Computationally modeling interactions and dynamics to promote the understanding of protein function

DISSERTATION

ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von
Patrick Löffler
aus Erlangen

April 2017

Promotionsgesuch wurde eingereicht am: 3. April 2017

Die Arbeit wurde angeleitet von: Prof. Dr. Rainer Merkl

Unterschrift:

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Florian Hartig
Erster Prüfer:	Prof. Dr. Rainer Merkl
Zweiter Prüfer:	Prof. Dr. Wolfram Gronwald
Dritter Prüfer:	Prof. Dr. Reinhard Sterner
Ersatzprüferin:	Prof. Dr. Christine Ziegler

Abstract

Understanding the structural and functional properties of molecular interactions on atomic-level is fundamental to elucidate biochemical phenomena such as protein interactions, protein-ligand binding, and enzymatic activities. For this purpose, molecular modeling provides a powerful toolkit allowing the observation of molecular interactions *in silico* through an approximation of the physics of Nature. In this work, we show how various methods of molecular modeling can be applied to analyze and understand different biological systems.

Depending on the biological problem to be solved, molecular modeling was performed either on static structures or by considering their dynamic properties. The basis for modeling were atomically-accurate crystal structures; if necessary, model generation was combined with homology modeling and molecular dynamics to account for mutational and conformational changes. In order to derive valuable insights from those models, their structural compositions were studied: By performing small-molecule docking, structure alignments, electrostatic surface calculations, analyses of intermolecular cavities and visual inspection, biological problems were studied qualitatively. The significance of our results was then substantiated through quantifying binding energies, interaction energies, and the spatial distribution of atoms.

A profound understanding of molecular interactions can be demonstrated via protein engineering, i.e. the targeted modification or design of proteins. So far, most computational protocols have used rigid structures for design which is a simplification because a protein's structure is more accurately specified by a conformational ensemble. As part of this work, a framework for computational protein design was developed that allows existing design protocols to make use of multiple design states, e.g. structural ensembles. An *in silico* assessment simulating ligand-binding design made clear that this new multi-state approach generates more reliably native-like sequences than a single-state approach. As a proof-of-concept, *de novo* retro-aldolase activity was introduced into a scaffold protein and nine variants were characterized experimentally. All variants displayed measurable catalytic activity, testifying to a high success rate for this novel concept of multi-state enzyme design.

Kurzfassung der Arbeit

Das Verständnis der strukturellen und funktionellen Eigenschaften molekularer Wechselwirkungen auf atomarer Ebene ist von grundlegender Bedeutung für die Aufklärung biochemischer Phänomene, wie beispielsweise Protein-Wechselwirkungen, Protein-Ligand-Bindung und enzymatischer Aktivität. Zu diesem Zweck ist die molekulare Modellierung ein leistungsfähiges Werkzeug, das die Beobachtung von molekularen Wechselwirkungen *in silico* durch eine Simulation der physikalischen Naturgesetze ermöglicht. In dieser Arbeit zeigen wir wie verschiedene Methoden der molekularen Modellierung angewendet werden können, um unterschiedliche biologische Systeme zu analysieren und zu verstehen.

Abhängig von dem zu lösenden biologischen Problem wurde die molekulare Modellierung entweder auf statischen Strukturen oder unter Berücksichtigung der dynamischen Eigenschaften von Strukturen durchgeführt. Als Grundlage für die Modellierung dienten Kristallstrukturen mit atomarer Auflösung; wenn nötig wurde die Generierung von Modellen mit Homologiemodellierung und Moleküldynamik kombiniert, um Mutations- und Konformationsänderungen zu berücksichtigen. Um aus diesen Modellen wertvolle Erkenntnisse zu gewinnen, wurden ihre strukturellen Zusammensetzungen analysiert: Durch die Anwendung von molekularem Docking, strukturellen Alignments, elektrostatischen Oberflächenberechnungen, Analysen von intermolekularen Hohlräumen und visueller Inspektion wurden biologische Probleme qualitativ untersucht. Die Signifikanz unserer Ergebnisse wurde anschließend durch die Quantifizierung von Bindungsenergien, Wechselwirkungsenergien und der räumlichen Verteilung von Atomen nachgewiesen.

Ein tiefgreifendes Verständnis der molekularen Wechselwirkungen kann durch Protein-Engineering, d.h. die gezielte Modifikation oder das Design von Proteinen, demonstriert werden. Bisher haben die meisten computergestützten Protokolle starre Strukturen für das Design verwendet. Diese stellen eine Vereinfachung dar, weil die Struktur eines Proteins genauer durch ein konformationelles Ensemble spezifiziert wird. Im Rahmen dieser Arbeit wurde ein Framework für computergestütztes Proteindesign entwickelt. Dieses ermöglicht bestehenden Designprotokollen, mehrere Zustände während dem Design zu berücksichtigen, z.B. strukturelle Ensembles. Eine *in silico* Bewertung von simulierter Ligandenbindung

verdeutlichte, dass dieser neuartige Multi-State-Ansatz zuverlässiger nativ-ähnliche Sequenzen generiert, als ein Single-State-Ansatz. Als Proof-of-Concept wurde mittels computergestütztem Design *de novo* Retro-Aldolase Aktivität in ein Gerüstprotein eingeführt. Neun Varianten wurden experimentell charakterisiert. Dabei zeigten alle Varianten messbare katalytische Aktivität, was für eine hohe Erfolgsquote dieses neuen Konzepts des Multi-State-Enzymdesigns spricht.

List of publications

This thesis is based on the following publications:

A) Loedige, I., Stotz, M., Qamar, S., Kramer, K., Hennig, J., Schubert, T., **Löffler, P.**, Längst, G., Merkl, R., Urlaub, H., & Meister, G. (2014). The NHL domain of BRAT is an RNA-binding domain that directly contacts the hunchback mRNA for regulation. *Genes Dev, 28(7), 749-764.*

B) Reisinger, B., Kuzmanovic, N., **Löffler, P.**, Merkl, R., König, B., & Sterner, R. (2014). Exploiting protein symmetry to design light-controllable enzyme inhibitors. *Angew Chem Int Ed Engl, 53(2), 595-598.*

C) Silberhorn, E., Schwartz, U., **Löffler, P.**, Schmitz, S., Symelka, A., de Koning-Ward, T., Merkl, R., & Längst, G. (2016). *Plasmodium falciparum* nucleosomes exhibit reduced stability and lost sequence dependent nucleosome positioning. *PLoS Pathog, 12(12), e1006080.*

D) Plach, M. G., **Löffler, P.**, Merkl, R., & Sterner, R. (2015). Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes. *Angew Chem Int Ed Engl, 54(38), 11270-11274.*

E) **Löffler, P.**, Schmitz, S., Hupfeld, E., Sterner, R., & Merkl, R. (2017). Rosetta:MSF: A modular framework for multi-state computational protein design. *Submitted for publication*.

During my academic time, I contributed to the following publications not part of the dissertation:

F) Busch, F., Rajendran, C., Mayans, O., **Löffler, P.**, Merkl, R., & Sterner, R. (2014). TrpB2 enzymes are o-phospho-l-serine dependent tryptophan synthases. *Biochemistry, 53(38), 6078-6083.*

G) Gao, J., Truhlar, D. G., Wang, Y., Mazack, M. J., **Löffler, P.**, Provorse, M. R., & Rehak, P. (2014). Explicit polarization: A quantum mechanical framework for developing next generation force fields. *Acc Chem Res, 47(9), 2837-2845.*

H) Hauptmann, J., Kater, L., **Löffler, P.**, Merkl, R., & Meister, G. (2014). Generation of catalytic human Ago4 identifies structural elements important for RNA cleavage. *Rna, 20(10), 1532-1538.*

Personal contributions

As stated below, parts of this thesis as well as figures have been published equally worded for the listed publication. Only those parts were reused where I contributed in writing.

Publication A)
Homology modeling, electrostatic surface calculations and structure-based multiple sequence alignments were performed by myself under supervision of Prof. Dr. Rainer Merkl. The phylogenetic analysis was done by Prof. Dr. Rainer Merkl. Parts of this publication were reused in Section 2.1 and Subsection 3.1.1.

Publication B)
Molecular dynamics simulations and binding energy computations were done by myself under supervision of Prof. Dr. Rainer Merkl. The data analysis was performed together with Dr. Bernd Reisinger. Parts of this publication were reused in Section 2.4 and Subsection 3.2.1.

Publication C)
Homology modeling, molecular dynamics simulations and the computation of total interaction energies were performed by myself. The residue-wise assessment of protein-DNA interactions was performed by myself and Samuel Schmitz. The computational work was supervised by Prof. Dr. Rainer Merkl. Parts of this publication were reused in Section 2.5 and Subsection 3.2.2.

Publication D)
Homology modeling, molecular dynamics simulations, and the computation of nucleophile channels and their spatial distribution were done by myself. The deduction of putative nucleophile trajectories and the data analysis were performed together

with Maximilian Plach. Prof. Dr. Rainer Merkl supervised the computational work. Parts of this publication were reused in Section 2.6 and Subsection 3.2.3.

Publication E)
The study was conceptualized by myself and Prof. Dr. Rainer Merkl. The software implementation, validation and benchmarking were done by myself and Samuel Schmitz. Computational design of retro-aldolases was performed by myself and Samuel Schmitz. All biochemical experiments were performed by Enrico Hupfeld. The publication was written by myself and Prof. Dr. Rainer Merkl; review and editing were done by myself, Samuel Schmitz, Enrico Hupfeld, Prof. Dr. Reinhard Sterner and Prof. Dr. Rainer Merkl. Prof. Dr. Reinhard Sterner and Prof. Dr. Rainer Merkl supervised this work. Parts of an early version of the manuscript were reused in chapters 1, 2, 3, and in the appendices.

Table of contents

Chapter 1

Introduction

In a simplistic view, life can be described as a hierarchical organization of complex biological structures and systems [Solomon et al., 2002, pages 9-10]: In the lowest level are atoms that covalently bind to give molecules. Groups of molecules interact with each other and form complexes. Functional groups of molecule complexes build up organelles that, in their entirety, define a cell. The functional unit of life - the cell - is further organized in multicellular organisms: Homogeneous cells form tissues that join in structural units to compose an organ; a complete group of organs assembles a higher order organism. This hierarchy can be extended to the system integrating all living beings and their relationships, the biosphere. Intriguingly, changes in the lower levels account for all effects on the higher lying levels. Theoretically, a change in one of the lower levels, for example a specific molecular interaction has the potential to change the entire biosphere.

It is thus one of the grand challenges of science to understand in detail the underlying molecular processes of life. An alternative to the variety of experimental methods available for this task is molecular modeling: It allows the researcher to perform experiments in the computer instead of the real world by making use of the vast amount of available experimental and theoretical data. This can save time, money and allows deep insights into molecular relations when an approximate model of the system of interest is available.

Essentially important for life are proteins, large macromolecules assembled from chains of amino acids. Proteins are the workhorses of a cell that perform the tasks of an organism that are specified by the information encoded in its genes [Lodish et al., 2000, Section 1.2]. In order to perform their native biological tasks, proteins usually fold into unique 3D structures allowing them to act as complex molecular machines that transport, convert, and bind molecules. Thus, a protein's native structure is the

key to fully understand these processes. Fortunately, molecular modeling of proteins is supported by a fast-growing database of experimentally-determined structures in atomic resolution, the Protein Data Bank (PDB) [Berman et al., 2000].

