.4c).

The superior catalytic properties of the **A2** GOx mutant over the **WT** enzyme could be explained by the differences in the His516 side chain dynamics. The origin of this flexibility in the **WT** enzyme may lay in the fact that a small cavity, close to the active site, allows plenty of room for the noncatalytic His516 conformation to occur (Figure 4.5a). In the **A2** mutant, where the M556V mutation is located in the cavity and causes small geometric perturbations to the adjacent residues, the cavity size is significantly decreased (Figure 4.5b), which prevents big structural rearrangements of the His516, thereby locking it to the catalytic conformation.[59]
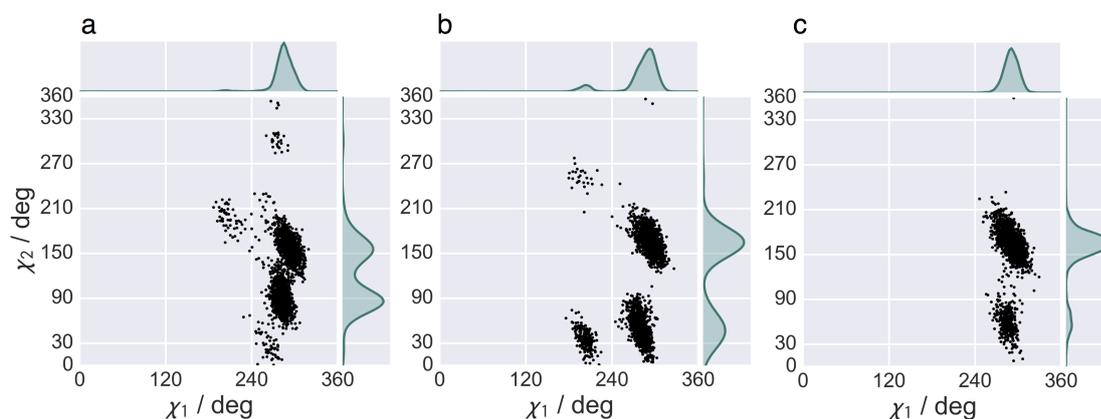
Figure 4.4: His516 side chain geometries in the (a) **WT**, (b) **P**, and (c) **A2** GOx obtained from HREX-MD simulations. Only one pronounced minimum can be identified for the χ_1 dihedral, while two different populations can be observed for the χ_2: the catalytic and noncatalytic conformations, with χ_2 dihedrals around 160 deg and 60 deg, respectively. *Modified with permission from ref. 59. Copyright 2017 American Chemical Society.*
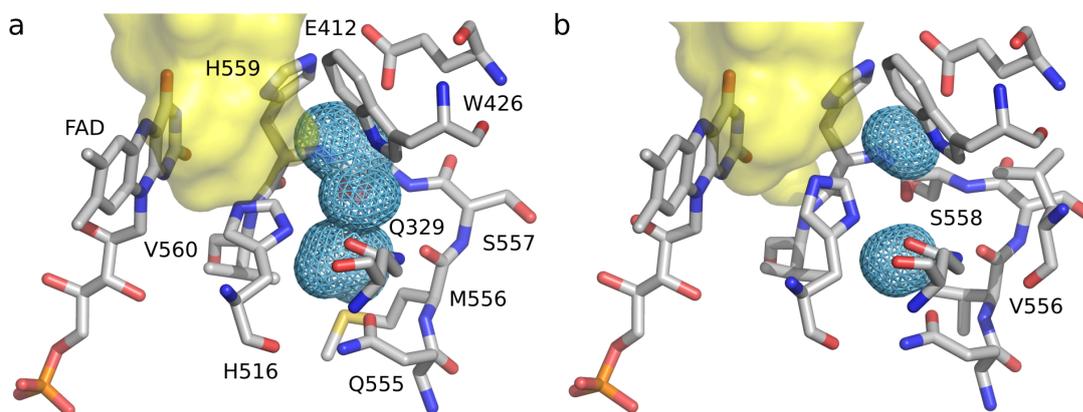
Figure 4.5: A small cavity (blue mesh) proximal to the active site (yellow surface) of the (a) **WT** GOx and (b) **A2** mutant. *Reprinted with permission from ref. 59. Copyright 2017 American Chemical Society.*

Finally, the US-MD simulations were able to corroborate the HREX-MD results (Figure 4.6). The relative positions and energies of the two minima on the free energy profile of the His516 dihedral motion are in good agreement between the two methods. However, HREX-MD does not sample, with sufficient frequency, the conformations at the transition state, so it is not a reliable method for the precise determination of the barrier energy.[59]
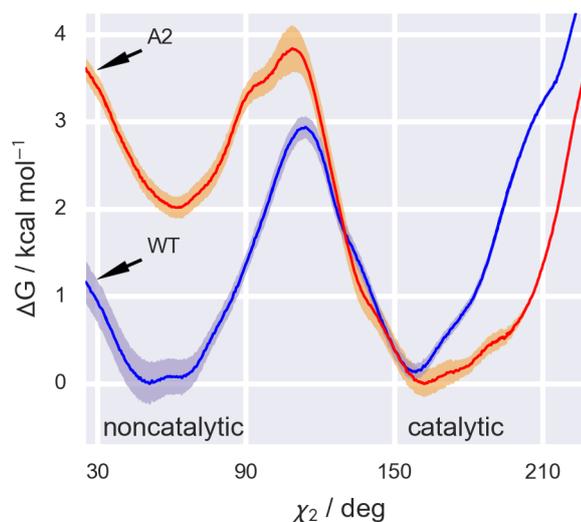
Figure 4.6: Free energy surface of the His516 rotation around the χ_2 angle of the **WT** GOx (blue) and its **A2** mutant (orange). *Modified with permission from ref. 59. Copyright 2017 American Chemical Society.*

4.2 Paper II: Validating side chain dynamics from MD

As mentioned in Chapter 3, many different force fields were developed to simulate proteins by MD. The main four groups of force fields currently used are AMBER, CHARMM, OPLS, and GROMOS, each of them providing many different versions, also called flavors in the remainder. As all these force fields were parameterized in unique ways, to represent certain experimental or QM observables, the results of MD simulations could be force field dependent. Therefore, to confirm that the force field choice for the **Paper I** was reasonable, an investigation on how well different force fields represent the side chain dynamics was performed.

Many force field benchmarks have already been published, often looking at the parameters describing the backbone motion. Side chain dynamics was never validated for all common force fields against a unique data set. Previous work on testing the performance of force fields for side chains was focused on the comparison of only a few force fields. Furthermore, different benchmarks used various experimental NMR observables, such as the scalar ^3J coupling constants, S^2 relaxation order parameters, and order parameter of the methyl symmetry axis, offering the results that were not directly comparable. Finally, many benchmarks were not performed on proteins but rather on short peptides.

4.2.1 Benchmark systems

Two small proteins, whose side chain preferences were well characterized by NMR, were used for the study: ubiquitin and GB3, which are shown in Figure 4.7. Bax and coworkers determined, based on ^3J and residual dipolar coupling (RDC) constants, the average rotamer

angles and their populations of Val, Ile, and Thr residues in ubiquitin, and of all non-Gly and non-Ala residues in GB3.
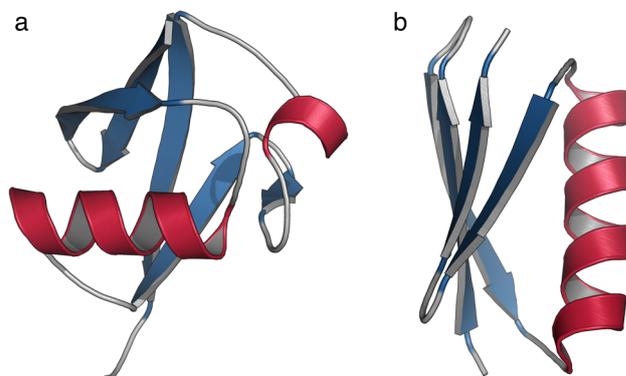
Figure 4.7: 3D structures of the proteins used in the study: (a) ubiquitin and (b) GB3. *Reproduced with permission from ref. 78. Copyright 2018 Wiley Periodicals, Inc.*

The combined analysis of almost 200 μs of MD simulations of ubiquitin and GB3 was performed for twelve force fields:

1. *AMBER* 03, 03*, 99SB-ILDN, 99SB*-ILDN, 14SB, and FB15

2. *CHARMM* 22, 22*, 27, and 36

3. *OPLS*-AA

4. *GROMOS* 54A7

4.2.2 Force field accuracy

Comparing the mean rotamer angles with the experimentally determined ones gives very high correlations (R^2 > 0.96) for all force fields (see Figure 4.8a for example results for AMBER 99SB*-ILDN), even for the GB3 simulation with the CHARMM 22 force field, where the protein unfolds. This observation indicates that the force fields are well parameterized to identify side chain minima on the potential energy surface.[78]

However, quantifying the free energies of the structures at minima (i.e., the relative rotamer populations) is significantly harder. Many force fields identify the most populated rotamer incorrectly, especially for side chains on the surface of the protein. Here, AMBER 99SB*-ILDN performs the best (Figure 4.8b), having only 7 outliers from 43 residues tested in GB3. Regarding performance, this force field is followed by AMBER 99SB-ILDN and 14SB, as well as CHARMM 36 force field. AMBER 99SB*-ILDN force field gives a very good agreement with experiments (Figure 4.8c), considering the populations of all rotamers from all residues, and it performs at a very similar level of accuracy as its star-uncorrected version AMBER 99SB-ILDN, as well as the new CHARMM 36. The only force field that performs even better in this regard is

the new AMBER 14SB force field.[78] These results indicate that the force field choice for **Paper I** is reasonable, providing trustworthy results.
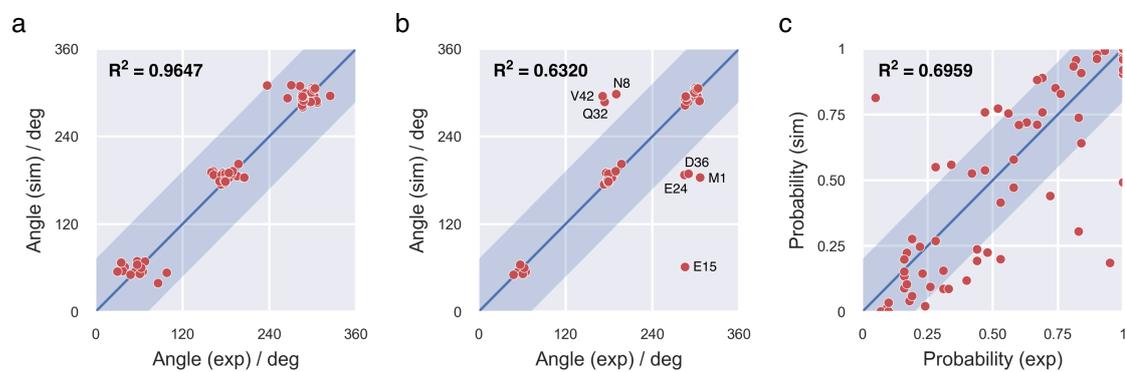
Figure 4.8: AMBER 99SB*-ILDN force field validation for the GB3 protein: (a) average rotamer angle, (b) the most populated rotamer, and (c) all rotamer populations. *Reproduced with permission from ref. 78. Copyright 2018 Wiley Periodicals, Inc.*

5 Enzyme selectivity

The cytochrome P450 (CYP) enzymes are hemoproteins widely spread in the tree of life.[79] Heme is bound to the protein through an axially positioned cysteine residue. If carbon monoxide (CO) binds as the sixth ligand of the protoporphyrin IX, CYPs show a characteristic Soret absorbance at 450 nm in the CO difference spectrum (Figure 5.1). This feature is useful for the analysis and quantification of CYPs and was also the origin of the superfamily's name.[80, 81]
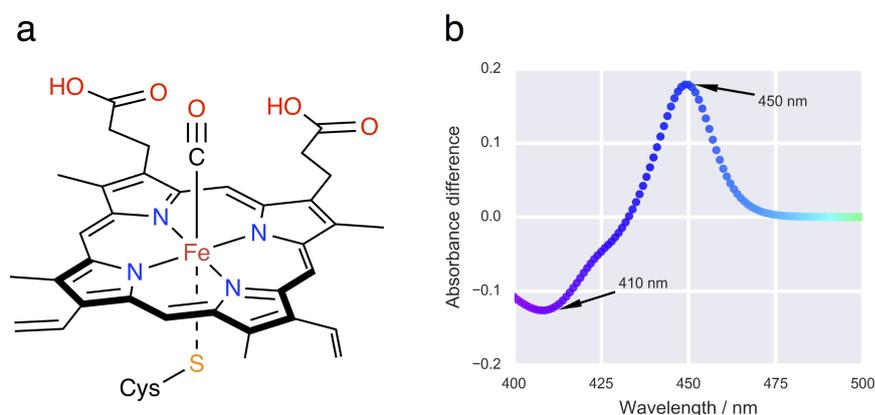
Figure 5.1: (a) The structure of iron(II) protoporphyrin IX coordinated with the axial cysteinate and carbonyl ligands. (b) Typical CO difference spectrum (*courtesy of A. Bokel*) observed for P450 enzymes shows peaks at 410 and 450 nm.

CYPs are responsible for the metabolism of xenobiotics (e.g., increasing their polarity by oxidation) and for the biosynthesis of signaling molecules (e.g., hormones).[80] Humans have around 60 different CYPs that are involved in the drug metabolism and detoxification. While mammalian and plant CYPs are attached to membranes,[82, 83] their microbial counterparts are soluble proteins.[84] In addition, many bacterial CYPs were shown to be promiscuous regarding the substrate,[85, 86] and combined with their high reactivity they became an attractive group of enzymes for many studies, including enzyme design for pharmaceutical and biotechnological purposes.

31

One of the most common reactions that CYPs catalyze is the monooxygenation (e.g., allylic hydroxylation). The cytochrome P450 uses molecular oxygen and the heme to catalyze the incorporation of one oxygen atom into a substrate molecule, while the other oxygen is reduced to water. The general catalytic cycle is shown in Figure 5.2.[87, 88] When the substrate binds to the resting state of the enzyme, heme goes through several preparation steps to become what is thought to be the catalytically competent state denoted compound I (cI).[89] The cI attacks the substrate, abstracting its hydrogen atom. The substrate radical further attacks the oxygen of the cI, which gives an impression that the oxygen atom got nested in the C–H bond.
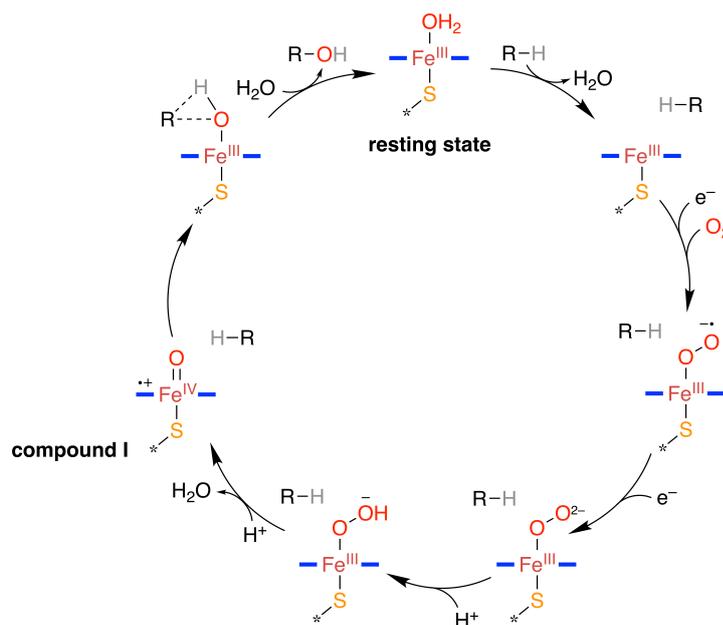
Figure 5.2: The catalytic cycle of P450 enzymes where the compound I is thought to be the catalytically competent species.

Among the P450 monooxygenases, BM3 (or CYP102A1) from *Bacillus megaterium* is one of the most commonly studied members that performs the sub-terminal fatty acid hydroxylation.[90, 91] Le-Huu et al. engineered P450 BM3 mutants that can selectively perform oxidation of the monocyclic diterpenoid β-cembrenediol (**1** in Figure 5.3).[92] In that study, Leu75, Val78, Phe87, and Ile263 were identified as positions for mutations based on the crystal structure examination and previous experimental knowledge (e.g., bulky Phe87 is known to extend close to the heme, thereby preventing the binding of large substrates).[86, 90]

5.1 Paper III: Product stereochemistry elucidation

Although Le-Huu et al. identified the main products of the P450 BM3 monooxygenation of **1**, the absolute stereochemistry of the hydroxylated products could not be determined based on the NMR data alone.[92] In certain cases when the assignment of all NMR signals to the specific nuclei is challenging, simulating a spectrum at the QM level can be particularly useful.

Recent improvements in the computer hardware, coupled with the advances in the molecular modeling methods, especially the development of the gauge-independent atomic orbital (GIAO) approach,[93] facilitate a reliable and affordable computational prediction of nuclear magnetic properties. The simulated NMR spectrum of a molecule gives an exact correspondence between each nucleus (typically H and C) and the spectral signals, which simplifies structure elucidation and stereochemical assignments.[94] Further comparison of an experimental NMR spectrum with simulated spectra of possible diastereomers allows determining the most likely stereochemical structure.[95]

5.1.1   Simulated ^{1}H and ^{13}C NMR spectra

Knowing the proper product stereochemistry is crucial, as the further P450 BM3 design towards a higher chemo- and regioselectivity for the oxidation of **1** should be based on the substrate binding preferences and previous knowledge of the reactivity. The stereochemical characterization was performed by analyzing both experimental and simulated NMR spectra. In addition to (7S,8S)-epoxy-β-cembrenediol (**2**), common products included 9-β-cembrenetriols (**3a–b**), 10-β-cembrenetriols (**4a–b**), β-cembrenediol-10-one (**4c**), and 9,10-β-cembrenetetraols (**5a–b**), as shown in Figure 5.3.[92,96] All possible hydroxylated isomers of **3**, **4**, and **5** were simulated to identify the most probable structures.

3a-b: R$_1$ = OH *(R/S)*, R$_2$ = H
4a-b: R$_1$ = H, R$_2$ = OH *(R/S)*
4c: R$_1$ = H, R$_2$ = O
5a-b: R$_1$ = OH *(S)*, R$_2$ = OH *(S/R)*
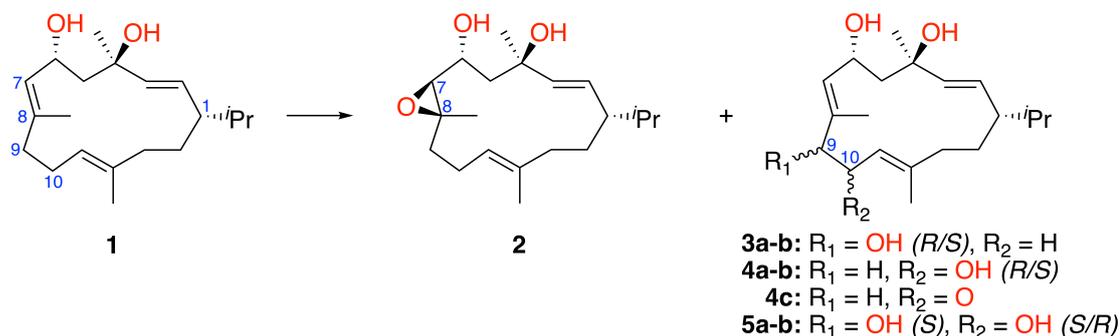
Figure 5.3: β-Cembrenediol (**1**) is a suitable substrate for P450 BM3 mutants (e.g., V78A/F87G, V78A/F87A, F87A/I263L), producing epoxy-, hydroxy-, and enone-derivatives (**2–5**) in one or two reaction steps.

Compounds **3a–b** are two epimers, 9R- and 9S-β-cembrenetriol, which had to be assigned to two experimentally identified products. In such cases, one can use the Bayesian CP3 method to compare experiments with simulations.[97] This approach was shown to perform significantly better than classical statistical methods. Consideration of both ^{1}H and ^{13}C chemical shifts established, with a probability of 92.5% , that compounds **3a** and **3b** correspond to the 9R- and 9S-epimer, respectively. The $^3J_{\mathrm{H,H}}$ coupling constants in chloroform could not be measured experimentally.

A similar problem occurred for compounds **4a–b**, with possible 10R- and 10S-epimers. The spectra of compounds **4a** and **4b** were shown to be the 10R- and 10S-epimer, respectively,

with 99.8% confidence. The coupling constants ($^3J_{9a,10}$, $^3J_{9b,10}$, and $^3J_{10,11}$) corroborate the conclusions from the chemical shift analysis.

A more complex situation was found for compounds **5a–b**, which involve two new stereocenters leading to four possible combinations: ($9R$,$10R$), ($9R$,$10S$), ($9S$,$10R$), and ($9S$,$10S$). The $^3J_{9,10}$ constant examination identified that isolated compounds **5a** and **5b** have very different values. The simulations showed that **5a** has *trans* geometry (i.e., ($9R$,$10R$) or ($9S$,$10S$)), which corresponds to the higher coupling constant of 7.6 Hz. On the other side, compound **5b** has the lower $^3J_{9,10}$ value of 3.3 Hz, corresponding to the *cis* geometry (i.e., ($9R$,$10S$) or ($9S$,$10R$)). Similar to CP3, a DP4 Bayesian method was developed for such cases where many sets of computed shifts need to be compared to one experimental set.[98] DP4 identified, with 99.9% probability, that **5a** is the ($9S$,$10S$) diastereomer. The analysis of compound **5b** suggested the ($9S$,$10R$) configuration, although with a somewhat lower probability of 76.3%. Having in mind that both **5a–b** are products of **3a** with $9R$ configuration, it can be expected that both products retain the same stereochemistry at the C9-atom.[†] This expectation supports the conclusion that ($9S$,$10R$) is the most likely configuration for **5b**.

5.2 Paper IV: Engineering chemo- and regioselectivity

The two main approaches in enzyme engineering are (1) directed evolution and (2) rational design.[99] While the former approach relies on randomness,[100–102] the latter requires detailed knowledge about the structure, dynamics, and mechanism of an enzyme for suggesting appropriate structural modifications.[103, 104] Therefore, identifying probable mutations that can lead to enhanced properties is not a trivial task. Finding the proper binding mode is a necessary initial step of a rational design protocol. Molecular docking is typically used to position a substrate in the active site of an enzyme. However, macrocycle docking is much harder to perform than the docking of typical acyclic substrates.[105] This chapter presents a workflow for computational P450 design.

β-Cembrenediol (**1**) is a challenging substrate that has several functional groups (including two hydroxy groups) and many possible hydroxylation sites (i.e., seven allylic and six non-allylic hydroxylation, and three epoxidation positions). The wild-type enzyme converts less than 2% of the substrate, but mutating the bulky Phe87 to Ala (Figure 5.4) increases the conversion rate threefold.[92] The additional V78A mutation further increases the conversion rate eight times, yet without high selectivity, giving a mixture of products **2**, **3a–b**, and **4a–b**.[92] Thus, this double mutant has promising properties and leaves space for further improvement. Molecular docking of **1** revealed many substantially different binding poses with similar energies, indicating that there is more than one preferential binding pose.[106]

[†]The notation at C9 changes due to the change in the Cahn-Ingold-Prelog priorities when C10 gets hydroxylated. However, the absolute orientation of the hydroxy group on this atom stays the same!
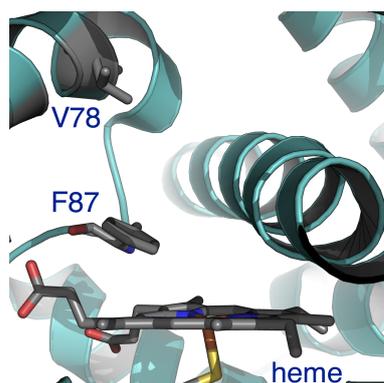
Figure 5.4: Structure of the active site of the wild-type P450 BM3 (PDB ID: 1JPZ). The bulky residues Val78 and Phe87, in the proximity to the heme prosthetic group, need to be mutated to accommodate the macrocyclic β-cembrenediol substrate.

5.2.1 β-Cembrenediol binding

Under the assumption that the product distribution is a direct outcome of how substrate binds to a CYP,[107–110] all binding poses of **1** in the active site of the V78A/F87A P450 BM3 mutant had to be identified. To this end, HREX-MD simulations were performed to ensure thorough sampling. For the analysis of the resulting simulations, several regions of the substrate were selected based on the experimentally observed products (i.e., atoms C7, C8, C9, and C10). The probability distributions were estimated for distances between the reaction centers on the substrate and the cI (represented by the axial O-atom) to identify the stable binding modes. Preferences in binding to one over the other reaction center of the substrate were investigated using bivariate probability distributions.

To describe the chemoselectivity of monooxygenation, for example, the probability distribution was studied for the distances between the heme oxygen and the centroids of the C7–C8 (epoxidation) and C9–C10 bonds (hydroxylation) of the substrate (Figure 5.5a). Four maxima could be identified on this surface. Two maxima, with the approximate coordinates of (4 Å, 4 Å) and (6 Å, 4 Å), were positioned sufficiently close for the chemical reaction to happen, hence the title—*productive* modes. At another two maxima, located at (5 Å, 7 Å) and (8 Å, 9 Å), the reactive centers of **1** are positioned very far away from the cI, preventing the reaction to occur in these binding modes. Therefore, such modes are denoted as *unproductive*.

Examining the productive modes in more detail, it is apparent that the maximum corresponding to C9–C10 bond centroid is broader and taller that of the C7–C8 bond. This fact indicates that the substrate spends more time in an orientation optimal for the hydroxylation at C9 or C10, and it can explain why around 70% of the products are hydroxylated and only 30% bear the epoxy group.[106] It is necessary to note here that such analysis is qualitative only, and further work would be needed to quantify such experimental observables (e.g., simulating the complete binding process and the reaction mechanisms, and identifying energies of all the reactant and product states and the barriers between them).
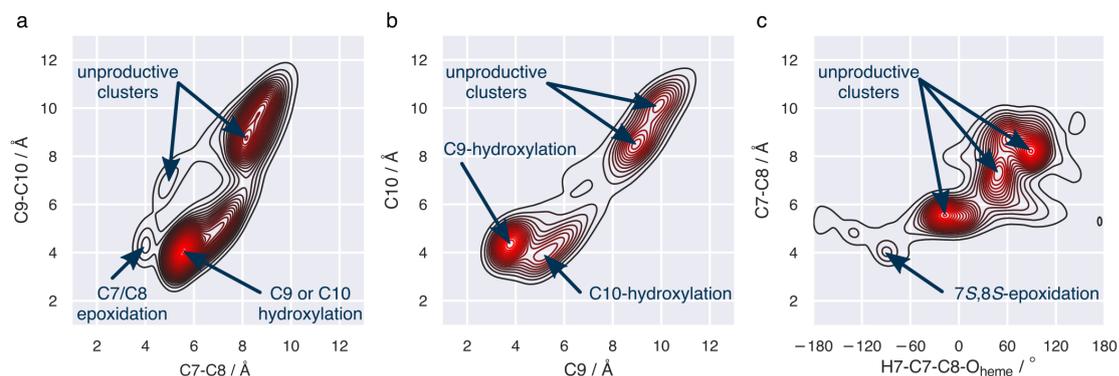
Figure 5.5: Bivariate probability distributions for β-cembrenediol binding to the V78A/F87A P450 BM3 mutant. Discriminating (a) 7,8-epoxidation *vs.* 9/10-hydroxylation, (b) 9- *vs.* 10-regioselective hydroxylation, and (c) stereoselectivity of 7,8-epoxidation. All distances are measured from the axial oxygen of the cI. *Reprinted with permission from ref. 106. Copyright 2018 American Chemical Society.*

Applying the bivariate probability distributions analysis, one can examine various other kinds of selectivity. For describing the regioselectivity of the hydroxylation reaction, the probability distributions was investigated using distances to the C9 and C10 atoms (Figure 5.5b). Again, two unproductive and two productive clusters were found, with the productive maxima located at (4 Å, 4 Å) and (5 Å, 4 Å). The tallest maximum represents binding geometrically suitable for 9-hydroxylation, explaining the 7:2 ratio of the 9- *vs.* 10-hydroxylated products.

Experimental findings suggest that the only epoxidation product has the 7S,8S configuration. For investigating the probability distributions for the stereoselectivity of epoxidation, the distance between the heme and the C7–C8 bond was one parameter, while the dihedral angle between the H7–C7–C8–O_{heme} was the other one (Figure 5.5c). The examination of this surface indicates that the cI attack could happen from either side of the double bond. However, the substrate is sufficiently close only when the orientation is suitable for the 7S,8S-epoxidation (maximum located at –90 deg, 4 Å).

5.2.2 Hotspot identification

For a particular type of selectivity for the transformation of **1** to be engineered into P450 BM3, two routes can be followed: to stabilize the productive substrate conformations and/or to destabilize the unproductive ones. To this end, MM with Poisson–Boltzmann surface area calculations were performed and the resulting energies decomposed on a per residue basis to quantify the interaction energies between the substrate, in different binding modes, and the protein. The interaction energies, in combination with the analysis of the hydrogen bonds between **1** and the enzyme, enabled the identification of the key residues for substrate binding and the selection of the mutagenesis hotspots.

Several mutations were proposed for these hotspots, based on their size, shape, and electrostatic complementarity to the binding mode that was to be stabilized (or destabilized). For example, Leu75 was identified to be a severely destabilizing residue for the 10-hydroxylation due to the shape incompatibility (Figure 5.6). Mutating this bulky residue to an alanine led to a fivefold regioselectivity increase for C10 oxidation, compared to the V78A/F87A P450 BM3 mutant. A small number of additional mutations was proposed based on this analysis and tested experimentally by Urlacher and co-workers. The summary of the product distribution of the selected P450 BM3 mutants is given in Figure 5.7.[106]
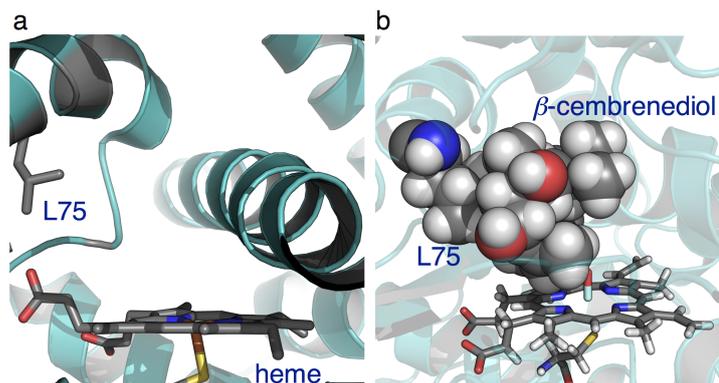
Figure 5.6: (a) The bulky Leu75 residue side chain is located close to the active site of P450 BM3. (b) Leu75 clashes with β-cembrenediol in the binding mode productive for the 10-hydroxylation.
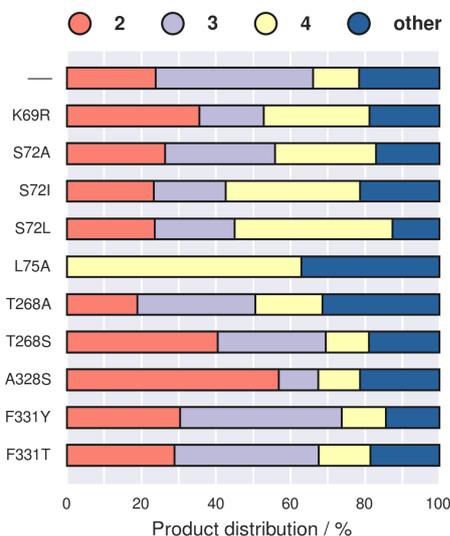
Figure 5.7: Distribution of the oxidized products of **1** by the V78A/F87A P450 BM3 mutant and the triple mutants designed in the present work. Product **3** includes both **3a** and **3b**, while product **4** combines **4a**, **4b**, and **4c**. *Reprinted with permission from ref. 106. Copyright 2018 American Chemical Society.*

6 Abiotic surface binders

Our contemporary world is dependent on synthetic polymers or plastics. A leader in the "plastic world" is poly(ethylene terephthalate) (PET) (Figure 6.1), whose annual production reaches 56 million tons.[111] The high chemical inertness of PET is a desirable property for many applications in the food packaging, textile, and medical industries.[112,113] This, however, presents a challenge to modifying or degrading PET.
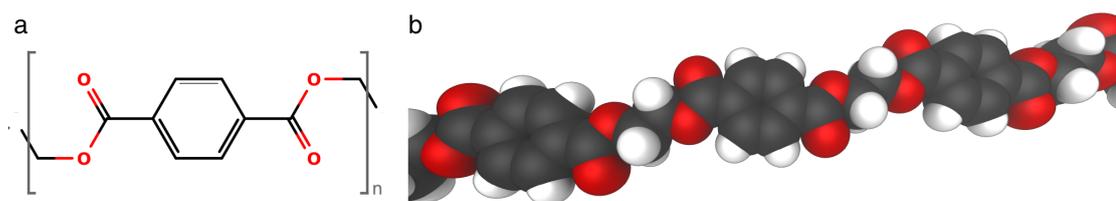
Figure 6.1: Chemical structure of poly(ethylene terephthalate): (a) monomer and (b) linear chain.

While PET can be mechanically recycled several times,[114] the complete chemical depolymerization (i.e., hydrolysis) requires harsh conditions.[115] Enzymatic PET degradation is currently gaining importance as an alternative to chemical treatment. Two PETase enzymes were found in a recently engineered bacterium.[116] The high PET hydrophobicity, however, presents a major challenge due to its low affinity for soluble enzymes.[117,118] A similar problem appears in the hydrolysis of natural polymers. In this case, esterases are often fused with short peptides that exhibit high affinity toward the polymer and can, thus, act as anchors.[119]

6.1   CBM–PET interface

Carbohydrate-binding modules (CBMs) are the noncatalytic domains of cellulases, hemicellulases or chitinases, that is, enzymes that degrade natural polysaccharides.[120,121] As their name suggests, CBMs enable enzymes to attach to the polysaccharide surface and stabilize the complex long enough to perform the hydrolysis locally.[122] CBMs can be classified into several types, folds, and families, where type A is characterized by a planar, surface-exposed aromatic

patch. The CBMs are clustered to families based on their amino acid sequence similarity, and the families 1, 2a, 3, 5, and 10 belong to the type A covering four different folds.[123]

The identification of suitable PET binders is necessary for establishing innovative PET applications. Certain CBMs were previously shown to be promiscuous with regard to the substrate they bind. The exposed aromatic residues of type A CBMs were postulated to be important for the CBM–PET interplay.[124] Further investigations were conducted to explore the nature of these interactions.[125]

6.1.1   Promiscuous CBMs

Several CBM sequences (type A, families 1, 2, 5, and 10), which are characterized by an exposed aromatic triad, were selected for further studies. Homology models were generated to obtain protein models in the absence of crystal or NMR structures. The stabilities of the resulting peptide folds, and especially of their aromatic triad, were further investigated with MD simulations in bulk water (Figure 6.2). The simulations revealed that the bigger peptides (i.e., families 2 and 5 containing 50–100 amino acid residues) are stable in water and maintain their folds (e.g., the β-sandwich fold of BaCBM2 and BsCBM2). These peptides show root-mean-square deviation (RMSD) values of 1–3 Å from the homology model, and stable values for the radius of gyration. Shorter peptides (i.e., families 1 and 10, having around 30 residues) are very flexible and typically do not maintain a specific fold, as indicated by the fluctuating radius of gyration values. The MD models are often very different from the homology models, as their RMSD values are often higher than 4 Å. This flexibility is not unexpected since short CBMs are parts of bigger enzymes, where their unique folds can be stabilized by intramolecular interactions.
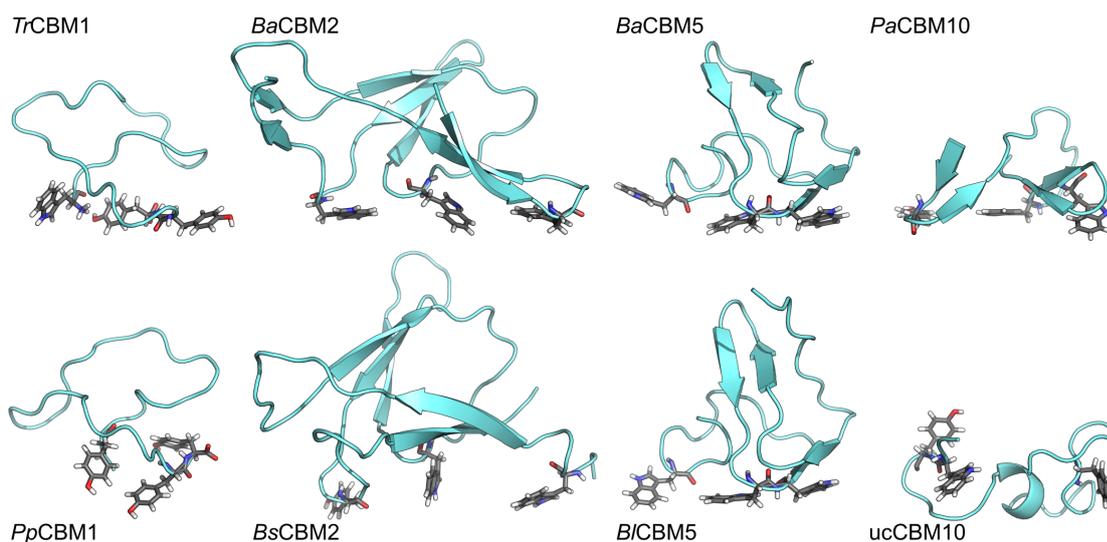
Figure 6.2: Representative structures of investigated CBMs from MD simulations. The aromatic triads anticipated from the homology modeling is highlighted in the sticks.