1.1 Proteins

Proteins account for about 20% of a cell's weight [Lodish et al., 2000, Section 1.2] and are arguably the functionally most versatile macromolecules. Just to name a few examples, motor proteins act as molecular motors that control a cell's logistics, structural proteins build up connective tissue, antibodies neutralize pathogens as part of the immune system, and enzymes are powerful biocatalysts that accelerate chemical reactions. To avoid waste of resources, protein production is tightly regulated by a number of processes [Kafri et al., 2016] and requires two major steps: During transcription, the information stored in a gene's DNA is transferred to another molecule, RNA. During translation, this RNA interacts with a specialized protein complex named ribosome. Ribosomes assemble the protein as a chain of covalently-linked amino acids by reading the genetic information from the RNA and recruiting transporter molecules providing the encoded amino acids.

1.1.1 Biochemistry

Proteins are polymeric structures that are usually built from a series of up to 20 different canonical L-α-amino acids referred to as amino acids in the following. Amino acids are composed of a backbone (a carbon atom C_α bonded to a hydrogen, an amino group, and a carboxyl group) and a side-chain part with an organic substituent (see Fig. 1.1) which makes up their physico-chemical properties that can be classified into four groups: Hydrophobic amino acids possess a non-polar side-chain and hydrophilic amino acids a polar side-chain; other amino acids have acidic or basic side-chains. The covalently-bonded chain of amino acids is named a protein and ordered from N- to C-terminus. Here, the amino group of a protein's first amino acid is referred to as the N-terminus and the carboxyl group of a protein's last amino acid as the C-terminus. Depending on their physico-chemical properties and thus on the information encoded in their genes, proteins adopt specific structures to perform their native functions.

1.1.2 Structure and function

Protein sequences were formed during evolution for hundreds of millions of years to possess optimal function in their environment. As described before, a protein's native function is associated with its structure: Upon production, proteins fold into complex 3D structures depending on their amino acid sequence and the environment. Three main types of higher-ordered structures exist, the α-helix, the β-strand and the turn (see Fig. 1.1). Those so-called secondary structure elements form up by hydrogen-bonding, an important electrostatic attraction of backbone atoms between oxygens and amide hydrogens [Hubbard and Kamran Haider, 2010]; α-helices and β-strands are energetic favorable because all of their hydrogen bond donors and acceptors are satisfied. A protein can consist of several domains, which are structurally-conserved units that fold, evolve, and function independently of the rest of the protein [Wetlaufer, 1973].

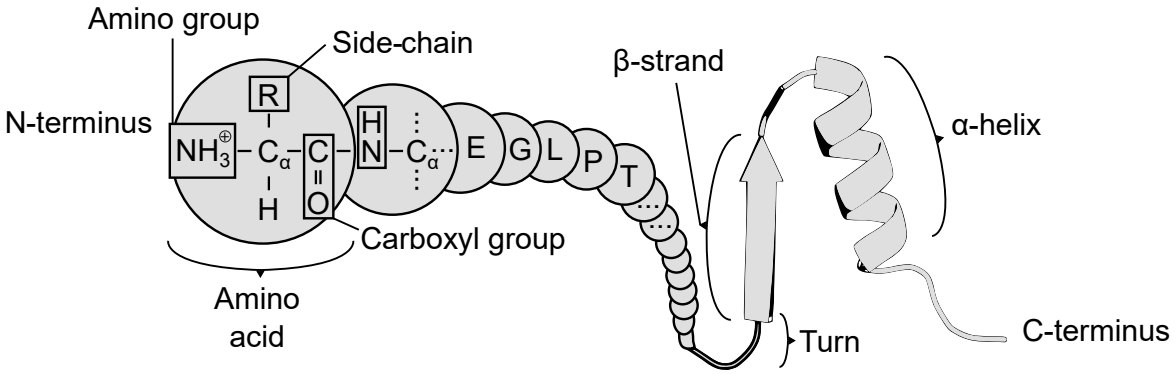


Fig. 1.1 **The structure of proteins.** The basic element of a protein is the amino acid, consisting of a central carbon (C_α) bonded to a hydrogen, an amino group, a carboxyl group, and an organic substituent (side-chain, R). The sequence of amino acids folds into higher order structure elements, such as α-helices, β-strands, and turns.

The rapid growth of structural data in the PDB has led to the development of tools for semi-automatic [Fox et al., 2014; Sillitoe et al., 2015] or manual [Murzin et al., 1995] structure annotation. By means of those tools, all experimentally-determined protein domains are identified based on structure comparison and grouped into so-called folds. A fold bundles all protein structures with the same major secondary structure arrangement and topology. Depending on the database, estimations for the total number of folds in Nature range from several hundred to several thousand [Govindarajan et al., 1999]. It is fascinating that this small number of protein folds is sufficient to manage the complexity of life. How is this possible?

The answer is that a protein's native state is not a single rigid structure but rather a dynamic entity that carries out the protein's broad spectrum of functions [Wei et al., 2016]. It is now widely accepted that proteins in solution exist as conformational ensembles with several populated substates [Ansari et al., 1985; Dill et al., 1997; Ferreiro et al., 2014] and that X-ray structures usually do not capture these substates [Fenwick et al., 2014]. The fluctuations observed for a protein ensemble are a thermodynamic phenomenon but are optimized to perform its biological function. Thus, some substates of a protein's ensemble may be more populated and the native states of proteins can move in a spectrum from being disordered and lacking a fixed 3D structure to a rather rigid 3D structure.

1.1.3 Enzymes

Dynamics are especially important for a certain class of proteins named enzymes [Agarwal, 2006]. Enzymes speed up chemical reactions by lowering their free energy of activation [Garcia-Viloca et al., 2004]; without enzymes, most biological processes in living organisms would not occur fast enough. A typical catalysis takes place in a buried pocket within an enzyme structure, referred to as active site. A model first suggested by Linus Pauling [Pauling, 1946] provided the conceptual foundation: An enzyme decreases the activation energy of a reaction by tightly binding the rate-limiting transition state (TS) structure. For a reaction to occur, a substrate molecule approaches the enzyme and binds to the active site. Next, TS-binding is facilitated to form an enzyme/TS complex. The substrate is turned into a product over one or more reactive intermediates. It is apparent that an enzyme's dynamics are essential to making this process efficient by providing the structural scaffold to bind the substrate, facilitate the formation of a TS and one or more reaction intermediates, and release the product. Two mechanisms have been proposed to play a role in performing the dynamic adaption of proteins in order to achieve a tight fit between the substrate and the enzyme - induced fit and conformational selection [D'Abramo et al., 2012; Hammes et al., 2009]. During induced fit [Koshland, 1958], the active site reshapes by interactions with the substrate until it is completely bound and in a precise position for catalysis. During conformational selection [Burgen, 1981], active and inactive enzyme conformations exist in equilibrium; when a substrate is present, it binds only to the active conformations and shifts the equilibrium towards those.

1.2 Molecular modeling of proteins

The award of the Nobel Prize in Chemistry 2013 to Martin Karplus, Michael Levitt, and Arieh Warshel "for the development of multiscale models for complex chemical systems", shows how molecular modeling has become an essential tool to complement experimental approaches by simulating the system of interest. Molecular modeling includes all methods that can be used to model, design, and understand the behavior of molecules. This section gives a general introduction about the tools for molecular modeling related to this thesis.

1.2.1 Protein structure models

To start modeling, a structure model is required. Evidently, the quality of *in silico* modeling highly relies on the accuracy of the model that describes the process to observe. For proteins and other macromolecules, the PDB provides essential information to generate an atomic-resolution structure model of a target protein. If this protein's structure is not available but its amino acid sequence, a structure model can be created by homology modeling (HM): HM exploits the fact that two proteins which are evolutionary-related (homologous) share a similar structure. Here, as little as 25% sequence identity suffice to assume the same fold but to build high quality homology models with atomic-accuracy generally more than 70% sequence identity is required. As mentioned in Subsection 1.1.2, the fold space of proteins is relatively limited and well-covered by experimental data [Sadowski and Taylor, 2009], so there is a high chance of finding template proteins homologous to the target protein and successfully building a homology model.

HM begins with identifying one or more homologous template structures by database search methods [Altschul et al., 1997; Söding et al., 2005]. Next, a selection of identified template sequences is aligned with the target sequence to produce the target-template alignment. In order to transform the gathered information into a 3D structure, a number of different model generation methods are available. HM usually starts with fragment assembly; here, fragments from structure homologs are treated as rigid-bodies and are combined to build the conserved structure core where the model accuracy is rather certain [Greer, 1981]. For sequence-variable and loop regions, where the structure accuracy based on the template protein is rather uncertain, segment matching [Levitt, 1992] can be applied to search the PDB for structures representing segments of the target sequence. Generally, several different methods are combined to give the best results. For example, HM also makes use

of methods for protein threading, which are usually applied to compute models when no homology can be detected [Yang et al., 2015]. Here, the geometry and arrangement of backbone and side-chain atoms is refined by profile comparison [Bowie et al., 1991], statistical potentials [Shen and Sali, 2006] or other methods [Wu and Zhang, 2008]. During HM, a multitude of models are generated. In order to choose a good model, model assessment algorithms such as VERIFY3D [Eisenberg et al., 1997] are utilized to determine a score for each model and select a final model. Some programs for HM, such as YASARA [Krieger et al., 2009] further combine the best parts of several models into a hybrid model.

When the determination of a model is not possible with standard methods, programs to fold protein structures based only on their amino acid sequence can be used [Rohl et al., 2004; Xu and Zhang, 2012]. However, these methods only work well for small proteins and provide rather poor accuracy. The field of structure prediction is continuously evolving and recently, new methods have emerged that combine protein folding with methods to predict residue-residue contacts [Marks et al., 2012; Ovchinnikov et al., 2015] which has led to the discovery of several new folds [Ovchinnikov et al., 2017].

1.2.2 Structure comparison

Comparing proteins by means of their structures has become a standard tool for molecular modeling. Its importance for revealing evolutionary relationships between proteins, predicting protein structures and protein functions [Hasegawa and Holm, 2009] has generated many different algorithms and procedures [Kufareva and Abagyan, 2012]. Due to the continuously growing number of available structural information, several methods for automated structure alignment methods exist, such as CE [Shindyalov and Bourne, 1998], DALI [Holm and Sander, 1993] or TM-ALIGN [Zhang and Skolnick, 2005] to search for structure homologs of a target protein in the whole PDB. However, there is no single proper metric for measuring the distance between protein structures. A simple and commonly used criterion to compare the structural similarity of two sets of atoms is the Root-Mean-Square Deviation (RMSD), defined as:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}\,, \tag{1.1}$$

where N is the number of amino acids and δ is the distance between the two atoms in the i-th pair. A relatively simple but important problem is to find the optimal

superposition of two sets of atoms, i.e. to minimize the RMSD. This problem can be solved by the Kabsch-algorithm [Kabsch, 1976], which computes the optimal rotation and translation of one set of atoms onto the other. However, the RMSD is length-dependent and not suitable for more complex tasks like comparing differently sized proteins. Here, other metrics such as the length-normalized TM-score [Zhang and Skolnick, 2004a] are to be preferred.

1.2.3 Protein-ligand docking

Protein-ligand docking (PLD) is a widely used technique to predict bound conformations of a small molecule (ligand) to a protein receptor. It is thus applied to study protein-ligand interactions which is especially interesting for drug discovery and pharmaceutical research where hundreds of thousands of candidates are virtually screened for activity [Kitchen et al., 2004]. Typically, PLD is performed in two steps, named sampling and scoring. During sampling, a number of protein-ligand conformations are generated at the given binding site. Scoring then predicts the binding energy of those conformations by physical and empirical energy functions. Typically, the top scoring hits (bound conformations) are identified by ranking all conformations and extracting those with the lowest binding energy.

Depending on the available computational resources and the problem size, parameters for sampling must be set adequately to achieve the best docking accuracy [Sousa et al., 2006]. Four main factors determine accuracy: Receptor flexibility, ligand flexibility, ligand sampling, and most importantly the energy function used for scoring. Receptor flexibility can be considered by **i**) allowing a certain amount of side-chain or backbone flexibility **ii**) the explicit consideration of multiple protein conformations as an ensemble **iii**) the subsequent structural refinement of docked conformations. On the other hand, ligand flexibility is coupled to ligand sampling to generate putative ligand conformations at the protein's binding site using different methods [Huang and Zou, 2010]: **i**) Shape matching is a quick method to find matches of the surface of the ligand to complement the molecular surface of the protein. **ii**) Systematic search algorithms explore a ligand's degrees of freedom by exhaustive searching, fragmentation or by considering an ensemble of ligand conformations. **iii**) Stochastic algorithms sample ligand conformations by making random steps in the conformational space and pursuing energetic favorable solutions.

Given a number of docked conformations of a ligand at a protein binding site, the assessment of binding strength is done via scoring functions. Typically, scoring functions for PLD are an empirically derived set of energy terms, including those for

electrostatic energies, hydrogen bonding, entropy, hydrophobicity, and more; this set of terms is weighted to reproduce experimentally measured binding affinities of a training set. Depending on the method, scoring functions can be arbitrary complex and computationally demanding, for example by considering charge polarization [Cho et al., 2005] in the docking process.

1.2.4 Molecular dynamics simulation of proteins

Molecular dynamics (MD) simulation allows to study particle motions of biochemical systems in full atomic detail as a function of time. The importance of directly observing protein dynamics have made MD simulations an indispensable tool for molecular modeling. This has contributed to replacing the early model of rigid protein structures with a dynamic model, in which intrinsic motions and conformational changes play an essential functional role [Karplus and McCammon, 2002]. Modeling a system of biomolecules requires to set two critical parameters: Simulation time and accuracy, which are mutually dependent. For example, atomically-accurate simulations with explicit water put severe constraints on the simulation timescale, limiting the simulation time to microseconds - even on a supercomputer. Although proceedings in the methodology and fast growth in computing power have led to protein simulations reporting at the millisecond timescales [Lane et al., 2013], much longer simulations are desirable. While local motions, such as atomic fluctuations, side-chain, and loop motions are observable in the femto- to nanosecond scale, larger motions like protein folding and unfolding require milliseconds up to minutes [Kubelka et al., 2004].

Providing adequate timescales and accuracy, Molecular Mechanics (MM) force fields are the method of choice for protein simulations. MM force fields rely on classical mechanics for modeling: In all-atom simulations as used in this thesis, each atom is a particle with a certain mass, which is assigned a van der Waals (vdW) radius as well as a constant net charge. Covalent atoms are treated as springs with equilibrium distances. Here, the parameters for different atom types and bonds are usually derived from experiment or quantum-mechanical calculations. Given an initial position as well as an initial velocity of all particles in the system, they are allowed to interact for a fixed amount of time, typically in the fs scale; consequential forces between particles and their potential energy are calculated by Newton's laws of motion [Newton, 1999] using the MM force field. The potential energy E of the

system is computed as the sum of two terms

$$E = E_{bonded} + E_{non\_bonded} \text{ ,} \tag{1.2}$$

where E_{bonded} comprises three types of bonded interactions:

$$E_{bonded} = E_{bond\_stretch} + E_{angle\_bend} + E_{dihedrals} \tag{1.3}$$

$E_{bond\_stretch}$ and E_{angle_bend} approximate the bond stretching and angle bending of covalently-bonded atoms by harmonic oscillators as a function of the bond length and valence angle, respectively. $E_{dihedrals}$ represents the dihedral terms, which have multiple minima and are typically approximated by a sum of cosine functions with several multiplicities and amplitudes [Levitt et al., 1995].

Non-bonded interactions between all atoms are divided into two terms:

$$E_{non\_bonded} = E_{vdW} + E_{electrostatics} \tag{1.4}$$

E_{vdW} approximates the vdW interactions that are typically modeled as a Lennard-Jones 6-12 potential; $E_{electrostatics}$ describes the electrostatic interactions between fixed point charges of particles, usually handled by the Coulomb potential (see Fig. 1.2). Due to their large number, a cutoff radius for both interaction types is defined to reduce computational costs.

After computing the forces between all atoms, atom positions and velocities can be updated. Because the movement of particles in MD simulations is an n-body problem which cannot be solved analytically, this step is performed by algorithms for numerical integration, such as the most commonly used velocity-verlet integration [Verlet, 1967]: Given all atom positions and randomly assigned initial velocities, a cycle starts by computing the forces acting on each atom using the force field. Based on the forces and masses of all atoms, their acceleration can be computed. Next, a certain timestep is chosen to determine the change in velocities. After updating the velocities, new atom positions can be determined, which in turn allows the recalculation of forces, beginning a new cycle. In order to set up a typical MD simulation many more parameters need to be defined, such as the simulation type, temperature, thermostat, and boundary conditions. In this thesis, NpT-ensembles were captured at room temperature by keeping the number of particles N, the pressure p, and the temperature T constant in order to emulate the given experimental conditions.

The popularity and the broad applicability of MD has led to the implementation of several software packages to model biomolecules. Just to name a few, CHARMM [Brooks et al., 1983] originates from Martin Karplus's group and is a historically grown and popular software package; GROMACS [Hess et al., 2008] is a popular software package known for its efficient implementation; YASARA [Krieger et al., 2002] is a proprietary software package providing easy setup and a graphical user interface. However, there is no single software package that performs well for all simulation tasks and thus all software packages have individual strengths and weaknesses.

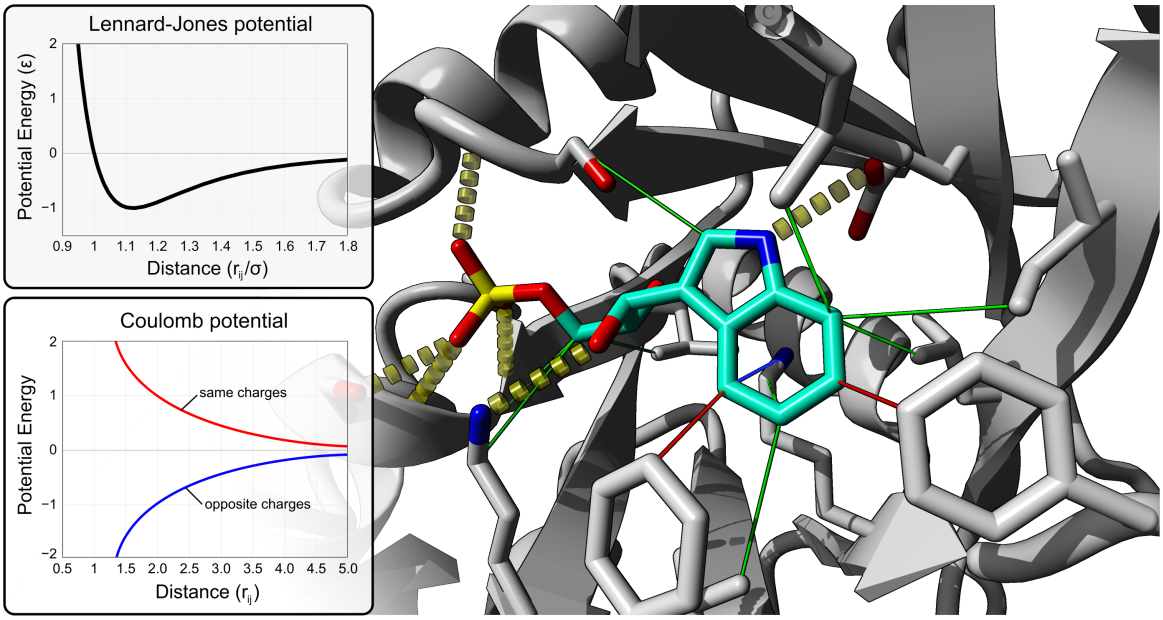


Fig. 1.2 **MD potentials and interactions**
(**Panel Lennard-Jones potential**) *vdW* interactions are described by the Lennard-Jones potential: $Pot_{LJ} = 4\ \epsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}]$. Here, the potential energy function of two atoms i, j at distance r_{ij}/σ_{ij} is defined by the depth of the energy well ϵ_{ij}, where r_{ij} is the distance between the two atoms and σ_{ij} is the distance at which the inter-particle potential is 0. (**Panel Coulomb potential**) Electrostatic interactions are described by the Coulomb potential $Pot_{Coulomb} = \frac{q_1 q_2}{4\ \pi\epsilon_0} \cdot \frac{1}{r_{ij}}$, where q_1 and q_2 are two charges, r_{ij} their distance, and ϵ_0 the electric constant. (**Right side**) Interactions at the active site of a protein computed by a force field and visualized with YASARA. Green lines: hydrophobic contacts. Red lines: Pi-Pi-stacking. Blue line: Cation-Pi stacking. Yellow dashed lines: Hydrogen bonds.

1.2.5 Computational protein design

Since the 1990s computational protein design (CPD) evolved into a powerful tool for protein engineering. For example, CPD has been successfully utilized to increase thermostability of proteins [Dantas et al., 2003; Malakauskas and Mayo, 1998; Shah et al., 2007] and to design new or altered binding specificities for metals [Marvin and Hellinga, 2001], DNA [Ashworth et al., 2006] or other ligands [Allert et al., 2004; Shifman and Mayo, 2003]. During the last decade, CPD was applied to even more challenging tasks like the design of novel protein-protein interfaces [Fleishman et al., 2011; Procko et al., 2014], *de novo* enzymes [Röthlisberger et al., 2008] or artificial folds not found in Nature [Hill et al., 2000; Kuhlman et al., 2003].

CPD aims at predicting the sequence folding to a target structure which is known as the inverse folding problem [Bowie et al., 1991]. The approach is built on Anfinsen's dogma [Epstein et al., 1963], proposing that proteins fold into the lowest energy conformation available to their sequences. In order to emulate the physical conditions of proteins, physicochemical potentials, energy functions, and statistical terms were developed specifically for protein design [Boas and Harbury, 2007]. Given such a mathematical description of a protein model, CPD becomes an optimization problem. However, its search space is unimaginably vast: For a small protein of length 50, 20^{50} possible design sequences exist. Thus, scoring all design solutions is not practical and CPD therefore relies on a multitude of algorithms to speed up the optimization.