6.1.2 Developing the ordered PET model

As protein properties can change at the interface with abiotic materials,[126, 127] the CBM behavior at the PET surface needed to be studied. Several MD simulations of PET are reported in the literature.[128–131] They, however, considered only PET and did not include any biomolecules.

The initial simulations in the present work were based on an amorphous PET model, which gave results that were not in agreement with the experimental observations. Namely, the high negative charge of the many PET termini in such a model suggested that *Ba*CBM2 binds PET mainly via several positively charged residues. The mutagenesis studies, however, confirmed that the *Ba*CBM2 mutant K29A/K32A/K48A/K52A binds PET with a similar strength as the wild-type peptide (Figure 6.3).
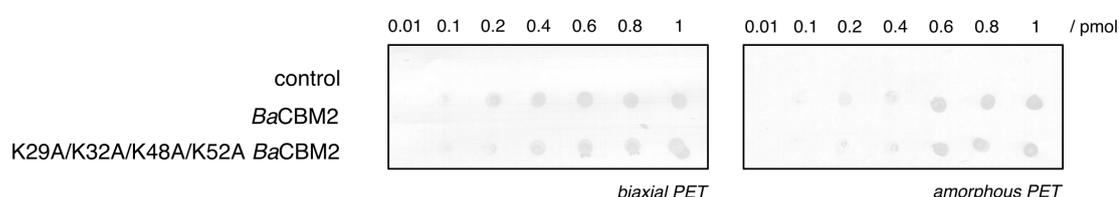
Figure 6.3: PET surface affinity assay indicates that the wild-type *Ba*CBM2 and K29A/K32A/K48A/K52A mutant bind PET with similar strengths. (*Courtesy of J. Weber*)

Therefore, the ordered (*"crystalline"*) PET model with a zero net charge was expected to give more reasonable results. The model was developed by applying the symmetry transformations on the crystal structure of a monomer.[132] One PET chain was composed of five monomers, and eight such chains were arranged to form a layer big enough to simulate the binding of CBMs. Five identical layers were stacked to form the surface, as depicted in Figure 6.4a. The PET surface was aligned to the x- and y-axes, where the first and last monomer of each chain were connected to each other by imposing periodic boundary conditions (PBC). The surface was immersed in a water box, and a CBM was freely positioned in the water phase, around 10 Å above the surface (along the z-axis), as shown in Figure 6.4b.

6.1.3 Structure and energetics at the interface

Three CBMs were selected for studying their binding to the PET interface: *Tr*CBM1, *Ba*CBM2, and *Ba*CBM5. The experimental investigation suggested that *Tr*CBM1 and *Ba*CBM5 do not bind PET, while *Ba*CBM2 possesses superior binding properties.

The *Ba*CBM5 simulations confirmed that the aromatic triad Trp27/Trp28/Trp40, as predicted by the homology modeling, stacks with the PET surface. The peptide structure is rigid (RMSD < 2 Å), with many hydrophobic residues located at the interface. Although PET is highly hydrophobic, it also has local polar regions (i.e., the ester groups). Therefore, the *Ba*CBM5 peptide is not compatible with the amphipathic nature of PET (Figure 6.5a), despite the aromatic triad. The *Ba*CBM5–PET energy remains constant over the simulation (Figure 6.6a),
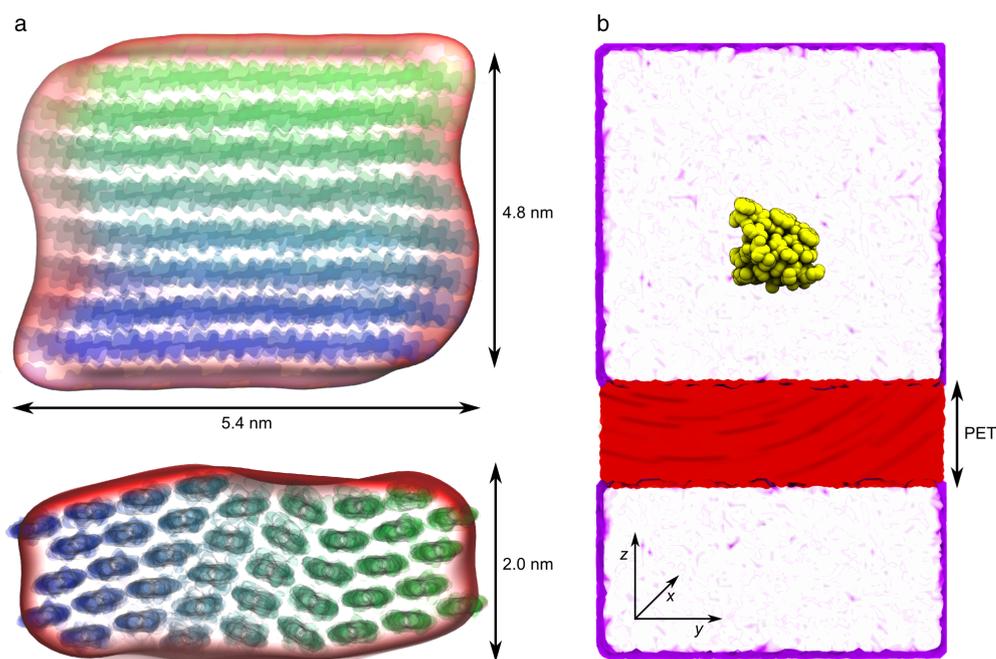
Figure 6.4: Ordered PET system developed to study peptide–PET interactions by MD simulations. (a) The size of the PET surface and (b) a typical simulation box.

indicating that although the peptide approaches the surface, it does not get stabilized there, leading to the low PET affinity identified experimentally.

The best binder, *Ba*CBM2, readily approaches PET, upon which the energy significantly drops (Figure 6.6b). The peptide fold is stable, but a certain level of internal dynamics is involved in the binding. As predicted by homology modeling, the aromatic triad consists of Trp10/Trp45/Tyr64, and unlike *Ba*CBM5 it is surrounded by a network of hydrophilic residues (Figure 6.5b). These polar interactions are significant for the binding, as indicated by the big energy difference compared to *Ba*CBM5 binding.

Finally, the *Tr*CBM1 peptide, identified as flexible in the bulk solvent, shows interesting properties at the PET interface. The original aromatic triad anticipated from the homology
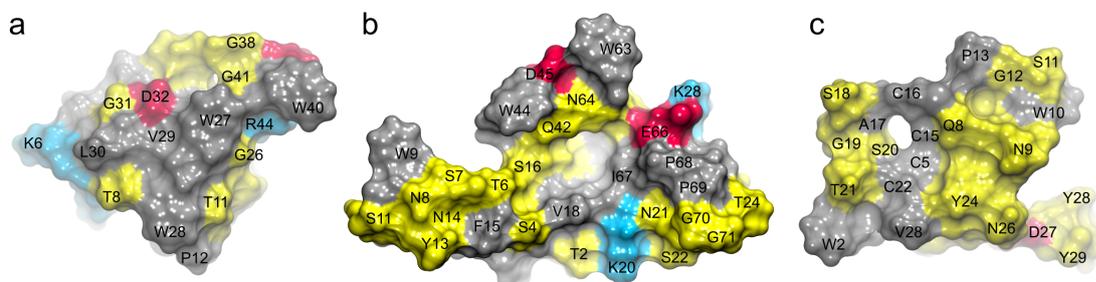
Figure 6.5: The CBM–PET interface of (a) *Ba*CBM5, (b) *Ba*CBM2, and (c) *Tr*CBM1.

model, Trp2/Tyr28/Tyr29, is not stable in MD simulations. Instead, a new triad is formed and consists of Trp2/Trp10/Tyr24. However, the contact surface of *Tr*CBM1 is much smaller than that of *Ba*CBM2. Also, the surface is very hydrophilic (Figure 6.5c) and not completely compatible with the amphipathic PET nature. The resulting peptide–PET energy is higher than that for *Ba*CBM5 but lower than that of *Ba*CBM2 (Figure 6.6c). Furthermore, Figure 6.6 demonstrates that the energy funnel is much steeper for the best binder than for *Tr*CBM1 or *Ba*CBM5.
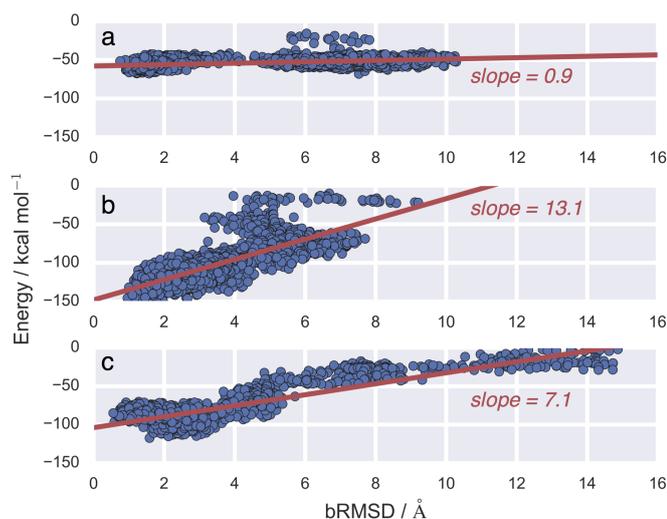
Figure 6.6: The CBM–PET binding funnels of (a) *Ba*CBM5, (b) *Ba*CBM2, and (c) *Tr*CBM1.

6.1.4   Folding at the interface

As previously noted, *Tr*CBM1 is a short peptide (30 residues), which is very flexible in water. Although CBM classification suggests that family 1 folds to a cystine knot,[123] the MD simulations in bulk water were unable to provide a well-defined peptide structure, which is evident from the RMSD and radius of gyration plots (Figures 6.7a). The RMSD reaches up to 8 Å from the homology model, and both the RMSD and radius of gyration oscillate significantly. No permanent intramolecular stabilizing interactions are observed in the peptide, and the structure is coiled except for the formation of a transient helix observed between 40–50 ns.

Moving from bulk water to the PET surface, however, induces a conformational change in *Tr*CBM1. As the originally proposed aromatic triad (Trp2/Tyr28/Tyr29) does not firmly interact with the surface, the peptide wiggles until the new triad (Trp2/Trp10/Tyr24) forms π-stacking interactions with PET. The established interactions force the peptide to fold to a stable β-sheet. The stability of this model is indicated by the constant RMSD of ~5 Å, and the radius of gyration of ~10.5 Å (Figure 6.7b).
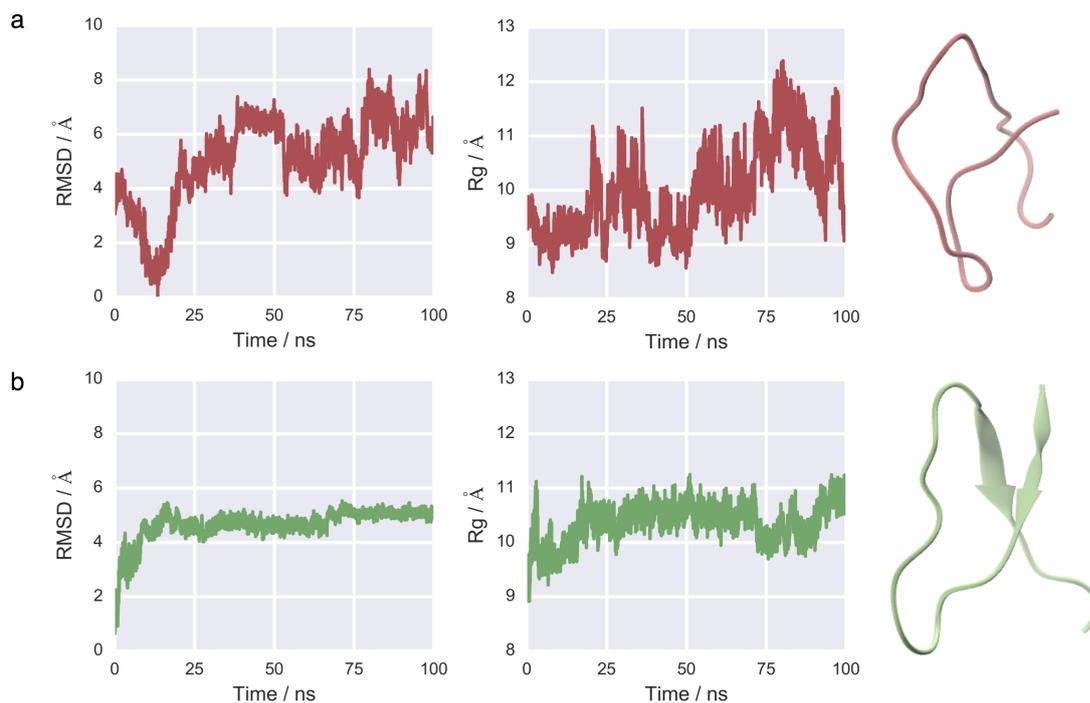
Figure 6.7: *Tr*CBM1 RMSD, radii of gyration, and representative structures in (a) bulk water and (b) on PET surface.

6.2   Further studies

Additional studies of the PET binders are being conducted in several directions. Based on the proposed binding model,[125] several *Ba*CBM2 mutations were selected to increase the PET affinity of this peptide, identified as the best binder in both simulations and experiments. The criteria for the mutations are twofold: extending the aromatic triad and optimizing the hydrophilic network at the interface.

The second direction dives deeper into the theoretical aspects of the peptide–PET binding. Since the experiments pointed out that Tyr might bind stronger to PET than Trp,[124] the QM level investigation is directed toward calculating the interaction energies between the aromatic amino acid residues and PET. Besides, enhanced sampling MD simulations are being performed to pull the peptides from the surface and calculate the free energies of binding, which should give better insight than potential energies reported in Figure 6.6.

The identification of novel peptides with potential PET-binding properties is of high industrial importance. Based on the theoretical knowledge gained in the previous projects, Monte Carlo simulations are being performed to find the optimal orientations of multiple aromatic residues on the PET surface. The protein data bank (PDB) search for the identified aromatic motifs might determine novel protein classes with a potentially high PET affinity.

7 Conclusions

Enzymes are biomacromolecules able to catalyze chemical reactions. For their mild reaction conditions regarding temperature, pressure, solvents, and pH range optimum, enzymes became popular catalysts in many industries. Enzyme kinetics and structural studies using, for example, NMR spectroscopy or X-ray crystallography, were able to propose mechanisms on how particular enzymes work. However, in some cases, it is difficult to make mechanistic assumptions due to the lack of the enzyme structure or dynamic effects that can be related to catalysis. Furthermore, for some enzymes, different reaction mechanisms could be imagined, and sometimes even very advanced experimental methods cannot distinguish between the correct and wrong reaction pathways.

Computational enzymology is an emerging field that provides complementary information to the structural and mechanistic biocatalysis. Molecular simulations can describe enzyme conformational ensembles, together with the underlying dynamics, enhancing the static image typically observed in X-ray crystallography. Moreover, simulations can describe substrate binding and model different reaction mechanisms to identify the most appropriate one. This thesis presents three topics related to the field of computational enzymology and discusses both novel findings on enzyme function and methodological advances in the field.

The active site preorganization is essential for many enzymes as it reduces the energy demanding reorientation of the protein elements. Current MD simulations confirmed indications that the active site His of GOx may be flexible. Furthermore, the enhanced sampling simulations were shown to be of great value to *in silico* enzymology, as HREX-MD and US-MD were able to quantify how much time does this His residue spend in each of the catalytic and noncatalytic states. The lower computational cost of HREX-MD over US-MD enabled to study the flexibility of the His residue over the laboratory evolution trajectory that increased the GOx catalytic activity. The main reason for the higher activity of the **A2** GOx pentamutant was the increased active site preorganization, caused by the lower flexibility of the His residue and its apparent lock in the catalytically active conformation.

Since MD simulations are based on empirically derived force fields, the outcome of a simulation can be force-field dependent. Therefore, the choice of the AMBER 99SB-ILDN force field for the GOx study had to be justified. To that end, the benchmark of twelve commonly used flavors of the AMBER, CHARMM, OPLS, and GROMOS force fields was performed for the side chain properties of amino acids. The benchmark was based on two small proteins (i.e., ubiquitin and GB3), for which the extensive experimental work was performed to quantify the side chain probabilities. The AMBER 99SB*-ILDN force field was one of the best performers, closely followed by the AMBER 99SB-ILDN, used in the GOx studies. The close resemblance of the two force fields should not be a surprise, as the star correction is applied to the backbone dihedral angles only. The only better-performing force field is the newly developed AMBER 14SB, which included a significant reparameterization for both the backbone and the side chain parameters.

Another important aspect of enzyme catalysis is selectivity/promiscuity. Enzyme promiscuity is crucial in evolution, as it can lead to peculiar enzymes accepting different substrates or catalyzing novel chemical reactions. Selectivity is, on the other hand, typically demanded under industrial settings, e.g., for organic synthesis, where a predictable outcome is desired. Many CYPs are promiscuous regarding the substrate they can accept and the position where the substrate will be oxidized. Some P450 BM3 mutants were shown not to be regioselective for the oxidation of the 14-membered macrocycle β-cembrenediol, typically giving a mixture of products. The absolute stereochemistries of the products were determined in combination of NMR and QM methods, which enabled to work on the further computational design toward the more selective P450 BM3 variants.

Substrate binding is a major step in the catalytic cycle of an enzyme, and the binding is often found to be the most important step for determining the regioselectivity of CYPs. This effect can be exploited to design mutants that would perform a selective oxidation. In the current work, the shift of regioselectivity of the β-cembrenediol oxidation was engineered into the V78A/F87A P450 BM3 variant. The enhanced sampling HREX-MD simulations were used to explore the conformational ensemble of the substrate in the active site of the enzyme, as the initial results obtained with docking and unbiased MD simulations were not consistent with the experimental observables. Based on the shape and chemical complementarity, point mutations were proposed to shift the regioselectivity of oxidation, and the quality of the model was confirmed experimentally.

Although most enzymes bind their substrates in the buried active sites, polysaccharide esterases developed carbohydrate binding modules to attach to the substrate (i.e., the polymer surface), to be able to cleave the ester linkages. Such modules could be fused with PET-degrading enzymes to assist with the binding and subsequent depolymerization of this relevant material. The combination of experimental screening and MD simulations enabled to identify promiscuous CBMs that can bind to PET. In addition, the simulations were able to determine the most important types of interactions leading to the desired promiscuity of these peptides. However, further QM and US-MD simulations are needed to provide a more

reliable description of the interaction, which could be confirmed by experimental testing of the proposed mutants.

The studies presented in the current thesis on ***Computational enzyme evolution and design*** illustrate the potential of MD simulations for enzyme evolution studies, and especially for enzyme design purposes. Although unbiased MD simulations can often identify the relevant regions on an energy surface, they are not potent enough to reliably quantify the probabilities between the different states. In such cases, HREX-MD and US-MD simulations provide enhanced sampling and thoroughly explore the conformational ensembles of enzymes or enzyme–substrate complexes. Finally, to simulate the underlying reaction mechanisms, QM/MM and empirical valence bond calculations should be used.

Bibliography

[1] E. Buchner: Alkoholische Gährung ohne Hefezellen. *Ber. Dtsch. Chem.* **1897**, *30*: 117–124.

[2] E. Buchner: Ueber zellenfreie Gährung. *Ber. Dtsch. Chem.* **1898**, *31*: 568–574.

[3] Nobelprize.org: The Nobel Prize in Chemistry 1907. www.nobelprize.org/nobel_prizes/chemistry/laureates/1907/ (Retrieved 08.06.2017).

[4] J. Sumner: The isolation and crystallization of the enzyme urease: Preliminary paper. *J. Biol. Chem.* **1926**, *69*: 435–441.

[5] J. Northrop: Crystalline pepsin. *Science* **1929**, *69*: 580.

[6] Nobelprize.org: The Nobel Prize in Chemistry 1946. www.nobelprize.org/nobel_prizes/chemistry/laureates/1946/ (Retrieved 08.06.2017).

[7] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, V. Shore: Structure of myoglobin: A three-dimensional fourier synthesis at 2 Å resolution. *Nature* **1960**, *185*: 422–427.

[8] M. Perutz, M. Rossmann, A. Cullis, H. Muirhead, G. Will, A. North: Structure of hæmoglobin: A three-dimensional fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **1960**, *185*: 416–422.

[9] C. Blake, D. Koenig, G. Mair, A. North, D. Phillips, V. Sarma: Structure of hen egg-white lysozyme: A three-dimensional fourier synthesis at 2 Å resolution. *Nature* **1965**, *206*: 757–761.

[10] L. Johnson, G. Petsko: David Phillips and the origin of structural enzymology. *Trends Biochem. Sci.* **1999**, *24*: 287–289.

[11] A. Warshel, M. Levitt: Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*: 227–249.

[12] A. Warshel: Energetics of enzyme catalysis. *Proc. Nat. Acad. Sci. USA* **1978**, *75*: 5250–5254.

[13] J. Knowles: Enzyme catalysis: not different, just better. *Nature* **1991**, *350*: 121–124.

[14] A. Warshel: Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **1998**, *273*: 27035–27038.

[15] C.-J. Tsai, A. del Sol, R. Nussinov: Allostery: Absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.* **2008**, *378*: 1–11.

[16] J. Richard, T. Amyes, B. Goryanova, X. Zhai: Enzyme architecture: on the importance of being in a protein cage. *Curr. Opin. Chem. Biol.* **2014**, *21*: 1–10.

[17] G. P. Lisi, J. P. Loria: Solution NMR spectroscopy for the study of enzyme allostery. *Chem. Rev.* **2016**, *116*: 6323–6369.

[18] Y. S. Kulkarni, Q. Liao, D. Petrović, D. M. Krüger, B. Strodel, T. L. Amyes, J. P. Richard, S. C. L. Kamerlin: Enzyme architecture: Modeling the operation of a hydrophobic clamp in catalysis by triosephosphate isomerase. *J. Am. Chem. Soc.* **2017**, *139*: 10514–10525.

[19] W. Heitler, F. London: Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Z. Phys.* **1927**, *44*: 455–472.

[20] D. R. Hartree: The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. *Proc. Cambr. Phil. Soc.* **1928**, *24*: 89–110.

[21] V. Fock: Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.* **1930**, *61*: 126–148.

[22] P. Hohenberg, W. Kohn: Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*: B864–B871.

[23] W. Kohn, L. J. Sham: Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*: A1133–A1138.

[24] Nobelprize.org: The Nobel Prize in Chemistry 1998. www.nobelprize.org/nobel_prizes/chemistry/laureates/1998/ (Retrieved 08.06.2017).

[25] H. M. Senn, W. Thiel: QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48*: 1198–1229.

[26] Nobelprize.org: The Nobel Prize in Chemistry 2013. www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/ (Retrieved 08.06.2017).

[27] S. Osuna, G. Jiménez-Osés, E. L. Noey, K. N. Houk: Molecular dynamics explorations of active site structure in designed and evolved enzymes. *Acc. Chem. Res.* **2015**, *48*: 1080–1089.

[28] R. C. Bernardi, M. C. Melo, K. Schulten: Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*: 872–877.

[29] L. Verlet: Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **1967**, *159*: 98–103.

[30] H. Berendsen, J. Postma, W. van Gunsteren, A. DiNola, J. Haak: Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*: 3684–3690.

[31] R. Hockney, J. Eestwood: Computer simulations using particles. New York, NY: Taylor & Francis Group, **1988**.

[32] G. R. Bowman: Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *J. Comput. Chem.* **2016**, *37*: 558–566.

[33] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski: Coarse-grained protein models and their applications. *Chem. Rev.* **2016**, *116*: 7898–7936.

[34] G. Torrie, J. Valleau: Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*: 187–199.

[35] A. Laio, M. Parrinello: Escaping free-energy minima. *Proc. Nat. Acad. Sci. USA* **2002**, *99*: 12562–12566.

[36] J.-Y. Yi, J. Bernholc, P. Salamon: Simulated annealing strategies for molecular dynamics. *Comput. Phys. Commun.* **1991**, *66*: 177–180.

[37] Y. Sugita, Y. Okamoto: Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*: 141–151.

[38] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman: The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*: 1011–1021.

[39] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*: 1087–1092.

[40] W. Hastings: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*: 97–109.

[41] A. Patriksson, D. van der Spoel: A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **2008**, *10*: 2073–2077.

[42] K. A. Beauchamp, Y.-S. Lin, R. Das, V. S. Pande: Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.* **2012**, *8*: 1409–1414.

[43] L. Wang, R. A. Friesner, B. J. Berne: Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **2011**, *115*: 9431–9438.

[44] G. Bussi: Hamiltonian replica exchange in GROMACS: a flexible implementation. *Mol. Phys.* **2014**, *112*: 379–384.

[45] S. W. May, S. R. Padgette: Oxidoreductase enzymes in biotechnology: Current status and future potential. *Nat. Biotechnol.* **1983**, *1*: 677–686.

[46] T. Yakushi, K. Matsushita: Alcohol dehydrogenase of acetic acid bacteria: Structure, mode of action, and applications in biotechnology. *Appl. Microbiol. Biotechnol.* **2010**, *86*: 1257–1265.

[47] O. de Smidt, J. C. du Preez, J. Albertyn: The alcohol dehydrogenases of *Saccharomyces cerevisiae*: A comprehensive review. *FEMS Yeast Res.* **2008**, *8*: 967–978.

[48] J. Strommer: The plant ADH gene family. *Plant J.* **2011**, *66*: 128–142.

[49] H. Jörnvall, J. O. Höög, H. von Bahr-Lindströ, B. L. Vallee: Mammalian alcohol dehydrogenases of separate classes: Intermediates between different enzymes and intraclass isozymes. *Proc. Nat. Acad. Sci. USA* **1987**, *84*: 2580–2584.

[50] P. J. Strong, H. Claus: Laccase: A review of its past and its future in bioremediation. *Crit. Rev. Environ. Sci. Technol.* **2011**, *41*: 373–434.

[51] J. F. Osma, J. L. Toca-Herrera, S. Rodríguez-Couto: Uses of laccases in the food industry. *Enzyme Res.* **2010**: 918761.

[52] F. Barrière, P. Kavanagh, D. Leech: A laccase–glucose oxidase biofuel cell prototype operating in a physiological buffer. *Electrochim. Acta* **2006**, *51*: 5187–5192.

[53] V. B. Urlacher, M. Girhard: Cytochrome P450 monooxygenases: an update on perspectives for synthetic application. *Trends Biotechnol.* **2012**, *30*: 26–36.

[54] S. B. Bankar, M. V. Bule, R. S. Singhal, L. Ananthanarayan: Glucose oxidase – An overview. *Biotechnol. Adv.* **2009**, *27*: 489–501.

[55] W. P. Dijkman, G. de Gonzalo, A. Mattevi, M. W. Fraaije: Flavoprotein oxidases: classification and applications. *Appl. Microbiol. Biotechnol.* **2013**, *97*: 5177–5188.

[56] A. Hernández-Ortega, F. Lucas, P. Ferreira, M. Medina, V. Guallar, A. T. Martínez: Role of active site histidines in the two half-reactions of the aryl-alcohol oxidase catalytic cycle. *Biochemistry* **2012**, *51*: 6595–6608.

[57] V. Leskovac, S. Trivić, G. Wohlfahrt, J. Kandrač, D. Peričin: Glucose oxidase from *Aspergillus niger*: the mechanism of action with molecular oxygen, quinones, and one-electron acceptors. *Int. J. Biochem. Cell Biol.* **2005**, *37*: 731–750.

[58] J. P. Roth, J. P. Klinman: Catalysis of electron transfer during activation of O_2 by the flavoprotein glucose oxidase. *Proc. Nat. Acad. Sci. USA* **2003**, *100*: 62–67.

[59] D. Petrović, D. Frank, S. C. L. Kamerlin, K. Hoffmann, B. Strodel: Shuffling active site substrate populations affects catalytic activity: The case of glucose oxidase. *ACS Catal.* **2017**, *7*: 6188–6197.

[60] J. Wang: Electrochemical glucose biosensors. *Chem. Rev.* **2008**, *108*: 814–825.

[61] A. Zebda, S. Cosnier, J. P. Alcaraz, M. Holzinger, A. Le Goff, C. Gondran, F. Boucher, F. Giroud, K. Gorgy, H. Lamraoui, P. Cinquin: Single glucose biofuel cells implanted in rats power electronic devices. *Sci. Rep.* **2013**, *3*: 1516.

[62] X. Yu, W. Lian, J. Zhang, H. Liu: Multi-input and -output logic circuits based on bio-electrocatalysis with horseradish peroxidase and glucose oxidase immobilized in multi-responsive copolymer films on electrodes. *Biosens. Bioelectron.* **2016**, *80*: 631–639.

[63] J. T. Holland, J. C. Harper, P. L. Dolan, M. M. Manginell, D. C. Arango, J. A. Rawlings, C. A. Apblett, S. M. Brozik: Rational redesign of glucose oxidase for improved catalytic function and stability. *PLoS ONE* **2012**, *7*: e37924.

[64] R. Ostafe, R. Prodanovic, J. Nazor, R. Fischer: Ultra-high-throughput screening method for the directed evolution of glucose oxidase. *Chem. Biol.* **2014**, *21*: 414–421.

[65] D. Petrović: Semi-rational design of glucose oxidase toward higher oxidative stability. Master's thesis, University of Belgrade, Faculty of Chemistry, Serbia, **2014**.

[66] Y. Horaguchi, S. Saito, K. Kojima, W. Tsugawa, S. Ferri, K. Sode: Engineering glucose oxidase to minimize the influence of oxygen on sensor response. *Electrochim. Acta* **2014**, *126*: 158–161.

[67] J. T. Holland, C. Lau, S. Brozik, P. Atanassov, S. Banta: Engineering of glucose oxidase for direct electron transfer via site-specific gold nanoparticle conjugation. *J. Am. Chem. Soc.* **2011**, *133*: 19262–19265.

[68] H. J. Hecht, H. M. Kalisz, J. Hendle, R. D. Schmid, D. Schomburg: Crystal structure of glucose oxidase from *Aspergillus niger* refined at 2.3 Å resolution. *J. Mol. Biol.* **1993**, *229*: 153–172.

[69] G. Wohlfahrt, S. Witt, J. Hendle, D. Schomburg, H. M. Kalisz, H. J. Hecht: 1.8 and 1.9 Å resolution structures of the *Penicillium amagasakiense* and *Aspergillus niger* glucose oxidases as a basis for modelling substrate complexes. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **1999**, *55*: 969–977.

[70] A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, M. H. M. Olsson: Electrostatic basis for enzyme catalysis. *Chem. Rev.* **2006**, *106*: 3210–3235.

[71] A. J. T. Smith, R. Müller, M. D. Toscano, P. Kast, H. W. Hellinga, D. Hilvert, K. N. Houk: Structural reorganization and preorganization in enzyme active sites: Comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J. Am. Chem. Soc.* **2008**, *130*: 15361–15373.

[72] A. V. Pisliakov, J. Cao, S. C. L. Kamerlin, A. Warshel: Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Nat. Acad. Sci. USA* **2009**, *106*: 17359–17364.

[73] M. Roca, M. Oliva, R. Castillo, V. Moliner, I. Tuñón: Do dynamic effects play a significant role in enzymatic catalysis? A theoretical analysis of formate dehydrogenase. *Chem. Eur. J.* **2010**, *16*: 11399–11411.

[74] A. D. Scouras, V. Daggett: The dynameomics rotamer library: Amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water. *Prot. Sci.* **2011**, *20*: 341–352.

[75] P. R. Kommoju, Z. Chen, R. C. Bruckner, F. S. Mathews, M. S. Jorns: Probing oxygen activation sites in two flavoprotein oxidases using chloride as an oxygen surrogate. *Biochemistry* **2011**, *50*: 5521–5534.

[76] M. Meyer, G. Wohlfahrt, J. Knäblein, D. Schomburg: Aspects of the mechanism of catalysis of glucose oxidase: A docking, molecular mechanics and quantum chemical study. *J. Comput.-Aided Mol. Des.* **1998**, *12*: 425–440.

[77] H. Yoshida, G. Sakai, K. Mori, K. Kojima, S. Kamitori, K. Sode: Structural analysis of fungus-derived FAD glucose dehydrogenase. *Sci. Rep.* **2015**, *5*: 13498.

[78] D. Petrović, X. Wang, B. Strodel: How accurately do force fields represent protein side chain ensembles? *Proteins: Struct., Funct., Bioinf.* **2018**, *doi: 10.1002/prot.25525.*

[79] D. Werck-Reichhart, R. Feyereisen: Cytochromes P450: a success story. *Genome Biol.* **2000**, *1*: 3003.

[80] I. G. Denisov, T. M. Makris, S. G. Sligar, I. Schlichting: Structure and chemistry of cytochrome P450. *Chem. Rev.* **2005**, *105*: 2253–2278.

[81] J. B. Schenkman, I. Jansson: Spectral analyses of cytochromes P450. In I. R. Phillips, E. A. Shephard, editors, Cytochrome P450 protocols., Totowa, NJ: Humana Press, **2006**, pages 11–18.

[82] S. D. Black: Membrane topology of the mammalian P450 cytochromes. *FASEB J.* **1992**, *6*: 680–685.

[83] S. Bak, F. Beisson, G. Bishop, B. Hamberger, R. Höfer, S. Paquette, D. Werck-Reichharta: Cytochromes P450. *The Arabidopsis Book* **2011**, *9*: e0144.

[84] S. L. Kelly, D. E. Kelly: Microbial cytochromes P450: biodiversity and biotechnology. Where do cytochromes P450 come from, what do they do and what can they do for us? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2013**, *368*.

[85] E. O'Reilly, M. Corbett, S. Hussain, P. P. Kelly, D. Richardson, S. L. Flitsch, N. J. Turner: Substrate promiscuity of cytochrome P450 RhF. *Catal. Sci. Technol.* **2013**, *3*: 1490–1492.

[86] A. Chefson, K. Auclair: Progress towards the easier use of P450 enzymes. *Mol. BioSyst.* **2006**, *2*: 462–469.

[87] B. Meunier, S. P. de Visser, S. Shaik: Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. *Chem. Rev.* **2004**, *104*: 3947–3980.

[88] S. Shaik, D. Kumar, S. P. de Visser, A. Altun, W. Thiel: Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. *Chem. Rev.* **2005**, *105*: 2279–2328.

[89] J. Rittle, M. T. Green: Cytochrome P450 compound I: Capture, characterization, and C-H bond activation kinetics. *Science* **2010**, *330*: 933–937.

[90] C. J. C. Whitehouse, S. G. Bell, L. L. Wong: P450$_{BM3}$ (CYP102A1): connecting the dots. *Chem. Soc. Rev.* **2012**, *41*: 1218–1260.

[91] A. W. Munro, D. G. Leys, K. J. McLean, K. R. Marshall, T. W. B. Ost, S. Daff, C. S. Miles, S. K. Chapman, D. A. Lysek, C. C. Moser, C. C. Page, P. L. Dutton: P450 BM3: the very model of a modern flavocytochrome. *Trends Biochem. Sci.* **2002**, *27*: 250–257.

[92] P. Le-Huu, T. Heidt, B. Claasen, S. Laschat, V. B. Urlacher: Chemo-, regio-, and stereoselective oxidation of the monocyclic diterpenoid β-cembrenediol by P450 BM3. *ACS Catal.* **2015**, *5*: 1772–1780.

[93] R. Ditchfield: Self-consistent perturbation theory of diamagnetism. I. A gauge-invariant LCAO method for N.M.R. chemical shifts. *Mol. Phys.* **1974**, *27*: 789–807.

[94] M. W. Lodewyk, M. R. Siebert, D. J. Tantillo: Computational prediction of ^{1}H and ^{13}C chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem. Rev.* **2012**, *112*: 1839–1862.

[95] G. Bifulco, P. Dambruoso, L. Gomez-Paloma, R. Riccio: Determination of relative configuration in organic compounds by NMR spectroscopy and computational methods. *Chem. Rev.* **2007**, *107*: 3744–3779.

[96] P. Le-Huu, D. Petrović, B. Strodel, V. B. Urlacher: One-pot, two-step hydroxylation of the macrocyclic diterpenoid β-cembrenediol catalyzed by P450 BM3 mutants. *ChemCatChem* **2016**, *8*: 3755–3761.

[97] S. G. Smith, J. M. Goodman: Assigning the stereochemistry of pairs of diastereoisomers using GIAO NMR shift calculation. *J. Org. Chem.* **2009**, *74*: 4597–4607.

[98] S. G. Smith, J. M. Goodman: Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: The DP4 probability. *J. Am. Chem. Soc.* **2010**, *132*: 12946–12959.

[99] M. Baker: Protein engineering: navigating between chance and reason. *Nat. Methods* **2011**, *8*: 623–626.

[100] F. H. Arnold, A. A. Volkov: Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.* **1999**, *3*: 54–59.

[101] C. Jäckel, P. Kast, D. Hilvert: Protein Design by Directed Evolution. *Annu. Rev. Biophys.* **2008**, *37*: 153–173.

[102] N. J. Turner: Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **2009**, *5*: 567–573.

[103] N. Pokala, T. M. Handel: Protein design—Where we were, where we are, where we're going. *J. Struct. Biol.* **2001**, *134*: 269–281.

[104] G. Kiss, N. Çelebi Ölçüm, R. Moretti, D. Baker, K. N. Houk: Computational enzyme design,. *Angew. Chem. Int. Ed.* **2013**, *52*: 5700–5725.

[105] H. Alogheli, G. Olanders, W. Schaal, P. Brandt, A. Karlén: Docking of macrocycles: Comparing rigid and flexible docking in Glide. *J. Chem. Inf. Model.* **2017**, *57*: 190–202.

[106] D. Petrović, A. Bokel, M. Allan, V. Urlacher, B. Strodel: Simulation-guided design of cytochrome P450 for chemo- and regioselective macrocyclic oxidation. *J. Chem. Inf. Model.* **2018**, *58*: 848–858.