For this task, heuristic algorithms do not guarantee to find the sequence providing the global minimum energy conformation for a structure but offer adequate solutions and fast computing times. Examples for heuristics in CPD include genetic algorithms (GA) [Wernisch et al., 2000], Monte Carlo (MC) simulated annealing [Kuhlman and Baker, 2000] and greedy algorithms [Nivón et al., 2014]. To solve the inverse folding problem for a given protein structure, amino acid rotational isomers (rotamers) are placed at each amino acid position in order to find an optimal set of rotamers that minimizes the global energy. Each rotamer is one specific side-chain conformation of a given amino acid picked from a large dataset of energetic favorable conformations, which further increases the dimensionality of the problem. In the interest of reducing the search space, the protein model is simplified: Instead of using a continuous model for rotamers, CPD relies on discretized rotamer libraries like the Dunbrack library [Shapovalov and Dunbrack, 2011]. Typically, rotamers are optimized using the MC method that combines statistical sampling with simulated annealing: In brief, the optimization starts from a random distribution d_1 of rotamers

with different amino acid identities and computes the energy of this state $E(d_1)$. Next, a number of rotamer changes are performed leading to d_2. The energetic change $\Delta E = E(d_1) - E(d_2)$ is calculated and the probability of accepting this state p_{accept} is computed by assessing the metropolis criterion:

$$p_{accept} = min(1, e^{-\frac{\Delta E}{kT}}) \tag{1.5}$$

where T is the temperature and k is Boltzmann's constant. When starting the computation at a high temperature, most MC steps are accepted which allows sampling of the energy landscape. Annealing forces the algorithm to accept rather energetic favorable substitutions and reject most energetic unfavorable ones, leading to convergence. Typically, the algorithm is started several times using different random seeds to explore the energy landscape without running into local minima.

Having become a standard tool in computational biology, software for computational protein design is offered by several different academic groups. The most flexible and widely used software suite for CPD is Rosetta [Leaver-Fay et al., 2011b]. It was implemented in a community approach and is maintained by a large user-base. Rosetta bundles many protocols for computational design and macromolecular modeling. However, several other software tools for CPD are actively developed: OSPREY [Gainza et al., 2013] includes a number of powerful algorithms meant for finding optimal solutions for CPD and aiming at further reducing the search space, such as dead-end elimination [Desmet et al., 1992] or A* [Roberts et al., 2015]; FOLDX is a protein design algorithm that makes use of an empirical force field. It is specialized in determining energetic effects of mutations and is a convenient tool to compute interaction energies. Apart from these few examples, many more programs, webservers, software suites, and scripts for CPD exist.

1.3 Aim of this thesis

Steadily growing computational resources and available experimental data have made experiments in the computer more appealing than ever. Particularly important are simulations, which offer an intuitive way of understanding Nature - by studying the molecular context. In structural bioinformatics, special attention is paid to proteins that perform a vast array of functions in living organisms.

The aim of this thesis was to explore the benefit of molecular modeling for the understanding of protein function. Because protein function and dynamics

are coupled, a particular focus was put on modeling the dynamic properties of proteins. Depending on the research question, its complexity, and the feasibility of the *in silico* approaches, the projects in this work were addressed by studying static models (Section 3.1) and their dynamic properties via MD simulations (Section 3.2). Logically, protein dynamics is important for another widely-used tool of protein modeling - computational protein design: In Section 3.3, a novel algorithm for CPD named MSF is illustrated. This algorithm allows the design of proteins considering their inherent conformational dynamics via a discrete conformational ensemble. We assess the performance of this algorithm on several benchmark datasets and show that the consideration of dynamics improves the design performance. In order to demonstrate the actual applicability of MSF, we developed a protocol for enzyme design based on conformational ensembles. Using this protocol, MSF was applied to computationally design *de novo* retro-aldolases based on a conformational ensemble of a suitable scaffold protein.

Chapter 2

Materials and Methods

This chapter describes the programs, algorithms, and methods used in this thesis. For improved readability, programs, protocols, and their options are designated as PROGRAM[:PROTOCOL[:OPTION]]. The first two sections describe methods, in which molecular modeling was applied in a static way without taking protein dynamics into account. Next, general parameters used for MD simulations are defined, followed by methods that involved the consideration of protein dynamics via MD. The last two sections outline the description of a novel method for multi-state computational protein design, its benchmarking, and finally the procedure applied to design *de novo* enzymes.

2.1 Homology modeling and electrostatic surface calculations of six-bladed NHL domains

The experimental work performed in publication **A)** was supported with a comprehensive computational analysis from our side: Homology modeling and electrostatic surface calculations were used to structurally characterize six-bladed NHL domains and the phylogeny of the family was derived from a structure-based multiple sequence alignment.

At the time of analysis, only three crystal structures had been determined experimentally: the peptidyl-α-hydroxyglycine α-amidating lyase from *Rattus norvegicus* (rnPAL, PDB ID 3fw0), a serine/threonine-protein kinase from *Mycobacterium tuberculosis* (mtPknD, PDB ID 1rwi), and the brain tumor NHL domain from *Drosophila melanogaster* (dmBrat, PDB ID 1q7f). To gain more insight into the structural differences of this protein family, comparative modeling was used. For this analysis, the

sequences of 22 six-bladed NHL domains were downloaded, including all human, fly, and worm proteins. Next, 15 homology models were built for each representative group by means of I-TASSER [Zhang, 2008] using default settings. In each case, the model with the best confidence score (*C-score*) was chosen, resulting in a mean *C-score* of 0.1 ± 1.1, which indicates a satisfactory structure quality for this analysis. Subsequently, surface electrostatic calculations were performed on all 3D structures with the Particle Mesh Ewald approach [Krieger et al., 2006] as implemented in YASARA (version 13.4.21) [Krieger et al., 2004] and by employing the YASARA:YAMBER3 force field in physiological pH. For the graphical representation shown in Results (Fig. 3.1), the solvent accessible surface was color-coded representing the local electrostatic potential. The darkest blue color represents a positive and the darkest red color a negative potential of 300 $^{kJ}/_{mol}$, respectively.

Next, the phylogeny was inspected: All 25 considered NHL domains possess a six-bladed β-propeller fold; however sequence similarity of the selected proteins is low, which makes a purely sequence-based derived phylogeny difficult. The average pairwise sequence identity value was $21 \pm 18\%$ as determined by EMBOSS:NEEDLE [Rice et al., 2000]. This is why a structure-based algorithm was chosen for constructing a multiple sequence alignment (MSA) needed for phylogenetic analysis. By means of CHIMERA (version 1.8.1) [Pettersen et al., 2004], all 18 structures were superimposed on the model of hsTrim71 from *Homo sapiens*, which yielded the lowest sum of RMSD values in an all against all superposition. Subsequently, CHIMERA:MATCH→ALIGN [Meng et al., 2006] was utilized to generate a MSA based on this superposition. Finally, the sequences of the seven remaining proteins with unknown 3D structure were included by applying MAFFT:ADD [Katoh and Frith, 2012; Katoh and Standley, 2013] with default parameters. This MSA was the basis to determine a neighbor-joining tree [Saitou and Nei, 1987] by means of SPLITSTREE4 [Huson and Bryant, 2008]; this analysis was performed by Prof. Dr. Rainer Merkl.

2.2 Docking of putative light-inducible inhibitors to β-galactosidase

The following methods are part of a collaboration with the groups of Prof. Dr. Burkhard König and Prof. Dr. Hans-Heiner Gorris to create photo-switchable inhibitors for β-galactosidase (β-gal):

In total, the 3D structures of 121 ligands were analyzed and docked, containing two substrates, a known inhibitor for β-gal, and 118 photoswitchable inhibitor designs. Docking was performed with flexible binding site residues and the consideration of multiple β-gal structures: Therefore, 55 crystal structures of β-gal from *Escherichia coli* were downloaded from the PDB, superpositioned and prepared for docking with YASARA (version 16.4.6). Next, the global search space for docking was limited to a cell of size 20 $Å^3$ centered at the substrate binding site of β-gal. The 13 residues closest to the substrate binding site, excluding residues with rigid side-chains (Ala/Gly), were defined as flexible. Docking was performed with YASARA:AUTODOCKVINA (version 1.1.2) [Trott and Olson, 2010] using all 121 ligands. Each ligand was docked in 24 individual runs on each of the 55 crystal structures. Some ligands were explicitly designed as cis isomers. To prevent cis-trans isomerism by the conformational sampling of the ligand performed during docking, the corresponding torsion angles allowing isomerization were frozen. Finally, results for each ligand were collected based on the best hit sorted by binding energy in any of the 55 crystal structures.

2.3 General parameters for molecular dynamics simulations

Here, an overview is given about the protocol used for MD simulations in the next sections. All MD simulations were performed with YASARA employing either the YAMBER3 or YASARA2 [Krieger et al., 2009] force field. Simulations were run at 298 *K* under periodic boundary conditions and with explicit water, using a multiple time step of 1 *fs* for intramolecular and 2 *fs* for intermolecular forces. If multiple individual simulations were performed, independent calculations were seeded by slightly changing the temperature ($\pm$0.01 *K*) for the respective next runs which reassigns the initial atom velocities. Lennard-Jones forces and long-range electrostatic interactions were treated with a 7.86 Å cutoff, the latter were calculated using the Particle Mesh Ewald method [Essmann et al., 1995]. Temperature was adjusted using a Berendsen thermostat based on the time-averaged temperature and simulations were carried out at constant pressure. For non-protein and non-nucleic acid molecules, the parameterization was performed using the AM1BCC protocol [Jakalian et al., 2002] that assigns atomic charges by applying additive bond charge corrections to semi-empirical AM1 [Dewar et al., 1985] atomic charge calculations.

MD simulations require the definition of a simulation cell that should be adequately sized to prevent self-interaction through periodic boundaries. Simulation cells were thus defined as 5 Å larger than the considered structure along each axis. Cells were filled with water to a density of 0.997 g/ml and counterions were added to a final concentration of 0.9% NaCl. Next, the protonation states of all molecules were assigned accordingly [Krieger et al., 2006]. Before capturing production runs, an equilibration run was performed unless otherwise noted to remove conformational stress. Prior to energetic analysis of structural snapshots, an energy minimization was done as follows: After removing conformational stress by a steepest descent minimization, the procedure continued by simulated annealing (time step 2 fs, atom velocities scaled down by 0.9 every 10th step) until convergence was reached, i.e., the energy improved by less than 0.05 $^{kJ}/_{mol}$ per atom during 200 steps.

2.4 Refining small-molecule ligand-protein interactions via molecular dynamics

In publication **B)**, several photoswitchable inhibitors targeting the phosphoribosyl isomerase A from *Mycobacterium tuberculosis* (mtPriA) were designed. However, the structural basis of the different binding affinities of the receptors remained unclear and thus, the experimental part was supported by a structural characterization of the strongest switching inhibitor *compound 6* via MD simulations.

Based on the crystal structure of mtPriA (PDB ID 3zs4), the original ligand PRFAR was removed and manually replaced by *compound 6* in the open and closed form. A good fit was provided through superposition of the phosphate binding pockets. Both the open and closed model were prepared for MD simulations with YASARA (version 13.4.21) as described in Section 2.3. In order to remove conformational stress, the equilibration was conducted in two phases: After a 100 ps equilibration with fixed protein coordinates, the liganded structure was equilibrated for 1 ns. The two equilibrated models of the open and closed form were subsequently used for the six following (three for each conformer) production MD simulations. Trajectories were sampled at intervals of 100 ps for a total of 10 ns for each model. Binding energies were obtained for each energy-minimized snapshot using YASARA'S integrated binding energy function that computes the energetic difference of the ligand at bound state and at infinite distance from its binding site. Representative structure models for each simulation were extracted and were based on the snapshots with

the best binding energy after energy minimization. The ligand binding energies and standard deviations given in Table 3.3 were calculated by using the full production trajectory.

2.5 Comparing differences in binding affinities of nucleosomal cores

Publication **C)** is a comparative study of human and *Plasmodium falciparum* nucleosomes. In this work, experimental findings suggested that the latter have a strongly reduced ability to recognize sequence-dependent nucleosome positioning signals. For a structural analysis, the DNA-protein interaction of both human and plasmodial nucleosomes was modeled, simulated, and compared by means of MD simulations.

All modeling tasks were performed with YASARA (version 16.4.6). First, the standard protocol of YASARA was used to create homology models of all histones and the complete nucleosome consisting of an octamer that had 146 *bp* of DNA wrapped around it. For each model of a nucleosomal complex, the input of YASARA was a multiple FASTA file with two DNA and eight protein sequences. The DNA sequences were two copies of the palindromic DNA fragment (146 *bp* long) from human X-chromosome alpha satellite DNA as found in the dataset with PDB ID 3afa. The protein sequences originated from the histones of *H. sapiens* or *P. falciparum*, respectively. The GenBank accession numbers for the human histones were AAA63191.1 (H2A), AAN59961.1 (H2B), NP_066403.2 (H3), NP_003539.1 (H4) and for the plasmodial histones AAA29612.1 (H2A), XP_001347738.1 (H2B), AAO23910.1 (H3), AAP45785.1 (H4). Due to their flexibility, the N- and C-termini of histones could not be resolved in X-ray structures and modeling the long flexible termini would result in structural uncertainty; therefore, their 3D-orientation was unclear. This is why the histone sequences were trimmed according to the resolved 3D structure reported in PDB ID 3afa. In order to determine the homology models for plasmodium, three rounds of PSI-BLAST [Altschul et al., 1997] restricted to PDB IDs were conducted and YASARA selected PDB IDs 3afa, 5av6, 3tu4, 3x1t and 2nqb as templates. These datasets represent the structures of nucleosomal core particles from different eukaryotic species. For the human template (3afa), all of the 740 target residues could be aligned to template residues; for these the sequence identity was 84%. After building models for each template, YASARA combined the best scoring fragments of all models to deduce a hybrid homology model. The resulting hybrid

model scored best and the internal quality assessment of YASARA determined an overall Z-score of 0.056, which indicates a model quality 0.056 standard deviations better than an average high-resolution X-ray structure. Note that model quality is most reliable for globular proteins and can be misleading for other protein types. Dihedrals and packing in 1D and 3D were rated as optimal by YASARA.

Next, the models were prepared for MD simulations with YASARA (see Section 2.3) and three simulations were run for each nucleosome complex. Both the human and plasmodial datasets thus consist of three MD trajectories each comprising 200 snapshots that represent varying poses of a 50 *ns* interval. To assess the protein-DNA interaction, FOLDX (version 4) [Guerois et al., 2002] was used to calculate a score for the interaction energy between individual histone cores and the DNA: The 200 snapshots with a time period of 250 *ps* were stored in PDB format and contained the complex plus all water molecules within a maximal distance of 3 Å to a protein or DNA molecule. Then, the side-chain orientation of all snapshots was optimized with FOLDX:REPAIRPDB to prepare the structures for the given force field. Subsequently, mean interaction energies between histone and DNA as well as their standard deviations were deduced with the FOLDX:ANALYZECOMPLEX command. A further analysis was carried out with the help of Samuel Schmitz to characterize the underlying protein-DNA interactions in detail: The 200 snapshots were used to deduce mean values of scores assessing the following interactions, which were determined in a residue-specific manner: π-π stacking, cation-π stacking, contacts, hydrophobic interactions, and hydrogen bond networks. For the first four interactions, scores were taken from the YASARA output. To score hydrogen bond networks, distances were analyzed between residues, DNA, and water molecules in a snapshot-specific manner. Thus, a graph was computed that consisted of nodes that represent putatively interacting atoms on the surface of the considered molecules and of edges modeling hydrogen bonds. An edge was inserted, if the distance between a donor and an acceptor atom was not larger than 2.5 Å. Based on this network, a score was computed for each path interconnecting a pair of atoms from DNA and a protein according to:

$$S_{path}(atom_i^k, atom_j^l) = 1/(edges(atom_i^k, atom_j^l) \cdot \#path\_ident\_len) \tag{2.1}$$

Here, $edges(atom_i^k, atom_j^l)$ is the number of edges interconnecting an atom k of residue i with atom l of nucleotide j and the normalization factor $\#path\_ident\_len$ is the number of paths with the same length observed in the full dataset. Thus, the score for a hydrogen-mediated interaction decreases with the number of involved

water molecules and results in a higher score for a more direct one. The maximal number of co-operative water molecules was limited to one and for each residue res_j, all S_{path}-values were summed up. For each of the averaged scores with noticeable amplitude, the log_2-value was plotted for corresponding residues of the histones from *H. sapiens* and *P. falciparum* together with the sequences by means of a CIRCOS graph [Krzywinski et al., 2009].

2.6 Analyzing molecular tunnels of chorismate-utilizing enzymes by simulation

In publication **D)**, biochemical studies showed that the primary metabolic enzyme anthranilate synthase (AS, subunits TrpE:TrpG) can be converted into the secondary metabolic enzyme isochorismate synthase (ICS) by introducing not more than two mutations. In order to deduce a relationship between structure and function, the structural characteristics of the wild-type and mutant proteins were analyzed *in silico*. First, structure models of TrpE:TrpG variants were generated and studied in MD simulations. Second, the MD trajectories were exploited to examine the putative substrate channels of all variants.

The crystal structure of TrpE in complex with TrpG from *Salmonella typhimurium* (stTrpE:stTrpG, PDB ID 1i1q) represents an unliganded, open, inactive T-state form [Morollo and Eck, 2001]. Therefore an stTrpE:stTrpG homology model was generated based on the crystal structure of the TrpE:TrpG complex from *Serratia marcescens* (PDB ID 1i7q), which resembles a ligand-bound form with a closed active site [Spraggon et al., 2001]. Modeling was performed with YASARA (version 14.7.17). The high similarity of target and template sequences argues in favor of a good 3D model: Sequence identity values determined by EMBOSS:NEEDLE were 71.3% (TrpE) and 79.8% (TrpG). Moreover, YASARA'S Z-scores were -0.462 (TrpE) and -0.352 (TrpG), indicating high model quality. Chorismate (CH) was placed in the active site of the stTrpE model, substituting the benzoate and pyruvate ligands present in 1I7Q; the RMSD for all matching atoms was 0.713 Å. Structures of mutant stTrpE variants were generated by *in silico* mutating residue positions 263 (Lys), 364 (Ala, Ile, Leu, Met), and 365 (Leu, Val, Ser, Ala) of the stTrpE homology model (see Fig. 3.8). Mutated residues were rotamer-optimized employing YASARA:SCWALL [Canutescu et al., 2003]. To remove conformational stress, all homology models were

equilibrated by means of a 100 *ps* simulation, resulting in equilibrated homology models (EHMs).

To identify the most likely paths of ammonia in wild-type stTrpE:stTrpG and of the nucleophile in complexes with mutated stTrpE variants, the respective EHMs were subjected to MD simulations and nucleophile channels were computed as follows: EHMs of wild-type stTrpE and of each stTrpE variant were simulated in three production MD runs (see Section 2.3). Trajectories were sampled at intervals of 10 *ps* for a total of 2 *ns*, resulting in 600 snapshots for each stTrpE variant. These structures were further energy-minimized prior to the computation of channels. Moreover, for visual inspection, average 3D models were generated for each MD trajectory as follows: An EHM served as a reference structure and average positions for all atoms were deduced after superimposing the protein structures from all snapshots. Nucleophile channels were computed utilizing MOLE (version 2.13.9.6) [Sehnal et al., 2013]. Default values were used except a probe radius of 2.14 Å, which is the size of ammonia in the given force field. The starting point was the all-atom centroid of the ligand glutamine and Cys87 in stTrpG that approximates the location where nascent ammonia is generated. The endpoint was the all-atom centroid of the ligand CH and Ala327 in stTrpE that approximates the location of the CH-C2 atom where the initial nucleophilic attack in the AS reaction occurs. For each of the 600 resulting channels per variant, the channels centerlines served to specify a putative nucleophile trajectory (PNT). As the MD simulations induced small translational and rotational movements of the stTrpE:stTrpG complexes, a direct comparison of related PNTs was not possible. To compensate for this effect, all PNTs were superimposed on the respective EHM [Kabsch, 1976] and the resulting PNT-bundles were analyzed further.

Visual inspection of PNT-bundles by means of PYMOL [DeLano, 2008] in the region near the CH ligand indicated a preference for two major paths (Fig. 3.8); one proceeding alongside Val265 and the other one alongside residues 365 and 425 (see Fig. 2.1). Due to their prevalence in Leu365 and Ala365 variants, these paths were termed L-path and A-path, respectively. The spatial distribution of PNTs observed in a variant was determined by counting the number of PNTs that proceed along the L-path or the A-path, as follows: First, for each PNT j, the segment with a distance of 3 to 7 Å from CH-C2 was identified. Due to the complexity of the local curvature of individual PNTs, these segments were represented by a different number of 3D coordinates. Thus, the coordinates $coord_k^{i,j}$ were binned in 16 shells i (thickness 1/4 Å, centered on CH-C2) and the vector $pv_k^{i,j}$ was computed that starts at $coord_k^{i,j}$ and ends

in CH-C2. Then, a plane P_j (with normal vector nv_j) was defined by CH-C2 and the C_α atoms of the PNT-lining residues Met364 and Thr425. Each $pv_k^{i,j}$ was multiplied with the normal vector nv_j; the sign of the scalar product $(+,-)$ $s_k^{i,j}$ indicates the position of $coord_k^{i,j}$ relative to P_j. For each shell i and PNT j, the mean was computed as:

$$\overline{s^{i,j}} = 1/l \sum_{k=1}^{l} s_k^{i,j} \,, \tag{2.2}$$

where l is the number of coordinates $coord_k^{i,j}$. Next, $\overline{s^{i,j}}$ was normalized to $[0,1]$; a value of 0 indicates that all PNT coordinates are located on the L-path side of P_j and a mean of 1 shows that all coordinates are on the A-path side of P_j. The shell-wise computed mean was computed as:

$$\overline{s^i} = 1/m \sum_{j=1}^{m} \overline{s^{i,j}} \,, \tag{2.3}$$

where m is the number of PNTs in this shell. $\overline{s^i}$ was then used to determine

$$\overline{s} = 1/n \sum_{i=1}^{n} \overline{s^i} \,, \tag{2.4}$$

where n is the number of shells, which is the percentage of all PNTs along the A-path in the 3 to 7 Å shell around CH-C2. For a graphical overview, compare Fig. 2.1. The fraction $\overline{s}$ was considered indicative for the prevalent localization of PNTs and proposes the overall putative nucleophile path of each variant (see Fig. 3.7 C)