[107] S. P. de Visser, F. Ogliaro, P. K. Sharma, S. Shaik: What factors affect the regioselectivity of oxidation by cytochrome P450? A DFT study of allylic hydroxylation and double bond epoxidation in a model reaction. *J. Am. Chem. Soc.* **2002**, *124*: 11809–11826.

[108] K. D. Dubey, B. Wang, S. Shaik: Molecular dynamics and QM/MM calculations predict the substrate-induced gating of cytochrome P450 BM3 and the regio- and stereoselectivity of fatty acid hydroxylation. *J. Am. Chem. Soc.* **2016**, *138*: 837–845.

[109] R. Ramanan, K. D. Dubey, B. Wang, D. Mandal, S. Shaik: Emergence of function in P450-proteins: A combined quantum mechanical/molecular mechanical and molecular dynamics study of the reactive species in the H_2O_2-dependent cytochrome $P450_{SP\alpha}$ and its regio- and enantioselective hydroxylation of fatty acids. *J. Am. Chem. Soc.* **2016**, *138*: 6786–6797.

[110] K. D. Dubey, B. Wang, M. Vajpai, S. Shaik: MD simulations and QM/MM calculations show that single-site mutations of cytochrome $P450_{BM3}$ alter the active site's complexity and the chemoselectivity of oxidation without changing the active species. *Chem. Sci.* **2017**, *8*: 5335–5344.

[111] The new plastics economy: Rethinking the future of plastics. www3.weforum.org/docs/ WEF_The_New_Plastics_Economy.pdf (Retrieved 08.06.2017).

[112] V. Siracusa, P. Rocculi, S. Romani, M. Dalla Rosa: Biodegradable polymers for food packaging: a review. *Trends Food. Sci. Technol.* **2008**, *19*: 634–643.

[113] G. Mensitieri, E. Di Maio, G. G. Buonocore, I. Nedi, M. Oliviero, L. Sansone, S. Iannace: Processing and shelf life issues of selected food packaging materials and structures from renewable resources. *Trends Food. Sci. Technol.* **2011**, *22*: 72–80.

[114] F. Awaja, D. Pavel: Recycling of PET. *Eur. Polym. J.* **2005**, *41*: 1453–1477.

[115] L. Bartolome, M. Imran, B. G. Cho, W. A. Al-Masry, D. H. Kim: Recent developments in the chemical recycling of PET. In D. Achilias, editor, Material recycling – Trends and perspectives, Rijeka, Croatia: InTech, **2012**, pages 65–84.

[116] S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura, K. Oda: A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **2016**, *351*: 1196–1199.

[117] G. M. Guebitz, A. Cavaco-Paulo: Enzymes go big: surface hydrolysis and functionalisation of synthetic polymers. *Trends Biotechnol.* **2008**, *26*: 32–38.

[118] W. Zimmermann, S. Billig: Enzymes for the biofunctionalization of poly(ethylene terephthalate). In G. S. Nyanhongo, W. Steiner, G. Gübitz, editors, Biofunctionalization of polymers and their applications, Berlin, Heidelberg: Springer, **2011**, pages 97–120.

[119] D. Ribitsch, A. O. Yebra, S. Zitzenbacher, J. Wu, S. Nowitsch, G. Steinkellner, K. Greimel, A. Doliska, G. Oberdorfer, C. C. Gruber, K. Gruber, H. Schwab, K. Stana-Kleinschek, E. H. Acero, G. M. Guebitz: Fusion of binding domains to *Thermobifida cellulosilytica* cutinase to tune sorption characteristics and enhancing PET hydrolysis. *Biomacromolecules* **2013**, *14*: 1769–1776.

[120] U. Bornscheuer, K. Buchholz, J. Seibel: Enzymatic degradation of (ligno)cellulose. *Angew. Chem. Int. Ed.* **2014**, *53*: 10876–10893.

[121] C. M. Payne, B. C. Knott, H. B. Mayes, H. Hansson, M. E. Himmel, M. Sandgren, J. Ståhlberg, G. T. Beckham: Fungal cellulases. *Chem. Rev.* **2015**, *115*: 1308–1448.

[122] S. Armenta, S. Moreno-Mendieta, Z. Sánchez-Cuapio, S. Sánchez, R. Rodríguez-Sanoja: Advances in molecular engineering of carbohydrate-binding modules. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*: 1602–1617.

[123] A. B. Boraston, D. N. Bolam, H. J. Gilbert, G. J. Davies: Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **2004**, *382*: 769–781.

[124] Y. Zhang, L. Wang, J. Chen, J. Wu: Enhanced activity toward PET by site-directed mutagenesis of *Thermobifida fusca* cutinase–CBM fusion protein. *Carbohydr. Polym.* **2013**, *97*: 124–129.

[125] J. Weber, D. Petrović, B. Strodel, S. Smits, S. Kolkenbrock, C. Leggewie, K.-E. Jaeger: Interactions of carbohydrate-binding modules with poly(ethylene terephthalate). *(in preparation)*.

[126] J. N. Mabry, M. Kastantin, D. K. Schwartz: Capturing conformation-dependent molecule–surface interactions when surface chemistry is heterogeneous. *ACS Nano* **2015**, *9*: 7237–7247.

[127] A. E. Nel, L. Mädler, D. Velegol, T. Xia, E. M. V. Hoek, P. Somasundaran, P. Klaessig, V. Castranova, M. Thompson: Understanding biophysicochemical interactions at the nano–bio interface. *Nat. Mater.* **2009**, *8*: 543–557.

[128] S. U. Boyd, R. H. Boyd: Chain dynamics and relaxation in amorphous poly(ethylene terephthalate): A molecular dynamics simulation study. *Macromolecules* **2001**, *34*: 7219–7229.

[129] E. R. Cruz-Chu, T. Ritz, Z. S. Siwy, K. Schulten: Molecular control of ionic conduction in polymer nanopores. *Faraday Discuss.* **2009**, *143*: 47–62.

[130] Q. Wang, D. J. Keffer, S. Petrovan, J. B. Thomas: Molecular dynamics simulation of poly(ethylene terephthalate) oligomers. *J. Phys. Chem. B* **2010**, *114*: 786–795.

[131] Q. Wang, D. J. Keffer, D. M. Nicholson, J. B. Thomas: Coarse-grained molecular dynamics simulation of polyethylene terephthalate (PET). *Macromolecules* **2010**, *43*: 10722–10734.

[132] Y. Fu, W. R. Busing, Y. Jin, K. A. Affholter, B. Wunderlich: Poly(ethylene terephthalate) fibers. 1. Crystal structure and morphology studies with full-pattern x-ray diffraction refinement. *Macromolecules* **1993**, *26*: 2187–2193.

Paper I

Shuffling active site substate populations affects catalytic activity: The case of glucose oxidase

by <u>D. Petrović</u>, D. Frank, S.C.L. Kamerlin, K. Hoffmann, and B. Strodel

ACS Catal. **2017**, *7*: 6188–6197

D.P. contributed to this paper by performing and analyzing all simulations. D.P. wrote the manuscript with inputs from the coauthors.

Shuffling Active Site Substate Populations Affects Catalytic Activity: The Case of Glucose Oxidase

Dušan Petrović,[†] David Frank,[‡,⊥] Shina Caroline Lynn Kamerlin,[§] Kurt Hoffmann,[*,‡] and Birgit Strodel[*,†,‖]

[†]Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany
[‡]Institute of Molecular Biotechnology, RWTH Aachen University, Worringerweg 1, 52074 Aachen, Germany
[§]Department of Cell and Molecular Biology, Uppsala University, BMC Box 596, S-751 24 Uppsala, Sweden
[‖]Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

ⓢ Supporting Information

ABSTRACT: Glucose oxidase has wide applications in the pharmaceutical, chemical, and food industries. Many recent studies have enhanced key properties of this enzyme using directed evolution, yet without being able to reveal why these mutations are actually beneficial. This work presents a synergistic combination of experimental and computational methods, indicating how mutations, even when distant from the active site, positively affect glucose oxidase catalysis. We have determined the crystal structures of glucose oxidase mutants containing molecular oxygen in the active site. The catalytically important His516 residue has been previously shown to be flexible in the wild-type enzyme. The molecular dynamics simulations performed in this work allow us to quantify this floppiness, revealing that His516 exists in two states: catalytic and noncatalytic. The relative populations of these two substates are almost identical in the wild-type enzyme, with His516 readily shuffling between them. In the glucose oxidase mutants, on the other hand, the mutations enrich the catalytic His516 conformation and reduce the flexibility of this residue, leading to an enhancement in their catalytic efficiency. This study stresses the benefit of active site preorganization with respect to enzyme conversion rates by reducing molecular reorientation needs. We further suggest that the computational approach based on Hamiltonian replica exchange molecular dynamics, used in this study, may be a general approach to screening in silico for improved enzyme variants involving flexible catalytic residues.

KEYWORDS: molecular dynamics, Hamiltonian replica exchange, X-ray, enzyme floppiness, active-site preorganization, side-chain dynamics, anticorrelated motions
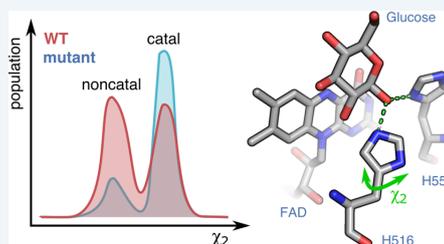
■ INTRODUCTION

Glucose oxidase (GOx) from *Aspergillus niger* is a β-D-glucose specific flavoprotein oxidase (EC 1.1.3.4) that efficiently catalyzes substrate conversion to D-glucono-δ-lactone. Due to its diverse potential applications in the fields of clinical, pharmaceutical, chemical and food industries, which reach far beyond the glucose biosensors typically used for blood sugar diagnostics, GOx has gained remarkable economic importance.[1] In this respect, flavoprotein oxidases are generally attractive biocatalysts due to their high regio- and stereoselectivity and the ability to use molecular oxygen as an oxidizing agent.[2] Furthermore, increasing the catalytic activity and stability of enzymes is a persisting necessity for many industrial applications.

The nondeterministic nature of evolution, both natural and directed, provides multiple uphill paths on the fitness landscape of an enzyme; most pathways, however, lead downhill.[3,4] As some enzymes sacrifice their catalytic power for metabolic control or live under low evolutionary pressure, the location of

a natural enzyme on the fitness landscape is not necessarily at the global optimum.[5,6] Another constraint in enzyme evolution is diminishing returns: as an enzyme approaches its theoretical limit on the landscape, mutations keep having smaller additive benefits. The gain in one property often has a high cost for another (e.g., the apparent stability−activity tradeoff), and nature usually does not pay the price of complete catalytic optimization.[7,8] GOx was postulated to be an "ideal enzyme" for biosensors because it fulfills three important criteria: high specificity, turnover, and stability.[9] Although GOx is several orders of magnitude less efficient than the "perfect enzyme" triosephosphate isomerase, where the reaction is diffusion limited,[5] GOx has a much higher rate constant than other oxidases, leading to its label "the 'Ferrari' of the oxidases".[10] The high efficiency and selectivity suggest that GOx is a highly-

evolved enzyme that lies close to its catalytic limit. This is supported by several mutagenesis studies that managed to achieve only marginal improvements of the kinetic properties of GOx (e.g., up to around 5 times higher k_{cat} or lower K_M).[11−15] However, despite the quite modest improvements in comparison to the wild-type enzyme (WT),[15] their impact in relation to the multimillion dollar industry involving applications of GOx is still important.

In this work, we study several improved GOx mutants that were recently derived using a combination of directed evolution and ultrahigh-throughput screening.[15] Since relatively unspectacular mutations, far from the active site, were responsible for the observed catalytic enhancement, we aimed to find out the underlying rationale for their improvement and studied them in more detail. GOx is described to operate by a ping-pong bi-bi mechanism, where the first step (reductive half-reaction) involves a concerted proton and hydride transfer from the anomeric carbon of glucose respectively to His516 and the N5 atom of the flavin adenine dinucleotide (FAD) cofactor (Figure 1a).[16] Although not directly involved in the reaction
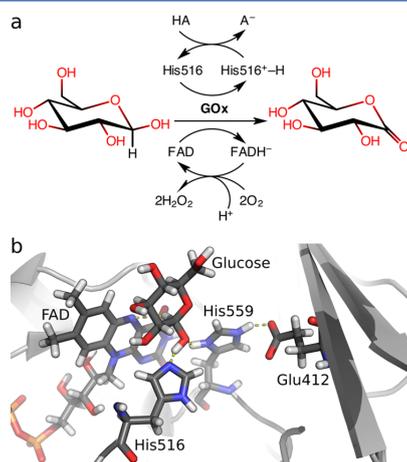

Figure 1. (a) In the reductive half-reaction, glucose binding is followed by concerted proton and hydride transfer from the C1 carbon of glucose to His516 and FAD, respectively. Electrons are then transferred, in the oxidative half-reaction, from reduced FAD to oxygen in two single-electron-transfer steps. (b) The active site of glucose oxidase from *A. niger* is buried in a pocket, and it is defined by Glu412, His516, His559, and FAD, which are shown as sticks, together with glucose, and colored by atom type (gray, C; blue, N; red, O; white, H; orange, P). The rest of the protein is shown in gray cartoon, and hydrogen bonds are indicated by dashed yellow lines.

mechanism, Glu412 and His559 are thought to act as a buffer for controlling the reactivity of the active site by maintaining the proper acidity (Figure 1b).[16] The protonation state of

His516 is crucial for the subsequent oxidative half-reaction,[17] as it leads to increased oxygen binding and reactivity via stepwise single-electron transfers.[10]

The catalytic ability of an enzyme originates mostly from the stabilization of the transition state geometry for ligand conversion,[18] where binding is based on shape and electrostatic complementarity.[19] There are three main factors that can lead to suboptimal enzyme kinetics: (1) a slow chemical step, (2) dissociation of the initial enzyme−substrate encounter complexes before the advanced enzyme−substrate complex is being established, and (3) floppiness, that is, the coexistence of multiple closely related enzyme substates (e.g., multiple side chain rotamers), of which only some are productive.[6,20] Floppiness increases the ratio of nonproductive to productive substates, leading to more futile enzyme−substrate encounters, which negatively affects enzymatic rates, an effect also observed by heating an enzyme.[21] Therefore, a preorganized and rigid active site (i.e., a single substate) leads to the most efficient chemical step.[6,22−26] The significance of productive and nonproductive substates in enzyme ensembles was recently shown, for example, for T4 lysozyme,[27] cyclophilin A,[28] and α-esterase 7.[29] As a consequence of this active site preorganization, the relative residue rigidity has often been found to be higher for the catalytic than for the noncatalytic amino acids.[22] Atomistic molecular dynamics (MD) simulations were shown to be a useful method for analyzing the floppiness of an enzyme and determining how mutations affect the enzyme substates.[30−32] Thus, they nicely complement the analysis of static crystal structures by also describing the underlying dynamics of the protein.

The aim of the current work is to understand the catalytic properties of GOx by studying several variants with improved catalytic activities and to create a basis for further improvement of the catalytic efficiency of this important enzyme. We report the very first crystal structures for GOx mutants, all bearing a molecule of oxygen in the active site. In addition to X-ray crystallography, we used MD simulations to corroborate our conclusions drawn from the crystal structures and to investigate the effects of the distant GOx mutations on the protein dynamics. Using Hamiltonian replica exchange MD, we further explored the conformational ensemble of the active site's His516 that was previously reported as flexible in the wild-type enzyme.[33,34] Our results provide structural and dynamic proofs that His516 is indeed flexible in the **WT**, where it can flip between catalytic and noncatalytic conformations, while in the most active mutant, **A2**, His516 is apparently locked in the catalytically active conformation.

■ MATERIALS AND METHODS

Purification and Deglycosylation. The GOx mutants **A2** and **F9** (Table 1) were expressed in *Pichia pastoris* strain KM71H (Invitrogen) according to the manufacturer's protocol. After 4 days of fermentation, the supernatant was concentrated

Table 1. Mutations Present in the Simulated GOx Variants

GOx	mutations					
P	T30V		I94V		A162T		
Pk	T30V		I94V		A162T	R537K	
Pv	T30V		I94V		A162T		M556V
A2	T30V		I94V		A162T	R537K	M556V
F9	T30V	R37K	I94V	V106I	A162T		M556V

to 20 mL on a Viva Flow 50 system (Sarotius) with a 50 kDa ultrafiltration membrane. The concentrate was dialyzed against 10 mM phosphate buffer (pH 6.0) overnight at 4 °C and loaded onto a 20 mL Fast Flow DEAE Sepharose column (GE Healthcare). The protein was purified using a linear gradient from 10 to 250 mM phosphate buffer (pH 6.0) over 12 column volumes. The GOx peaks were pooled together and concentrated to 1 mL using 10 kDa ultrafiltration columns. The enzyme solution was dialyzed against 50 mM sodium acetate buffer (pH 5.5) overnight at 4 °C.

GOx deglycosylation was performed by incubating the protein solution with Endo H enzyme (30 U mg^{-1}, NEB) for 20 h at 37 °C. The deglycosylated samples were loaded on a 120 mL Hi Load Superdex 75 gel filtration column using 10 mM sodium acetate buffer (pH 5.5) supplemented with 50 mM NaCl. The fractions with GOx activity were collected and concentrated to 25 mg mL^{-1} on a 10 kDa ultrafiltration column (Millipore).

GOx Crystallization. The concentrated solution was filtered through a 0.1 μm centrifugal filter (Millipore), and crystal growth conditions were screened initially using Hampton Screens I and II with the vapor diffusion sitting drop method on TAORAD crystallization plates. The first screening revealed that 1,4-dioxane is suitable to promote the crystallization of GOx, and optimal conditions were 100 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, pH 7.0) in 40% 1,4-dioxane or 100 mM HEPES (pH 7.5) in 45% 1,4-dioxane. Crystals were picked from the droplets using cryo-loops and equilibrated stepwise for several seconds in the crystallization buffer containing increasing concentrations of polyethylene glycol (PEG400) or glycerol in order to cryoprotect the crystal before flash-freezing in a liquid nitrogen stream at −173 °C. The data sets were collected in house, using a Bruker FR591 rotating anode X-ray generator and a Mar345dtb detector.

The collected data were analyzed and processed using the software iMOSFLM, Pointless and scaled using SCALA, all belonging to the CCP4 suite.[35] In order to solve structures of the **F9** and **A2** mutants, molecular replacement was carried out using the structure of the wild-type GOx from *A. niger* (PDB ID: 3QVP).[36] All residues mutated in **A2** or **F9** were replaced by alanine residues in 3QVP for the model generation using Chainsaw. Molecular replacement was done using Molrep, and final refinement was carried out by iterative steps of modeling/refinement cycles with WinCoot[37] and Refmac5.[35] The Ramachandran plot analysis was performed with Rampage.[38]

Molecular Dynamics. Molecular dynamics simulations were performed for the **WT** (PDB ID: 1CF3),[33] **P**, **A2** (PDB ID: 5NIT), and **F9** (PDB ID: 5NIW) GOx variants. Two additional mutants were considered, **Pk** and **Pv**, to examine the effect of single mutations on **P** (Table 1). In the absence of crystal structures, starting coordinates of the **P**, **Pk**, and **Pv** mutants were obtained by reverting the corresponding residues in **A2** GOx to a rotamer present in the wild-type enzyme. All structures were simulated in the oxidized form as holoenzyme (GOx + FAD) in complex with β-D-glucopyranose. The missing heavy atoms in the **A2** and **F9** structures were built using MODELER 9.14.[39]

Glucose was docked into the active site using AutoDock Vina.[40] A binding mode positioned on the *re* face of FAD (Figure 1b) was chosen for further modeling, as this is the most reasonable mode according to the catalytic mechanism and as it resembles the previously proposed substrate position.[33] In this orientation, the glucose H atom at the anomeric carbon C1 is directed toward the N5 atom of the isoalloxazine moiety of FAD, while the hydroxyl hydrogen from the same glucose carbon is oriented toward His516 and His559.

The β-D-glucopyranose and FAD topologies were created using ACPYPE[41] and Antechamber.[42] The glucose structure was optimized with Gaussian 09[43] at the B3LYP/6-31G* level of theory, followed by the calculation of restrained electrostatic potential (RESP) charges at the HF/6-31G* level. The FAD charges were obtained from Todde et al.[44]

MD simulations were performed using the GROMACS 4.6.7 suite,[45] with the Amber 99SB-ILDN force field[46] and TIP3P explicit water.[47] Hydrogen atoms were added, and the protonation states of all titratable residues were assigned on the basis of a PROPKA 3.1 analysis[48] corresponding to a pH of 5.5, which is optimal for GOx activity. A disulfide bridge was defined between the Cys164 and Cys206 side chains. The protein was centered in a truncated octahedral box, at least 10 Å away from each of the box edges, and solvated with around 22000 water molecules. The net charge of the system was neutralized with sodium ions. The system was minimized in two stages: an initial minimization with steepest descent (maximum force of 500 kJ mol^{-1} nm^{-1}), followed by a minimization with the conjugate gradient algorithm (maximum force of 100 kJ mol^{-1} nm^{-1}).

Periodic boundary conditions were applied, and electrostatic interactions were treated using the particle mesh Ewald method.[49] The cutoff distance for the short-range nonbonded interactions was 12 Å. An integration step of 2.0 fs was used, and bonds were constrained with the LINCS algorithm.[50] The minimized system was gradually heated and equilibrated at 25 °C for 100 ps in the NVT ensemble with the protein and ligands restrained using a positional restraint force constant of 1000 kJ mol^{-1} nm^{-2}. Following the equilibration under a constant volume, two stages of NpT equilibration were carried out. In the first phase, a 2 ns equilibration was performed with restraints on the protein and ligands. A second, 8 ns equilibration followed, with restrained protein backbone and FAD motion while glucose was free to move.

Production MD simulations were carried in the NpT ensemble for 100 ns (three independent simulations were performed for each GOx variant), collecting coordinates of the system every 20 ps. The modified Berendsen (v-rescale) thermostat[51] and the Parrinello−Rahman barostat[52] were employed. The production MD sampling time accumulated over all GOx variants amounted to 1.8 μs.

Hamiltonian Replica Exchange MD. The Hamiltonian replica exchange MD (HREX-MD) simulations were performed using GROMACS 4.6.7 in combination with the Plumed 2.1 plugin,[53] as implemented by Bussi.[54] The same conditions were applied as in the standard MD simulations. Four replicas were simulated for each GOx variant, where only the energy terms (i.e., the Hamiltonian) affecting His516 were scaled. The Hamiltonian scaling factors were exponentially distributed between 1.00 and 0.67 (exact scaling factors were 1.000, 0.874, 0.763, and 0.667), which corresponds to temperatures between 25 and 174 °C. The exchange of replicas was attempted every 4 ps during the 50 ns simulations. The exchange acceptance ratio was 30−70% in all HREX-MD simulations; only the **Pv** variant had an exchange rate of ~15%. The sampling time accumulated over all HREX-MD simulations was 1.2 μs. Structures sampled for the HREX-

MD simulation with the unperturbed Hamiltonian were used for the analysis.

Umbrella Sampling MD. The umbrella sampling (US-MD) simulations were performed using GROMACS 5.1.2 in combination with the Plumed 2.2 plugin. The same conditions were applied as in the previous MD simulations. The sampling was performed for the varying χ_2 dihedral of His516 in the range of 25–235°, over 26 windows; exact χ_2 restraining values are shown in Table S1 in the Supporting Information. The χ_1 dihedral was restrained to 285° with a weak force of 50 kJ mol^{-1} rad^{-1} to keep this angle in the g^- geometry. Each window was simulated for 50 ns. Dihedral angles were written every 0.5 ps, and the first 5 ns was discarded for the potential of mean force (PMF) calculations. The PMF was examined for the **WT** and **A2** GOx, amounting to 2.6 μs US-MD sampling time. The PMF was calculated using the weighted histogram analysis method (WHAM),[55] where the error was estimated using the blocking procedure.[56] Briefly, each window from a US-MD simulation was split into 10 segments (i.e., blocks) of increasing length, ranging from 1 ns for the shortest block up to the full 45 ns per window for the longest block. The PMF was calculated for each block, and all PMFs were aligned to the final point at $\chi_2 = 235°$. The blocked standard error was calculated for each window using data from all 10 blocks.

Data Analysis. GROMACS tools, VMD 1.9.1[57] and MATLAB R2015b were used for the trajectory analysis. The dynamic cross-correlated motion analysis was performed in R 3.2.5 using the Bio3D package.[58] The active site volumes were calculated with POVME.[59] PyMOL[60] and Chimera[61] were used for figure rendering.

RESULTS AND DISCUSSION

Crystallization and Structure Determination. In previous work,[15] several GOx mutants with improved activity and stability were identified. The **A2** mutant shows the highest catalytic activity, while the mutant **F9** has the highest thermal stability. The first crystallization experiments were carried out using glycosylated GOx expressed in *P. pastoris*. Hampton Screens I and II were tested, but no promising crystallization conditions were found. Deglycosylation has been previously shown to be important for crystallization,[62] as the process of crystallization demands highly uniform macromolecules. The glycosylation in *P. pastoris* is characterized by a uniform *N*-acetylglucosamine (NAG) core glycosylation with NAG-β(1,4)-NAG, followed by a heterogeneous glycosylation of a high mannose (β-D-mannose, BMA) content.[63] In order to obtain uniform GOx molecules, the carbohydrate moieties were removed by enzymatic hydrolysis using glycosidases.

The crystallization experiments with the deglycosylated GOx were successful for the **A2** and **F9** GOx variants. The first screening using Hampton Screens I and II revealed that 1,4-dioxane is suitable to promote the crystallization of GOx. Further fine screening with different buffers, pHs, and 1,4-dioxane concentrations indicated that the crystallization works best using 100 mM HEPES (pH 7.0) in 40% 1,4-dioxane or 100 mM HEPES (pH 7.5) in 45% 1,4-dioxane. Regarding incompatibilities of classic crystallization plates, vapor diffusion crystallization was performed using TAORAD crystallization plates in their sitting drop configuration. The **A2** and **F9** GOx crystals grew in the form of long thick needles and showed an intense yellow color. The crystal growth took three to 5 days at room temperature and yielded crystals in the $P3_21$ space group (Table 2). The cell content analysis gave a probability of 0.99 that one molecule is present per asymmetric unit with a water content of 57%.

Table 2. Data Collection and Refinement Statistics (Molecular Replacement)^a

	A2 (5NIT)	F9 (5NIW)
Data Collection		
space group	$P3_21$	$P3_21$
cell dimens		
a, b, c (Å)	128.7, 128.7, 77.7	128.1, 128.1, 77.7
α, β, γ (deg)	90.0, 90.0, 120.0	90.0, 90.0, 120.0
resolution range (Å)	45.3–1.9	42.0–1.8
R_{merge} (%)	13.6 (4.2/76)	15.7 (3.8/82)
$I/\sigma(I)$	12.6 (34.8/2.6)	12.1 (29.8/2.8)
completeness (%)	98.1 (99.8/88.1)	98.3 (99.8/88.7)
redundancy	8.2 (10.2/6.6)	11.1 (11.1/10.4)
Refinement		
resolution (Å)	1.86	1.80
no. of unique rflns	61296 (2106/7940)	67817 (2333/8856)
R_{work}/R_{free} (%)	16.5/20.4	15.4/19.1
no. of water molecules	427	364
B factors		
protein	19.6	24.2
FAD	15.6	19.4
water	28.9	33.3
RMS deviations		
bond lengths (Å)	0.019	0.022
bond angles (deg)	1.94	2.20
Ramachandran plot		
favored region	561	557
allowed region	18	17
outlier region	0	0

^{a}Values in parentheses are for the lowest- and highest-resolution shells.

The *N*-linked Asn glycosylation is related to specific motifs, i.e., Asn-X-Ser or Asn-X-Thr, where X can be any amino acid except Pro.[64] Eight possible glycosylation sites (Asn43, Asn89, Asn161, Asn168, Asn258, Asn355, Asn388, and Asn473) are present in the GOx sequence. In the crystal structures, GOx was Asn-glycosylated at all sites, except Asn43. Interpretable electron density was observed for mutant **A2** at four sites (89, 161, 355, and 388) and in mutant **F9** at six sites (89, 161, 258, 355, 388, and 473), whereas the two additional sites showed a less pronounced electron density. In most cases, a single NAG moiety remained at each of these positions, while the deglycosylation removed the other initially present carbohydrates. At Asn89, the electron density indicates the presence of the core glycosylation, Asn89-NAG-NAG-BMA, which was not pruned by Endo H due to steric hindrance. Asn89 is located at the homodimeric interface of the GOx dimer, and BMA-rich glycosylation protrudes out of the dimer cleft. As the glycosylation is involved in the intermolecular interactions, it promotes the dimer state.

The cocrystallization approaches with D-glucose and D-glucal did not yield crystals. Soaking experiments were also unsuccessful and caused crystal degradation or did not yield visible ligand electron densities.

Protein Flexibility and Dynamics. To test the influence of the mutations on the protein dynamics, we performed MD simulations of several GOx variants. Figure S1 in the Supporting Information, which shows the root-mean-square

Table 3. Selected GOx Variants: Wild Type (WT), Parent (P), and Two Well-Performing Mutants[a]

GOx	K_M (mM)[b]	k_{cat} (s^{-1})[b]	k_{cat}/K_M (mM^{-1} s^{-1})[b]	$t_{1/2}$ (min)[b]	$\langle V_{as} \rangle$ (Å^3)
WT	28.26 ± 1.15	189.38 ± 8.94	6.7	10.50 ± 0.71	261.9 ± 103.6
P	14.98 ± 0.51	291.82 ± 10.10	19.5	9.00 ± 0.70	238.3 ± 77.8
A2	18.54 ± 0.57	498.34 ± 15.12	26.9	11.74 ± 0.30	188.7 ± 65.2
F9	19.76 ± 0.54	345.16 ± 14.79	17.5	15.75 ± 0.71	239.3 ± 73.0

[a]The enzyme kinetics was measured at pH 5.5 and the thermal stability was estimated on the basis of the half-life ($t_{1/2}$) at 60 °C.[15] The average active site volume, $\langle V_{as} \rangle$, significantly decreases with increasing efficiency. [b]Data reproduced from Ostafe et al.[15]
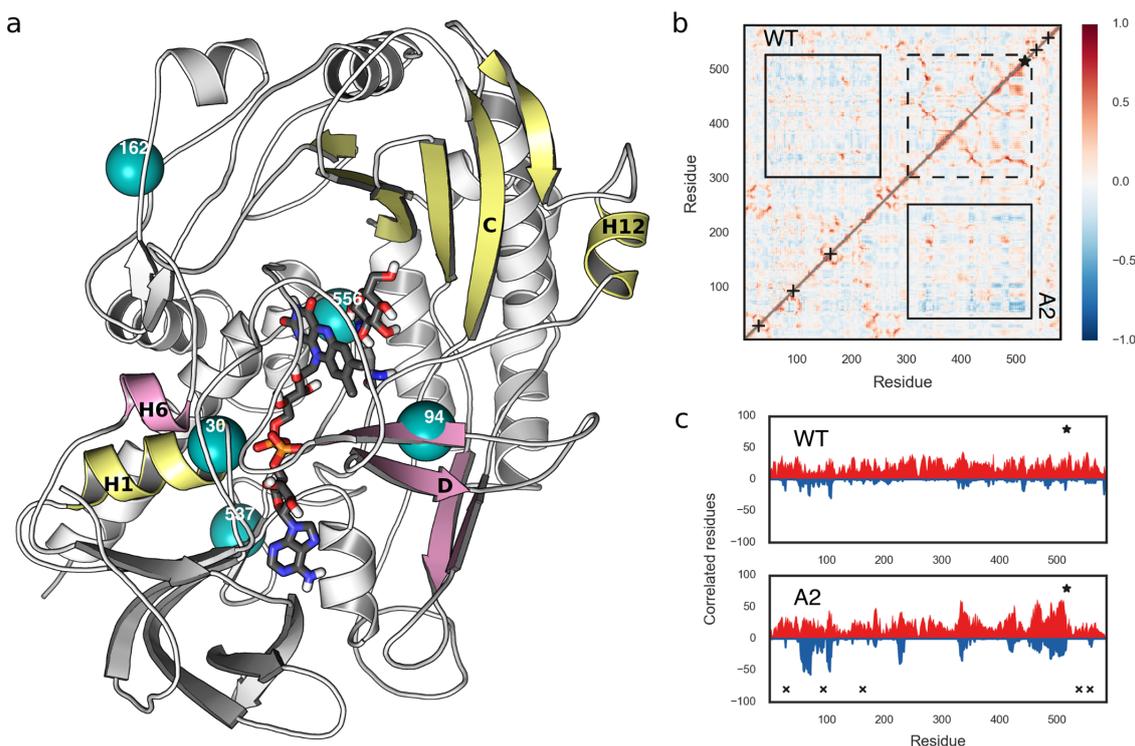

Figure 2. (a) **A2** GOx crystal structure with glucose docked into the active site. The secondary structures involved in the anticorrelated motions are shown in yellow and pink, and the positions of mutations are designated by cyan spheres. His516, FAD, and glucose are shown by sticks and colored by atom type (see Figure 1 for the color code). (b) Dynamic cross-correlation maps (DCCMs) of the **WT** GOx (top half) and the **A2** mutant (bottom half). The most discriminating regions are indicated by black rectangles: solid lines for anticorrelated and dashed lines for correlated motions. DCCMs for all GOx variants are given in Figure S3 in the Supporting Information. (c) Per-residue count of the correlated and anticorrelated motions (based on a cutoff of ±0.3) in the **WT** and **A2** GOx. Correlated and anticorrelated motions are shown in red and blue, respectively. The positions of the mutations are represented with × and that of His516 with ★. The count plots for all variants are given in Figure S4 in the Supporting Information.

deviation of the protein backbone motion, indicates that the **WT** and GOx mutants are stable during the 100 ns trajectories. A more detailed analysis of the enzyme dynamics reveals that the laboratory evolution of GOx led to a slight decrease in residue flexibility, especially in the active site region. This can be inferred from the associated standard deviations reflecting the change of the active site volumes (Table 3). Namely, the **A2** mutant has a much smaller deviation than the **WT** enzyme, indicating a less flexible active site in **A2**. On the other hand, the residual root-mean-square fluctuations (Figure S2 in the Supporting Information) reveal a notable destabilization of the β-sheet D (residues 77−81, 93−97, 434−438, and 448−451) in the glucose binding domain of all mutants, which is caused by the I94V mutation lying at this β-sheet (Figure 2a). This

mutation is present in all mutants, including the parent **P**. The **F9** variant is much less flexible than the other tested mutants, which has a positive effect on its thermostability but makes the enzyme less efficient than **P** (Table 3). These results are in agreement with Fraser et al., who suggested that, although mutations should be directed toward a more rigid active site, second-shell residues should be flexible to ensure the efficiency of the numerous steps involved in catalysis.[65]

Another aspect of enzyme dynamics is the correlated nature of residue motion that facilitates many biochemical processes.[66] Anticorrelated motions were previously related to enhanced catalysis in several enzymes.[67,68] While we observe a general increase in both correlated and anticorrelated motions over the course of the GOx evolution (Figure 2b,c and Figures S3 and

S4 in the Supporting Information), anticorrelated motions are particularly discriminating among the studied GOx mutants and the **WT**. Starting from the parent mutant **P**, the aforementioned I94V on β-sheet D, together with the T30V mutation located at helix H1, which is close to the phosphate groups of FAD, plays a very important role for anticorrelated motions in GOx. Namely, the motions of β-sheet D and α-helix H6 are anticorrelated to the motions of β-sheet C (glucose binding domain) and α-helices H1 (FAD binding domain) and H12, which extends to the active site's His516 (Figure 2a). This kind of ordered motion is, to a varying degree, visible in all variants. The R537K mutation, although located on the surface, has a positive influence on the magnitude of both correlated and anticorrelated motions in the **Pk** and **A2** variants (Figures S3 and S4). It further strengthens the anticorrelated motions already observed in the **P** mutant, while its effects on the correlated motions are especially high for the β-sheet C of the glucose binding domain (residues 211−213, 330−338, 347−353, 409−416, 420−427, and 484−489). From the **Pv**, **A2**, and **F9** variants it can be seen that the M556V mutation, which is close to the active site, exhibits a positive effect on the anticorrelated and, even more, on the correlated motions of the same region as influenced by R537K (i.e., the β-sheet C; see Figures S3 and S4).

In the pentamutant **A2**, the mutations work together to considerably enhance both correlated and anticorrelated motions, as shown in Figure 2b,c. A principal component analysis of the fluctuations of the pairwise distances between the residues performing highly anticorrelated motions in **A2** indicates that such motions contribute to the creation of a tighter active site in this variant (Figure S5 in the Supporting Information), which increases the probabilities of the contacts that directly stabilize the substrate in the proper position for the reaction to take place. This leads to the optimal orientation of the reactive atoms, which enhances catalysis and also lowers the K_M value 1.5-fold. It should be noted that the nature of glucose binding by the **WT** and mutant GOx does not change, as the same residues are always involved (Figure S6 in the Supporting Information). However, this figure also shows that the mutations changed the priorities of certain residues in stabilizing the substrate in the active site.

His516 Conformational Ensemble in GOx Crystals. The structures of **A2** (PDB ID: 5NIT) and **F9** (PDB ID: 5NIW) are similar to those of the wild-type GOx. Major differences in these structures exist only at the active site. For the first time, we see an important electron density situated between His516 and FAD and interpret it as a molecule of oxygen (Figure 3). A water molecule, present in all *A. niger* GOx structures apart from 1GAL, bridges the Nε of His516 and the N5 atom of FAD (HOH1000 in **A2**) and is between 2.74 and 2.89 Å distant from His516.