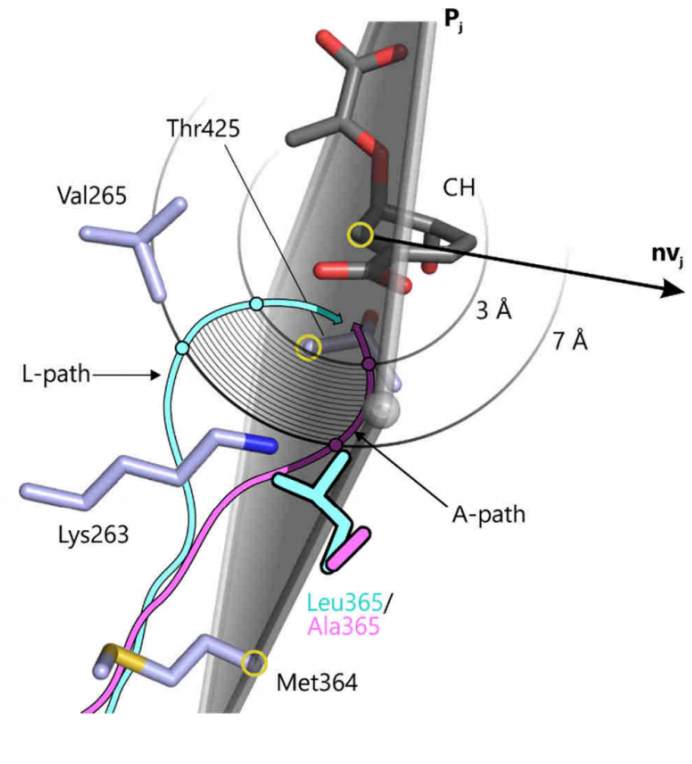


Fig. 2.1 **Principles used to quantify the spatial distribution of PNTs.** Graphical representation of L- and A-path (cyan and magenta, respectively) as well as of the plane P_j and its normal vector nv_j used for quantifying the spatial distribution of PNTs. P_j was specified by CH-C2 and the Cα atoms of residues 364 and 425 (yellow circles). For binning of PNT coordinates, sixteen 1/4 Å shells (black circles) spanning a segment of 3 to 7 Å from CH-C2 were defined. This figure was reused and modified based on [Plach et al., 2015] with permission from *John Wiley and Sons*.

2.7 ROSETTA:MSF: a modular framework for multi-state protein design

In publication **E)**, a novel framework for multi-state design (MSD) of proteins named ROSETTA:MSF was implemented. In contrary to single-state design (SSD), MSD allows to simultaneously consider multiple states for design, e.g. an ensemble. The compilation of benchmark datasets and assessment of design performance are outlined in this section and the procedure of applying this framework to design *de novo* retro-aldolases is delineated in the next section.

ROSETTA:MSF was implemented as part of Samuel Schmitz's master project and provides multi-state functionality for enzyme (MSF:GA:ENZDES) and protein-protein interface design (MSF:GA:ANCHORED). It has been integrated as an additional protocol into Rosetta and is purely written in C++98; development, testing, and benchmarking was done for Rosetta weekly release 2015.19.57819. During benchmarking, the performance of the default single-state protocols of Rosetta for enzyme design (ENZDES) and protein-protein interface design (ANCHORED) was compared to the performance using the multi-state protocols of ROSETTA:MSF.

2.7.1 Compilation of benchmark datasets

In total, four benchmark datasets were compiled: *hIFABP* and *BR_EnzBench* were used to compare SSD with MSD performance for protein-ligand binding; *MD_EnzBench* serves to test an alternative sampling method for the use in protein-ligand binding design. *BR_IfaceBench* was generated to analyze the SSD and MSD performance for protein-protein interface design.

Compilation of benchmark datasets for ligand-binding design

Dataset *hIFABP* with PDB ID 2mji contains ten conformers of human intestinal fatty acid binding protein (hIFABP) and the bound ligand ketorolac; this ensemble has been deduced by means of solution NMR [Patil et al., 2014]. The set was downloaded from PDB and the ligand was parameterized using ROSETTA:MOLFILE-TO-PARAMS [Davis and Baker, 2009]. Subsequently, each of the ten conformations was energy-minimized via ROSETTA:FASTRELAX with backbone constraints. To obtain consistent design and repack shells, the shells determined by ROSETTA:ENZDES:AUTODETECT for each conformation were merged.

Two subsets of the *scientific sequence recovery benchmark* of Rosetta [Nivón et al., 2014] were generated that contain 20 specifically prepared conformations of 16 proteins $prot(k)$ with bound ligand. In order to exclude an erroneous conformational sampling, missing residues were reconstructed by using YASARA:LOOP_MODELING [Canutescu and Dunbrack, 2003] and the respective native sequences. For convenience, all proteins were renumbered so that the first residue starts with a residue number of one. Additionally, all ligands were removed prior to the conformational sampling of the resulting apoproteins. The dataset *BR_EnzBench* was created by using the BACKRUBENSEMBLE method of the BACKRUB server [Lauck et al., 2010] to compute a conformational ensemble of 20 structures for each apoprotein. The second benchmark dataset *MD_EnzBench* was deduced from MD simulations of length 10 *ns* generated with YASARA (version 14.7.17) and the YAMBER3 force field which had been parameterized to produce crystal structure-like protein coordinates [Krieger et al., 2004]. For each of the 16 apoproteins, 1000 conformations were sampled at an interval of 10 *ps*. After sampling, the native ligands were reintroduced in all conformations of both subsets by means of PYMOL:SUPERPOSE and the respective apoproteins. For the corresponding holoproteins of *BR_EnzBench* and *MD_EnzBench*, the same design and repack shells were utilized. These were determined protein-wise for each of the *BR_EnzBench* conformations by means of ROSETTA:ENZDES:AUTODETECT and merged. In all conformations, design shell residues were replaced with alanine and prior to design, all conformations were energy-minimized by means of ROSETTA:FASTRELAX with backbone constraints. Parameters for MD simulations are specified in Section 2.3; the protocol for energy-minimization (see Section B.1) and the composition of the design and repack shell for the *hIFABP* and *EnzBench* datasets (see Subsection B.2) are listed in Appendix B.

Compilation of a benchmark dataset for anchored protein interface design

The benchmark dataset *BR_IfaceBench* consisting of 16 protein complexes was generated based on the original dataset used to benchmark ANCHORED [Lewis and Kuhlman, 2011]. To avoid erroneous conformational sampling, residues missing in the crystal structures were reconstructed using YASARA:LOOP_MODELING with the corresponding native sequences. For convenience, all complexes were renumbered so that the first residue starts with a residue number of one. Next, a conformational ensemble of 20 conformations was generated by applying the BACKRUBENSEMBLE method of the backrub server on each protein complex. The design and repack shell residues were adopted from the original benchmark dataset and are listed in

Subsection B.2. Prior to design, all design shell residues were mutated to alanine and all conformations were energy-minimized with backbone constraints (see Section B.1).

2.7.2 Assessing design performance

As defined in this section, the design performance on the four benchmark datasets was assessed by the total score (*ts*) given in Rosetta Energy Units (*REU*), the native sequence recovery (*nsr*), and the native sequence similarity recovery (*nssr*).

The native sequence similarity recovery

The *nssr* is defined analogously to the *nsr* but considers amino acid similarities instead of identities. Thus, for a given pair of residues aa_1, aa_2 the *nssr* value was deduced from the scores of the BLOSUM62-matrix [Henikoff and Henikoff, 1992] according to:

$$nssr(aa_1, aa_2) = \begin{cases} 1, & \text{if } BLOSUM62(aa_1, aa_2) > 0 \\ 0, & \text{else} \end{cases} \tag{2.5}$$

For a given pair of sequences seq_1, seq_2 of length n, the *nssr* value was determined as a mean value deduced for residue pairs $seq_1[i]$, $seq_2[i]$:

$$nssr(seq_1, seq_2) = \frac{1}{n} \sum_{i=1}^{n} nssr(seq_1[i], seq_2[i]) \tag{2.6}$$

For a given set of designed sequences $ds = \{seq_1, ..., seq_n\}$ and a native sequence seq_{nat}, the value $nssr(ds, seq_{nat})$ was computed according to:

$$nssr(ds, seq_{nat}) = \frac{1}{n} \sum_{i=1}^{n} nssr(seq_i, seq_{nat}) \tag{2.7}$$

Computing scores for single-state performance assessment

Single-state ENZDES and ANCHORED were applied to each initial conformation l of a protein k part of their corresponding *Benchmark* dataset, where K is the number of proteins and L the number of conformations per protein. Using the default MC optimization, sequences $seq_{k,l}(i)$ were generated by means of i randomly seeded $runs_{k,l}(i)$. In order to control the convergence of the design process and for

performance comparison, the $seq^*_{k,l}(i)$ with the best total score (ts) were chosen from $seq_{k,l}(1,...,i)$ for each k,l, and each i. Finally, the mean of the $K \cdot L$ ts values was determined as a measure of design quality $ts^{Benchmark}_{SSD}(i)$ reached in i SSD runs:

$$ts^{Benchmark}_{SSD}(i) = \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} ts(seq^*_{k,l}(i)) \tag{2.8}$$

Here, i is the number of runs, K is the number of proteins in dataset *Benchmark*, and L is the number of conformations. Design performance in terms of recovery was determined by comparing the protein-specific native sequence seq^k_{nat} with the designed sequences and computing the value of $score \in \{nsr, nssr\}$:

$$score^{Benchmark}_{SSD}(i) = \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} score(seq^*_{k,l}(i), seq^k_{nat}) \tag{2.9}$$

Computing scores for multi-state performance assessment

To realize the MSD approach, Rosetta's genetic algorithm was adapted for use in MSF: Briefly, the GA imitates the process of natural selection by maintaining a population of design sequences that are evolved for a number of generations, while the selection pressure of a fitness function eliminates less optimal solutions. The first generation consists of a given seed sequence and a user-defined number of mutants each with a randomly introduced single point mutation. During each generation cycle j, half of the population was replaced with sequences $seq(j)$ generated by means of single point mutations and recombination. The replaced sequences were those with worst fitness values $fitness(seq(j))$ which were computed according to:

$$fitness(seq(j)) = \frac{1}{N} \sum_{n=1}^{N} ts_n(seq(j)) \tag{2.10}$$

Here, N is the number of states (e.g. conformations of a given protein or protein-protein complex) and $ts_n(seq(j))$ is the application-specific Rosetta total score for a sequence $seq(j)$ given a state n. For MSD using MSF:GA:ENZDES or MSF:GA:ANCHORED, ensembles are required and thus, all conformations were divided into M ensembles which were taken as states for multi-state design. To consider the same number of sequences for performance comparison with SSD, T top scoring sequences were selected for each ensemble from the population of MSD sequences so that $L = M \cdot T$. Design performance in terms of energy was then

computed as

$$ts_{MSD}^{Benchmark}(j) = \frac{1}{K \cdot M \cdot T} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{t=1}^{T} fitness(seq_{k,m,t}(j)) \,, \tag{2.11}$$

Design quality measured as sequence recovery was computed as the value of $score \in \{nsr, nssr\}$:

$$score_{MSD}^{Benchmark}(j) = \frac{1}{K \cdot M \cdot T} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{t=1}^{T} score(seq_{k,m,t}(j), seq_{nat}^{k}), \tag{2.12}$$

Protein-ligand binding benchmark *hIFABP*

For SSD, ENZDES was applied to each of the ten initial conformations $conf(l)(1 \leq l \leq 10)$. Using the parameter set *ps_enzdes* (see Subsection B.3), sequences $seq_l(i)$ were generated by means of 1000 randomly seeded $runs_l(i)(1 \leq i \leq 1000)$. Finally, $ts_{SSD}^{hIFABP}(i)$ (Equation (2.8)) and $nsr/nssr_{SSD}^{hIFABP}(i)$ (Equation (2.9)) were determined as measures of design quality reached in i SSD runs.