In all of the structures from *A. niger*, the side chain of His516 populates the broadly defined (g^-, Nt) rotamer with dihedral angles 240° < χ_1 < 360° and 150° < χ_2 < 210° (Table 4). In the 1CF3 structure, where the His516 side chain deviates the most from the center of the (g^-, Nt) rotamer population, the water molecule in the active site follows the His516 motion toward the tip of the oxygen molecule that is present in the **A2** and **F9** structures (Figure 4). Apart from movements within the (g^-, Nt) rotamer observed in the wild-type *A. niger* structures, the (g^-, Ng^+) rotamer (240° < χ_1 < 360° and 30° < χ_2 < 90°) is structurally documented in GOx of *Penicillium amagasakiense* (PDB ID: 1GPE,[33] Table 4), where a water molecule is
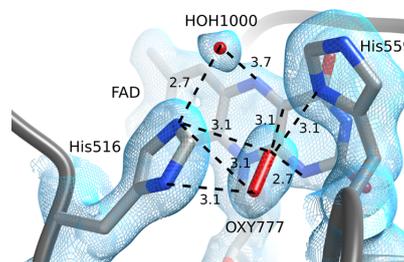

Figure 3. Structural view of the FAD *re* face of the **A2** mutant, showing the catalytically important His516 and His559 residues. A water molecule (HOH1000) interacts with His516 and is oriented toward the N5 nitrogen of FAD. An oxygen molecule (OXY777) is well centered with respect to His516. The electron density is shown as a cyan mesh, and important distances (in Å) are indicated by black dashed lines.

Table 4. Overview of *A. niger* and *P. amagasakiense* GOx and *A. flavus* GDH Crystal Structures Showing the Distribution of His516 Side Chain Dihedral Angles, the Number of Active Site Water Molecules in the PDB File, and the Distance between the Nε Atom of His516 and the Oxygen Atom of This Water Molecule

PDB ID	χ_1 (deg)	χ_2 (deg)	crystal water	HOH−His516 (Å)
1GAL	257	225	*a*	
1CF3	254	194	710	2.98
3QVP	291	185	1094	2.77
3QVR	295	195	1200	2.89
5NIT	293	197	1000	2.75
5NIW	288	199	1000	2.79
1GPE	284	64	837	2.70[b]
4YNT	277	201	798	2.68[c]
4YNU	284	197	*d*	*d*

[a]The absence of water might be due to the low resolution of the crystal structure. [b]Equivalent to His520 of *P. amagasakiense* GOx. [c]Equivalent to His505 of *A. flavus* GDH. [d]In 4YNU, the active site water is replaced by gluconolactone, whose O1 atom is positioned 2.79 Å from the His residue.
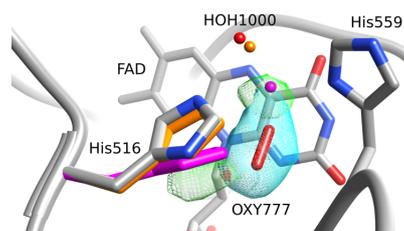
Figure 4. Structural view of the FAD *re* face of the **F9** mutant, showing the electron density of the oxygen molecule (cyan mesh) as well as the remaining positive electron density (green). Aligned to the His516 side chain of **F9** (colored by atom type), one can see the side chains of *A. niger* wild-type structures 3QVP (orange) and 1CF3 (magenta). Also shown are the oxygen atoms of the corresponding water molecules that are equivalent to HOH1000 in **F9**.

bridging His520 and His563, which are equivalent to His516 and His559, respectively, in GOx of *A. niger*. The (g^-, Nt) rotamer of His516 is the geometry necessary for the proton transfer from glucose to occur and will be therefore denoted

catalytic conformation. The same conformation of this conserved residue can be observed in many members of the superfamily, e.g., in aryl-alcohol oxidase,[69] cholesterol oxidase,[70] and cellobiose dehydrogenase,[71] and in a recently crystallized glucose dehydrogenase (GDH) from *Aspergillus flavus* (35% sequence identity with *A. niger* GOx).[72] GDH is oxygen-independent, yet it preserves the catalytic conformation in the unliganded state (PDB ID: 4YNT) and with gluconolactone (PDB ID: 4YNU) in the active site. QM/MM (quantum mechanics/molecular mechanics) calculations confirm that this conformation is the one present during catalysis in aryl-alcohol oxidase.[73] In the (g^-, Ng^+) conformation, His516 has moved away from the substrate, making the active site geometrically and chemically unsuitable for the concerted proton and hydride transfer (Figure S7 in the Supporting Information). Thus, this conformation is called *noncatalytic* henceforth.

The conformation of His516 in the **A2** and **F9** structures is similar to those in *A. niger* GOx structures with PDB codes 3QVP and 3QVR. While the His516 conformation of 3QVR is almost identical with that in **A2** and **F9**, that of 3QVP is slightly shifted toward the conformations found in the 1CF3 and 1GAL structures (Figure 4). The structures 3QVP and 3QVR resulted from an attempt to investigate the oxygen-binding site using chloride ions as oxygen substitutes,[36] a method postulated to be an alternative to the approach using xenon for the identification of potential oxygen-binding sites. However, despite the similarity of conformations adopted by His516 to those in **A2** and **F9**, no oxygen was reported in 3QVP and 3QVR. Instead, a water molecule was placed at the position occupied by the center of the oxygen molecule in **A2** and **F9**.

In order to further investigate the active site, the electron densities of five wild-type GOx structures (1GAL, 1CF3, 3QVP, 3QVR, and 1GPE) were re-examined using structure factors from the PDB database. All structures show at least some positive and/or negative electron density near His516 (Figure S8 in the Supporting Information). The angular displacement of His516 from the conformation observed in the **A2** mutant inversely follows the quality of the electron density around this residue and ends up with a partially missing electron density for the most deviating (g^-, Nt) structures (1CF3 and 1GAL). This last observation was at the origin to indicate that the His516 side chain is flexible.[33,34] A pH-induced conformational flexibility due to a different protonation state of His516 can be excluded, since crystals of the **A2** and **F9** mutants were grown at pH 7.0−7.5 and those of the 3QVP and 3QVR structures at pH 6.9 and 5.1, respectively, whereas crystals for the 1CF3 structure were obtained at an intermediate pH of 5.6. The 3QVR structure, with its His516 conformation closest to that observed in **A2**, shows only minor positive electron density on both sides of the water molecule, which was placed at the site of oxygen in **A2**, and may indicate the presence of oxygen already in this structure.

An interesting observation is the well-positioned oxygen in **A2** and **F9** with respect to the π-orbital system of His516 (Figure 3). As the oxygen reduction is spin forbidden by the triplet nature of molecular oxygen, a catalytic effect for the triplet−singlet transition might rely on the orbital coupling between oxygen and His516. Since the crystallographic electron density represents a mean observation of the conformational substates adopted by a protein and considering that the most active mutant **A2** shows neither positive nor negative electron density around His516, its conformation can be seen as very well defined, corresponding to a pure catalytic conformer of

His516. A more dynamic situation is observed for the **F9** mutant, which mainly adopts the catalytic conformation as in **A2**. However, modest positive electron densities at the active site of **F9** GOx indicate that the His516, with its water molecule, also samples small amounts of the displaced (g^-, Nt) conformation observed in the 1CF3 structure (Figure 4).

His516 Conformational Dynamics from Simulations. In order to quantify the flexibility of the side chain of His516, we performed standard and enhanced MD simulations. Our initial MD simulations of *A. niger* wild-type GOx showed that χ_1 is conserved to g^- geometry (240−360°; Figure S9 in the Supporting Information). The χ_2 dihedral samples two minima, Ng^+ (30−90°) and Nt (150°−210°), neither of which corresponds to the most stable conformation in the His rotamer library. However, the noncatalytic (g^-, Ng^+) geometry is 4 times more probable in the backbone-independent rotamer library than the catalytic (g^-, Nt) conformation.[74] In **WT**, His516 is free to flip to a small cavity located in the vicinity of the active site. The M556V mutation in **A2**, which resides at the border of this cavity, significantly decreases the cavity size, making His516 sterically hindered (Figure 5). It is important to note that valine is the most common residue found at this position in the consensus of glucose oxidase sequences.[15]
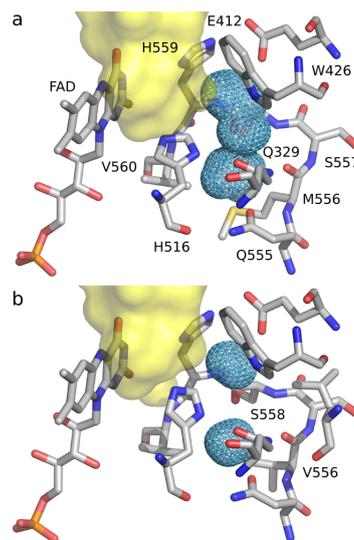
Figure 5. Cavity (light blue mesh) located in the vicinity of the active site (yellow surface) of (a) **WT** GOx and (b) **A2** mutant. The M556V mutation returns this residue to its consensus sequence, which significantly decreases the cavity size preventing His516 flipping.

The subsequent Hamiltonian replica exchange MD simulations, performed to quantify the flexibility of the His516 side chain, revealed that the g^- geometry is indeed dominant in all GOx variants. In **WT**, the catalytic (Nt) and noncatalytic (Ng^+) conformations are quite equally distributed (Figure 6). The **P** mutant introduces a clear separation between the two conformations while simultaneously enriching the catalytic form, and R537K and M556V further reduce the noncatalytic geometry. The synergy of these effects conserves His516 mostly in the catalytic conformation in **A2**, increasing k_{cat} 2.6 times and its efficiency 4-fold in comparison to that in **WT**. The
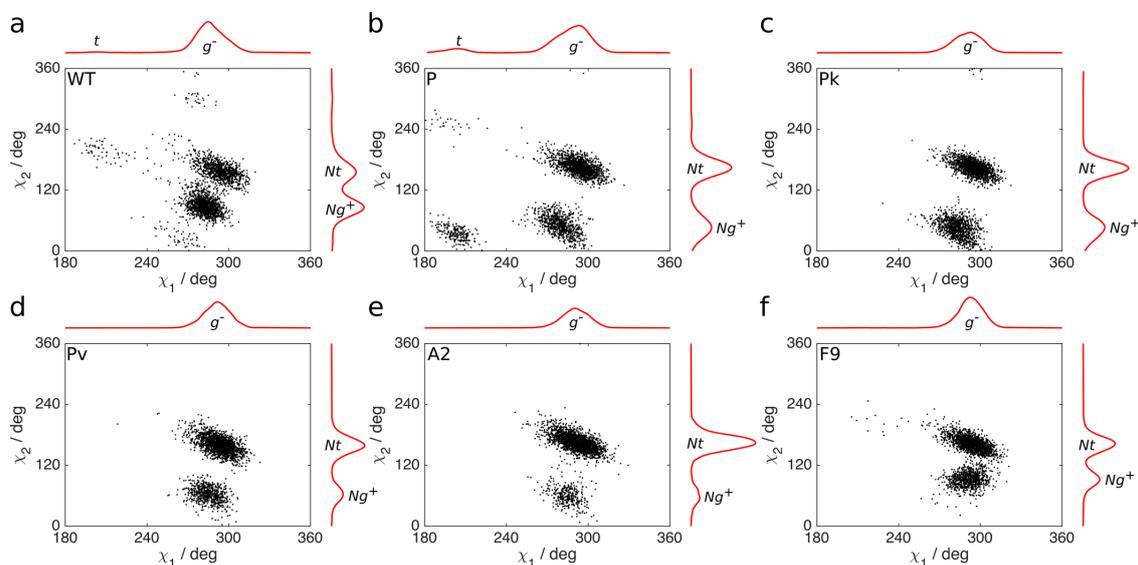
Figure 6. His516 side chain dihedral angles (χ_1 and χ_2) distribution: (a) **WT** GOx, (b) parent mutant **P**, (c) **Pk**, (d) **Pv**, (e) **A2**, and (f) **F9**. The χ_1 dihedral has a clear preference for g^- geometry. The χ_2 dihedral prefers either the catalytic Nt or noncatalytic Ng^+ geometries in the different GOx variants. The normalized integrated distributions are shown as red curves on either side of the panels.

F9 mutant bears the M556V mutation and, therefore, resembles the **Pv** variant, where the catalytic conformation is energetically more favorable. However, the absence of R537K and the addition of two other mutations (R37K/V106I) narrows the separation between the two conformations, making it more similar to **WT**. This indicates a lower energy barrier for the transition between the catalytic and noncatalytic states and, hence, slower catalysis than for the **A2** mutant (increase in k_{cat} 1.8 times and in efficiency 2.6 times in comparison to **WT**).

To properly quantify the energy barrier between the catalytic and noncatalytic His516 states, we performed umbrella sampling MD simulations of the **WT** and **A2** variants in χ_2 space. The global minimum of **WT** is at 60° (Ng^+), and it is only 0.2 kcal mol^{-1} more stable than the minimum at 160° (Nt), which corresponds to an equilibrium mixture of 60:40 of noncatalytic to catalytic conformations at room temperature (Figure 7). Furthermore, having a rather high energy barrier of 2.9 kcal mol^{-1} for the Ng^+ to Nt transition means that a significant amount of time is lost on making GOx conformationally fit for catalysis, indicating that **WT** GOx is not an optimal catalyst. In **A2**, on the other hand, the catalytic conformation is 30 times more probable, as it is 2.0 kcal mol^{-1} more stable than the noncatalytic form. Furthermore, the energy barrier for the conversion of the noncatalytic to the catalytic state is significantly lower (1.8 kcal mol^{-1}) than for **WT**. Thus, the catalytic conformation can be achieved much more easily than for **WT** while the transformation back to the Ng^+ conformation is slow due to the barrier of 3.8 kcal mol^{-1}.

The extensive US-MD simulations corroborate the relative energies of the minima determined from the HREX-MD simulations. Umbrella sampling, however, performs better in estimating barrier heights. On the other hand, HREX-MD is a very convenient and cost-effective technique and can thus represent an excellent screening method for identifying good enzyme designs that involve potentially flexible active site
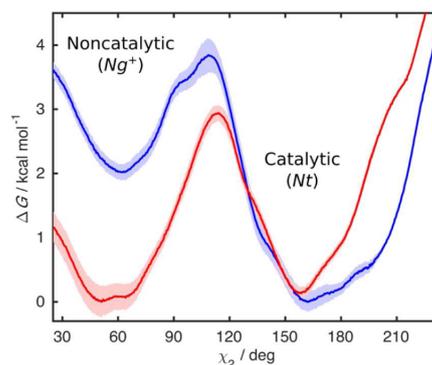
Figure 7. Free energy for the rotation around the χ_2 dihedral angle of His516: (red) **WT** GOx; (blue) **A2** mutant. The shaded areas around the free energy profiles represent the errors estimated using the blocking procedure.

residues. Furthermore, it is fast enough to be used for guiding directed evolution experiments.

■ **CONCLUSIONS**

Glucose oxidase is an important industrial catalyst for which many mutations were proposed to enhance various properties. However, not much is known about the mode of action of these mutations. In order to fill this gap, we solved the first crystal structures of GOx mutants from *A. niger* and performed an extensive molecular dynamics investigation based on a total of 5.6 μs simulation time to correlate mutations with kinetic data. The crystal structures of the mutants **A2** and **F9** revealed molecular oxygen to be present at the active site and suggest that the side chain of His516, which is of utmost importance for

the enzymatic reaction, is preorganized in the catalytic conformation and less flexible than in the wild-type GOx.

In the MD simulations, the most active mutant (**A2**) shows significant anticorrelated motions between secondary structure elements caused by the T30V and I94V mutations and both correlated and anticorrelated motions resulting from the R537K and M556V mutations. This long-range dynamic effect reduces the volume of the active site, which has a positive influence on catalytic efficiency. From all GOx variants studied here, **A2** possesses the tightest and least floppy active site, where protein contacts stabilize the optimal geometry of glucose for its interconversion to gluconolactone. Our MD simulations thus confirm the observation from the crystal structures that His516 is flexible in the **WT** and more rigid in the mutants. Furthermore, we find that His516 can flip between the two substates, catalytic and noncatalytic. To study the relative populations of the two substates and barriers between them, we employed Hamiltonian replica exchange and umbrella sampling MD simulations. While both substates are equally populated in the **WT** enzyme, the most favorable conformation of His516 in the **A2** mutant is the catalytic form. This results from the M556V mutation that reduces the size of a cavity in the vicinity of the active site and therefore restrains the movements of His516. As the turnover number of the discussed GOx variants is already very high (and probably very close to the theoretical limit), further design should be directed toward mutations that could provide higher binding affinities for glucose: mutations either in the first shells around the active site or at further positions (e.g., at the protein surface) that could positively modulate the correlated and anticorrelated motions.

From our study, we find that US-MD performs much better in estimating barrier heights, but both US-MD and HREX-MD are equally good for predicting positions and relative populations of the enzyme substates. Considering the relatively low computational cost and ease of use of HREX-MD simulations, we conclude that this method represents an attractive tool for in silico screening of enzyme variants involving flexible residues in the active sites.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscatal.7b01575.

Umbrella sampling MD setup, additional figures showing electron densities for His516 of various wild-type and mutant GOx structures, and further MD simulation results (PDF)

AUTHOR INFORMATION

Corresponding Authors
*E-mail for K.H.: kurt.hoffmann@rwth-aachen.de.
*E-mail for B.S.: b.strodel@fz-juelich.de.

ORCID

Dušan Petrović: 0000-0002-1834-7358
Shina Caroline Lynn Kamerlin: 0000-0002-3190-1173
Birgit Strodel: 0000-0002-8734-7765

Present Address
¹D.F.: Aquila Biolabs GmbH, Arnold-Sommerfeld-Ring 2, 52499 Baesweiler, Germany.

Notes
The authors declare no competing financial interest.

REFERENCES

(1) Bankar, S. B.; Bule, M. V.; Singhal, R. S.; Ananthanarayan, L. *Biotechnol. Adv.* **2009**, *27*, 489−501.
(2) Dijkman, W. P.; de Gonzalo, G.; Mattevi, A.; Fraaije, M. W. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 5177−5188.
(3) Tracewell, C. A.; Arnold, F. H. *Curr. Opin. Chem. Biol.* **2009**, *13*, 3−9.
(4) Harms, M. J.; Thornton, J. W. *Nat. Rev. Genet.* **2013**, *14*, 559−571.
(5) Knowles, J. R.; Albery, W. J. *Acc. Chem. Res.* **1977**, *10*, 105−111.
(6) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. *Biochemistry* **2011**, *50*, 4402−4410.
(7) Arnold, F. H.; Wintrode, P. L.; Miyazaki, K.; Gershenson, A. *Trends Biochem. Sci.* **2001**, *26*, 100−106.
(8) Tokuriki, N.; Jackson, C. J.; Afriat-Jurnou, L.; Wyganowski, K. T.; Tang, R.; Tawfik, D. S. *Nat. Commun.* **2012**, *3*, 1257.
(9) Wilson, R.; Turner, A. P. F. *Biosens. Bioelectron.* **1992**, *7*, 165−185.
(10) Mattevi, A. *Trends Biochem. Sci.* **2006**, *31*, 276−283.
(11) Zhu, Z.; Momeu, C.; Zakhartsev, M.; Schwaneberg, U. *Biosens. Bioelectron.* **2006**, *21*, 2046−2051.
(12) Holland, J. T.; Lau, C.; Brozik, S.; Atanassov, P.; Banta, S. *J. Am. Chem. Soc.* **2011**, *133*, 19262−19265.
(13) Holland, J. T.; Harper, J. C.; Dolan, P. L.; Manginell, M. M.; Arango, D. C.; Rawlings, J. A.; Apblett, C. A.; Brozik, S. M. *PLoS One* **2012**, *7*, e37924.
(14) Arango Gutierrez, E.; Mundhada, H.; Meier, T.; Duefel, H.; Bocola, M.; Schwaneberg, U. *Biosens. Bioelectron.* **2013**, *50*, 84−90.
(15) Ostafe, R.; Prodanovic, R.; Nazor, J.; Fischer, R. *Chem. Biol.* **2014**, *21*, 414−421.
(16) Leskovac, V.; Trivić, S.; Wohlfahrt, G.; Kandrač, J.; Peričin, D. *Int. J. Biochem. Cell Biol.* **2005**, *37*, 731−750.
(17) Roth, J. P.; Klinman, J. P. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 62−67.
(18) Herschlag, D.; Natarajan, A. *Biochemistry* **2013**, *52*, 2050−2067.
(19) Hanoian, P.; Liu, C. T.; Hammes-Schiffer, S.; Benkovic, S. *Acc. Chem. Res.* **2015**, *48*, 482−489.
(20) Agarwal, P. K.; Doucet, N.; Chennubhotla, C.; Ramanathan, A.; Narayanan, C. *Methods Enzymol.* **2016**, *578*, 273−297.
(21) Elias, M.; Wieczorek, G.; Rosenne, S.; Tawfik, D. S. *Trends Biochem. Sci.* **2014**, *39*, 1−7.
(22) Smith, A. J. T.; Müller, R.; Toscano, M. D.; Kast, P.; Hellinga, H. W.; Hilvert, D.; Houk, K. N. *J. Am. Chem. Soc.* **2008**, *130*, 15361−15373.
(23) Warshel, A. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 5250−5254.
(24) Pisliakov, A. V.; Cao, J.; Kamerlin, S. C. L.; Warshel, A. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 17359−17364.
(25) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev.* **2006**, *106*, 3210−3235.

(26) Roca, M.; Oliva, M.; Castillo, R.; Moliner, V.; Tuñón, I. *Chem. - Eur. J.* **2010**, *16*, 11399−11411.

(27) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. *Nature* **2011**, *477*, 111−114.

(28) Keedy, D. A.; Kenner, L. R.; Warkentin, M.; Woldeyes, R. A.; Hopkins, J. B.; Thompson, M. C.; Brewster, A. S.; Van Benschoten, A. H.; Baxter, E. L.; Uervirojnangkoorn, M.; McPhillips, S. E.; Song, J.; Alonso-Mori, R.; Holton, J. M.; Weis, W. I.; Brunger, A. T.; Soltis, S. M.; Lemke, H.; Gonzalez, A.; Sauter, N. K.; Cohen, A. E.; van den Bedem, H.; Thorne, R. E.; Fraser, J. S. *eLife* **2015**, *4*, e07574.

(29) Mabbitt, P. D.; Correy, G. J.; Meirelles, T.; Fraser, N. J.; Coote, M. L.; Jackson, C. J. *Biochemistry* **2016**, *55*, 1408−1417.

(30) Osuna, S.; Jiménez-Osés, G.; Noey, E. L.; Houk, K. N. *Acc. Chem. Res.* **2015**, *48*, 1080−1089.

(31) Ruscio, J. Z.; Kohn, J. E.; Ball, K. A.; Head-Gordon, T. *J. Am. Chem. Soc.* **2009**, *131*, 14111−14115.

(32) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700−5725.

(33) Wohlfahrt, G.; Witt, S.; Hendle, J.; Schomburg, D.; Kalisz, H. M.; Hecht, H. J. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 969−977.

(34) Hecht, H. J.; Kalisz, H. M.; Hendle, J.; Schmid, R. D.; Schomburg, D. *J. Mol. Biol.* **1993**, *229*, 153−172.

(35) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 235−242.

(36) Kommoju, P.-R.; Chen, Z.; Bruckner, R. C.; Mathews, F. S.; Jorns, M. S. *Biochemistry* **2011**, *50*, 5521−5534.

(37) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 486−501.

(38) Lovell, S. C.; Davis, I. W.; Arendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 437−450.

(39) Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.

(40) Trott, O.; Olson, A. J. *J. Comput. Chem.* **2009**, *31*, 455−461.

(41) Sousa da Silva, A. W.; Vranken, W. F. *BMC Res. Notes* **2012**, *5*, 367.

(42) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision A.02*; Gaussian, Inc., Wallingford, CT, 2009.

(44) Todde, G.; Hovmöller, S.; Laaksonen, A.; Mocci, F. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 2353−2363.

(45) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845−854.

(46) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.

(47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(48) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.

(49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(50) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.

(51) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(52) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182.

(53) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.

(54) Bussi, G. *Mol. Phys.* **2014**, *112*, 379−384.

(55) Grossfield, A. *WHAM: the weighted histogram analysis method, v. 2.0.9.*

(56) Grossfield, A.; Zuckerman, D. M. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23−48.

(57) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.

(58) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. *Bioinformatics* **2006**, *22*, 2695−2696.

(59) Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A. *J. Mol. Graphics Modell.* **2011**, *29*, 773−776.

(60) *The PyMOL Molecular Graphics System, Version 1.6*; Schrödinger LLC.

(61) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(62) Kalisz, H. M.; Hecht, H.-J.; Schomburg, D.; Schmid, R. D. *J. Mol. Biol.* **1990**, *213*, 207−209.

(63) Bretthauer, R. K.; Castellino, F. J. *Biotechnol. Appl. Biochem.* **1999**, *30*, 193−200.

(64) Mellquist, J. L.; Kasturi, L.; Spitalnik, S. L.; Shakin-Eshleman, S. H. *Biochemistry* **1998**, *37*, 6833−6837.

(65) Bhabha, G.; Biel, J. T.; Fraser, J. S. *Acc. Chem. Res.* **2015**, *48*, 423−430.

(66) Fenwick, R. B.; Orellana, L.; Esteban-Martín, S.; Orozco, M.; Salvatella, X. *Nat. Commun.* **2014**, *5*, 4070.

(67) Ma, H.; Szeler, K.; Kamerlin, S. C. L.; Widersten, M. *Chem. Sci.* **2016**, *7*, 1415−1421.

(68) Luo, J.; Bruice, T. C. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 13152−13156.

(69) Fernández, I. S.; Ruíz-Dueñas, F. J.; Santillana, E.; Ferreira, P.; Jesús Martínez, M.; Martínez, A. T.; Romero, A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 1196−1205.

(70) Li, J.; Vrielink, A.; Brick, P.; Blow, D. M. *Biochemistry* **1993**, *32*, 11507−11515.

(71) Martin Hallberg, B.; Henriksson, G.; Pettersson, G.; Divne, C. *J. Mol. Biol.* **2002**, *315*, 421−434.

(72) Yoshida, H.; Sakai, G.; Mori, K.; Kojima, K.; Kamitori, S.; Sode, K. *Sci. Rep.* **2015**, *5*, 13498.

(73) Hernández-Ortega, A.; Lucas, F.; Ferreira, P.; Medina, M.; Guallar, V.; Martínez, A. T. *Biochemistry* **2012**, *51*, 6595−6608.

(74) Scouras, A. D.; Daggett, V. *Protein Sci.* **2011**, *20*, 341−352.

Paper II

How accurately do force fields represent the protein side chain ensembles?

by D. Petrović, X. Wang, and B. Strodel

Proteins: Struct., Funct., Bioinf. **2018**, *doi: 10.1002/prot.25525*

How accurately do force fields represent protein side chain ensembles?

Dušan Petrović,[1,2] Xue Wang,[1,3] and Birgit Strodel[1,3,*]

[1] Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany

[2] Department of Cell and Molecular Biology, Uppsala University, BMC Box 596, 751 24 Uppsala, Sweden

[3] Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

* Corresponding author: Dr. Birgit Strodel, Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany, b.strodel@fz-juelich.de

Dušan Petrović

ABSTRACT: Although the protein backbone is the most fundamental part of the structure, the fine-tuning of side-chain conformations is important for protein function, for example, in protein-protein and protein-ligand interactions, and also in enzyme catalysis. While several benchmarks testing the performance of protein force fields for side chain properties have already been published, they often considered only a few force fields and were not tested against the same experimental observables; hence, they are not directly comparable. In this work, we explore the ability of twelve force fields, which are different flavors of AMBER, CHARMM, OPLS, or GROMOS, to reproduce average rotamer angles and rotamer populations obtained from extensive NMR studies of the ^3J and residual dipolar coupling constants for two small proteins: ubiquitin and GB3. Based on a total of 196 μs sampling time, our results reveal that all force fields identify the correct side chain angles, while the AMBER and CHARMM force fields clearly outperform the OPLS and GROMOS force fields in estimating rotamer populations. The three best force fields for representing the protein side chain dynamics are AMBER 14SB, AMBER 99SB*-ILDN, and CHARMM36. Furthermore, we observe that the side chain ensembles of buried amino acid residues are generally more accurately represented than those of the surface exposed residues.

2

76

INTRODUCTION

A protein structure can be divided into two major parts: the backbone and the side chains. While the backbone forms the secondary and tertiary structures of a protein, the side chains distinguish the amino acids from one another and therefore determine the protein structure. Moreover, the side chain ensembles and dynamics can often be crucial for the fine-tuning of protein behavior in, e.g., protein–ligand[1,2] and protein–protein interactions,[3] and for events like enzyme catalysis requiring stringent geometrical arrangements.[4–7] Side chains are usually more flexible than the backbone, and they can easily transition between several rotamers on the pico- to nanosecond timescales, which form a network of microstates.[8] Some side chains, however, shuffle their conformations on the micro- to millisecond timescales, which can define protein macrostates.[9,10] The side chain motion can be described in terms of χ dihedral angles and all residues, apart from Gly and Ala, have one to five such angles depending on the structural complexity of an amino acid. The χ_1 angle is defined by the N-Cα-Cβ-Cγ atoms, except for residues branching at the γ-position (Ile, Val, and Thr) or those with a heteroatom (Cys and Ser), where the fourth atom is instead Cγ_1/Oγ_1/Sγ/Oγ. In χ_1-space, side chains are typically stable in the staggered conformation, clustering around $\pm 60°$ or $180°$, as shown for Thr in Figure 1. The exception is Pro, which can take only g^+/g^- conformers due to steric restrictions as a result of the fusion of the side chain to the N atom of the backbone creating pyrrolidine.[11]

X-ray crystallography can identify alternative side chain rotamers from the conformational averaging observed in the electron density.[12] However, the cryogenic temperatures typically required in crystallography can significantly impact the side chain conformational ensembles, and recent crystal structures resolved at room temperature shed light on other, previously hidden conformations.[10,13] Further information on the dynamics at room temperature can be obtained with complementary methods, such as nuclear magnetic resonance (NMR)

3

spectroscopy or molecular dynamics (MD) simulations.[10,14–17] Other computational methods for the prediction of side chain geometries include Monte Carlo simulations[18] and bioinformatics analysis using neural networks.[19]

MD simulations have gained a significant importance in the fields of biochemistry and biophysics.[20,21] Biomolecules, such as proteins, nucleic acids, lipids, or carbohydrates are often treated at the atomic level, where force fields (FFs) are used to calculate the potential energy of a system. Although the mathematical models underlying these biomolecular FFs resemble each other, their parameters were derived in various fashions, often to reproduce different experimental or quantum mechanical properties.[22] Many protein FFs are currently available, though some flavors of AMBER, CHARMM, OPLS, and GROMOS are the most popular choices. Several studies were conducted to validate if FFs perform well in reproducing specific experimental observables. Some of these benchmarks were focused on probing the NMR observables of backbone structure and dynamics, such as chemical shifts or three-bond scalar coupling (^3J) and residual dipolar coupling (RDC) constants.[22–25] Another study compared the structural space sampled during MD simulations directly in order to quantify the overlap between the conformational ensembles.[26] Keller and co-workers used Markov state models to analyze how accurately different FFs represent dynamic properties of peptides.[27] Our and other groups published FF benchmarks for the simulation of intrinsically disordered proteins.[28–31]

Regarding the performance of FFs with respect to side chains, some of them were tested for reproducing the χ_1 propensities in dipeptides,[32] and tetrapeptides benchmarked against Protein Data Bank (PDB) based coil libraries,[33,34] where major differences were found between AMBER 03 and OPLS-AA, and both FFs were not able to correctly predict the χ_1 distributions. FFs were also tested on reproducing the NMR order parameter of the methyl symmetry axis in several small proteins.[35] A constant improvement in the modeling of side

4

78

chains with current FFs can be seen from several papers describing the development of novel FFs. AMBER 99SB-ILDN was shown to perform better than the ILDN-uncorrected variant in reproducing ^3J constants for Ile, Leu, Asp, and Asn residues.[36] AMBER 14SB outperforms the older AMBER flavors in reproducing experimental scalar couplings for three small proteins.[37] For ubiquitin, ^3J coupling constants and S^2 relaxation order parameters are better described by CHARMM 36 than by the older FF flavor, CHARMM 27.[38] Finally, the new OPLS-AA/M force field also shows a significant improvement in estimating rotamer populations.[39]

Unlike the previous FF benchmarks that compared only a small number of FFs and did not test the same observables and experimental data sets (i.e., their results are not directly comparable), we present a novel benchmark of the common flavors of the AMBER, CHARMM, OPLS, and GROMOS FFs. We test their performances against unique NMR data sets for the side chain ensembles of two small proteins: ubiquitin and the third IgG-binding domain of protein G (GB3), which are depicted in Figure 2 (PDB ID: 1UBQ[40] and 2OED[41]). Our benchmark study is based on thirteen different FFs and a total of 196 μs MD sampling, which should capture all protein dynamics as these two proteins show little motion beyond the microsecond timescale.[23,42]

METHODS

Benchmark systems. Ubiquitin and GB3 were selected as test systems for benchmarking the side chain dynamics with several commonly used FF flavors. For both proteins, 10 μs MD trajectories based on AMBER 03,[43] 03*,[43,44] 99SB-ILDN,[36,45] 99SB*-ILDN,[36,44,45] CHARMM 22,[46] 22*,[46–48] 27,[46,48] and OPLS-AA[49] with TIP3P[50] (or CHARMM modified TIP3P[46]) water model were kindly provided by D. E. Shaw Research.[23] In addition, we examined 10 μs MD trajectories of both proteins simulated with AMBER 99SB-ILDN and the TIP4P-D water model,[51] also provided by D. E. Shaw Research.[51] We further performed additional 2 μs

5

simulations with AMBER 14SB,[37] FB15,[52] CHARMM 36,[38] and GROMOS 54a7[53] for both proteins.

MD setup. The ubiquitin and GB3 topologies were created using the coordinates from the PDB entries 1D3Z[54] and 1P7E,[41] respectively. The AMBER 14SB topologies were created using *tleap* from AMBER,[55] with the TIP3P water model,[50] and were converted to the GROMACS format using *parmed*. The AMBER FB15, CHARMM 36, and GROMOS 54a7 topologies were built in GROMACS 5.1.1,[56] with the TIP3P-FB,[57] CHARMM modified TIP3P,[46] and SPC[58] solvent models, respectively.

The protein was centered in a cubic box with an edge length of 58 Å, and the total charge was neutralized with sodium ions. The system was minimized for 2,000 steps (or the convergence criterion of 500 kJ mol^{-1} nm^{-1}) using the steepest descent algorithm. During the subsequent equilibration MD run, all protein atoms were restrained to their positions with a force constant of 1,000 kJ mol^{-1} nm^{-2}. The system was heated over 200 ps to 300 K and simulated at 300 K for an additional 200 ps, using the v-rescale thermostat.[59] The restraints were then scaled down to 5 kJ mol^{-1} nm^{-2} over 1 ns NPT equilibration, and an additional 1 ns unrestrained NPT equilibration was performed, where the Berendsen barostat[60] was used to maintain the pressure at 1 bar.

The production MD was run for 2 μs in the NVT ensemble using a 2 fs time step integration, and the coordinates were saved every 20 ps. Periodic boundary conditions were imposed on the system, and the electrostatic interactions were treated with the particle mesh Ewald method.[61] The short-range Coulomb and Lennard-Jones interactions were calculated with a cutoff of 9.5 Å. The bond lengths were constrained with the LINCS algorithm. The protein stability during the MD simulation was tested based on the radius of gyration (Rg), the backbone root-mean-square deviation (RMSD), and the Cα root-mean-square fluctuations

6

80

(RMSF), which were calculated using GROMACS tools. In addition, the χ_1 dihedral RMSD (RMSDih) was calculated for the whole protein as:

$$RMSDih = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\arctan\left[\sin\left(\chi_i^t - \chi_i^0\right), \cos\left(\chi_i^t - \chi_i^0\right)\right]\right)^2}$$

where χ_i^0 and χ_i^t are the χ_1 angles at the initial frame and at time t, respectively.

Force field comparison. The χ_1 dihedrals were directly measured from MD simulations, and the selected FFs were benchmarked against experimentally determined χ_1 conformational distributions, which were obtained by analyzing multiple ^3J and RDCs. The average χ_1 angles and corresponding rotamer populations were determined by Bax and co-workers for the methyl-bearing residues (Val, Ile, and Thr) of ubiquitin,[8] and recently for all residues but Gly and Ala of GB3.[62] From the MD simulations, the χ_1 angles were calculated with the MDtraj library[63] and divided into three bins representing the rotamers g^+ ($0º \leq \chi_1 < 120º$), t ($120º \leq \chi_1 < 240º$), and g^- ($240º \leq \chi_1 < 360º$). The average angle per rotamer of a residue was calculated as mean of all data points belonging to the corresponding bin, and the rotamer population was calculated as the ratio of the number of states in the bin in question and the total number of conformations saved during that MD trajectory. To evaluate the quality of each force field versus the experimental data set, we calculated the coefficient of determination (R^2) for a linear fit and the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|$$

where n is the number of data points, while y_i and f_i are the actual experimental and simulated outcomes, respectively.

7

Additionally, we determined the side-chain ^3J coupling constants using the Karplus parameters developed by Schmidt et al. (for $^3J_{H\alpha,H\beta}$ and $^3J_{C',C\gamma}$ for all side chains bearing these bonds)[64] and Bax et al. (for the γ-methyl groups in Val, Ile, and Thr),[8] and then calculated the RMSD from the experimentally measured ^3J constants for ubiquitin and GB3, compiled by Jorgensen et al.,[39] to further estimate the quality of the investigated FFs.

RESULTS

Ubiquitin. Ubiquitin is a protein consisting of 76 amino acid residues that has many important functions in cellular processes.[65] Its small size makes ubiquitin a very good model system for many structural investigations. Therefore, ubiquitin has been very well characterized using both liquid and solid state NMR[8,66–72] as well as MD simulations,[42,73–76] which makes it a good benchmark system for our study.