For MSD, all $N = 10$ conformations were considered as states of one ensemble ($M = 1$) and MSF:GA:ENZDES was executed for 800 generations j (i.e. design cycles) on a population consisting of 210 sequences with parameter set *ps_msf_enzdes* (see Subsection B.3). The initial population was seeded with the native sequence of 2mji. The sequences representing a generation j were ranked with respect to *ts* values and the $T = 10$ top scoring sequences $seq_m^t(j)(1 \leq t \leq 10)$ were stored in order to allow for the subsequent performance comparison. Finally, $ts_{MSD}^{hIFABP}(j)$ (Equation (2.11)) and $nsr/nssr_{MSD}^{hIFABP}(j)$ (Equation (2.12)) were determined as measures of design quality achieved after j generations.

Protein-ligand binding benchmark BR_EnzBench

For SSD, ENZDES was applied to each of the $L = 20$ initial conformations $conf(l)(1 \leq l \leq 20)$ of each $prot(k)(1 \leq k \leq 16)$ from *BR_EnzBench*. Using default MC optimization and parameter set *ps_enzdes* (see Subsection B.3), sequences $seq_{k,l}(i)$ were generated by means of 1000 randomly seeded $runs_{k,l}(i)(1 \leq i \leq 1000)$. Then, mean performance reached in i SSD runs was determined as $ts_{SSD}^{BR\_EB}(i)$ (Equation (2.8)) and $nsr/nssr_{SSD}^{BR\_EB}(i)$ (Equation (2.9)). For a protein-specific comparison, final scores were determined for each $prot(k)$ from runs $i = 1000$.

To assess the performance of MSD, each of the $L = 20$ conformations of a $prot(k)$ was assigned to an ensemble $ens_m^k (1 \leq m \leq 4)$ consisting of $N = 5$ conformations each. These five conformations were considered as states and MSF:GA:ENZDES was executed for 600 generations on a population consisting of 210 sequences with parameter set *ps_msf_enzdes* (see Subsection B.3). The initial population was seeded with an all-alanine sequence. The sequences representing a generation j were ranked with respect to *ts* values and the $T = 5$ top scoring sequences $seq_{k,m,t}(j)(1 \leq t \leq 5)$ were stored in order to allow for the subsequent performance comparison. Mean performance values achieved after j MSD generations were determined as $ts_{MSD}^{BR\_EB}(j)$ based on Equation (2.11) and $nsr/nssr_{MSD}^{BR\_EB}(j)$ using Equation (2.12). To score MSD performance reached for one $prot(k)$, the score values were taken from the final generation $j = 600$.

Protein-ligand binding benchmark *MD_EnzBench*

ENZDES was applied to each of the $L = 1000$ conformations $conf(l)(1 \leq l \leq 1000)$ of each $prot(k)(1 \leq k \leq 16)$ from *MD_EnzBench*. Using default MC optimization and parameter set *ps_enzdes* (see Subsection B.3), sequences $seq_{k,l}(1)$ were generated by means of one randomly seeded $run_{k,l}$ for each protein k and conformation l. The mean performance $nssr_{SSD}^{MD\_EB}(1)$ was computed using Equation (2.9).

Protein-protein interface design benchmark *BR_IfaceBench*

For SSD, ANCHORED was applied to each of the $L = 20$ initial conformations $conf(l)(1 \leq l \leq 20)$ of each protein $complex(k)(1 \leq k \leq 16)$ from *BR_IfaceBench*. Sequence optimization was performed with the default MC protocol utilizing parameter set *ps_anchored* (see Subsection B.3) and generating sequences $seq_{k,l}(i)$ by means of eight randomly seeded $runs_{k,l}(i)(1 \leq i \leq 8)$. The mean performance reached in i SSD runs was determined as $ts_{SSD}^{BR\_IB}(i)$ (Equation (2.8)) and $nsr/nssr_{SSD}^{BR\_IB}(i)$ (Equation (2.9)). For a complex-specific comparison, final scores were determined for each $complex(k)$ after $i = 8$ runs.

To assess the performance of MSD, each of the $L = 20$ conformations of a $complex(k)$ was assigned to an ensemble $ens_m^k (1 \leq m \leq 4)$ consisting of $N = 5$ conformations considered as states. MSF:GA:ANCHORED was executed for 1000 generations on a population consisting of 50 sequences using parameter set *ps_msf_anchored_perturb* (see Subsection B.3). For this set of parameters, a coarse optimization was performed; the initial population was seeded with an all-alanine

sequence and the final generation served to seed the population for a refinement run. Afterwards, refinement was done by executing MSF:GA:ANCHORED for 500 generations on a population of 50 sequences using parameter set *ps_msf_anchored_refine* (see Subsection B.3). The sequences representing a generation j were ranked with respect to *ts* values and the five top scoring sequences $seq_{k,m,t}(j)(1 \leq t \leq 5)$ were stored in order to allow for the subsequent performance comparison. Mean performance values achieved after j MSD generations were determined as $ts_{MSD}^{BR\_EB}(j)$ using Equation (2.11) and $nsr/nssr_{MSD}^{BR\_EB}(j)$ using Equation (2.12). To score the final MSD performance reached for one $complex(k)$, score values were taken from the last generation $j = 500$ of the refinement run.

2.7.3 Characterization of ligand-binding design

The performance differences of the MSD implementation MSF:GA:ENZDES and the default SSD implementation of ENZDES for benchmark dataset *BR_EnzBench* were characterized in more detail. For this analysis, the results were regrouped and the amino acid composition of the designed sequences was studied.

Grouping ensembles on MSD performance

The 20 conformations of a given $prot(k)$ from *BR_EnzBench* belong to one of four ensembles $ens_1^k - ens_4^k$. The performance values $nssr_{MSD}(ens_m^k)$ were determined for each ensemble m and each $prot(k)$ according to:

$$nssr_{MSD}(ens_m^k) = \frac{1}{5}\sum_{t=1}^{5} nssr(seq_{k,m}^t(600), seq_{nat}^k) \tag{2.13}$$

Here, seq_{nat}^k is the native sequence of $prot(k)$. The values $nssr_{MSD}(ens_m^k)$ were used for a ranking $ens_{rank=u}^k(1 \leq u \leq 4)$ of the four ensembles such that $ens_{rank=1}^k$ is the one with the lowest $nssr_{MSD}(ens_m^k)$ value and $ens_{rank=4}^k$ possesses the largest one. Having ranked the ensembles of all $prot(k)$, sets of ensembles were created such that the set $ES_1 = \bigcup_{k=1}^{16} ens_{rank=1}^k$ contained the worst performing ensembles and $ES_4 = \bigcup_{k=1}^{16} ens_{rank=4}^k$ those that performed best; the intermediates with $rank = 2$ and $rank = 3$ performed accordingly. For these four sets ES_i, boxplots of the corresponding $nssr_{SSD}$ and $nssr_{MSD}$ values were determined.

Choosing sequences for the analysis of the sequence differences

In order to assess the amino acid composition of the ENZDES outcome, the 42 $seq_{k,l}(1,...,1000)$ with optimal *ts* values were identified for each of the 20 conformations of all $prot(k) \in BR\_EnzBench$. For these 16×840 sequences seq^k_{SSD}, the values $nssr(seq^k_{SSD}[i], seq^k_{nat}[i])$ were determined (Equation (2.5)) by comparing all designed and respective native (*nat*) residues of the design shell. The distribution $nssr_{SSD}(aa_j)$ represents for all amino acids aa_j their recovered similarity at all design shell positions. To assess the amino acid composition for the MSF:GA:ENZDES outcome, the $16 \times 4 \times 210$ sequences seq^k_{MSD} of the final populations (i.e. all $seq_{k,m}(600)$) generated for the four ensemble groups of each $prot(k) \in MD\_EnzBench$ were used to determine the values $nssr(seq^k_{MSD}[i], seq^k_{nat}[i])$. The distribution $nssr_{MSD}(aa_j)$ represents for all amino acids aa_j their recovered similarity at all design shell positions.

2.8 Multi-state design of retroaldolases

As a proof of concept, ROSETTA:MSF was applied to *de novo* design retro-aldolases in a multi-state manner. The following section describes the sampling of the scaffold protein to obtain structural ensembles which were subsequently used for design. After evaluating all design solutions, nine variants were chosen for biochemical characterization. Most variants showed weak solubility levels but could be solubilized with the help of stabilizing mutations predicted *in silico*. Described methods are supplemented by a protocol with additional details in Appendix C.

2.8.1 Scaffold sampling and multi-state design

The scaffold protein indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus* (ssIGPS, PDB ID 1a53) was downloaded from PDB and the ligand IGPS was removed. To generate a structural ensemble, three MD simulations were performed with the apoprotein for 10 *ns* by means of YASARA (version 14.7.17) and the YAMBER3 force field. Using DURANDAL:SMART-MODE:SEMI-AUTO [0.03..0.20], the snapshots of each trajectory were clustered individually and four conformations were chosen from the largest cluster. These 12 conformations and the crystal structure of 1a53 were used for matching the transition state and grafting the theozyme of the retroaldol reaction [Bjelic et al., 2014] by means of ROSETTA:MATCH [Zanghellini et al., 2006]. Each of the resulting matched transition states (*mTS*) consisted of a catalytic triad Lys_i-$[Asp, Glu]_j$-$[Ser, Thr]_k$ at three residue positions i, j, k that occurs in one of

the 13 conformations. Ensembles ens_{mTS} of mTS used as input for MSF:GA:ENZDES were generated as follows: First, mTS were discarded that were classified as weak TS binder or TS destabilizer. For example, matches with catalytic residues near the protein surface were eliminated. Second, mTS were grouped according to the composition and localization of their catalytic triad. Those ensembles were selected that contained the largest number of the 13 conformations. Third, ens_{mTS} were assessed with respect to the structural similarity of the superposed theozymes. In total, 23 ensembles ens_{mTS} containing four up to 13 conformations were chosen. For each ens_{mTS}, the design and repack shells were defined by merging the outcome of ROSETTA:ENZDES:AUTODETECT for all corresponding conformations and MSF:GA:ENZDES was executed on a population of 210 sequences that were seeded with the native sequence of ssIGPS. At convergence, the design process was stopped, which was the case after 97 to 710 generations. Parameters of MD simulations, parameters of ROSETTA:MATCH, the specification of TS, and the content of all ens_{mTS} are listed in Appendix C.

2.8.2 Evaluation of design solutions

After MSD of retro-aldolases, the designs were filtered by *ts* values and active-site geometry. The best 100 designs were selected for 10 *ns* MD simulations in water and for one conformation of each design ensemble, 100 snapshots were generated. Two simulations were performed; the first one was based on the enzyme/TS complex. As a control, the second MD was based on the enzyme/substrate complex and the substrate methodol was created by deleting the lysine-substrate bond of the TS. For each trajectory, catalytic distances, angles and torsion angles were plotted as boxplots and analyzed (see Appendix C).