Methyl groups usually appear in the hydrophobic protein core, and they are often used in NMR spectroscopy to probe protein structure and dynamics.[77] In combination with labeling, these residues are sensitive reporters even in big proteins.[78,79] In the present study we focus on four Val, seven Ile, and six Thr residues from the ubiquitin primary sequence, for which extensive NMR data are available, together with the estimated rotamer populations.[8] The χ_1 dihedrals were measured from the MD trajectories and assigned to one of the three rotamers: g^+, t, or g^-. We investigate whether a FF can reproduce both the average angle of each rotamer of a residue and the rotamer populations as observed experimentally (Figure 3).

For the eight FFs, which were released between 1998 and 2011 and are included in our benchmark, high correlations with experiment—with MAEs of around 10°—are obtained for the average rotamer angles (Table 1). Here, CHARMM 27 (Figure 3a) slightly outperforms the other FFs, yet the differences are minor. However, the second test, on quantifying rotamer populations, points to big differences among the FFs (Table 1). As in the previous test, CHARMM 27 performs the best, with a very high correlation between experiment and

8

simulation (R^2 = 0.9544) and an MAE of only 6% (Figure 3b). The second-best FF is AMBER 99SB*-ILDN with an R^2 of 0.9268 and MAE of 8%. OPLS-AA and AMBER 03, on the other hand, are clearly underperforming, with MAEs of 18% and 23%, respectively. A clear trend of improvement in rotamer populations for the residues under study (i.e., Val, Ile, and Thr) can be observed for the star-corrected FFs (i.e., AMBER 03*, 99SB*-ILDN, and CHARMM 22*) over their uncorrected counterparts. The star-correction was not made to the side chains but to the backbone, for improving the α-helical propensities of polypeptides,[44] as many of the older force fields were found to sample the α-helical region too much.[24]

Several FFs were published after 2011, which we denote as *post-2011* FFs here and of which we included following FFs in the current benchmark: GROMOS 54a7 released in 2011,[53] CHARMM 36 from 2013,[38] and AMBER 14SB[37] and FB15[52] from 2015 and 2017, respectively. The results in Table 1 show that for ubiquitin AMBER14 SB performs better than the older AMBER flavors, with R^2 = 0.9381 and an MAE of only 8%. Regarding the CHARMM FFs one can conclude that CHARMM 27 performs slightly better than CHARMM 36, yet the differences in their performance are minor. GROMOS 54a7, on the other hand, is not able to correctly reproduce the rotamer populations as R^2 = 0.5769 and an MAE of 19% show, which are similar to the corresponding values obtained with OPLS-AA.

GB3. The other test case in our benchmark is the third IgG-binding domain of protein G. GB3 is a 56 residue protein extensively studied experimentally by NMR[41,62,80–83] and, hence, often used for FF validation studies.[23,26,36,37,39,52] In a recent publication, Bax and co-workers estimated the conformational distribution of side chains in GB3 using a model-free analysis of RDCs.[62] Since average rotamer χ_1 angles and populations were determined for all residues but Gly and Ala, which lack the χ_1 dihedral, this data set represents an invaluable material for our FF benchmark. The residues L12, K19, and T25 were omitted from the analysis as no NMR results are available for their χ_1 angles.[62] As for ubiquitin, we investigate whether a FF can

9

reproduce the average angle of each rotamer in a residue and the corresponding rotamer populations, and, in addition, determine whether it correctly identifies the most stable rotamer (Figure 4).

While most *post-2011* FFs provided stable trajectories of GB3 (Figure 5a, average RMSD < 1 Å), GROMOS 54a7 did not. From three GROMOS 54a7 simulations initiated with different random seeds, in two of them the RMSD raised to more than 5 Å after 0.15 and 0.65 μs (Figure S9). Only the third simulation was relatively stable, where the average RMSD remained at ~2.5 Å. We included only this GROMOS 54a7 trajectory in our FF benchmark.

Similar to the previous test case, all FFs yield exceptionally high correlations with the experimentally measured average rotamer angles of GB3 (Figure 4a). CHARMM 22* and AMBER 03* slightly outperform the other FFs with regard to the overall correlation with experiment, while AMBER 14SB produces the smallest average error (Table 2). We further evaluated if a FF can correctly identify the lowest free energy rotamer (i.e., the top rotamer). To this end, we checked if there is a correlation between the angles of the most populated rotamers identified experimentally and from simulation, respectively. At this stage, the performance of the FFs started to considerably differentiate. AMBER 99SB*-ILDN clearly outperforms all the other older FFs, while among the older CHARMM flavors CHARMM 27 performs better than the other two. Our additional MD simulations indicate that the *post-2011* AMBER and CHARMM FFs also perform quite well in this test.

Finally, we tested how accurately the rotamer populations are predicted (Figure 4c). Like for ubiquitin, the star correction to the FFs has a positive influence, although its magnitude is not as high as in the previous test case. Once again, AMBER 99SB*-ILDN significantly outperforms the other older FFs (R^2 = 0.6959 and MAE = 14%), followed by CHARMM 22* (R^2 = 0. 5163 and MAE = 19%) and CHARMM 27 (R^2 = 0.4797 and MAE = 18%). OPLS-AA, on the other hand, clearly underperforms in this test with R^2 ~ 0.1 and an MAE of more

10

84

than 30%. CHARMM 22 has an even smaller R^2 value because GB3 unfolded during that simulation. The *post-2011* CHARMM and AMBER FFs show a significant improvement over their older flavors. The biggest improvement is seen for AMBER 14SB, with R^2 of 0.8017 and MAE of only 12%, which also produced the smallest deviation from experiment for the average rotamer angle. Thus, this FF represents the clear winner in our benchmark. Another *post-2011* FF, namely GROMOS 54a7, on the other hand, performed poorly in this benchmark as it has a rather low R^2 of ~0.2 and high MAE of 26% for the rotamer populations.

^3J coupling constant analysis. The alternative approach to estimate the FF quality is to directly compare experimentally measured side-chain ^3J coupling constants with those estimated from MD simulations using the Karplus relation. To that end, we used more than 200 unique ^3J coupling constants, i.e., $^3J_{H\alpha,H\beta}$, $^3J_{C',C\gamma}$, as well as the γ-methyl group coupling constants ($^3J_{C',C\gamma}$ and $^3J_{N,C\gamma}$) to calculate the RMSD of simulations from the experimental values (Table 3). The conclusions drawn from the ^3J coupling analysis support those obtained from the rotamer population analysis. While CHARMM 22 and OPLS-AA lead to particularly high RMSD values (2.9 Hz and 2.4 Hz for ubiquitin and GB3, respectively, modeled with CHARMM 22, and 2.7 Hz and 2.0 Hz for ubiquitin and GB3, respectively, modeled with OPLS-AA), the previously identified most accurate FFs perform much better in this test, too. Namely, the RMSD values for CHARMM 36 are only 1.9 Hz and 1.8 Hz for ubiquitin and GB3, respectively. Furthermore, the RMSD values are 2.1 Hz and 1.5 Hz for the two proteins modeled with the AMBER 99SB-ILDN and 99SB*-ILDN, and only 2.0 Hz and 1.4 Hz for ubiquitin and GB3, respectively, simulated with AMBER 14SB.

Water model impact. In recent years it has become apparent that the quality of a FF does not only depend on the protein FF itself but also on the water model. For instance, the water model can have considerable effects on properties such as the compactness of a protein, which

11

plays especially important roles for proteins that can be partially or completely disordered. It was found that water models such as TIP3P typically used in MD simulations significantly underestimate London dispersion interactions, and to correct for this shortcoming the protein-water interactions were reparameterized leading to new water models such as TIP4P-D.[51,84] In order to test whether the water model has an effect on the performance of a protein FF for representing side chains, we analyzed 10 μs of MD simulations of ubiquitin and GB3 modeled with AMBER 99SB-ILDN and TIP4P-D and compared these results with those from the simulations using the same protein FF but in combination with TIP3P. As Tables 1-3 show, the combination AMBER 99SB-ILDN/ TIP3P produces quite good but also not the best results so that the usage of TIP4P-D could in principle lead to both an improvement or deterioration of the performance. Another advantage of using this combination is that TIP4P-D has not been used for the training of AMBER 99SB-ILDN, which should allow us to disentangle the performance of a protein FF from that of the water model. However, the results for AMBER 99SB-ILDN/TIP4P-D in Tables 1–3 reveal that the water model has more or less no effect on the performance of AMBER 99SB-ILDN for representing side-chain ensembles as the results for TIP3P and TIP4P-D are rather similar. This is different to the observations made for the backbone since Shaw and co-workers had found[51] that while the native states of both ubiquitin and GB3 were stable on the 10 μs time scale, both proteins were more dynamic than when simulated with TIP3P and in somewhat worse agreement with NMR data. While this loss in performance for the protein backbone due to TIP4P-D does not lead to a worsening of the side-chain representation, the better description of the protein–water interactions with TIP4P-D does also not improve the modeling of the side chain ensembles as our results indicate.

χ_1 convergence. While both ubiquitin and GB3 are reported to show little motion at the long timescales,[23,42] we analyzed the convergence of all MD simulations using χ_1 RMSDih.

Figure 5b shows average RMSDih for ubiquitin and GB3 using the twelve tested FFs and indicates that less accurate FFs show more χ_1 dynamics and converge more slowly since side chains are in a changing environment, e.g., for CHARMM 22 where GB3 unfolded. FFs showing smaller deviations from experiments also show lower side-chain dihedral dynamics, and the sampling usually converges within 1-2 μs, as indicated in Figures S12-S16.

DISCUSSION AND CONCLUSION

As already pointed out, for both test cases (i.e., ubiquitin and GB3) the average rotamer angles are well reproduced, clustering around ±60º and 180º. Both older and *post-2011* FFs maintain these angles with MAEs on the order of 9–13 Å, yielding high correlations to experiment with R^2 values of 0.96 or higher. Even when the protein unfolds as GB3 did after 1 μs in the simulation with CHARMM 22, the side chain conformations are still correctly modeled (R^2 = 0.9724 and MAE = 11.2°). This indicates that the FFs were parameterized to properly identify minima on the energy surface.

More challenging is to develop the proper FF parameters for correctly modeling the relative free energies (i.e., populations) of the rotamers. Our test whether the FFs can properly identify the most stable rotamers in GB3 indicates that AMBER 99SB*-ILDN performs the best, and it is closely followed by AMBER 99SB-ILDN, AMBER 14SB, and CHARMM 36. GROMOS 54a7 and OPLS-AA are at the bottom of this evaluation as they identified for more than 35% of the tested side chains the wrong top rotamer. Since the protein unfolded with CHARMM 22, only about 50% of the residues stayed in the configuration preferred in the folded state of GB3. It can also be noted that the star-corrected FFs (i.e., AMBER 03*, AMBER 99SB*-ILDN and CHARMM 22*) perform better than their uncorrected counterparts.

13

Beyond the identification of the top rotamer, the most rigorous test is to check how well the FFs can estimate the rotamer populations. The case of ubiquitin showed that AMBER 99SB-ILDN and 99SB*-ILDN as well as CHARMM 22* and 27 reproduce the populations of the methyl-bearing residues Val, Ile, and Thr with a high correlation of $R^2 > \sim 0.9$. Most of the Val and Ile side chains are located in the protein interior. However, for the surface exposed Val70 these four FFs fail to correctly model the rotamer populations, producing population deviations for the t and g^- rotamers. On the other side, even though all six Thr side chains are surface exposed, their populations are well modeled by the four FFs. The population plots for GB3 (Figure S5) are noisier and have generally lower R^2 values than for ubiquitin (Figure S2), simply because the results for 43 residues of 13 types are shown. From the older FFs, it is again AMBER 99SB-ILDN and 99SB*-ILDN as well as CHARMM 22* and 27, which give the best agreement with experiment, yielding R^2 values on the order of 0.5–0.7. The novel CHARMM 36 FF significantly improves the rotamer populations compared to the older CHARMM versions. The most significant improvement is, however, obtained with AMBER 14SB, where only a few rotamer populations fall outside of the ±20% confidence interval, giving rise to a correlation of $R^2 > 0.8$. GB3 contains 15 residue types, as it has no Ser, Cys, Pro, Arg, and His residues. There are one Trp, three Phe, and three Tyr residues, which are all located in the protein interior and their rotamer populations are generally well represented by six of the FFs: AMBER 99SB-ILDN, 99SB*-ILDN, and 14SB, and CHARMM 22*, 27, and 36. The side chain ensembles of the other buried residues (Ile, Leu, and Val) are also well described by these six FFs, which is not always the case for the surface exposed residues Val21, and Val41. The representation of the surface exposed residues is generally somewhat less accurate with these six FFs. Besides the N-terminal Met1, where, instead of the g^- rotamer, the t rotamer is often identified as the most stable rotamer, several charged residues (Lys, Asp, and Glu) are also often found as outliers.

14

In order to understand the origin of why surface-exposed and charged residues are in some cases not correctly modeled, we did some further testing. First, it should be noted that the sampling of the side chains has converged for all FFs (Figures S12-S16), which rules out the possibility that the low performance for particular residues is due to insufficient sampling. Second, we did not find a general correlation between residue flexibility and the performance of a FF for representing side chains. To this end, we calculated the Cα RMSF of each residue and calculated the Pearson correlation coefficient between these fluctuations and the RMSD between the ^3J coupling constants calculated for the side chains and those from experiment. All correlation coefficients are close to zero, indicating that there is no correlation. This means that FFs are in general able to also correctly reproduce the side-chain ensembles of flexible residues. Thus, the failure for some surface-exposed residues is likely to be due to the fact that they interact with water and not with other protein residues. However, as our results for TIP4P-D show, also the usage of an improved water model does not lead to a further improvement of the side-chain modeling.

In summary, our study revealed that, from the twelve investigated FFs, AMBER and CHARMM flavors clearly outperform OPLS-AA and GROMOS 54a7. From the AMBER FFs, flavors 99SB*-ILDN and 14SB are the clear winners. Regarding CHARMM, we find that CHARMM 22* and 27 perform in a similar capacity in estimating the side chain rotamer populations, while the new flavor, CHARMM 36, shows a clear improvement over the older versions, what is also supported by the analysis of the the ^3J coupling constants. We demonstrate that improved backbone parameters are sometimes enough for generating better side chain ensembles, as seen, for example, for the star-corrected FFs versus their star-uncorrected counterparts. Furthermore, the FFs that show less backbone dynamics (low backbone RMSD in Figure 5a), often also show less dihedral dynamics (low χ_1 RMSDih in Figure 5b), and these are the FFs that perform the best in estimating rotamer populations. The

15

most significant improvement for the side chain ensembles is, however, observed for the FFs for which not only the backbone but also the side chain torsion parameters have been optimized for reproducing either quantum mechanical energy surfaces or NMR observables, or both. The best example is AMBER 14SB, which comes top in our benchmark, that included adjustments to the backbone parameters and a complete refit of the side chain dihedral parameters.[37] Other examples that contain improvements in both backbone and side chains parameters, and which perform well in our side chain benchmark, are AMBER99SB*-ILDN and CHARMM 36. It should be noted that these two FFs have also been demonstrated to be good choices for modeling protein secondary and tertiary structures as well as conformational dynamics.[23,28,29,38,85] Moreover, it is worth noticing that, while for the parameterization of CHARMM 36 ^3J coupling constants of unfolded ubiquitin and GB3 were used, the parameterization of the χ_1 potentials in AMBER 14SB did not make use of these two proteins. From this fact we can conclude that the FF side-chain parameters are perfectly transferable from (unfolded) proteins included in the parameterization set to other (folded) proteins. Another finding of our study is that the side chains of buried residues are often more accurately represented than those of surface-exposed residues, possibly due to the confinement of buried residues in the protein interior, with more directed interactions helping to stabilize certain rotamers. For the surface-exposed residues, one may need to reparameterize protein–water interactions in order to reproduce the rotamer energies of the free amino acids. Nonetheless, our study demonstrates that the FFs with improved side chain parameters indeed allow to use atomistic MD simulations in applications where side chain conformations and flexibilities are pivotal, like in protein-ligand/protein interactions or catalysis.[7]

16

90

ACKNOWLEDGEMENTS

17

REFERENCES

(1) Najmanovich R, Kuttner J, Sobolev V, Edelman M. Side-chain flexibility in proteins upon ligand binding. *Proteins Struct Funct Genet* 2000;39:261–268.

(2) Gaudreault F, Chartier M, Najmanovich R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* 2012;28:i423–i430.

(3) Watkins AM, Bonneau R, Arora PS. Side-chain conformational preferences govern protein–protein interactions. *J Am Chem Soc* 2016;138:10386–10389.

(4) Lassila JK. Conformational diversity and computational enzyme design. *Curr Opin Chem Biol* 2010;14:676–682.

(5) Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN. Computational enzyme design. *Angew Chem Int Ed* 2013;52:5700–5725.

(6) Bhabha G, Biel JT, Fraser JS. Keep on moving: discovering and perturbing the conformational dynamics of enzymes. *Acc Chem Res* 2015;48:423–430.

(7) Petrović D, Frank D, Kamerlin SCL, Hoffmann K, Strodel B. Shuffling active site substate populations affects catalytic activity: the case of glucose oxidase. *ACS Catal* 2017;7:6188–6197.

(8) Chou JJ, Case DA, Bax A. Insights into the mobility of methyl-bearing side chains in proteins from $^3J_{CC}$ and $^3J_{CN}$ couplings. *J Am Chem Soc* 2003;125:8959–8966.

(9) Smith CA, Ban D, Pratihar S, Giller K, Schwiegk C, de Groot BL, Becker S, Griesinger C, Lee D. Population shuffling of protein conformations. *Angew Chemie Int Ed* 2015;54:207–210.

(10) Keedy DA, Kenner LR, Warkentin M, Woldeyes RA, Hopkins JB, Thompson MC, Brewster AS, Van Benschoten AH, Baxter EL, Uervirojnangkoorn M, McPhillips SE, Song J, Alonso-Mori R, Holton JM, Weis WI, Brunger AT, Soltis SM, Lemke H,

18

92

Gonzalez A, et al. Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *eLife* 2015;4:e07574.

(11)   Towse C-L, Rysavy SJ, Vulovic IM, Daggett V. New dynamic rotamer libraries: Data-driven analysis of side-chain conformational propensities. *Structure* 2016;24:187–199.

(12)   Miao Z, Cao Y. Quantifying side-chain conformational variations in protein structure. *Sci Rep* 2016;6:37024.

(13)   Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T. Hidden alternative structures of proline isomerase essential for catalysis. *Nature* 2009;462:669–673.

(14)   Chi CN, Vögeli B, Bibow S, Strotz D, Orts J, Güntert P, Riek R. A structural ensemble for the enzyme cyclophilin reveals an orchestrated mode of action at atomic resolution. *Angew Chemie Int Ed* 2015;54:11657–11661.

(15)   McGowan LC, Hamelberg D. Conformational plasticity of an enzyme during catalysis: Intricate coupling between cyclophilin a dynamics and substrate turnover. *Biophys J* 2013;104:216–226.

(16)   Esteban-Martín S, Fenwick RB, Salvatella X. Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdiscip Rev Comput Mol Sci* 2012;2:466–478.

(17)   Holliday MJ, Camilloni C, Armstrong GS, Vendruscolo M, Eisenmesser EZ. Networks of dynamic allostery regulate enzyme function. *Structure* 2017;25:276–286.

(18)   Bhowmick A, Head-Gordon T. A Monte Carlo method for generating side chain structural ensembles. *Structure* 2015;23:44–55.

(19)   Carbonell P, del Sol A. Methyl side-chain dynamics prediction based on protein structure. *Bioinformatics* 2009;25:2552–2558.

(20)   Karplus M, Lavery R. Significance of molecular dynamics simulations for life sciences. *Isr J Chem* 2014;54:1042–1051.

19

93

(21) De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *J Med Chem* 2016;59:4035–4061.

(22) Beauchamp KA, Lin Y-S, Das R, Pande VS. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 2012;8:1409–1414.

(23) Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. *PLoS One* 2012;7:e32131.

(24) Best RB, Buchete N-V, Hummer G. Are current molecular dynamics force fields too helical? *Biophys J* 2008;95:L07–L09.

(25) Li D-W, Brüschweiler R. Certification of molecular dynamics trajectories with NMR chemical shifts. *J Phys Chem Lett* 2010;1:246–248.

(26) Martín-García F, Papaleo E, Gomez-Puertas P, Boomsma W, Lindorff-Larsen K. Comparing molecular dynamics force fields in the essential subspace. *PLoS One* 2015;10:e0121114.

(27) Vitalini F, Mey ASJS, Noé F, Keller BG. Dynamic properties of force fields. *J Chem Phys* 2015;142:084101.

(28) Uluca B, Viennet T, Petrović D, Shaykhalishahi H, Weirich F, Gönülalan A, Strodel B, Etzkorn M, Hoyer W, Heise H. DNP-enhanced MAS NMR: A tool to snapshot conformational ensembles of α-synuclein in different states. *Biophys J* 2018;114:1614–1623.

(29) Carballo-Pacheco M, Strodel B. Comparison of force fields for Alzheimer's aβ$_{42}$: A case study for intrinsically disordered proteins. *Protein Sci* 2017;26:174–185.

(30) Palazzesi F, Prakash MK, Bonomi M, Barducci A. Accuracy of current all-atom force-fields in modeling protein disordered states. *J Chem Theory Comput* 2015;11:2–7.

20

94

(31) Rauscher S, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmüller H. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J Chem Theory Comput* 2015;11:5513–5524.

(32) Vymětal J, Vondrášek J. Critical assessment of current force fields. Short peptide test case. *J Chem Theory Comput* 2013;9:441–451.

(33) Jiang F, Han W, Wu Y-D. Influence of side chain conformations on local conformational features of amino acids and implication for force field development. *J Phys Chem B* 2010;114:5840–5850.

(34) Jiang F, Han W, Wu Y-D. The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. *Phys Chem Chem Phys* 2013;15:3413–3428.

(35) O'Brien ES, Wand AJ, Sharp K. On the ability of molecular dynamics force fields to recapitulate NMR derived protein side chain order parameters. *Protein Sci* 2016;25:1156–1160.

(36) Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99sb protein force field. *Proteins: Struct, Func, Bioinf* 2010;78:1950–1958.

(37) Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99sb. *J Chem Theory Comput* 2015;11:3696–3713.

(38) Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 2013;34:2135–2145.

(39) Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J Chem Theory Comput* 2015;11:3499–3509.

(40) Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution.

21

J Mol Biol 1987;194:531–544.

(41)   Ulmer TS, Ramirez BE, Delaglio F, Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J Am Chem Soc* 2003;125:9179–9191.

(42)   Lindorff-Larsen K, Maragakis P, Piana S, Shaw DE. Picosecond to millisecond structural dynamics in human ubiquitin. *J Phys Chem B* 2016;120:8313–8320.

(43)   Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 2003;24:1999–2012.

(44)   Best RB, Hummer G. Optimised molecular dynamics force fields applied to the helix−coil transition of polypeptides. *J Phys Chem B* 2009;113:9004–9015.

(45)   Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins Struct Funct Bioinf* 2006;65:712–725.

(46)   MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998;102:3586–3616.

(47)   Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* 2011;100:L47–L49.

(48)   Mackerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput*

96

Chem 2004;25:1400–1415.

(49) Kaminski GA, Friesner RA, Tirade-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001;105:6474–6487.

(50) Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79:926.

(51) Piana S, Donchev AG, Robustelli P, Shaw DE. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* 2015;119:5113–5123.

(52) Wang L-P, McKiernan KA, Gomes J, Beauchamp KA, Head-Gordon T, Rice JE, Swope WC, Martínez TJ, Pande VS. Building a more predictive protein force field: A systematic and reproducible route to Amber-FB15. *J Phys Chem B* 2017;121:4023–4039.

(53) Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, Mark AE, van Gunsteren WF. Definition and testing of the GROMOS force-field versions 54a7 and 54b7. *Eur Biophys J* 2011;40:843–856.

(54) Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 1998;120:6836–6837.

(55) Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. *J Comput Chem* 2005;26:1668–1688.

(56) van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *J Comput Chem* 2005;26:1701–1718.

23

(57)   Wang L-P, Martinez TJ, Pande VS. Building force fields: An automatic, systematic, and reproducible approach. *J Phys Chem Lett* 2014;5:1885–1891.

(58)   Berweger CD, van Gunsteren WF, Müller-Plathe F. Force field parametrization by weak coupling. Re-engineering SPC water. *Chem Phys Lett* 1995;232:429–436.

(59)   Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys* 2007;126:014101.

(60)   Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984;81:3684.

(61)   Darden T, York D, Pedersen L. Particle mesh ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 1993;98:10089.

(62)   Li F, Grishaev A, Ying J, Bax A. Side chain conformational distributions of a small protein derived from model-free analysis of a large set of residual dipolar couplings. *J Am Chem Soc* 2015;137:14798–14811.

(63)   McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang L-P, Lane TJ, Pande VS. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys J* 2015;109:1528–1532.

(64)   Pérez C, Löhr F, Rüterjans H, Schmidt JM. Self-consistent Karplus parametrization of ^{3}J couplings depending on the polypeptide side-chain torsion χ_1. *J Am Chem Soc* 2001;123:7081–7093.

(65)   Monia BP, Ecker DJ, Crooke ST. New perspectives on the structure and function of ubiquitin. *Nat Biotechnol* 1990;8:209–215.

(66)   Massi F, Grey MJ, Palmer AG. Microsecond timescale backbone conformational dynamics in ubiquitin studied with NMR r1 relaxation experiments. *Protein Sci* 2005;14:735–742.

(67)   Seidel K, Etzkorn M, Heise H, Becker S, Baldus M. High-resolution solid-state NMR

24

studies on uniformly [^{13}C,^{15}N]-labeled ubiquitin. *ChemBioChem* 2005;6:1638–1647.

(68)   Zech SG, Wand AJ, McDermott AE. Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. *J Am Chem Soc* 2005;127:8618–8626.

(69)   Farès C, Lakomek N-A, Walter KFA, Frank BTC, Meiler J, Becker S, Griesinger C. Accessing ns–µs side chain dynamics in ubiquitin with methyl RDCs. *J Biomol NMR* 2009;45:23–44.

(70)   Salvi N, Ulzega S, Ferrage F, Bodenhausen G. Time scales of slow motions in ubiquitin explored by heteronuclear double resonance. *J Am Chem Soc* 2012;134:2481–2484.

(71)   Ban D, Funk M, Gulich R, Egger D, Sabo TM, Walter KFA, Fenwick RB, Giller K, Pichierri F, de Groot BL, Lange OF, Grubmüller H, Salvatella X, Wolf M, Loidl A, Kree R, Becker S, Lakomek N-A, Lee D, et al. Kinetics of conformational sampling in ubiquitin. *Angew Chemie Int Ed* 2011;50:11437–11440.

(72)   Lakomek N-A, Penzel S, Lends A, Cadalbert R, Ernst M, Meier BH. Microsecond dynamics in ubiquitin probed by solid-state ^{15}N nmr spectroscopy $r_{1\rho}$ relaxation experiments under fast mas (60-110 kHz). *Chem - A Eur J* 2017;23:9425–9433.

(73)   Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 2008;320:1471–1475.

(74)   Peters JH, de Groot BL. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS Comput Biol* 2012;8:e1002704.

(75)   Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding.

25

Proc Natl Acad Sci U S A 2013;110:5915–5920.

(76) Ganoth A, Tsfadia Y, Wiener R. Ubiquitin: molecular modeling and simulations. *J Mol Graph Model* 2013;46:29–40.

(77) Tugarinov V, Kay LE. Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. *ChemBioChem* 2005;6:1567–1577.

(78) Otten R, Chu B, Krewulak KD, Vogel HJ, Mulder FAA. Comprehensive and cost-effective NMR spectroscopy of methyl groups in large proteins. *J Am Chem Soc* 2010;132:2952–2960.

(79) Godoy-Ruiz R, Guo C, Tugarinov V. Alanine methyl groups as NMR probes of molecular structure and dynamics in high-molecular-weight proteins. *J Am Chem Soc* 2010;132:18340–18350.

(80) Nadaud PS, Helmus JJ, Jaroniec CP. ^{13}C and ^{15}N chemical shift assignments and secondary structure of the B3 immunoglobulin-binding domain of streptococcal protein G by magic-angle spinning solid-state NMR spectroscopy. *Biomol NMR Assign* 2007;1:117–120.

(81) Vögeli B, Yao L, Bax A. Protein backbone motions viewed by intraresidue and sequential HN–Hα residual dipolar couplings. *J Biomol NMR* 2008;41:17–28.

(82) Yao L, Ying J, Bax A. Improved accuracy of ^{15}N–^{1}H scalar and residual dipolar couplings from gradient-enhanced IPAP-HSQC experiments on protonated proteins. *J Biomol NMR* 2009;43:161–170.

(83) Orts J, Vögeli B, Riek R, Güntert P. Stereospecific assignments in proteins using exact NOEs. *J Biomol NMR* 2013;57:211–218.

(84) Best RB, Zheng W, Mittal J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J Chem Theory Comput* 2014;10:5113–5124.

26

(85)   Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, MacKerell AD. Optimisation
       of the additive CHARMM all-atom protein force field targeting improved sampling of
       the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J Chem Theory Comput*
       2012;8:3257–3273.

27

Dušan Petrović

FIGURE 1 Threonine rotamers: *gauche⁻* (g^-), *trans* (t), and *gauche⁺* (g^+). The χ_1 angle is defined by the N-Cα-Cβ-Oγ_1 atoms.

FIGURE 2 Structures of (a) ubiquitin and (b) GB3 protein.

FIGURE 3 The side chain ensembles of ubiquitin simulated with CHARMM 27: (a) average rotamer angle and (b) rotamer populations. The simulated values are plotted against the corresponding NMR results. The blue diagonal represents the ideal correlation, and the confidence interval of ±20% is shown shaded in blue. Val, Ile, and Thr residues are represented by ■, ▲, and ▼, respectively. The results for all force fields are shown in Figures S1 and S2.

FIGURE 4 The side chain ensembles of GB3 simulated with AMBER 99SB*-ILDN: (a) average rotamer angle, (b) top rotamer angle, and (c) rotamer population. The simulated values are plotted against the corresponding NMR results. The blue diagonal represents the ideal correlation, and the confidence interval of ±20% is shown shaded in blue. The results for all force fields are given in Figures S3–S5.

FIGURE 5 The mean (a) backbone RMSD and (b) RMSDih for χ_1 angles of ubiquitin and GB3, with the standard deviation shown as error bars. The time evolution of ubiquitin and GB3 RMSD and Rg for the *post-2011* FFs is shown in Figures S7–S8 and Figures S9–S11, respectively. The time evolution of the RMSDih is shown in Figures S12–S16.

102

TABLE 1 Force field benchmark for the methyl-bearing residues (Val, Ile, and Thr) of ubiquitin.

Force field	Average rotamer angle		Rotamer population	
	R^2	MAE	R^2	MAE
AMBER 03	0.9856	10.7°	0.3772	0.229
AMBER 03*	0.9853	10.9°	0.4061	0.218
AMBER 99SB-ILDN (TIP3P)	0.9864	10.0°	0.8958	0.095
AMBER 99SB-ILDN (TIP4P-D)	0.9870	10.5°	0.9227	0.079
AMBER 99SB*-ILDN	0.9857	10.7°	0.9268	0.083
AMBER 14SB	0.9759	10.9°	0.9381	0.075
AMBER FB15	0.9774	11.5°	0.7542	0.210
CHARMM 22	0.9871	10.0°	0.6878	0.137
CHARMM 22*	0.9878	10.0°	0.9032	0.086
CHARMM 27	0.9881	9.8°	0.9544	0.065
CHARMM 36	0.9836	9.9°	0.9205	0.087
OPLS-AA	0.9872	9.1°	0.5769	0.178
GROMOS 54a7	0.9799	12.4°	0.5284	0.189

103

Dušan Petrović

TABLE 2 Force field benchmark for non-Gly/Ala residues in GB3.

Force field	Average rotamer angle		Rotamer population	
	R^2	MAE	R^2	MAE
AMBER 03	0.9711	10.4°	0.3176	0.225
AMBER 03*	0.9725	10.2°	0.3672	0.213
AMBER 99SB-ILDN (TIP3P)	0.9649	11.2°	0.6948	0.135
AMBER 99SB-ILDN (TIP4P-D)	0.9663	10.9°	0.5823	0.164
AMBER 99SB*-ILDN	0.9647	11.2°	0.6959	0.137
AMBER 14SB	0.9701	9.8°	0.8017	0.117
AMBER FB15	0.9671	11.2°	0.5587	0.186
CHARMM 22	0.9724	11.3°	0.0429	0.327
CHARMM 22*	0.9728	10.5°	0.5163	0.190
CHARMM 27	0.9717	10.3°	0.4797	0.181
CHARMM 36	0.9673	10.4°	0.6298	0.161
OPLS-AA	0.9585	12.5°	0.1026	0.329
GROMOS 54a7	0.9649	12.8°	0.2046	0.255

104

TABLE 3 The RMSD values (± standard error of the mean, in Hz) between experimental and calculated ^3J coupling constants of ubiquitin and GB3.

Force field	Ubiquitin				GB3		
	$^3J_{H\alpha,H\beta}$	$^3J_{C',C\gamma}$	$^3J_{C'/N,C\gamma}$ methyl	overall	$^3J_{H\alpha,H\beta}$	$^3J_{C'/N,C\gamma}$ methyl	overall
AMBER 03	3.3 ± 0.2	1.4 ± 0.2	1.5 ± 0.1	2.4 ± 0.1	3.1 ± 0.6	1.3 ± 0.1	1.9 ± 0.2
AMBER 03*	3.3 ± 0.2	1.4 ± 0.2	1.5 ± 0.1	2.4 ± 0.1	3.0 ± 0.5	1.3 ± 0.1	1.8 ± 0.2
AMBER 99SB-ILDN (TIP3P)	2.8 ± 0.1	1.3 ± 0.1	1.4 ± 0.1	2.1 ± 0.1	2.0 ± 0.3	1.2 ± 0.1	1.5 ± 0.1
AMBER 99SB-ILDN (TIP4P-D)	2.8 ± 0.2	1.3 ± 0.1	1.4 ± 0.1	2.1 ± 0.1	2.3 ± 0.2	1.3 ± 0.1	1.6 ± 0.1
AMBER 99SB*-ILDN	2.8 ± 0.2	1.3 ± 0.1	1.4 ± 0.1	2.1 ± 0.1	2.0 ± 0.3	1.2 ± 0.1	1.5 ± 0.1
AMBER14SB	2.7 ± 0.2	1.2 ± 0.1	1.4 ± 0.1	2.0 ± 0.1	1.9 ± 0.2	1.2 ± 0.1	1.4 ± 0.1
AMBER FB15	3.5 ± 0.2	1.5 ± 0.1	1.5 ± 0.2	2.5 ± 0.1	1.9 ± 0.2	1.3 ± 0.1	1.5 ± 0.1
CHARMM 22	4.2 ± 0.2	1.7 ± 0.1	1.4 ± 0.1	2.9 ± 0.1	4.8 ± 0.4	1.2 ± 0.1	2.4 ± 0.3
CHARMM 22*	3.0 ± 0.2	1.3 ± 0.1	1.4 ± 0.2	2.2 ± 0.1	2.7 ± 0.4	1.2 ± 0.1	1.7 ± 0.2
CHARMM 27	3.1 ± 0.2	1.4 ± 0.1	1.4 ± 0.1	2.3 ± 0.1	2.7 ± 0.4	1.2 ± 0.1	1.7 ± 0.2
CHARMM 36	2.5 ± 0.1	1.2 ± 0.1	1.4 ± 0.2	1.9 ± 0.1	3.0 ± 0.5	1.2 ± 0.1	1.8 ± 0.2
OPLS-AA	3.8 ± 0.2	1.6 ± 0.1	1.5 ± 0.1	2.7 ± 0.1	3.4 ± 0.5	1.3 ± 0.1	2.0 ± 0.2
GROMOS 54a7	—[a]	1.4 ± 0.1	1.5 ± 0.1	—	—	1.3 ± 0.1	—

[a] No data available for Hα-Hβ coupling due to the united-atom nature of the GROMOS FF.

48
49
50
51
52
53
54
55
56
57
58
59
60
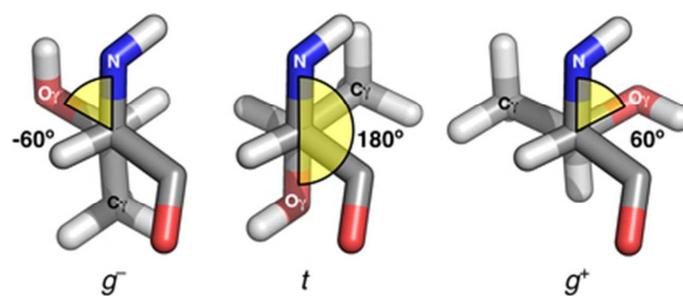
105

Dušan Petrović

Figure 1.Threonine rotamers: *gauche⁻* (*g⁻*), *trans* (*t*), and *gauche⁺* (*g⁺*). The χ_1 angle is defined by the N-C_α-C_β-$O_{\gamma 1}$ atoms.
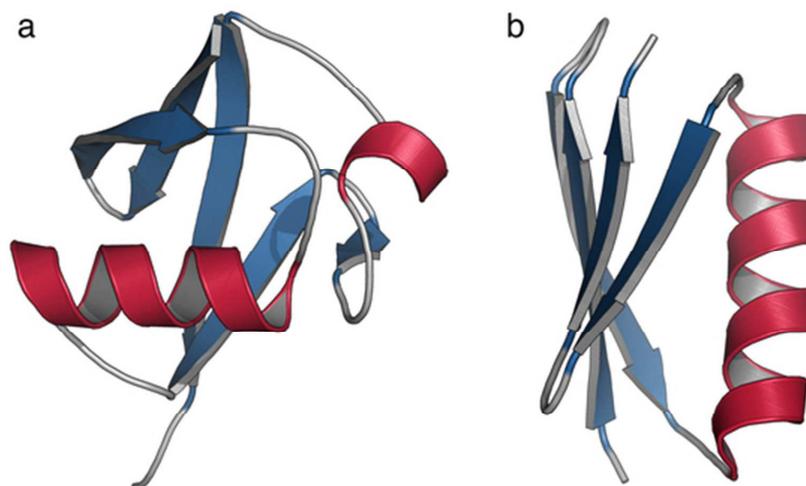
34x15mm (300 x 300 DPI)

106

Figure 2. Structures of (a) ubiquitin and (b) GB3 protein.

46x27mm (300 x 300 DPI)
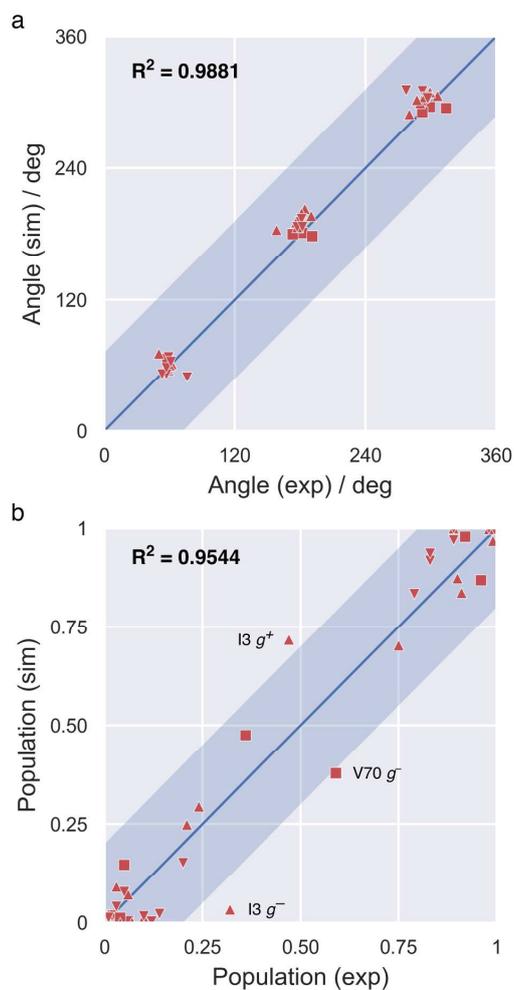
Figure 3.The side chain ensembles of ubiquitin simulated with CHARMM 27: (a) average rotamer angle and (b) rotamer populations. The simulated values are plotted against the corresponding NMR results. The blue diagonal represents the ideal correlation, and the confidence interval of ±20% is shown shaded in blue. Val, Ile, and Thr residues are represented by ■, △, and ▽, respectively. The results for all force fields are shown in Figures S1 and S2.

129x247mm (300 x 300 DPI)

108

Figure 4.The side chain ensembles of GB3 simulated with AMBER 99SB*-ILDN: (a) average rotamer angle, (b) top rotamer angle, and (c) rotamer population. The simulated values are plotted against the corresponding NMR results. The blue diagonal represents the ideal correlation, and the confidence interval of ±20% is shown shaded in blue. The results for all force fields are given in Figures S3–S5.

194x555mm (300 x 300 DPI)
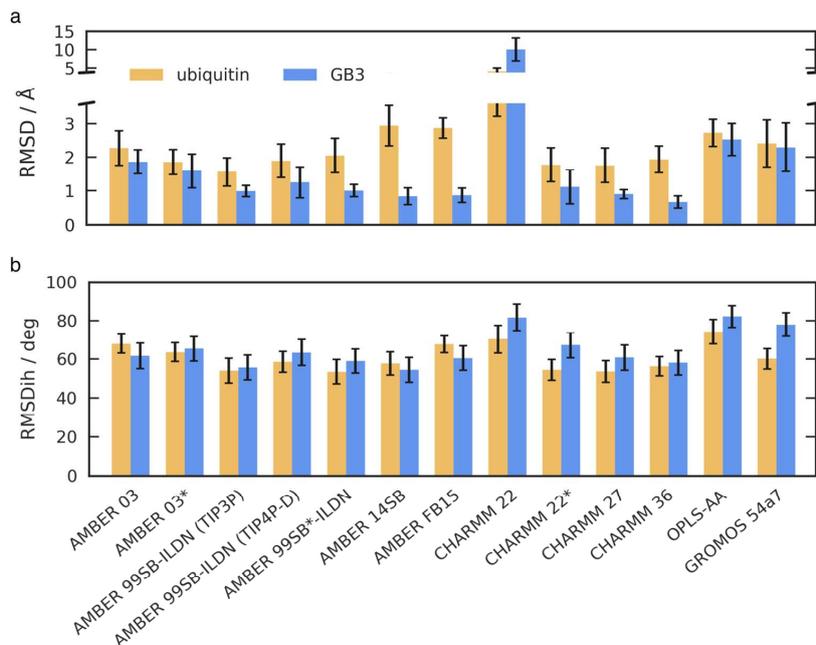
Figure 5.The mean (a) backbone RMSD and (b) RMSDih for χ_1 angles of ubiquitin and GB3, with the standard deviation shown as error bars. The time evolution of ubiquitin and GB3 RMSD and Rg for the post-2011 FFs is shown in Figures S7–S8 and Figures S9–S11, respectively. The time evolution of the RMSDih is shown in Figures S12–S16.

138x108mm (300 x 300 DPI)

110

Paper III

**One-pot, two-step hydroxylation of the macrocyclic diterpenoid
β-cembrenediol catalyzed by P450 BM3 mutants**

by P. Le-Huu, D. Petrović, B. Strodel, and V.B. Urlacher

*D.P. contributed to this paper by performing and analyz-
ing quantum mechanical calculations and comparing
experimental and theoretical NMR spectra. D.P. also as-
sisted in writing of the manuscript.*

One-Pot, Two-Step Hydroxylation of the Macrocyclic Diterpenoid β-Cembrenediol Catalyzed by P450 BM3 Mutants

Priska Le-Huu,[a] Dusan Petrovic,[b] Birgit Strodel,[b, c] and Vlada B. Urlacher*[a]

Cytochrome P450 monooxygenases (P450s) are involved in the biosynthesis of a wide range of bioactive secondary metabolites. They often introduce several oxy functionalities at different positions of a substrate through multiple steps and produce a range of oxidized derivatives. Herein, we describe a one-pot two-step hydroxylation of the diterpenoid β-cembrenediol isolated from the plant *Nicotiana tabacum*. This 14-membered macrocycle shows neuroprotective effects and is, along with its oxidized derivatives, of pharmaceutical interest.
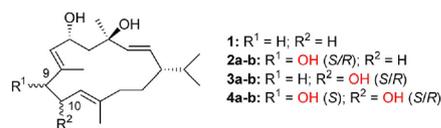
Sequential hydroxylations catalyzed by the regioselective P450 BM3 mutants F87A/I263L and V78A/F87G yielded the epimeric (9S,10R/S)-β-cembrenetetraols with a diastereomeric ratio of 48:52. The replacement of the mutant V78A/F87G with L75A/V78A/F87G in the second step improves the diastereomeric ratio up to 10:90. Absolute configurations of the newly introduced hydroxy groups were determined by quantum-mechanical calculations of NMR spectra.

Introduction

Cytochrome P450 monooxygenases (EC 1.14.-.-; P450s) are involved in the biosynthesis of various natural metabolites and catalyze oxidation reactions at late stages. Thereby, P450 catalysis leads to the diversification of natural products, for example, terpenoids.[1] Depending on the complexity of the final compound, P450s are able to introduce several oxy functionalities, which require multiple reaction steps.[2] Several papers reported reactions in which one P450 monooxygenase uses more than one molecule of atmospheric O_2 to catalyze sequential oxidations at one position (e.g., to an alcohol and further to an aldehyde and a carboxylic acid).[3] There are many biosynthetic gene clusters known in which one or several P450s oxidize a substrate sequentially at different positions. However, reports on the reconstitution of such P450-catalyzed reactions in vitro or in recombinant organisms are scarce. Among them are, for example, the hydroxylation of progesterone at three distinct positions reconstituted in *S. cerevisiae* and the multi-step hydroxylation of vitamin D3 catalyzed by a single monooxygenase, namely, CYP105A1 from *Streptomyces griseolus*.[4]

We report herein the development of a two-step hydroxylation of the 14-membered macrocyclic diterpenoid β-cembrenediol (1). Cembranoids constitute a versatile class of natural metabolites, the members of which are found mainly in plants and marine invertebrates.[5,6] Their structures differ in the number and kind of additional oxy functionalities.[6] One of the most abundant cembranoids found in the plant *Nicotiana tabacum* is 1 (Scheme 1), which could be isolated from the cuticular wax of the leaves and flowers.[7] Next to its natural role as

[a] Dr. P. Le-Huu, Prof. Dr. V. B. Urlacher
Institute of Biochemistry
Heinrich-Heine University Düsseldorf
Universitätsstrasse 1, Bldg. 26.42.U1
40225 Düsseldorf (Germany)
E-mail: vlada.urlacher@uni-duesseldorf.de

[b] D. Petrovic, Prof. Dr. B. Strodel
Institute of Complex Systems: Structural Biochemistry (ICS-6)
Forschungszentrum Jülich GmbH
52425 Jülich (Germany)

[c] Prof. Dr. B. Strodel
Institute of Theoretical and Computational Chemistry
Heinrich-Heine University Düsseldorf
Universitätsstrasse 1
40225 Düsseldorf (Germany)

Supporting information for this article can be found under:
http://dx.doi.org/10.1002/cctc.201600973.

1: R^1 = H; R^2 = H
2a-b: R^1 = OH (S/R); R^2 = H
3a-b: R^1 = H; R^2 = OH (S/R)
4a-b: R^1 = OH (S); R^2 = OH (S/R)

Scheme 1. Compound 1 and its hydroxylated derivatives: 2a,b, 3a,b, and 4a,b.

a key flavor ingredient in tobacco plants,[8] several biological activities of 1 have been found, which include anti-inflammatory,[9] anticancer,[10] and neuroprotective effects.[11] Compound 1 is a noncompetitive inhibitor of nicotinic acetylcholine receptors[12] and has already tested positively as a neuroprotective drug against the neurotoxic effect of the organophosphate diisopropylfluorophosphate.[13] Furthermore, the neuroprotective effect of 1 was also demonstrated in rodent ischemic stroke models.[14] As a result of their pharmaceutical potential, cembranoids have been described as promising lead compounds.[15] Oxidized derivatives of 1 have been synthesized

chemically[16] or through biotransformations[17] and have been investigated for their biological activities. The protective activity of 9-OH- and 10-OH-β-cembrenetriols against diisopropyl-fluorophosphate neurotoxicity has been reported.[18] Thus, it would be interesting to investigate if multiple oxidations of **1** lead to new metabolites with meaningful biological activities.

Although the *N. tabacum* P450 CYP71D16 has been identified to catalyze the hydroxylation of the cembrenediol precursor (1*S*,2*E*,7*E*,11*E*)-cembra-2,7,11-triene-4-ol regio- and stereoselectively at position C-6,[19] no naturally occurring P450 is known to oxidize cembrenediol or any of its oxidized metabolites. As bacterial P450s can be produced in recombinant hosts efficiently, they are often used in protein engineering studies to achieve new activities.[20] One of the best-studied monooxygenases is P450 BM3 (CYP102A1) from *Bacillus megaterium*.[21] P450 BM3 is a catalytically self-sufficient enzyme because of its fused reductase domain that transfers electrons from the cofactor NADPH to the heme Fe.[22] Although naturally it performs a fatty acid hydroxylation, numerous P450 BM3 variants have been developed to oxidize bulky substrates.[23] Recently, we engineered the F87A/I263L mutant that enables the hydroxylation of **1** at position C-9 with a regioselectivity of 92 %.[24] This reaction furnished the epimeric β-cembrenetriols **2 a,b** with a diastereomeric ratio of 89:11, one of which, (9*R*)-β-cembrenetriol (**2 a**), was isolated with a yield of 57 %.[24]

Although the F87A/I263L mutant was already known for the first reaction step, in this study we applied first-sphere active-site mutagenesis to construct P450 BM3 mutants for the second hydroxylation step to introduce a fourth hydroxy group into the diterpenoid skeleton. Sequential catalysis by two P450 BM3 variants facilitated a one-pot, two-step hydroxylation of **1** at positions C-9 and C-10 for the production of diastereomeric β-cembrenetetraols **4 a,b** (Scheme 1). The stereochemistry of all of the newly introduced hydroxy groups was assigned by comparing the experimental NMR spectra with the quantum mechanical predictions of diastereomeric NMR spectroscopic properties.

Results and Discussion

Screening for chemo-, regio-, and stereoselective P450 BM3 variants that hydroxylate (9R)-β-cembrenetriol

After we chose the F87A/I263L mutant as one of the catalysts for the multistep reaction, we constructed P450 BM3 variants to oxidize (9*R*)-β-cembrentriol (**2 a**) selectively. For the generation of P450 BM3 variants, we focused on amino acid substitutions at positions in the active site that had a significant impact on the selective oxidations of **1**.[24] As **2 a** is a bulky molecule, a set of 12 single to triple mutants was constructed in which the single mutants comprised either alanine or glycine at position 87 and the double and triple mutants had additional alanine or glycine substitutions at positions 75 and/or 78, respectively (Table 1).

The screening was aimed to identify chemo-, regio-, and stereoselective mutants that produce only one main product: **2 a**. The reaction mixtures were analyzed by LC MS.

Wild-type P450 BM3 and the F87A mutant (Table 1, generation 1a) showed no activity on **2 a**. Double and triple mutants of F87A (generation 2a 3a) were either inactive or led to the formation of several products with mostly unknown structures. In contrast to the F87A mutant, the F87G mutant converted 24 % of **2 a** (Table 1, generation 1b). Thereby, **2 a** was hydroxylated at position C-10 with a regioselectivity of 91 % to the epimeric (9*S*,10*S*)-β-cembrenetetraol (**4 a**)[25] and (9*S*,10*R*)-β-cembrenetetraol (**4 b**)[25] with a diastereomeric ratio (dr) of 64:36 in favor of **4 a**. Double mutants of the F87G-based generation (2b) showed a further increase in activity (28 54 % conversion). The addition of alanine or glycine at position 75 to the F87G mutant decreased the regioselectivity for C-10 and led to the formation of additional products. Conversely, the addition of

Table 1. Conversion and product distribution of **2 a** oxidation by P450 BM3 variants.[a]										
Generation	Amino acid substitution			Conv. [%]	Product distribution[c] [%]						dr[d] (4a/4b)
	L75	V78	F87		rt=4.5 min[e]	rt=8.6 min[e]	rt=8.8 min[e]	4 a	4 b	Other products[f]	
1 a			A	n.d.[b]							
2 a	G		A	35±4	18	55	11		7	9	>1:99
2 a	A		A	15±3	31	44			25		>1:99
2 a		G	A	n.d.[b]							
2 a		A	A	n.d.[b]							
3 a	A	A	A	54±9	4	43	8	-	38	7	>1:99
1 b			G	24±3	9			58	33		64:36
2 b	G		G	54±11	35	31	17	8	8		49:51
2 b	A		G	48±10	37	22	7	9	25		27:73
2 b		G	G	28±5				35	61	4	36:64
2 b		A	G	48±8				49	49	2	50:50
3 b	A	A	G	47±5		6	2	9	75	8	11:89

[a] Mean values and standard deviations are calculated from three separate experiments. [b] Three mutants showed conversions < 10 % (n.d.=not detectable) without product formation. [c] Differences from 100 % products in total come from rounding. [d] The dr of the products **4 a** (rt=10.3 min) and **4 b** (rt=10.5 min) is shown. [e] Products with a retention time (rt) of 4.5, 8.6, and 8.8 min had a mass equivalent to derivatives of **2 a** with one additional hydroxy or epoxy group. [f] Small amounts of other products were detected that had a mass equivalent to one additional hydroxy, epoxy, or enone group compared to **2 a**.

114

alanine or glycine at position 78 to F87G increased the regioselectivity for C-10 (products **4a,b**) up to 96 98%. Although the mutant V78A/F87G converted 48% of **2a** to **4a,b** without a stereopreference, the mutant V78G/F87G converted 28% of **2a** to **4a,b** with a dr of 36:64. The triple mutant L75A/V78A/F87G (generation 3a) catalyzed the hydroxylation of **2a** at position C-10 with a lower regioselectivity (84%) but with an improved diastereoselectivity in favor of **4b** (11:89 dr; Table 1).

One-pot two-step hydroxylation of β-cembrenediol

In the next step we established a one-pot reaction that combined two P450 BM3 mutants for the two-step hydroxylation of **1**. All reactions were performed as a simple one-pot process, without the isolation and purification of intermediates. The selected mutants F87A/I263L and V78A/F87G were added either simultaneously or sequentially in both orders.

After the simultaneous addition, **1** was converted mainly to the epimeric (10R/S)-β-cembrenetriols (**3a,b**; 52% yield; Figure 1A), similar to the reaction catalyzed by V78A/F87G only. The desired products, (9S,10R/S)-β-cembrenetetraols (**4a,b**), were produced with a yield of 15% after 17 h; extended reaction time did not improve the yield. These results suggest that V78A/F87G first produced the (10R/S)-β-cembrenetriols (**3a,b**), which are not accepted by F87A/I263 for further hydroxylation, and only minor amounts of **1** were hydroxylated first at position C-9 by the F87A/I263L mutant. These results can be explained by a lower activity of the mutant F87A/I263L towards **1** (17% conversion) compared to mutants V78A/F87G (67% conversion) or L75A/V78A/F87G (73% conversion).[24] Additionally, 30% of other products, the structures of which have not been identified yet, were formed in the simultaneous reaction mode. According to our MS measurements, the predominant product results from the further oxidation of an alcohol to an enone, which has also been observed during the oxidation of **1** by the mutant V78A/F87G.[24]

The sequential reaction mode with the V78A/F87G mutant as the starting enzyme (reaction mode Ia) produced again mainly **3a,b** (63% yield in total; Figure 1B). As expected, the addition of F87A/I263L mutant after 17 h did not lead to the hydroxylation at position C-9. In a control experiment in which the L75A/V78A/F87G mutant (with a similar regioselectivity to the V78A/F87G mutant but different stereoselectivity) was applied as the first and F87A/I263L as the second enzyme (Table S1, reaction mode Ib) similar results to reaction mode Ia were obtained. Clearly, steric restrictions in the active site of the F87A/I263L mutant hamper the oxidation of **3a,b**, which is analogous to the simultaneous reaction mode.

We used the F87A/I263L mutant as the starting enzyme in a sequential reaction mode (IIa) to convert **1** almost completely (97%) after 17 h to **2a** (80% yield) and **2b** (9% yield; 90:10 dr; Figure 1C). After the addition of the V78A/F87G mutant, 89% of **2a** was converted in 7 h to form (9S,10S)-β-cembrenetetraol (**4a**; 40% yield) and (9S,10R)-β-cembrenetetraol (**4b**; 43% yield; 48:52 dr; Figure S1).

As expected, the F87A/I263L mutant alone was not able to produce the desired products **4a,b** in a control reaction within
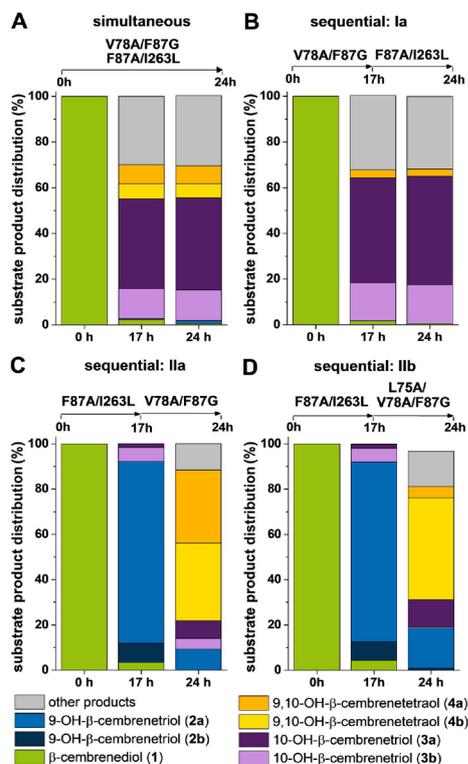

Figure 1. Substrate/product distribution during the two-step hydroxylation of **1**. A) P450 BM3 mutants F87A/I263L and V78A/F87G were added simultaneously. B D) Sequential reactions. B) The reaction was catalyzed by mutant V78A/F87G first (0 17 h) and mutant F87A/I263L second (17 24 h). C D) The reactions were catalyzed by mutant F87A/I263L first (0 17 h) and mutant C) V78A/F87G or D) L75A/V78A/F87G second (17 24 h). Other products are the products whose structures have not yet been identified. According to MS, these products result from a further oxidation of an alcohol to an enone (such as β-cembrenediol-one and β-cembrenetriol-one) or are β-cembrene-tetraols with other retention times than **4a,b**. Differences from 100% substrate and products in total occur because of rounding. In A and B, mean values are calculated from two, and in C and D from three separate experiments. Numerical values and standard deviations are given in Table S1 S2.

24 h (Table S2), which proved that C-10 hydroxylation was catalyzed by the V78A/F87G mutant. Thus, sequential catalysis performed with the F87A/I263L mutant (first) and V78A/F87G mutant (second) enabled the production of **4a,b** with a total yield of 66%.

The replacement of the V78A/F87G mutant by the L75A/V78A/F87G mutant in the second step (reaction mode IIb) led to 77% conversion of **2a** after 7 h to give **4a** (6% yield) and **4b** (56% yield) with an improved dr of 10:90 (Figure 1D).

Structure elucidation of oxidized β-cembrenediol derivatives

The NMR spectra of **2**–**4** were not sufficient to assign the configurations of the new stereocenters at C-9 or/and C-10. Therefore, we supplemented this study by performing a quantum mechanical prediction of the NMR spectra of **2a,b**, **3a,b**, and **4a,b** and comparing them with the experimental spectra.

Both epimeric (9R)- and (9S)-β-cembrenetriols (**2a,b**) were simulated and compared with the experimental spectra of **2a** and **2b**,[24] and a generally good agreement was found (Tables S3–S5). To assign the stereochemistry to this pair of epimers, we used the CP3 approach developed by Smith and Goodman that relies on Bayesian statistics.[26] With a reasonably high probability (92.5%, based on both ^{1}H and ^{13}C NMR spectroscopic data), we conclude that **2a** has a 9R configuration and **2b** has a 9S configuration. In general, the $^3J_{H,H}$ coupling constants could not be resolved very well in chloroform, but $^3J_{9,10a}$ agrees with the assignments based on the CP3-predicted probabilities (Table S6).

The experimental NMR shifts of **3a,b** (complete NMR spectroscopic data are given in Figures S2–S11) correlate well with the simulated values (Tables S7–S9). The exceptionally high CP3 probability (99.8%, based on both ^{1}H and ^{13}C NMR spectroscopic data) implies the 10R configuration in **3a** and 10S in **3b**. The consideration of proton–proton coupling constants (especially $^3J_{9a,10}$, $^3J_{9b,10}$, and $^3J_{10,11}$ in the region of interest) corroborates the chemical shift analysis (Table S10).

For the structural identification of **4a,b**, we performed the sequential reaction with the mutants F87A/I263L and V78A/F87G on a larger scale (Scheme 2). Products **4a,b** were isolated using semipreparative HPLC in one (**4a**) or two (**4b**) chromatographic steps with isolated yields of 38% (**4a**) and 11% (**4b**).
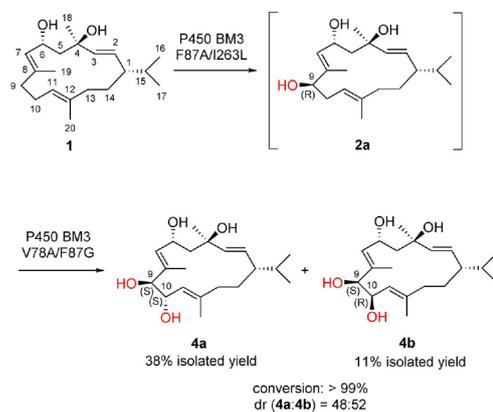
Scheme 2. P450 BM3-catalyzed one-pot, two-step hydroxylation of **1** on a large scale for product preparation. The conversion value and dr refer to the scaled-up experiments. The crude product was analyzed using LC–MS to determine the conversion and dr. The calculation of the conversion value is based on the ratio of all product peaks to the sum of product and substrate peaks. The isolated yield of **4a** was determined after one chromatographic step, and the yield of **4b** after two chromatographic steps.

The oxidation products were analyzed using NMR spectroscopy and MS. The ^{1}H and ^{13}C NMR chemical shifts were assigned fully by evaluating the ^{1}H,^{1}H-COSY, ^{1}H,^{13}C-HSQC, and ^{1}H,^{13}C-HMBC spectra (Figure S12–S21). Compounds **4a,b** were confirmed to be the diastereomeric (1S,2E,4R,6R,7E,11E)-2,7,11-cembratriene-4,6,9,10-tetraols. The two newly introduced hydroxy groups at C-9 and C-10 generated two additional stereocenters, which theoretically, could result in four different stereoisomers (9R-10R, 9R-10S, 9S-10R, and 9S-10S). We simulated all of them, although we expected the conversion of **2a** (which has a 9R configuration) to result in 9S products.[25] Compounds **4a,b** showed different vicinal coupling constants between protons H-9 and H-10 (**4a**: $^3J_{9,10} = 7.6$ Hz and **4b**: $^3J_{9,10} = 3.3$ Hz), according to their different dihedral angles. The comparison of the experimental and simulated $^3J_{9,10}$ coupling constants showed a *trans* geometry (9R-10R or 9S-10S) for **4a**, whereas the much lower value for **4b** originates from a *cis* geometry (9R-10S or 9S-10R; Table S11). This was taken into account when we compared the experimental and simulated chemical shifts (Tables S12–S14). Instead of CP3, we used another function by Smith and Goodman: DP4.[27] This function allowed us to compare several simulated spectra against one experimental dataset. DP4 revealed, with utmost probability (99.9%, based on both ^{1}H and ^{13}C NMR spectroscopic data), that **4a** has (9S,10S) configuration, whereas the stereochemical assignment identified **4b** as the (9S,10R) diastereomer, however, with a lower probability (76.3%, based on both ^{1}H and ^{13}C NMR spectroscopic data). The expected 9S configuration was fulfilled and **4a,b** were confirmed to be epimers.

Conclusions

A one-pot, two-step hydroxylation of the macrocyclic diterpenoid β-cembrenediol (**1**) was established using two P450 BM3 mutants, one for each reaction step. Although the F87A/I263L mutant was already known to hydroxylate **1** at position C-9, the V78A/F87G and L75A/V78A/F87G mutants were designed to hydroxylate the intermediate compound (9R)-β-cembrenetriol (**2a**) at position C-10 to produce (9S,10R/S)-β-cembrenetetraols (**4a,b**). Although substrates **1** and **2a** contain 16 potential oxidation sites, the F87A/I263L and L75A/V78A/F87G mutants displayed a high chemo-, regio-, and stereoselectivity for the hydroxylation of **1** or its oxidized derivative **2a**, respectively. To establish a one-pot, two-step hydroxylation, the sequential use of the P450 BM3 mutants was a prerequisite. The produced β-cembrenediol derivatives (9S,10R/S)-β-cembrenetetraols (**4a,b**) are, to the best of our knowledge, described for the first time. We interpreted the experimental NMR spectra supported by quantum mechanical calculations to determine the absolute configuration of **4a,b**. This is a useful tool that might help to elucidate other complex structures.

As **1** is known to possess a variety of biological activities,[14, 28] both oxidized derivatives **4a,b** may have new or enhanced biological activities, which will be investigated in further studies. The two-step, chemo-, regio-, and stereoselective hydroxylation of **1** described here highlights its usefulness for the biomimetic oxidation of terpenes.[29] As late-stage oxida-

tions are still a challenging task for chemical catalysts,[30] engineered P450 monooxygenases as biocatalysts for multiple oxidations of complex terpenoids have a great potential and could be applied for the efficient diversification of complex terpenoids.

Experimental Section

Analytical-scale experiments: The screening of **2a**-oxidizing P450 BM3 variants was executed in a total volume of 0.25 mL potassium phosphate buffer (50 mм, pH 7.5) using *Escherichia coli* cell-free extracts. Reactions contained P450 BM3 (2.5 μм), NADPH (200 μм), and **2a** (100 μм, dissolved in ethanol, 1% v/v). For cofactor regeneration, glucose (20 mм) and glucose dehydrogenase (GDH, 3 UmL⁻¹) were applied. Catalase from bovine liver (600 UmL⁻¹) was added to remove hydrogen peroxide, which can be produced in uncoupling reactions. Screening reactions were performed for 17 h at 25 °C. Subsequently, the internal standard cyclotetradec-4-ene-1,1-diylbis(methylene)diacetate (50 μм) was added, and the reaction mixtures were extracted twice with ethyl acetate. The combined organic layers were concentrated under reduced pressure and resolved in methanol for LC MS analysis. After the initial screening, the products were identified based on their *m/z* values. Conversion values are based on substrate consumptions with respect to a control reaction in which we used cell-free extract of *E. coli* cells that were transformed with an empty vector. Product distribution is derived from the detected peak areas. Mean values and standard deviations are calculated from three separate experiments.

Two-step hydroxylation reactions were performed in potassium phosphate buffer (100 mм, pH 7.5) at 25 °C under luffing shaking. Reactions contained 2.5 μм of each purified P450 BM3 mutant, NADPH (200 μм), and **1** (150 μм, dissolved in ethanol, 1% v/v), glucose (20 mм), GDH (3 UmL⁻¹), and catalase from bovine liver (600 UmL⁻¹). Reactions were catalyzed by the mutants F87A/I263L and V78A/F87G or L75A/V78A/F87G in different reaction modes. In the simultaneous reaction mode, both mutants were added at the start of the reaction, whereas in the sequential reaction mode one mutant was added at the beginning and the other one after 17 h in both possible orders. A control reaction was performed for 24 h using the mutant F87A/I263L as the only catalyst. All reactions were supplemented with glucose (20 mм), GDH (3 UmL⁻¹), and catalase (600 UmL⁻¹) after 17 h. For each reaction mode, one reaction solution was prepared. Sample volumes of 250 μL were taken before the addition of the first enzyme, before the addition of the second enzyme (after 17 h), and after 24 h. After the addition of the internal standard cyclotetradec-4-ene-1,1-diylbis(methylene)diacetate (50 μм), samples were extracted twice with ethyl acetate. The combined organic layers were concentrated under reduced pressure and resolved in methanol for LC MS analysis. Substrate product ratios were based on detected peak areas, and the sum of the substrate and product peaks was set to 100%. Conversion values of each reaction step and yields of each product were calculated with respect to the previous sample. Mean values and standard deviations were calculated from 2 3 separate experiments.

Scale-up experiments: The two-step hydroxylation for product preparation was performed similar to the reactions on the analytical scale with **1** (150 μм), NADPH (150 μм), and 2.0 μм of each P450 BM3 mutant. Several reactions with a total volume of 20 mL each were executed. The reactions were started by addition of mutant F87A/I263L. After 17 h, the reactions were supplemented with the mutant V78A/F87G, GDH (3 UmL⁻¹), glucose (20 mм), and

catalase (600 UmL⁻¹) and incubated further for 24 h. Afterwards the reaction mixtures were extracted with ethyl acetate. The organic layers of all reactions were combined and concentrated under reduced pressure. The products were isolated by semipreparative HPLC. The twofold hydroxylation of **1** (40 mg) resulted in **4a** as a yellow oil (17 mg; 38%) and **4b** (11 mg; 25%) with a LC MS purity of 84%. After a second chromatography step, **4b** was obtained as a colorless oil (5 mg; 11%) with a LC MS purity of >99%.

Structure elucidation: (1*S*,2*E*,4*R*,6*R*,7*E*,9*S*,10*S*,11*E*)-2,7,11-Cembratriene-4,6,9,10-tetraol (**4a**): HRMS (ESI): *m/z*: calcd for C₂₀H₃₄NaO₄: 361.2349 [*M*+Na]⁺; found: 361.2354; ¹H NMR (600 MHz, CD₃OD): δ = 0.83 (d, *J* = 6.8 Hz, 3H, H-17), 0.87 (d, *J* = 6.8 Hz, 3H, H-16), 1.33 (s, 3H, H-18), 1.36 1.45 (m, 1H, H-14ₐ), 1.48 1.55 (m, 1H, H-15), 1.59 1.66 (m, 1H, H-14ₐ), 1.68 (d, *J* = 1.3 Hz, 3H, H-20), 1.72 (dd, *J* = 13.8, 9.6 Hz, 1H, H-5ₐ), 1.74 1.80 (m, 1H, H-1), 1.89 (d, *J* = 1.4 Hz, 3H, H-19), 1.98 2.01 (m, 1H, H-5ᵦ), 1.97 2.04 (m, 1H, H-13ᵦ), 2.06 2.14 (m, 1H, H-13ₐ), 3.83 (d, *J* = 7.6 Hz, 1H, H-9), 4.39 (dd, *J* = 8.4, 7.6 Hz, 1H, H-10), 4.78 (td, *J* = 9.6, 1.5 Hz, 1H, H-6), 5.20 (dq, *J* = 8.4, 1.3 Hz, 1H, H-11), 5.23 (dd, *J* = 15.7, 9.1 Hz, 1H, H-2), 5.36 5.39 (m, 1H, H-7), 5.38 ppm (d, *J* = 15.7 Hz, 1H, H-3); ¹³C NMR (151 MHz, CD₃OD): δ = 17.39 (C-20), 18.27 (C-19), 19.85 (C-16), 20.74 (C-17), 29.31 (C-18), 29.34 (C-14), 34.88 (C-15), 39.46 (C-13), 48.02 (C-1), 54.19 (C-5), 64.97 (C-6), 71.53 (C-10), 72.04 (C-4), 79.13 (C-9), 125.06 (C-11), 131.23 (C-2), 132.33 (C-7), 137.53 (C-3), 137.77 (C-8), 140.54 ppm (C- 12).

(1*S*,2*E*,4*R*,6*R*,7*E*,9*S*,10*R*,11*E*)-2,7,11-Cembratriene-4,6,9,10-tetraol (**4b**): HRMS (ESI): *m/z*: calcd for C₂₀H₃₄NaO₄: 361.2349 [*M*+Na]⁺; found: 361.2345; ¹H NMR (600 MHz, CD₃OD): δ = 0.84 (d, *J* = 6.5 Hz, 3H, 17-H), 0.86 (d, *J* = 6.5 Hz, 3H, 16-H), 1.27 1.35 (m, 1H, 14-Hₐ), 1.36 (s, 3H, 18-H), 1.48 1.54 (m, 2H, 15-H, 1-H), 1.67 (s, 6H, 19-H, 20-H), 1.65 1.71 (m, 1H, 14-Hᵦ), 1.84 (dd, *J* = 14.1, 8.6 Hz, 1H, 5-Hₐ), 1.97 (td, *J* = 13.4, 4.1 Hz, 1H, 13-Hₐ), 2.02 (dd, *J* = 14.1, 1.6 Hz, 1H, 5-Hᵦ), 2.09 2.15 (m, 1H, 13-Hᵦ), 4.12 (dq, *J* = 3.3, 0.8 Hz, 1H, 9-H), 4.41 (dd, *J* = 7.3, 3.3 Hz, 1H, 10-H), 4.75 (ddd, *J* = 10.1, 8.5, 1.6 Hz, 1H, 6-H), 5.21 (dd, *J* = 15.6, 9.0 Hz, 1H, 2-H), 5.25 (d, *J* = 7.3 Hz, 1H, 11-H), 5.35 (d, *J* = 15.6 Hz, 1H, 3-H), 5.73 ppm (d, *J* = 10.1 Hz, 1H, 7-H); ¹³C NMR (151 MHz, CD₃OD): δ = 14.89 (C-19), 15.09 (C-20), 20.01 (C-16), 20.83 (C-17), 28.41 (C-14), 29.01 (C-18), 34.67 (C-15), 37.59 (C-13), 47.84 (C-1), 54.38 (C-5), 64.59 (C-6), 68.40 (C-10), 72.02 (C-4), 77.62 (C-9), 124.68 (C-11), 130.54 (C-7), 130.77 (C-2), 135.38 (C-8), 137.96 (C-3), 141.32 ppm (C-12).

Molecular modeling: The conformational ensemble of the possible stereoisomers of **2a,b**, **3a,b**, and **4a,b** was produced using OMEGA v. 2.5.1.4,[31] which uses the MMFF94s force field[32] in vacuo. The initial structures were generated from fragment libraries, followed by a knowledge-based exhaustive scan of rotatable bonds. Conformers of a given compound (within 5 kcal mol⁻¹ of the lowest energy conformer) were subjected to further quantum mechanical geometry optimization. All quantum mechanical calculations were performed using Gaussian 09.[33] The geometry optimization was conducted at the B3LYP/6-31+G(d,p) level of theory.[34] Between one and three stable conformers were found for each of the modeled compounds (Cartesian coordinates of these conformers can be found in the Supporting Information). Analysis of the vibrational frequencies confirmed that all optimized structures correspond to the actual energy minimum. For low-energy conformers (within 2.5 kcal mol⁻¹ of the global minimum), further NMR calculations were performed. The appropriate solvent (chloroform for **2**, methanol for **3** and **4**) was represented by the polarizable continuum model (PCM).[35] The NMR shielding tensors were calculated by the gauge-independent atomic orbital (GIAO) method[36] at the

mPW1PW91/6 311+G(2d,p) level of theory.[37] Shielding tensors were averaged over all symmetry-related nuclei, and chemical shifts were calculated using a linear regression method [Eq. (1)]:

$$\delta = \frac{b - \sigma}{-m} \tag{1}$$

in which δ is a chemical shift, σ is a shielding tensor, and the intercept (b) and slope (m) depend on the level of theory used.[38] Chemical shifts were scaled empirically to remove the systematic errors of the method and reported as the Boltzmann-averaged values over all low-energy conformers [Eq. (2)]:

$$\delta_{\text{calc}} = \frac{\sum_i \delta_i e^{-\frac{E_i}{RT}}}{\sum_i e^{-\frac{E_i}{RT}}} \tag{2}$$

in which δ_i is the chemical shift in conformer i, E_i is the free energy of conformer i relative to the global minimum, R is the molar gas constant (1.987×10^{-3} kcal K^{-1} mol^{-1}), and T is the temperature (298.15 K). To assign configurations based solely on chemical shifts, we used the well-established CP3[26] and DP4[27] tools that rely on Bayesian statistics. $^3J_{\text{H,H}}$ coupling constants were calculated in the gas phase at the B3LYP/6-31G(d,p) level of theory, and after empirical scaling,[39] the results were represented as Boltzmann averages.

Acknowledgements

[1] a) T. Brück, R. Kourist, B. Loll, ChemCatChem 2014, 6, 1142 1165; b) I. Pateraki, A. M. Heskes, B. Hamberger in Biotechnology of Isoprenoids (Eds.: J. Schrader, J. Bohlmann), Springer, Heidelberg, 2015, pp. 107 139; c) Y.-J. Zhao, Q.-Q. Cheng, P. Su, X. Chen, X.-J. Wang, W. Gao, L.-Q. Huang, Appl. Microbiol. Biotechnol. 2014, 98, 2371 2383.

[2] V. B. Urlacher, S. Schulz in Cascade biocatalysis: integrating stereoselective and environmentally friendly reactions (Eds.: S. Riva, W.-D. Fessner), Wiley-VCH, Weinheim, 2014, pp. 87 132.

[3] a) S. Brown, M. Clastre, V. Courdavault, S. E. O Connor, Proc. Natl. Acad. Sci. USA 2015, 112, 3205 3210; b) C. M. Bertea, J. R. Freije, H. van der Woude, F. W. A. Vertappen, L. Perk, V. Marquez, J.-W. de Kraker, M. A. Posthumus, B. J. M. Jansen, A. de Groot, M. C. R. Franssen, H. J. Bowmeester, Planta Med. 2005, 71, 40 47; c) D. Morrone, X. Chen, R. M. Coates, R. J. Peters, Biochem. J. 2010, 431, 337 344; d) K. H. Teoh, D. R. Polichuk, D. W. Reed, G. Nowak, P. S. Covello, FEBS Lett. 2006, 580, 1411 1416.

[4] a) C. Duport, R. Spagnoli, E. Degryse, D. Pompon, Nat. Biotechnol. 1998, 16, 186 189; b) K. Hayashi, K. Yasuda, Y. Yogo, T. Takita, K. Yasukawa, M. Ohta, M. Kamakura, S. Ikushiro, T. Sakaki, Biochem. Biophys. Res. Commun. 2016, 473, 853 858; c) K. Hayashi, K. Yasuda, H. Sugimoto, S. Ikushiro, M. Kamakura, A. Kittaka, R. L. Horst, T. C. Chen, M. Ohta, Y. Shiro, T. Sakaki, FEBS J. 2010, 277, 3999 4009; d) N. Sawada, T. Sakaki,

S. Yoneda, T. Kusudo, R. Shinkyo, M. Ohta, K. Inouye, Biochem. Biophys. Res. Commun. 2004, 320, 156 164; e) F. M. Szczebara, C. Chandelier, C. Villeret, A. Masurel, S. Bourot, C. Duport, S. Blanchard, A. Groisillier, E. Testet, P. Costaglioli, G. Cauet, E. Degryse, D. Balbuena, J. Winter, T. Achstetter, R. Spagnoli, D. Pompon, B. Dumas, Nat. Biotechnol. 2003, 21, 143 149.

[5] N. Yan, Y. Du, X. Liu, H. Zhang, Y. Liu, P. Zhang, D. Gong, Z. Zhang, Ind. Crops Prod. 2016, 83, 66 80.

[6] B. Yang, X.-F. Zhou, X.-P. Lin, J. Liu, Y. Peng, X.-W. Yang, Y. Liu, Curr. Org. Chem. 2012, 16, 1512 1539.

[7] D. L. Roberts, R. L. Rowland, J. Org. Chem. 1962, 27, 3989 3995.

[8] I. Wahlberg, C. R. Enzell, Beitr. Tabakforsch. Int. 1984, 12, 93 104.

[9] E. Olsson, A. Holth, E. Kumlin, L. Bohlin, I. Wahlberg, Planta Med. 1993, 59, 293 295.

[10] a) Y. Saito, Y. Tsujino, H. Kaneko, D. Yoshida, S. Mizusaki, Agric. Biol. Chem. 1987, 51, 941 943; b) Y. Saito, H. Takizawa, S. Konishi, D. Yoshida, S. Mizusaki, Carcinogenesis 1985, 6, 1189 1194.

[11] a) P. Ferchmin, J. Hao, D. Perez, M. Penzo, H. M. Maldonado, M. T. Gonzalez, A. D. Rodriguez, J. de Vellis, J. Neurosci. Res. 2005, 82, 631 641; b) V. A. Eterovic, D. Pérez, A. H. Martins, B. L. Cuadrado, M. Carrasco, P. Ferchmin, Toxicol. In Vitro 2011, 25, 1468 1474.

[12] P. Ferchmin, R. J. Lukas, R. M. Hann, J. D. Fryer, J. B. Eaton, O. R. Pagán, A. D. Rodríguez, Y. Nicolau, M. Rosado, S. Cortés, V. A. Eterovic, J. Neurosci. Res. 2001, 64, 18 25.

[13] a) P. A. Ferchmin, D. Perez, B. L. Cuadrado, M. Carrasco, A. H. Martins, V. A. Eterovic, Neurochem. Res. 2015, 40, 2143 2151; b) P. Ferchmin, M. Andino, R. Reyes Salaman, J. Alves, J. Velez-Roman, B. Cuadrado, M. Carrasco, W. Torres-Rivera, A. Segarra, A. H. Martins, J. Eun Lee, V. A. Eterovic, Neurotoxicology 2014, 44, 80 90.

[14] A. H. Martins, J. Hu, Z. Xu, C. Mu, P. Alvarez, B. D. Ford, K. El Sayed, V. A. Eterovic, P. A. Ferchmin, J. Hao, Neuroscience 2015, 291, 250 259.

[15] P. Ferchmin, O. R. Pagán, H. Ulrich, A. C. Szeto, R. M. Hann, V. A. Eterovic, Toxicon 2009, 54, 1174 1182.

[16] a) H. N. Baraka, M. A. Khanfar, J. C. Williams, E. M. El-Giar, K. A. El Sayed, Planta Med. 2011, 77, 467 476; b) K. A. El Sayed, S. Laphookhieo, M. Yousaf, J. A. Prestridge, A. B. Shirode, V. B. Wali, P. W. Sylvester, J. Nat. Prod. 2008, 71, 117 122; c) I. Wahlberg, A. Eklund in Progress in the Chemistry of Organic Natural Products (Eds.: W. Herz, G. Kirby, R. E. Moore, W. Steglich, C. H. Tamm), Springer-Verlag, New York, 1992, pp. 142 293.

[17] a) J. Arnarp, W. L. A. Chu, C. R. Enzell, G. M. Hewitt, J. P. Kutney, K. Li, R. K. Milanova, H. Nakata, A. Nasiri, Y. Okada, Acta Chem. Scand. 1993, 47, 683; b) K. A. El Sayed, S. Laphookhieo, H. N. Baraka, M. Yousaf, A. Hebert, D. Bagaley, F. A. Rainey, A. Muralidharan, S. Thomas, G. V. Shah, Bioorg. Med. Chem. 2008, 16, 2886 2893.

[18] V. A. Eterovic, A. Del Valle-Rodriguez, D. Pérez, M. Carrasco, M. A. Khanfar, K. A. El Sayed, P. A. Ferchmin, Bioorg. Med. Chem. 2013, 21, 4678 4686.

[19] a) E. Wang, S. Gan, G. J. Wagner, J. Exp. Bot. 2002, 53, 1891 1897; b) E. Wang, G. J. Wagner, Planta 2003, 216, 686 691; c) E. Wang, R. Wang, J. DeParasis, J. H. Loughrin, S. Gan, G. J. Wagner, Nat. Biotechnol. 2001, 19, 371 374.

[20] H. M. Girvan, A. W. Munro, Curr. Opin. Chem. Biol. 2016, 31, 136 145.

[21] a) S. T. Jung, R. Lauchli, F. H. Arnold, Curr. Opin. Biotechnol. 2011, 22, 809 817; b) C. J. C. Whitehouse, S. G. Bell, L.-L. Wong, Chem. Soc. Rev. 2012, 41, 1218 1260.

[22] a) L. O. Narhi, A. J. Fulco, J. Biol. Chem. 1987, 262, 6683 6690; b) L. O. Narhi, A. J. Fulco, J. Biol. Chem. 1986, 261, 7160 7169.

[23] a) S. Kille, F. E. Zilly, J. P. Acevedo, M. T. Reetz, Nat. Chem. 2011, 3, 738 743; b) J. N. Kolev, K. M. O Dwyer, C. T. Jordan, R. Fasan, ACS Chem. Biol. 2014, 9, 164 173; c) J. C. Lewis, S. M. Mantovani, Y. Fu, C. D. Snow, R. S. Komor, C.-H. Wong, F. H. Arnold, ChemBioChem 2010, 11, 2502 2505; d) J. Reinen, G. Postma, C. Tump, T. Bloemberg, J. Engel, N. P. E. Vermeulen, J. N. M. Commandeur, M. Honing, Anal. Bioanal. Chem. 2016, 408, 1425 1443; e) X. Ren, J. A. Yorke, E. Taylor, T. Zhang, W. Zhou, L. L. Wong, Chem. Eur. J. 2015, 21, 15039 15047; f) A. Seifert, S. Vomund, K. Grohmann, S. Kriening, V. B. Urlacher, S. Laschat, J. Pleiss, ChemBioChem 2009, 10, 853 861.

[24] P. Le-Huu, T. Heidt, B. Claasen, S. Laschat, V. B. Urlacher, ACS Catal. 2015, 5, 1772 1780.

118

[25] The spatial orientation of the groups on C-9 stays the same; however, the notation changes because of the different Cahn Ingold Prelog priorities upon C-10 hydroxylation.

[26] S. G. Smith, J. M. Goodman, *J. Org. Chem.* **2009**, *74*, 4597 4607.

[27] S. G. Smith, J. M. Goodman, *J. Am. Chem. Soc.* **2010**, *132*, 12946 12959.

[28] K. A. El Sayed, P. W. Sylvester, *Expert Opin. Invest. Drugs* **2007**, *16*, 877 887.

[29] M. Razzak, J. K. de Brabander, *Nat. Chem. Biol.* **2011**, *7*, 865 875.

[30] a) Y. Ishihara, P. S. Baran, *Synlett* **2010**, *12*, 1733 1745; b) M. C. White, *Synlett* **2012**, *23*, 2746 2748.

[31] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, M. T. Stahl, *J. Chem. Inf. Model.* **2010**, *50*, 572 584.

[32] T. A. Halgren, *J. Comput. Chem.* **1999**, *20*, 720 729.

[33] Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.

[34] a) A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098 3100; b) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785 789.

[35] G. Scalmani, M. J. Frisch, *J. Chem. Phys.* **2010**, *132*, 114110.

[36] R. Ditchfield, *Mol. Phys.* **1974**, *27*, 789 807.

[37] C. Adamo, V. Barone, *J. Chem. Phys.* **1998**, *108*, 664 675.

[38] G. K. Pierens, *J. Comput. Chem.* **2014**, *35*, 1388 1394.

[39] M. W. Lodewyk, C. Soldi, P. B. Jones, M. M. Olmstead, J. Rita, J. T. Shaw, D. J. Tantillo, *J. Am. Chem. Soc.* **2012**, *134*, 18550 18553.

119

Paper IV

Simulation-guided design of cytochrome P450 for chemo- and regioselective macrocyclic oxidation

by D. Petrović, A. Bokel, M. Allan, V.B. Urlacher, and B. Strodel

D.P. contributed to this paper by designing the study and performing the simulations. Data analysis was jointly carried out by D.P. and M.A. (under the supervision of D.P.). D.P. wrote the manuscript with inputs from the coauthors.

Simulation-Guided Design of Cytochrome P450 for Chemo- and Regioselective Macrocyclic Oxidation

Dušan Petrović,[*,†,⊥] Ansgar Bokel,[‡] Matthew Allan,[†,§] Vlada B. Urlacher,[‡] and Birgit Strodel[*,†,∥]

[†]Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany

[‡]Institute of Biochemistry, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

[§]Schreyer Honors College, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

[∥]Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

S *Supporting Information*

ABSTRACT: Engineering high chemo-, regio-, and stereo-selectivity is a prerequisite for enzyme usage in organic synthesis. Cytochromes P450 can oxidize a broad range of substrates, including macrocycles, which are becoming popular scaffolds for therapeutic agents. However, a large conformational space explored by macrocycles not only reduces the selectivity of oxidation but also impairs computational enzyme design strategies based on docking and molecular dynamics (MD) simulations. We present a novel design workflow that uses enhanced-sampling Hamiltonian replica exchange (HREX) MD and focuses on quantifying the substrate binding for suggesting the mutations to be made. This computational approach is applied to P450 BM3 with the aim to shift regioselectively toward one of the numerous possible positions during *β*-cembrenediol oxidation. The predictions are experimentally tested and the resulting product distributions validate our design strategy, as single mutations led up to 5-fold regioselectivity increases. We thus conclude that the HREX-MD-based workflow is a promising tool for the identification of positions for mutagenesis aiming at P450 enzymes with improved regioselectivity.
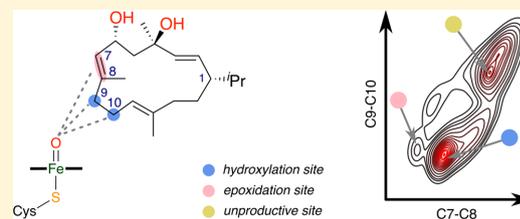
INTRODUCTION

The selective oxidation of an unactivated C−H bond for many years presented a major problem for chemists.[1,2] Even today, an environmentally friendly chemo-, regio-, and stereoselective oxidation of hydrocarbons is a challenging task.[3,4] Nature, however, has evolved umpteen enzyme catalysts for this purpose.[5,6] The heme-containing monooxygenases (i.e., cytochromes P450 or CYPs) use the electrophilic oxyferryl radical of heme, commonly known as the *compound I* (cI),[7−9] to insert one atom of molecular oxygen into a hydrocarbon skeleton, such as a fatty acid or steroid. A CYP from *Bacillus megaterium*, namely CYP102A1 or P450 BM3, typically performs subterminal hydroxylation of saturated and hydroxylation/epoxidation of (poly)unsaturated fatty acids. Due to a number of attractive features, P450 BM3 is commonly used as the CYP scaffold to engineer catalysts for oxidation of novel substrates.[6,7,10]

To bind unnatural substrates, especially bulky ones, enzymes often require an active site overhaul.[11] Exchanging Phe87, which is located in the active site, with a residue with a smaller side chain allows for the extension of the P450 BM3 binding pocket and enables a bulky substrate to bind in the vicinity of heme.[12] On the other hand, engineering overly spacious active sites can lead to a high substrate mobility that is often reflected as enzyme promiscuity.[13,14] In a simplified view of P450

engineering, there should be enough space in the active site for a substrate to bind in a desired, transition-state-like pose, but not more than that, which would allow alternative binding modes.[15] In addition to the shape and size match, the proper electrostatic complementarity between the substrate and the active site and the existence of a geometrically precise H-bond network can lead to better-performing enzymes.[16,17]

The mutation of P450 BM3 first-shell residues around a substrate has a strong impact on the substrate selectivity and activity.[18] Many P450 BM3 variants have been engineered for the selective oxidation of different classes of organic molecules, such as alkanes (e.g., *n*-octane and cyclododecane),[19] polycyclic aromatic hydrocarbons (e.g., naphthalene[20] and anthracene[21]), alkaloids (e.g., thebaine and dextromethorphan),[22] steroids (e.g., testosterone[23] and norandrostenedione[24]), and terpenes (e.g., geranylacetone,[25] (+)-valencene,[26] and artemisinin[27]). Macrocycles have been recognized as potential scaffolds for therapeutic agents,[28−30] and we recently studied the P450 BM3-catalyzed transformations of (1*S*,2*E*,4*R*,6*R*,7*E*,11*E*)-4,6-dihydroxycembra-2,7,11-triene or *β*-cembrenediol (**1**)[31] and its oxidation products (**2−4**)[32] (Figure 1). A monocyclic diterpenoid isolated from *Nicotiana tabacum*, **1**, is characterized
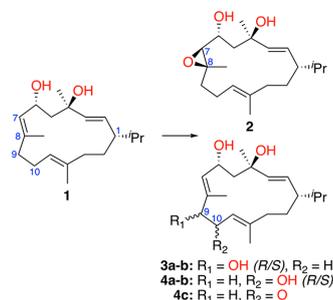
Article

Figure 1. Common oxidation products of β-cembrenediol (**1**) by P450 BM3 mutants.[31,32]

by a 14-membered macrocyclic ring. β-Cembrenediol (**1**) is a challenging substrate for selective oxidation as well as for rational design, as it is flexible and bears seven potential allylic and six nonallylic hydroxylation sites, as well as three epoxidation sites. While the conversion of **1** by the wild-type P450 BM3 is lower than 2%, the Phe87 mutations create more spacious active sites leading to higher conversion rates (i.e., 6% for the F87A and 59% for the F87G mutation).[31] However, the increased conversion rate of **1** by the F87G mutant comes at the cost of low chemo-, regio-, and stereoselectivities. Additional mutations to either F87A or F87G P450 BM3 mutants typically increase the conversion rates, with unpredictable implications on the product selectivity.[31]

In this study, we present a computational workflow to predict positions influencing the regioselectivity of P450 BM3 for oxidation of the flexible macrocycle β-cembrenediol, which bears numerous potential oxidation sites. The P450 BM3 double mutant V78A/F87A was used as a parent enzyme for improvement due to its decent conversion rate of **1** but poor chemo-, regio-, and stereoselectivities.[31] Shaik and co-workers, among others, argued that the selectivity of CYP-catalyzed monooxygenation is, in fact, a combination of *electronic* (i.e., substrate activation) and *classical* factors (i.e., substrate binding and interactions with the active site pocket).[33] While electronic factors can explain why some potential oxidation sites are nonreactive (e.g., high activation energies),[34] substrate binding is a very important factor for determining regio- and stereoselectivities of CYPs with fast cI activation of a substrate.[33,35−38] To this end, we aimed to test if solely optimizing the binding of β-cembrenediol would lead to changes in regioselectivity. Our strategy involves substrate docking by Hamiltonian replica exchange (HREX) molecular dynamics (MD) simulations to elucidate how **1** binds to the enzyme, which allowed us to adjust the shape and electrostatic complementarity at selected binding hotspots by introducing mutations with the aim to shift product distribution. The quality of our models is validated by experimentally measured product distributions.

■ METHODS

Unliganded Enzyme MD. The starting structure of the parent enzyme (V78A/F87A P450 BM3) was prepared from the 1.65 Å resolution crystal structure of the wild-type enzyme (PDB ID: 1JPZ[39]) by truncating the side chains of Val78 and Phe87 to alanine. The 1JPZ crystal structure contains *N*-palmitoylglycine, which was removed prior to simulations. The mutated protein was relaxed in MD simulations performed

in GROMACS 5.0.4,[40] using the Amber 99SB*-ILDN force field[41,42] with TIP3P water.[43] The heme parameters for the cI state were adopted from Cheatham and co-workers.[44] The protein was centered in a dodecahedral box and positioned at least 10 Å away from any box edge. The protonation states were assigned to all titratable protein residues based on the PROPKA 3.1[45] analysis at a pH of 7.5. The total charge was neutralized by adding sodium ions to the solvated system, followed by minimization of the system in two steps: steepest descent (force convergence criterion of 500 kJ mol^{-1} nm^{-1}) and conjugate gradient (force convergence criterion of 100 kJ mol^{-1} nm^{-1}). The system was modeled under periodic boundary conditions where electrostatic interactions were treated with the particle mesh Ewald method.[46] The short-range nonbonded interactions were calculated within a cutoff of 9 Å. An integration step of 2.0 fs was used, while bond lengths were constrained using the LINCS algorithm.[47] The minimized system was gradually heated to 298 K and equilibrated using the v-rescale thermostat[48] for 200 ps, with the protein Cα atoms restrained using a positional restraint force constant of 1000 kJ mol^{-1} nm^{-2}. An NPT equilibration was subsequently carried out for 2.8 ns, over which the restraint forces were gradually reduced to 10 kJ mol^{-1} nm^{-2}, and the pressure was kept constant at 1 bar using the Berendsen barostat.[49] The 200 ns unrestrained production MD was performed in the NPT ensemble using the Parrinello−Rahman barostat[50] at 1 bar and the v-rescale thermostat at 298 K. The coordinates of the system were collected every 10 ps. The unliganded enzyme was stable during both MD simulations (Figure S1). All simulation snapshots were clustered based on the backbone conformation, using the Daura algorithm,[51] to produce representative enzyme structures. The structures of the two most populated clusters were used for the subsequent simulations.

Molecular Docking. The conformational ensemble of **1** was generated using the OMEGA 2.5.1.4 software,[52] which employs fragment libraries with a subsequent exhaustive knowledge-based scan of rotatable bonds. Energies were estimated using the MMFF94s force field *in vacuo*.[53] Conformations with root-mean-square deviations (RMSDs) of atomic coordinates of less than 0.5 Å were considered to be identical and only the lowest energy conformer from the set of identical structures was kept. A set of 20 conformers was generated with energies up to 42 kJ mol^{-1} higher than the lowest energy conformer, which were subsequently employed in rigid docking using the FRED 3.0.1 tool from the OEDocking suit.[54] The receptor active site was defined as a 20 Å × 20 Å × 20 Å box that encompasses cI. For each of the two protein models obtained from clustering, 100 binding poses of **1** were generated. From these, four binding modes, with different geometries but similar binding affinities were selected for further MD simulations.

Enzyme−Substrate MD. The topology of **1** was created using ACPYPE[55] and Antechamber.[56] The restrained electrostatic potential (RESP) charges, shown in Figure S2, were determined at the HF/6-31G*//B3LYP/6-31G* level of theory in Gaussian 09.[57] The MD simulations of the enzyme−substrate systems were prepared and run in the same manner as the MD simulations of the unliganded enzyme. Each of the four production simulations was 200 ns long.

The Hamiltonian replica exchange simulations were initiated from the same starting structures as the enzyme−substrate complex MD simulations. HREX-MD was performed in GROMACS 4.6.7 patched with the PLUMED 2.1 plugin.[58,59]

124

To enhance the conformational sampling of β-cembrenediol, six replicas were simulated for each of the four binding modes, where the energy term (i.e., the Hamiltonian) for the substrate's nonbonded interactions was scaled to effectively correspond to temperatures between 300 and 600 K (the exact λ scaling factors were 1.00, 0.87, 0.76, 0.66, 0.57, and 0.50). The chosen λ range of 1.0−0.5 should be sufficient to enhance the conformational sampling of β-cembrenediol, while at the same time achieving optimal exchange rates (i.e., 30−70%) with a low number of replicas. In the GROMACS/PLUMED implementation of HREX-MD,[59] charges of all "hot" atoms (i.e., the substrate in the present case) were scaled by a factor $\sqrt{\lambda}$, while the depth of the Lennard-Jones potentials (ε) and the torsional potentials were scaled by a factor λ.

The exchange between replicas was attempted every 2 ps, giving exchange rates of around 55%. Coordinates were saved every 5 ps during the 115 ns simulation per replica, giving a total of ~2.8 μs sampling time (6 replicas × 115 ns × 4 binding modes). The initial 15 ns of each replica were discarded as equilibration, and data analysis was performed on the replicas with the unperturbed Hamiltonian (i.e., the scaling factor of 1). Binding density surfaces were produced from the HREX-MD data, showing the binding preferences of **1** along selected enzyme−substrate distances or angles.

MM/PBSA Calculations. After the alignment of the protein backbone, the snapshots from the combined HREX-MD trajectories were clustered based on the orientation of **1** in the active site using the Daura algorithm with a cutoff of 1.5 Å. The representative structures of each cluster were projected onto the binding density surfaces and the clusters at or close to binding density maxima considered for further analysis. For each of the selected six clusters, a set of 30 short MD simulations was performed for the molecular mechanics/Poisson−Boltzmann surface area (MM/PBSA) calculations. In each simulation, the initial velocities were randomized and the system was equilibrated in the NPT ensemble for 0.5 ns, followed by a 2 ns production sampling run. We confirmed that each simulation sampled only a narrow region around its initial binding mode. The MM/PBSA calculations were performed for each cluster on a combined 60 ns trajectory (30 simulations of 2 ns) with the g_mmpbsa tool,[60] assuming a solute dielectric constant of 4. The binding free energy was decomposed on a per-residue basis.

Simulation Analysis. Data analysis was performed using GROMACS 5.0.4 tools, MDAnalysis,[61] and VMD.[62] Shown structures were rendered with PyMOL.[63]

Experimental Reagents. β-Cembrenediol was purchased from Santa Cruz Biotechnology (Dallas, USA). Ethyl acetate and methanol were from Fisher Chemical (Waltham, USA) in either analytical grade (ethyl acetate) or certified ACS reagent grade (methanol).

Cloning and Expression. The pET28a plasmid (Novagen) with the integrated gene encoding P450 BM3 (GenBank J04832) from *B. megaterium* was already available in our laboratory. Mutants of P450 BM3 were created following a modified protocol of Edelheit et al.[64] and expressed as described elsewhere.[31] Glucose dehydrogenase (GDH) from *B. megaterium* (gdhIV, GenBank D10626) was used for NADPH cofactor regeneration, which was available in our laboratory in pET22b vector. Expression was done using *E. coli* BL21(DE3) cells. First, precultures comprising 5 mL Lysogeny Broth (LB) medium, containing 100 μg mL^{-1} ampicillin, were inoculated with single colonies and incubated overnight at 37

°C and 180 rpm. For the main cultures, 400 mL TB medium in 2 L flasks, supplemented with 100 μg mL^{-1} ampicillin, were inoculated with the preculture to O.D.$_{600}$ 0.05 and incubated at 37 °C and 180 rpm until O.D.$_{600}$ 0.6−0.8 was reached. Gene expression was then induced by adding 0.25 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and cultures were further kept at 25 °C and 140 rpm for 20 h. Harvested cells were resuspended in 50 mM potassium phosphate buffer (pH 7.5, 0.5 M NaCl) supplemented with 100 μM phenylmethylsulfonyl fluoride (PMSF) and disrupted by sonification on ice. Afterward, cell debris was removed by centrifugation.

Enzyme Assays. The concentrations of all P450 BM3 variants were calculated from CO difference spectra using the extinction coefficient $\varepsilon_{450-490}$ = 91 mM^{-1}.[65] Since the spectra were not recorded under the exclusion of oxygen, reduction with sodium dithionite followed after bubbling with CO. Absorbance spectra were recorded from 400−500 nm at least three times successively. The NADP$^+$ reduction activity of GDH was measured by the increase of absorbance at 340 nm (ε_{340} = 6.22 mM^{-1} cm^{-1}) over a period of 1 min. The reaction solutions contained 100 mM glucose and 100 μL of diluted, cell-free sample in 900 μL 50 mM potassium phosphate buffer, pH 7.5. The reaction was started by adding 100 μL NADP$^+$ (1 mM).

LC/MS Analysis of Product Distribution. For the experimental validation of the reaction selectivity of the suggested P450 BM3 mutants, the following reaction setup in a total volume of 500 μL was used. The reaction mixture contained 100 μM β-cembrenediol (**1**) dissolved in ethanol (reaction finally contained 2% (v/v) ethanol), 2.5 μM P450 BM3 mutants, and 200 μM NADPH. The latter was regenerated by addition of 5 U mL^{-1} GDH and 20 mM glucose. Removal of hydrogen peroxide, which might arise from uncoupling reactions, was accomplished by adding 600 U mL^{-1} catalase from bovine liver (Sigma-Aldrich). The reactions were performed in potassium phosphate buffer (50 mM, pH 7.5) at 25 °C and 500 rpm (ThermoShaker) for 17 h. Prior the extraction (twice with 300 μL ethyl acetate), 25 μM dioctyl phthalate was added as internal standard. The combined ethyl acetate phases were evaporated to dryness and resolved in methanol for liquid chromatography/mass spectrometry (LC/MS) analysis.

LC/MS analysis was carried out on a Prominence/LCMS 2020 instrument (Shimadzu, Duisburg, Germany) equipped with a Chromolith RP-18e 100-4.6 column (Merck, Darmstadt, Germany). Oven temperature was set to 30 °C and a flow rate of 1 mL min^{-1} was chosen. The following solvent gradient of 0.1% formic acid (A) and methanol (B) was used: increasing B from 50% to 70% B within the first 25 min, and further increasing it to 100% within another 10 min, finally holding for 2 min. Mass spectroscopy was done in the positive scan mode measuring electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI) in the dual ionization mode at the same time.[32]

To minimize the cell extract background, negative controls were done with crude extract of *E. coli* carrying the pET28a empty vector (vector does not contain the CYP102A1 encoding gene). The products were identified by comparison of retention times of the corresponding peaks and m/z values with published data.[31] The conversion was calculated using a calibration curve with dioctyl phthalate as internal standard. The product distribution was calculated based on the observed product peak areas under the assumption that the ionization of

the metabolites is similar. All reactions were performed in triplicate.

RESULTS

β-Cembrenediol Docking. The first objective of this study was to investigate how **1** binds to the parent enzyme, bearing in mind that the binding mode can dictate reaction selectivity. A typical approach to study a small molecule binding to a protein is molecular docking. Substrate docking and classical MD simulations have been successfully used to predict the selectivity of P450 BM3 for alkanes and fatty acids, see e.g., ref 66, but fatty acids are native substrates for P450 BM3 and there are several substrate-bound crystal structures that can assist in docking. A combination of short MD simulations (3 ns) to study binding and quantum mechanical (QM) calculations to obtain transition state geometries and reaction barriers was used to understand the reaction selectivity of P450 2D6 for the antipsychotic drug thioridazine.[67] More exhaustive sampling (500 ns) was recently used by Houk and co-workers to engineer the desired reactivity in PikC monooxygenase for menthol derivatives.[68]

Macrocycle docking is, however, a more challenging task because of the larger conformational space than that accessible to small organic molecules.[69] This problem usually requires a thorough conformational analysis of the substrate and docking of a predefined substrate conformational library to a rigid protein.[70] To this end, we performed the conformational analysis of **1** using OMEGA. We recently showed that this procedure gives reasonable coordinates for molecules **2** and **3**, which enabled us to study their NMR spectra.[32] Furthermore, one of the conformers of **1** that was generated in this work corresponds to the crystal structure of β-cembrenediol (Figure S3).[71]

To account for protein flexibility in substrate docking,[72,73] we introduced the mutations to P450 BM3 (V78A and F87A), performed MD simulations to relax the protein, and clustered the conformations to identify representative structures. The two most populated clusters were selected, and the pregenerated conformational ensemble of **1** was docked into each instance of the rigid parent enzyme. Figure 2 shows that the regiopreferences from docking, which we define here based on the closest contact of the axial oxygen of cI and the carbon atoms of the substrate, are not fully corresponding to the
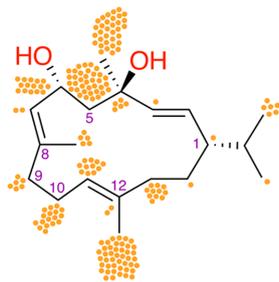

Figure 2. Rigid ensemble docking of **1** to the parent enzyme models. Data points indicate the closest contact between the axial oxygen of cI and C atoms of **1** in the binding modes resulting from docking. Only poses with FRED scores[52] ≤ −4 are shown. Most of the observed docking poses would not lead to oxidation of C atoms that were identified experimentally, i.e., C7/C8, C9, and C10 atoms.

experimentally identified products of **1** by the parent enzyme (i.e., P450 BM3 V78A/F87A).[31] Karlén et al. identified three requirements in rigid ensemble docking that commonly lead to inaccurate results if one or more of these requirements are not fulfilled: (1) proper *conformational analysis of the substrate*, and especially the generation of "protein-bound-like" conformations, (2) exhaustive *search of the substrate binding poses*, and (3) accurate *scoring* and ordering of the binding modes.[70] Since the discrepancies between simulated and experimental data can result from inaccuracies at any of these steps in β-cembrenediol docking, we turned to MD simulations to more thoroughly sample the conformational space of both the protein and the substrate.

To study the substrate dynamics, we selected four binding modes of **1** obtained from docking. The selected binding modes, numbered 1−4 in Figure S4, had similar docking scores but substantially different orientations in the active site. For structurally complex substrates, at commonly employed nanosecond time scales, unbiased MD simulations are incapable of crossing high free energy barriers that connect several metastable substrate binding modes. This is confirmed by the results obtained from our 200 ns MD simulations of the enzyme−substrate complex (see the ESI for details, including Figures S5 and S6). To overcome the barriers and describe the complete binding process, μs-long unbiased MD simulations are needed.[74] An alternative is applying enhanced-sampling MD approaches, e.g., adaptive biasing force MD,[75] alchemical transformations,[76] or computationally expensive 2D umbrella sampling (US) MD.[77] We increased the conformational sampling of **1** using HREX-MD, which we have recently shown to provide a level of conformational sampling comparable to that of US-MD.[78] While the substrate was very stable in the unbiased simulations (i.e., rather constant RMSD, Figure S5), it was quite flexible in the HREX-MD simulations, which is indicated by noisier RMSD profiles exceeding 7 Å from the initial structure. In addition, the substrate often visited three to four different regions in the RMSD space of each simulation (Figure S7), indicating that **1** was exhaustively exploring the active site in the HREX-MD simulations.

Binding Density Surfaces. To rationalize the selectivity of the parent enzyme based on binding preferences, we chose any two oxidation products and used their corresponding reaction coordinates to investigate the binding density surface (BDS), which we define as the 2D representation of two probability density functions. For the selection of the reaction coordinates we considered that the allylic C9- or C10-hydroxylation of **1** by the parent enzyme follows a mechanistic pathway that begins with the abstraction of a hydrogen atom from β-cembrenediol by the axial oxygen of cI,[8,9] while the initial cI attack for epoxidation is on the C7−C8 π-bond of the substrate.[79] To investigate the regiopreference of the substrate binding, we assumed that the corresponding position of **1** (i.e., C7, C8, C9, or C10) needs to be sufficiently close to the active oxygen species for the reaction to occur. Mulholland et al. suggested a common cutoff of 4 Å for the $C_{substrate}$−O_{cI} contact in the reactant state, which would lead to a model that resembles the transition state as closely as possible.[80] In addition, Houk et al. showed, based on QM calculations, that the $H_{substrate}$−O_{cI} distance is often ∼2.5 Å longer in the reactant than at the transition state leading to hydroxylation.[68]

We assumed that the chemo- and regioselectivity of C7,C8-epoxidation versus C9/C10-hydroxylation would be deter-
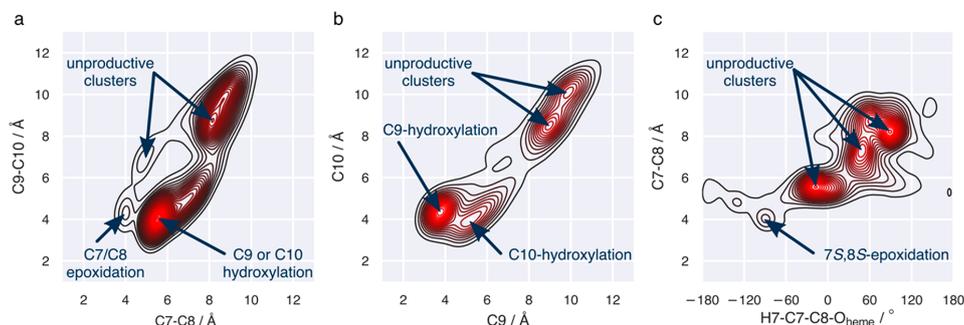
Article

Figure 3. Binding density surfaces underpinning chemo-, regio-, and stereoselectivity of the parent enzyme with **1**. (a) Chemo- and regioselectivity are preferential for C9/C10-hydroxylation rather than for C7/C8-epoxidation. (b) Hydroxylation at C9 is preferred to that at C10. (c) C7/C8-epoxidation is *S,S*-stereoselective.

mined by the distance from the cI oxygen atom to the centroids (i.e., the centers of mass) of the C7−C8 and C9−C10 bonds, respectively (Figure 3a). Several local maxima can be identified on the corresponding BDS of the parent enzyme, two of which place carbon atoms of the substrate sufficiently close to the cI oxygen to lead to product formation. Binding is more favorable for C9/C10-hydroxylation, as it is characterized by a broader and higher maximum, which is supported experimentally by the high yield (>60%) of hydroxylation products.[31] Although other maxima on the plot exist, they place the relevant atoms of **1** too far from cI to allow proton abstraction and are thus denoted *unproductive* for the specified reactivity.

After showing that the BDS approach resembles the experimental preference for hydroxylation over epoxidation, we investigated its ability to predict regioselectivity of hydroxylation of the parent enzyme at C9 vs C10. Two productive and two unproductive maxima could be identified on this BDS (Figure 3b). The maximum corresponding to C10-hydroxylation is lower and narrower than the one for the C9-hydroxylation, indicating a geometrically smaller window of opportunity for C10 to react. The preference for C9-hydroxylation on the BDS agrees qualitatively with the experimentally observed product distribution for P450 BM3 V78A/F87A.

The BDS approach also allows to describe the stereoselectivity of the hydroxylation and epoxidation reactions. For epoxidation, the dihedral angle criterion distinguishes the oxygen to attack from either the *re-* or *si-* side of the substrate (Figure 3c). However, only one maximum on this BDS is at a distance short enough for the reaction to occur, which corresponds to the sterically less restricted 7*S*,8*S*-epoxidation product (**2**). For understanding the stereoselectivity of the C9- and C10-hydroxylations, we constructed the BDSs using distances from the *pro-R* and *pro-S* H atoms to cI (Figure S8).

Selection of Mutagenesis Hotspots from Binding Free Energy Analysis. Once we generally understood how **1** binds to the parent enzyme, we identified hotspot residues whose mutation would lead to a shift in selectivity. For this, we focused on two main properties: enzyme−substrate interaction energies and H-bond networks. However, constructing one binding surface that encompasses all reaction coordinates for chemo-, regio-, and stereoselectivity would require a highly multidimensional analysis and would not be practical. In order to simultaneously consider all the coordinates, we needed to identify structural and geometric similarities between different

positions on the BDSs (Figures 3 and S8). To this end, we performed geometrical clustering of substrate positions over the HREX-MD simulation trajectory and projected the six dominant cluster representatives to BDSs shown in Figure S9.

The interaction free energies between the enzyme and substrate were investigated for different clusters using the MM/PBSA method. To this end, we ran a set of short MD simulations starting from each of the six clusters and ensured that they remained in their respective binding modes (Figure S10). From these simulations, we calculated the binding free energies and decomposed them per residue to identify important amino acids capable of stabilizing or destabilizing the substrate binding mode in each cluster. Although the MM/PBSA method is not the most accurate method for calculating the binding free energies, it works sufficiently well (see e.g., ref 81) and it is faster than other, more precise methods (e.g., alchemical transformations),[82,83] allowing it to be incorporated in an enzyme design strategy. The selection strategy for mutagenesis hotspots, based on the MM/PBSA results, is described in detail in the ESI. Figure 4 summarizes the identified stabilizing and destabilizing residues for several potential selectivity goals: regioselective C10-hydroxylation vs C7,C8-epoxidation and stereoselective C9-hydroxylation.

The Euler diagrams (Figure 4a,b) help to visualize that a hotspot for a given goal cannot be stabilizing and destabilizing at the same time. However, the hotspots for one goal do not necessarily depend on those for another. For example, while C7,C8-epoxidation is stabilized by Met185, destabilized by Ile263, and unaffected by Ala330, these three residues all stabilize C10-hydroxylation (Figure 4a). Interestingly, for the stereoselectivity of C9-hydroxylation, all positions that stabilize the *R*-epimer are destabilizing for the *S*, and *vice versa* (Figure 4b). Figure 4c compares all four goals at the same time and was used to make suggestions for mutation.

Identification of Promising Mutations. To identify the most beneficial mutations for each hotspot, we considered the structural and physicochemical properties of alternative amino acids at the hotspots. In testing for steric interactions, we performed *in silico* mutations. In the case where all rotamers (taken from the Dynameomics rotamer library[84]) of an amino acid were clashing with the substrate or the rest of the protein, that mutation was discarded. If we, for example, wanted to mutate Ala to an amino acid able to form an H-bond, we tested those amino acids with the matching properties. Based on these considerations, the most promising one or two mutations for a
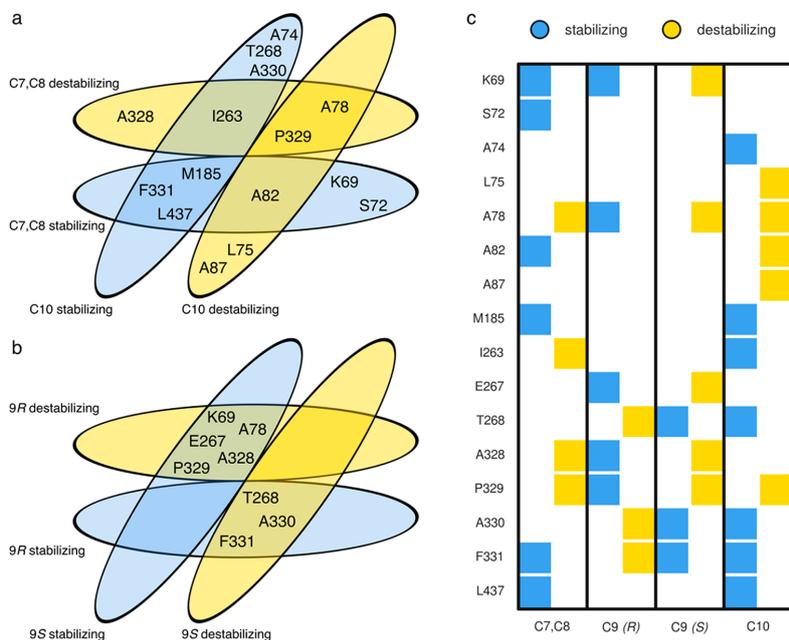
Article

Figure 4. Identification of the V78A/F87A P450 BM3 amino acid residues that stabilize or destabilize β-cembrenediol for a desired selectivity. Euler diagrams for (a) C7,C8-epoxidation vs C10-hydroxylation and (b) 9R- vs 9S-hydroxylation. (c) Combined diagram connecting all four goals.

Table 1. Selected Hotspots and Mutations for the V78A/F87A P450 BM3 Mutant

hotspot	mutation	goal	reason
Lys69	Arg	reduce C9	identified as stabilizing for C7,C8-epoxidation
			known to coordinate the propionate groups of heme
			K69R maintains heme coordination and at the same time destabilizes **1** in a mode productive for C9-hydroxylation
Ser72	Ala, Ile, Leu	increase C10	identified as stabilizing for C7,C8-epoxidation
			too far from the substrate to form H-bond (Figure 5a)
			adding a bulky hydrophobic residue (e.g., Leu) fills the void between the enzyme and substrate and locks it in a binding mode productive for C10-hydroxilation (Figure 5b)
Leu75	Ala	increase C10	identified as destabilizing only for the C10-hydroxylation
			clashes with the substrate in its productive conformation for C10-hydroxylation (Figure 5c)
			L75A mutant removes that clash and anchors the substrate in the appropriate position, which increases the likelihood of this substrate conformation to be sampled
Thr268	Ala, Ser	increase C10 C7,C8	identified as stabilizing for C10-hydroxylation as the methyl group stabilizes the necessary binding mode of **1**
			T268A mutation preserves the methyl group but removes the H-bonding capability, favoring C10-hydroxylation
			T268S mutation preserves the H-bond network with the axial oxygen of cI (Figure 5d) and OH-group on C6 of the substrate but removes the anchoring methyl group, which stabilizes the mode for epoxidation
Ala328	Ser	increase C7,C8	identified as destabilizing for epoxidation due to shape and polarity mismatch, i.e., methyl group positioned between the OH groups of the substrate (Figure 5e)
			A238S mutation introduces an H-bond to the OH-group of C6 of the substrate (Figure 5f), which is suitable for epoxidation
			also suitable for 9S-hydroxylation, but it geometrically prevents 9R-hydroxilation
Phe331	Tyr, Thr	increase C10	identified as stabilizing for epoxidation and hydroxylation
			located close to the C6 OH-group of the substrate
			mutation to a polar residue stabilizes **1** by an H-bond in a position suitable for C10-hydroxilation

certain goal were suggested per hotspot, which are summarized in Table 1.

Experimental Validation of Suggested Mutations. The conversions of **1** and the oxidation product distributions were measured for the parent enzyme (P450 BM3 V78A/F87A mutant) and the triple mutants, suggested in Table 1, to investigate the quality of the computational redesigns. For the parent mutant we observed the products **2** (24%), **3** (42%),

and **4** (12%), as well as a 22% of other products (Figure 6). Comparing the experimentally observed distributions of the products **2**, **3**, and **4** of the double and triple mutants (Figure 6) with our computational predictions (Table 1), we notice a qualitatively good agreement.

We should first note that for six out of ten tested triple mutants (i.e., V78A/F87A/S72A, V78A/F87A/L75A, V78A/F87A/T268A, V78A/F87A/T268S, V78A/F87A/F331Y, and
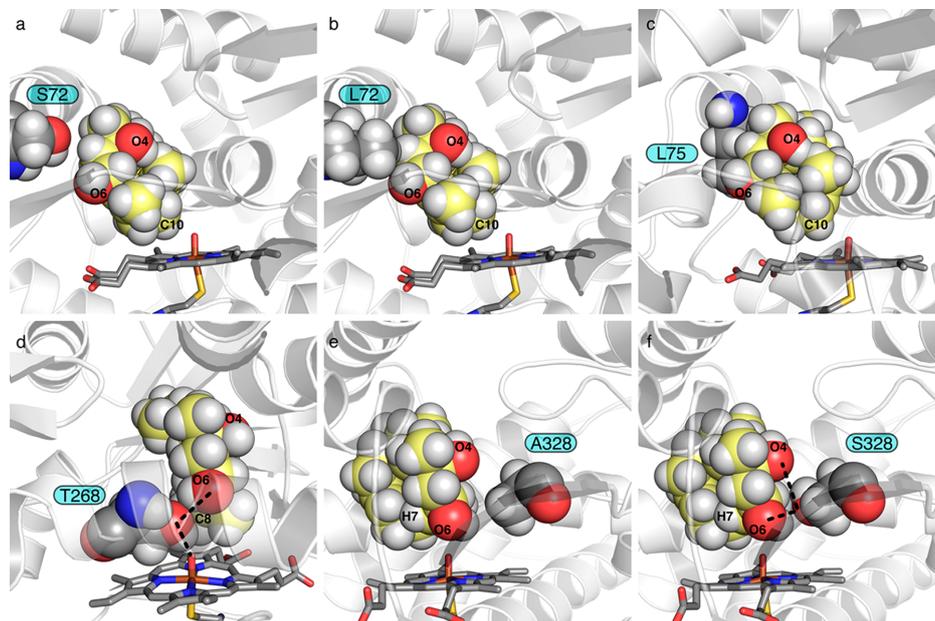
128

Figure 5. β-Cembrenediol−P450 BM3 interactions. The protein is shown in gray; protein hotspots and the substrate are shown as spheres; the C atoms of **1** are colored in light yellow; O and N atoms are colored in red and blue, respectively. H-bonds are indicated by dashed black lines. (a) S72 is too far away to form H-bonds with the substrate. (b) The S72L mutant could fill the void present in the productive mode for C10-hydroxylation. (c) L75 clashes with the substrate binding mode for C10-hydroxylation. (d) T268 forms an H-bond network with the substrate and cI. (e) A328 has a polarity mismatch for C7,C8-epoxidation. (f) The A328S mutation can form an H-bond with the substrate.
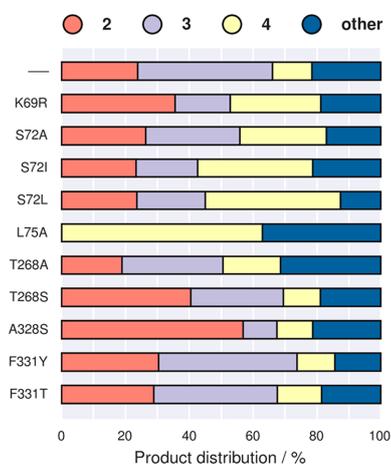

Figure 6. Experimental distribution of β-cembrenediol (**1**) oxidation products **2**−**4** for the parent enzyme and its designed mutants. For the detailed product distribution and conversion rates, consult Table S1.

V78A/F87A/F331T) we observed a similar level of substrate conversion of around 80−99%, indicating that those mutations are not detrimental for the activity. The remaining four mutants (i.e., V78A/F87A/K69R, V78A/F87A/S72I, V78A/F87A/S72L, and V78A/F87A/A328S) had significantly lower conversion rates (11−24%), suggesting that further optimization is needed. When comparing the ratio of a certain product

formed with the parent and triple mutant, for six of the ten suggested mutations we obtain at least 2-fold gain in the product intended by the given goal (i.e., V78A/F87A/K69R, V78A/F87A/S72A, V78A/F87A/S72I, V78A/F87A/S72L, and V78A/F87A/L75A for the C-10 hydroxylation (**4**) and V78A/F87A/A328S for the C7,C8-epoxidation (**2**)). While the K69R mutation was identified as beneficial for C10-hydroxylation, this mutant has a very low conversion rate; mutations at this position usually deteriorate catalytic parameters.[22,85] Another mutation that significantly decreases substrate conversion is A328S. This mutation, however, introduces a new H-bond between the enzyme and substrate and relatively enriches the C7,C8-epoxidation product in the mixture. In addition, this binding mode impedes the abstraction of the *pro-R* H atom at C9, preventing the formation of the **3a** product for this mutant. However, the benefit of the shift in product distribution for this mutant is significantly challenged by the low substrate conversion. The L75A mutation increases regioselective C10-hydroxylation over 5-fold (counting products **4a**, **4b**, and **4c**), completely removing the C9-hydroxylation and C7,C8-epoxidation products. This mutant also forms 37% of other uncharacterized products, identified at lower retention times. Two mutants showed modest improvements, enriching C10-hydroxylation (V78A/F87A/T268A) and C7,C8-epoxidation (V78A/F87A/T268S). Finally, only two mutants (V78A/F87A/F331Y and V78A/F87A/F331T) did not lead to a noteworthy change in the product distribution. Phe331 is located in the vicinity of the OH-group on C6 of the substrate, and these mutations were anticipated to form an H-bond with **1**, which apparently did not happen.

■ DISCUSSION AND CONCLUSIONS

P450 BM3 Design Workflow. Some CYPs show a remarkably high promiscuity in substrate binding. While in nature enzyme promiscuity is of great importance because it can lead to the evolution of new function, selectivity is one of the critical properties demanded from enzymes as biocatalysts in synthetic chemistry.[14] In this work, we addressed the challenging task of shifting product distribution for oxidation of macrocyclic substrates by P450 BM3 variants, by developing an *in silico* engineering methodology based on the assumption that the binding of the substrate determines reaction selectivity,[33,35−37] which is summarized in Figure 7.

In previous studies, substrate binding was investigated using various computational methods, including docking as well as unbiased and enhanced MD simulations.[66−68,75,77] Macrocycles are often more flexible than other substrates and, hence,
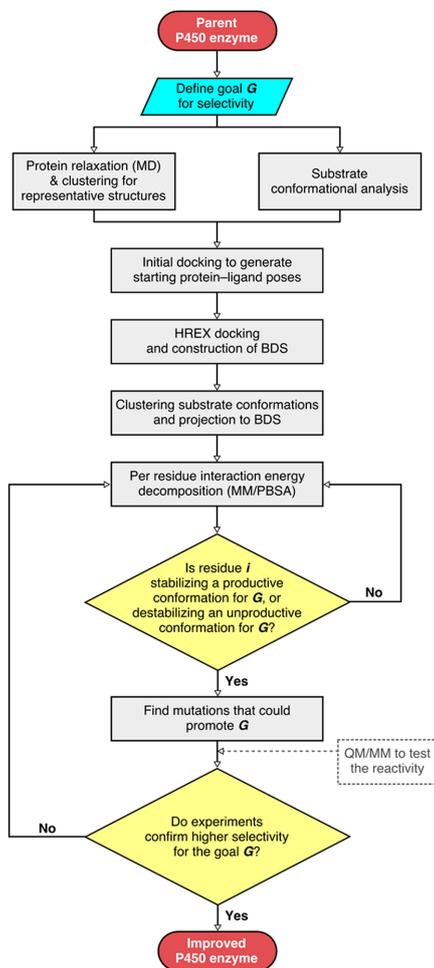

Figure 7. Flowchart overview for the P450 design process used in the present study. A possibility where quantum mechanics/molecular mechanics (QM/MM) calculations could be included is shown in dotted gray field.

particularly complicated to model.[69,70] To this end, we employed HREX-MD simulations to thoroughly explore the conformational ensemble of **1** in the active site of the parent enzyme (i.e., P450 BM3 V78A/F87A).[31,32] Based on the HREX-MD data, we constructed binding density surfaces along reaction coordinates representing the different chemo- and regioselectivities, which enabled us to identify productive an unproductive orientations of the substrate for a particular reaction. These were then submitted to MM/PBSA calculations for the mutagenesis hotspot selection, and mutations at hotspots were proposed to further stabilize the productive binding mode, or destabilize the unproductive ones.

The suggested mutations were evaluated through experiments and the resulting product distributions validated our computational workflow. Out of the ten predicted mutations, eight showed enhanced selectivity for either the C7,C8-epoxidation (**2**) or the C10-hydroxylation (**4**). In particular, the L75A/V78A/F87A mutant converts >99% of the substrate and mainly produces C10-hydroxylation products **4**, while **2−3** were not detected (the remainder are unidentified products). The V78A/F87A/A328S mutant increases the selectivity for **2** 2-fold; yet, at the cost of lowering the substrate conversion rate. However, it should be noted that enzyme activity was not considered during our design protocol, a problem which would require further QM/MM calculations.

Further Directions. At the current state, the BDS approach gives only a qualitative description of the binding. The conformational space would need to be sampled even more thoroughly than we already did with HREX-MD to obtain accurate binding free energies for each substrate orientation, which is beyond the scope of this manuscript. We use HREX-MD to study relative populations and propose that this limited approach is appropriate for enzyme design purposes, where identifying that a mutation leads to an increase (or decrease) in selectivity is more important than exactly quantifying this change with very expensive and time-consuming computational methods. Mulholland and co-workers successfully applied a similar approach using MD simulations to determine the distance between the substrate and residues important for its positioning to infer about reactivity and specificity of glutamate mutase.[86] However, it should be noted that, while in certain cases the substrate proximity is the leading factor in determining P450 selectivity, the proximity alone is not always enough to rationalize enzyme selectivity.[34,87,88] In such instances, further QM/MM[80] or empirical valence bond (EVB)[89] calculations would be necessary to investigate the effects of mutations on reactivity and selectivity.

The primary goal of the present study was to shift product distribution, which this *in silico* protocol demonstrated. Considering the ease of use and relatively low computational cost of HREX-MD simulations[78] (compared to other enhanced-sampling MD methods or QM/MM), the presented workflow could be incorporated in enzyme design strategies to reduce the screening efforts. While promising already, its performance could be further improved by involving QM/MM calculations to investigate chemical steps, which would help in selecting the mutations with high conversion rates (i.e., low activation barriers). The present study focuses on inverting the already observed product ratio; however, in future work it would be interesting to explore β-cembrenediol conformations currently designated unproductive, which would potentially introduce new reactivities. As other potentially reactive C-atoms were found sufficiently close to cI, addition of QM/MM

calculations would reveal if high activation barriers prevent such reactions from occurring, which could guide further P450 design strategies.

■ ASSOCIATED CONTENT

⑤ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00043.

Additional methodology description, including RESP charges for β-cembrenediol and mutagenesis hotspot selection. Supporting simulation results: RMSD and distance profiles, additional BDS profiles, and clustering for MM/PBSA. Supporting experimental validation results on substrate conversion by the designed mutants (PDF)

■ AUTHOR INFORMATION

Corresponding Authors
*E-mail: dusan.petrovic@icm.uu.se (D.P.).
*E-mail: b.strodel@fz-juelich.de (B.S.).

ORCID ⓘ
Dušan Petrović: 0000-0002-1834-7358
Birgit Strodel: 0000-0002-8734-7765

Present Address
⊥D.P.: Department of Cell and Molecular Biology, Uppsala University, BMC Box 596, 751 24 Uppsala, Sweden.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Newhouse, T.; Baran, P. S. If C−H Bonds Could Talk: Selective C−H Bond Oxidation. Angew. Chem., Int. Ed. 2011, 50, 3362−3374.
(2) Hartwig, J. F. Catalyst-Controlled Site-Selective Bond Activation. Acc. Chem. Res. 2017, 50, 549−555.
(3) Roduner, E.; Kaim, W.; Sarkar, B.; Urlacher, V. B.; Pleiss, J.; Gläser, R.; Einicke, W.-D.; Sprenger, G. A.; Beifuß, U.; Klemm, E.; Liebner, C.; Hieronymus, H.; Hsu, S.-F.; Plietker, B.; Laschat, S. Selective Catalytic Oxidation of C−H Bonds with Molecular Oxygen. ChemCatChem 2013, 5, 82−112.
(4) Hönig, M.; Sondermann, P.; Turner, N. J.; Carreira, E. M. Enantioselective Chemo- and Biocatalysis: Partners in Retrosynthesis. Angew. Chem., Int. Ed. 2017, 56, 8942−8973.
(5) Lewis, J. C.; Coelho, P. S.; Arnold, F. H. Enzymatic Functionalization of Carbon−Hydrogen Bonds. Chem. Soc. Rev. 2011, 40, 2003−2021.
(6) Fasan, R. Tuning P450 Enzymes as Oxidation Catalysts. ACS Catal. 2012, 2, 647−666.

(7) Warman, A. J.; Roitel, O.; Neeli, R.; Girvan, H. M.; Seward, H. E.; Murray, S. A.; McLean, K. J.; Joyce, M. G.; Toogood, H.; Holt, R. A.; Leys, D.; Scrutton, N. S.; Munro, A. W. Flavocytochrome P450 BM3: An Update on Structure and Mechanism of a Biotechnologically Important Enzyme. Biochem. Soc. Trans. 2005, 33, 747−753.
(8) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 Enzymes: Their Structure, Reactivity, and Selectivity. Chem. Rev. 2010, 110, 949−1017.
(9) Shaik, S.; Chen, H.; Usharani, D.; Thiel, W. QM/MM Studies of Structure and Reactivity of Cytochrome P450 Enzymes: Methodology and Selected Applications. In Drug metabolism prediction; Kirchmair, J., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2014; pp 133−178.
(10) Lewis, J. C.; Arnold, F. H. Catalysts on Demand: Selective Oxidations by Laboratory-Evolved Cytochrome P450 BM3. Chimia 2009, 63, 309−312.
(11) Ebert, M. C.; Pelletier, J. N. Computational Tools for Enzyme Improvement: Why Everyone Can − and Should − Use Them. Curr. Opin. Chem. Biol. 2017, 37, 89−96.
(12) Whitehouse, C. J. C.; Bell, S. G.; Wong, L.-L. P450_{BM3} (CYP102A1): Connecting the Dots. Chem. Soc. Rev. 2012, 41, 1218−1260.
(13) Toscano, M. D.; Woycechowsky, K. J.; Hilvert, D. Minimalist Active-Site Redesign: Teaching Old Enzymes New Tricks. Angew. Chem., Int. Ed. 2007, 46, 3212−3236.
(14) Wang, J.; Li, G.; Reetz, M. T. Enzymatic Site-Selectivity Enabled by Structure-Guided Directed Evolution. Chem. Commun. 2017, 53, 3916−3928.
(15) Huang, W.-C.; Westlake, A. C. G.; Maréchal, J.-D.; Joyce, M. G.; Moody, P. C. E.; Roberts, G. C. K. Filling a Hole in Cytochrome P450 BM3 Improves Substrate Binding and Catalytic Efficiency. J. Mol. Biol. 2007, 373, 633−651.
(16) Hammes-Schiffer, S. Catalysts by Design: The Power of Theory. Acc. Chem. Res. 2017, 50, 561−566.
(17) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; Baker, D. Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. Nature 2013, 501, 212−216.
(18) Urlacher, V. B.; Schmid, R. D. Protein Engineering of the Cytochrome P450 Monooxygenase from Bacillus megaterium. Methods Enzymol. 2004, 388, 208−224.
(19) Weber, E.; Seifert, A.; Antonovici, M.; Geinitz, C.; Pleiss, J.; Urlacher, V. B. Screening of a Minimal Enriched P450 BM3 Mutant Library for Hydroxylation of Cyclic and Acyclic Alkanes. Chem. Commun. 2011, 47, 944−946.
(20) Whitehouse, C. J. C.; Bell, S. G.; Tufton, H. G.; Kenny, R. J. P.; Ogilvie, L. C. I.; Wong, L.-L. Evolved CYP102A1 (P450_{BM3}) Variants Oxidise a Range of Non-Natural Substrates and Offer New Selectivity Options. Chem. Commun. 2008, 966−968.
(21) Appel, D.; Lutz-Wahl, S.; Fischer, P.; Schwaneberg, U.; Schmid, R. D. A P450 BM-3 Mutant Hydroxylates Alkanes, Cycloalkanes, Arenes and Heteroarenes. J. Biotechnol. 2001, 88, 167−171.
(22) Lewis, J. C.; Mantovani, S. M.; Fu, Y.; Snow, C. D.; Komor, R. S.; Wong, C.-H.; Arnold, F. H. Combinatorial Alanine Substitution Enables Rapid Optimization of Cytochrome P450_{BM3} for Selective Hydroxylation of Large Substrates. ChemBioChem 2010, 11, 2502−2505.
(23) Kille, S.; Zilly, F. E.; Acevedo, J. P.; Reetz, M. T. Regio- and Stereoselectivity of P450-Catalysed Hydroxylation of Steroids Controlled by Laboratory Evolution. Nat. Chem. 2011, 3, 738−743.
(24) Venkataraman, H.; te Poele, E. M.; Rosłoniec, K. Z.; Vermeulen, N.; Commandeur, J. N. M.; van der Geize, R.; Dijkhuizen, L. Biosynthesis of a Steroid Metabolite by an Engineered Rhodococcus erythropolis Strain Expressing a Mutant Cytochrome P450 BM3 Enzyme. Appl. Microbiol. Biotechnol. 2015, 99, 4713−4721.
(25) Watanabe, Y.; Laschat, S.; Budde, M.; Affolter, O.; Shimada, Y.; Urlacher, V. B. Oxidation of Acyclic Monoterpenes by P450 BM-3 Monooxygenase: Influence of the Substrate E/Z-Isomerism on

Enzyme Chemo- and Regioselectivity. *Tetrahedron* **2007**, *63*, 9413−9422.

(26) Sowden, R. J.; Yasmin, S.; Rees, N. H.; Bell, S. G.; Wong, L.-L. Biotransformation of the Sesquiterpene (+)-Valencene by Cytochrome P450$_{cam}$ and P450$_{BM-3}$. *Org. Biomol. Chem.* **2005**, *3*, 57−64.

(27) Dietrich, J. A.; Yoshikuni, Y.; Fisher, K. J.; Woolard, F. X.; Ockey, D.; McPhee, D. J.; Renninger, N. S.; Chang, M. C. Y.; Baker, D.; Keasling, J. D. A Novel Semi-Biosynthetic Route for Artemisinin Production Using Engineered Substrate-Promiscuous P450 BM3. *ACS Chem. Biol.* **2009**, *4*, 261−267.

(28) Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nat. Rev. Drug Discovery* **2008**, *7*, 608−624.

(29) Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.* **2014**, *57*, 278−295.

(30) Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54*, 1961−2004.

(31) Le-Huu, P.; Heidt, T.; Claasen, B.; Laschat, S.; Urlacher, V. B. B. Chemo-, Regio-, and Stereoselective Oxidation of the Monocyclic Diterpenoid β-Cembrenediol by P450 BM3. *ACS Catal.* **2015**, *5*, 1772−1780.

(32) Le-Huu, P.; Petrović, D.; Strodel, B.; Urlacher, V. B. One-Pot, Two-Step Hydroxylation of the Macrocyclic Diterpenoid β-Cembrenediol Catalyzed by P450 BM3 Mutants. *ChemCatChem* **2016**, *8*, 3755−3761.

(33) de Visser, S. P.; Ogliaro, F.; Sharma, P. K.; Shaik, S. What Factors Affect the Regioselectivity of Oxidation by Cytochrome P450? A DFT Study of Allylic Hydroxylation and Double Bond Epoxidation in a Model Reaction. *J. Am. Chem. Soc.* **2002**, *124*, 11809−11826.

(34) Oláh, J.; Mulholland, A. J.; Harvey, J. N. Understanding the Determinants of Selectivity in Drug Metabolism Through Modeling of Dextromethorphan Oxidation by Cytochrome P450. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6050−6055.

(35) Dubey, K. D.; Wang, B.; Shaik, S. Molecular Dynamics and QM/MM Calculations Predict the Substrate-Induced Gating of Cytochrome P450 BM3 and the Regio- and Stereoselectivity of Fatty Acid Hydroxylation. *J. Am. Chem. Soc.* **2016**, *138*, 837−845.

(36) Ramanan, R.; Dubey, K. D.; Wang, B.; Mandal, D.; Shaik, S. Emergence of Function in P450-Proteins: A Combined Quantum Mechanical/Molecular Mechanical and Molecular Dynamics Study of the Reactive Species in the H$_2$O$_2$-Dependent Cytochrome P450$_{SPα}$ and Its Regio- and Enantioselective Hydroxylation of Fatty Acids. *J. Am. Chem. Soc.* **2016**, *138*, 6786−6797.

(37) Dubey, K. D.; Wang, B.; Vajpai, M.; Shaik, S. MD Simulations and QM/MM Calculations Show That Single-Site Mutations of Cytochrome P450 BM3 Alter the Active Site's Complexity and the Chemoselectivity of Oxidation Without Changing the Active Species. *Chem. Sci.* **2017**, *8*, 5335−5344.

(38) Poulos, T. L. Heme Enzyme Structure and Function. *Chem. Rev.* **2014**, *114*, 3919−3962.

(39) Haines, D. C.; Tomchick, D. R.; Machius, M.; Peterson, J. A. Pivotal Role of Water in the Mechanism of P450BM-3. *Biochemistry* **2001**, *40*, 13456−13465.

(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(41) Best, R. B.; Hummer, G. Optimised Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.

(42) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.

(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.

(44) Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E. Quantum Mechanically Derived AMBER-Compatible Heme Parameters for Various States of the Cytochrome P450 Catalytic Cycle. *J. Comput. Chem.* **2012**, *33*, 119−133.

(45) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.

(46) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(47) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.

(48) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 14101.

(49) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics With Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684.

(50) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182.

(51) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

(52) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(53) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.

(54) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578−596.

(55) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE — AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 367.

(56) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian, Inc., Wallingford CT, 2009.

(58) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.

(59) Bussi, G. Hamiltonian Replica Exchange in GROMACS: A Flexible Implementation. *Mol. Phys.* **2014**, *112*, 379−384.

(60) Kumari, R.; Kumar, R.; Lynn, A. G_mmpbsa —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **2014**, *54*, 1951−1962.

(61) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(62) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

132

(63) *The PyMOL Molecular Graphics System*, Version 1.6; Schrödinger, LLC.

(64) Edelheit, O.; Hanukoglu, A.; Hanukoglu, I. Simple and Efficient Site-Directed Mutagenesis Using Two Single-Primer Reactions in Parallel to Generate Mutants for Protein Structure-Function Studies. *BMC Biotechnol.* **2009**, *9*, 61.

(65) Omura, T.; Sato, R. The Carbon Monoxide-Binding Pigment of Liver Microsomes: II. Solubilization, Purification, and Properties. *J. Biol. Chem.* **1964**, *239*, 2379−2385.

(66) Feenstra, K. A.; Starikov, E. B.; Urlacher, V. B.; Commandeur, J. N. M.; Vermeulen, N. P. E. Combining Substrate Dynamics, Binding Statistics, and Energy Barriers to Rationalize Regioselective Hydroxylation of Octane and Lauric Acid by CYP102A1 and Mutants. *Protein Sci.* **2007**, *16*, 420−431.

(67) Sasahara, K.; Mashima, A.; Yoshida, T.; Chuman, H. Molecular Dynamics and Density Functional Studies on the Metabolic Selectivity of Antipsychotic Thioridazine by Cytochrome P450 2D6: Connection with Crystallographic and Metabolic Results. *Bioorg. Med. Chem.* **2015**, *23*, 5459−5465.

(68) Narayan, A. R. H.; Jiménez-Osés, G.; Liu, P.; Negretti, S.; Zhao, W.; Gilbert, M. M.; Ramabhadran, R. O.; Yang, Y.-F.; Furan, L. R.; Li, Z.; Podust, L. M.; Montgomery, J.; Houk, K. N.; Sherman, D. H. Enzymatic Hydroxylation of an Unactivated Methylene C−H Bond Guided by Molecular Dynamics Simulations. *Nat. Chem.* **2015**, *7*, 653−660.

(69) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(70) Alogheli, H.; Olanders, G.; Schaal, W.; Brandt, P.; Karlén, A. Docking of Macrocycles: Comparing Rigid and Flexible Docking in Glide. *J. Chem. Inf. Model.* **2017**, *57*, 190−202.

(71) Begley, M. J.; Crombie, L.; McNamara, D.; Firth, D. F.; Smith, S.; Bevan, P. C. Cembrenediols in the Curing of Tobacco. X-Ray Crystal Structures of β-Cembrenediol and α-Cembreneketol. *Phytochemistry* **1988**, *27*, 1695−1703.

(72) Elokely, K. M.; Doerksen, R. J. Docking Challenge: Protein Sampling and Molecular Docking Performance. *J. Chem. Inf. Model.* **2013**, *53*, 1934−1945.

(73) Motta, S.; Bonati, L. Modeling Binding with Large Conformational Changes: Key Points in Ensemble-Docking Approaches. *J. Chem. Inf. Model.* **2017**, *57*, 1563−1578.

(74) Buch, I; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184−10189.

(75) Ebert, M. C. C. J. C.; Guzman Espinola, J.; Lamoureux, G.; Pelletier, J. N. Substrate-Specific Screening for Mutational Hotspots Using Biased Molecular Dynamics Simulations. *ACS Catal.* **2017**, *7*, 6786−6797.

(76) Capoferri, L.; Leth, R.; ter Haar, E.; Mohanty, A. K.; Grootenhuis, P. D. J.; Vottero, E.; Commandeur, J. N. M.; Vermeulen, N. P. E.; Jørgensen, F. S.; Olsen, L.; Geerke, D. P. Insights Into Regioselective Metabolism of Mefenamic Acid by Cytochrome P450 BM3 Mutants through Crystallography, Docking, Molecular Dynamics, and Free Energy Calculations. *Proteins: Struct., Funct., Genet.* **2016**, *84*, 383−396.

(77) Yang, Y.; Wong, S. E.; Lightstone, F. C. Understanding a Substrate's Product Regioselectivity in a Family of Enzymes: A Case Study of Acetaminophen Binding in Cytochrome P450s. *PLoS One* **2014**, *9*, e87058.

(78) Petrović, D.; Frank, D.; Kamerlin, S. C. L.; Hoffmann, K.; Strodel, B. Shuffling Active Site Substate Populations Affects Catalytic Activity: The Case of Glucose Oxidase. *ACS Catal.* **2017**, *7*, 6188−6197.

(79) Vaz, A. D. N.; McGinnity, D. F.; Coon, M. J. Epoxidation of Olefins by Cytochrome P450: Evidence from Site-Specific Mutagenesis for Hydroperoxo-Iron as an Electrophilic Oxidant. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 3555−3560.

(80) Lonsdale, R.; Houghton, K. T.; Żurek, J.; Bathelt, C. M.; Foloppe, N.; de Groot, M. J.; Harvey, J. N.; Mulholland, A. J. Quantum Mechanics/Molecular Mechanics Modeling of Regioselectivity of Drug Metabolism in Cytochrome P450 2C9. *J. Am. Chem. Soc.* **2013**, *135*, 8001−8015.

(81) Bonomo, S.; Jørgensen, F. S.; Olsen, L. Dissecting the Cytochrome P450 1A2- and 3A4-Mediated Metabolism of Aflatoxin B1 in Ligand and Protein Contributions. *Chem. - Eur. J.* **2017**, *23*, 2884−2893.

(82) Shirts, M. R.; Mobley, D. L.; Brown, S. P. Free-Energy Calculations in Structure-Based Drug Design. In *Drug design: Structure- and ligand-based approaches*; Merz, K. M., Ringe, D., Reynolds, C. H., Eds.; Cambridge University Press: Cambridge, 2010; pp 61−86.

(83) Aldeghi, M.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Statistical Analysis on the Performance of Molecular Mechanics Poisson−Boltzmann Surface Area Versus Absolute Binding Free Energy Calculations: Bromodomains as a Case Study. *J. Chem. Inf. Model.* **2017**, *57*, 2203−2221.

(84) Scouras, A. D.; Daggett, V. The Dynameomics Rotamer Library: Amino Acid Side Chain Conformations and Dynamics from Comprehensive Molecular Dynamics Simulations in Water. *Protein Sci.* **2011**, *20*, 341−352.

(85) Renata, H.; Lewis, R. D.; Sweredoski, M. J.; Moradian, A.; Hess, S.; Wang, Z. J.; Arnold, F. H. Identification of Mechanism-Based Inactivation in P450-Catalyzed Cyclopropanation Facilitates Engineering of Improved Enzymes. *J. Am. Chem. Soc.* **2016**, *138*, 12527−12533.

(86) Román-Meléndez, G. D.; von Glehn, P.; Harvey, J. N.; Mulholland, A. J.; Marsh, E. N. G. Role of Active Site Residues in Promoting Cobalt−Carbon Bond Homolysis in Adenosylcobalamin-Dependent Mutases Revealed Through Experiment and Computation. *Biochemistry* **2014**, *53*, 169−177.

(87) Lodola, A.; Mor, M.; Zurek, J.; Tarzia, G.; Piomelli, D.; Harvey, J. N.; Mulholland, A. J. Conformational Effects in Enzyme Catalysis: Reaction via a High Energy Conformation in Fatty Acid Amide Hydrolase. *Biophys. J.* **2007**, *92*, L20−22.

(88) Lodola, A.; Sirirak, J.; Fey, N.; Rivara, S.; Mor, M.; Mulholland, A. J. Structural Fluctuations in Enzyme-Catalyzed Reactions: Determinants of Reactivity in Fatty Acid Amide Hydrolase from Multivariate Statistical Analysis of Quantum Mechanics/Molecular Mechanics Paths. *J. Chem. Theory Comput.* **2010**, *6*, 2948−2960.

(89) Amrein, B. A.; Steffen-Munsberg, F.; Szeler, I.; Purg, M.; Kulkarni, Y.; Kamerlin, S. C. L. *CADEE*: Computer-Aided Directed Evolution of Enzymes. *IUCrJ* **2017**, *4*, 50−64.

133