2.8.3 *In silico* stabilization

Variant RA_MSD2 was chosen for solubilization experiments and all six conformations $conf(l)$ of the corresponding ensemble were submitted to the PROSS server [Goldenzweig et al., 2016], which was used with default settings allowing mutations at all positions. For each input $conf(l)$, PROSS provided seven mutated sequences $mut\_seq_l(i)(1 \leq i \leq 7)$ containing an increasing number of putatively stabilizing mutations. For each i (degree of stabilization), a MSA that contained all sequences $mut\_seq_l(i)$ computed for all $conf(l)$ was generated and WEBLOGO [Crooks et al., 2004] was utilized to determine a sequence logo. Finally, consensus residues de-

duced from the sequence logos were accepted as mutations at sites that did not interfere with the catalytic center. All sequence logos are listed in Appendix C, Fig. C.3.

2.8.4 Cloning, gene expression, protein purification, and activity assay

Biochemical experiments were performed by Philipp Bittner, Katharina Satzinger and Franziska Funke under supervision of Enrico Hupfeld. Synthesis of S-methodol used for activity assays was done by Paul Gehrtz and Ivana Fleischer. Here, only a brief description of the biochemical procedure is given, while the full details will be published: genes for the designed retro-aldolases were optimized for codon usage and ordered as synthetic gene strings from Life Technologies. Next, genes were cloned into pET28a and pMalC5T plasmids; both vectors fuse an N-terminal his_6-tag to the target protein, while pMalC5T also adds maltose-binding protein (MBP). Cells were grown in Luria broth and at a cell density of $OD_{600} = 0.5$, protein production was induced by adding isopropyl-β-thiogalactopyranoside. After growing over night at 20 °C, cells were harvested by centrifugation, cell pellets were resuspended in Tris/HCl buffer (pH 7.5), and lysed by sonication. Soluble protein was purified by nickel chelate affinity chromatography and eluted with Tris/HCl (pH 7.5) using an imidazole gradient. Fractions containing sufficiently pure protein were pooled and excess imidazole was removed by dialysis against a Tris/HCl buffer containing NaCl. Protein concentrations were determined by absorbance spectroscopy using extinction coefficients determined by the EXPASY:PROTPARAM web-tool. Retro-aldolase activity of the designs was measured at 25 °C in Tris/HCl (pH 7.5) containing NaCl and dimethyl sulfoxide by following the formation of the fluorescent product 6-methoxy-2-naphthaldehyde from non-fluorescent S-methodol (70% ee). Fluorescence was measured in a plate reader in black 96 well micro plates and the product concentrations were determined with the help of a calibration curve. To determine conversion rates, each measurement was repeated four times and for the determination of the k_{cat}/K_M value, points were measured as triplicates. ssIGPS and the solubility tag MBP served as negative controls and showed no detectable activity.

Chapter 3

Results and Discussion

The aim of this thesis was to solve different problems in structural biology by means of molecular modeling. Depending on the complexity of the problem and the available resources, molecular modeling was either performed on static structures (Section 3.1) or applied in combination with MD simulations to consider conformational dynamics (Section 3.2). The most computationally demanding application studied here is computational protein design based on structural information from conformational ensembles: In Section 3.3, the implementation, benchmarking, and characterization of a novel algorithm is outlined; this method allows a protein designer to mimic conformational flexibility, which increased the design performance in *in silico* benchmark datasets. As a proof of concept, the algorithm was applied to computationally design *de novo* retro-aldolases (see Section 3.4) based on a conformational ensemble of the scaffold protein.

3.1 Molecular modeling based on static structures

The Nobel Prize discovery of the double-helix structure of DNA by James Watson and Francis Crick in 1953 was based on an image taken by Raymond Gosling and Rosalind Franklin showing the X-ray diffraction pattern of DNA. Providing key information for the anti-parallel helical nature, the image helped Watson and Crick to calculate the 3D model of DNA [Watson and Crick, 1953]. Although their resulting model was completely static and did not reflect the molecular dynamics of DNA interactions with water, ions, proteins, and RNA in living cells, it provided important information to explain many *in vivo* functions of DNA. Today, molecular modeling using static structures is still extremely important because it is easy to perform,

consumes considerably less computational resources and relies on a continuously growing structural database. In this section, the contributions to two projects based on analyses of static structures will be illustrated.

3.1.1 Modeling six-bladed NHL domains predicts putative RNA binding

The family of TRIpartite Motif (TRIM) proteins is found in all multi-cellular organisms. It is known to be functional diverse, fulfilling tasks such as transcriptional regulation and controlling a broad range of biological processes associated with innate immunity. A prominent member is the brain tumor (Brat) protein of *D. melanogaster* which plays a central role in repressing hunchback mRNA, a transcription factor controlling thorax development, during early embryonic stages [Sonoda and Wharton, 2001]. Brat consists of a N-terminal TRIM domain and a C-terminal NHL domain. Complete lack of Brat function was shown to cause tumor growth in the larval brain induced by the failure of neuronal progenitor cells to exit proliferation [Arama et al., 2000]. The NHL domain of Brat is of critical importance: Flies with deletions or single point mutations in this domain are characterized by diverse mutant phenotypes, for example abdominal segmentation defects. The crystal structure of Brat-NHL was solved and revealed a six-bladed β-propeller [Edwards et al., 2003]. Mutagenesis studies showed that the top surface of Brat-NHL interacts with Pumilio (Pum), which is an important mediation factor for the recruitment of Brat to hunchback mRNA. Interestingly, mutations preventing the interaction of Brat with Pum are also the cause of the brain tumor phenotype. Previous studies suggested that the RNA-dependent interaction between Brat-NHL and the RNA-binding domain of Pum is the nature of a protein-protein interaction between RNA-bound Pum and Brat [Edwards et al., 2003; Sonoda and Wharton, 2001].

Publication **A)** provided the basis for a different model: Through various biochemical methods, it was shown that Brat and Pum contact RNA independent of each other and translational repression by Brat can occur independent of Pum. To consolidate the hypothesis, the data was supported by a computational analysis of the TRIM-NHL protein family from our side. Typically, RNA-binding proteins bind RNA via positive electrostatic surfaces that complement the negatively charged phosphate groups of RNA. To give an overview about the RNA binding potential of NHL domains, we analyzed the electrostatic surface potential of Brat-homologs: First, homology models of known six-bladed NHL domains were built as described

in Section 2.1. These models were the basis for electrostatic surface calculations to estimate the degree of putative RNA binding. Next, the phylogeny was studied: Due to the low sequence identity of all homologs, a structure-based algorithm was utilized to determine a neighbor-joining tree. Finally, the data was visualized as an unrooted tree, where the leafs of representative proteins show the corresponding color-coded surface charges on the structure (see Fig. 3.1). The philogenetic tree splits into several groups, while two bigger groups are well separated: Putative RNA binders and enzymes split with bootstrap values of 85 and 100, respectively. In this figure, putative RNA binders are marked as those associated with RNA interactions in the literature [Loedige et al., 2014]. These proteins possess at least patches of positively charged top surfaces. On the other hand and as expected, the group of NHLs with known enzymatic activity display an overall negatively charged surface suggesting no RNA binding. The analysis also highlighted putative novel RNA-binding proteins, such as hsTRIM32 and hsTRIM56. Taken together, the data indicates that RNA binding is a common feature of TRIM-NHL proteins and their functional diversity suggests that they have a distinct set of RNA-binding partners.

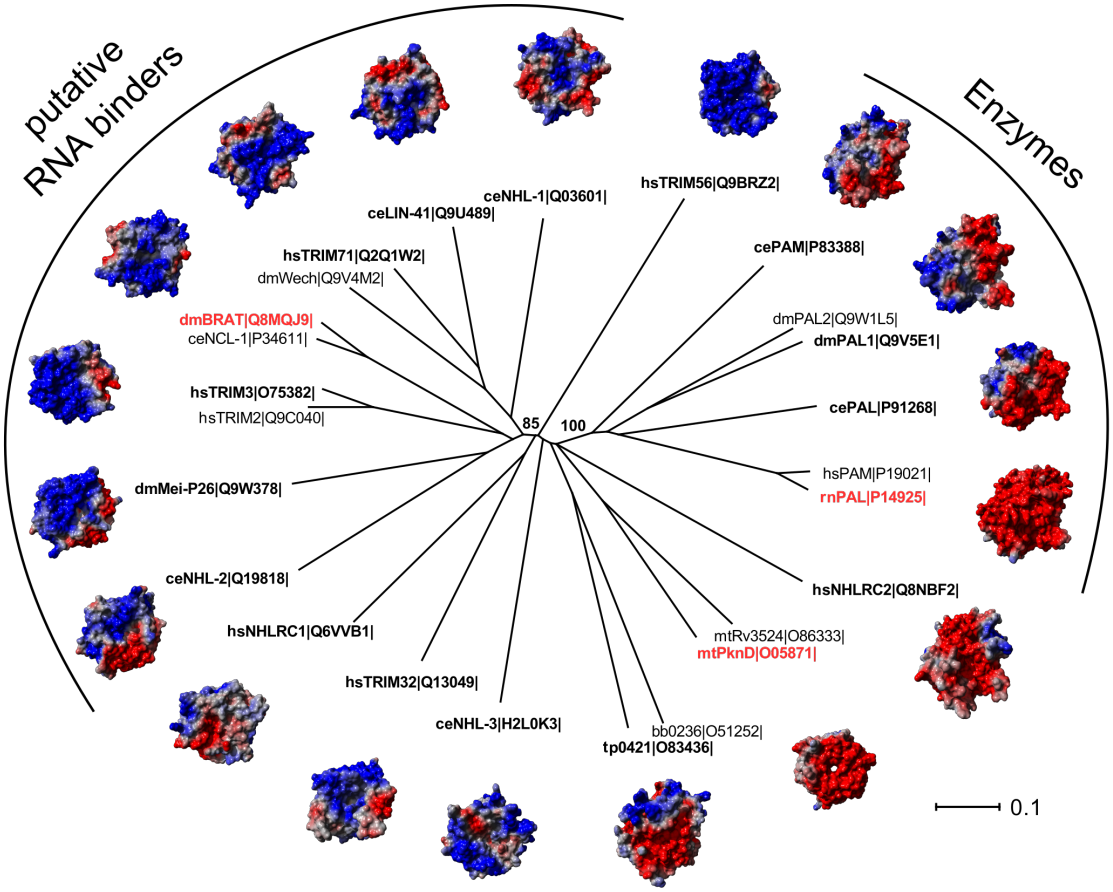


Fig. 3.1 **Phylogeny and surface electrostatics of six-bladed NHL domains**. The neighbor-joining tree was derived from a structure-based multiple sequence alignment as described in Section 2.1. Characteristic proteins are represented by their known 3D structure (name in red) or by homology models (name in bold and black). The subfamilies of putative RNA binders and of enzymes are separated from the rest by a bootstrap value of 85% or 100%, respectively. For each structure, Particle Mesh Ewald long-range electrostatic calculations were performed in YASARA and used to color-code the solvent-accessible surface: A negative charge is indicated by a red surface and a positive charge is indicated by a blue surface. Abbreviations for protein names and species are given next to the Uniprot ID. (bb) *Borrelia burgdorferi*; (ce) *Caenorhabditis elegans*; (dm) *Drosophila melanogaster*; (hs) *Homo sapiens*; (mt) *Mycobacterium tuberculosis*; (rn) *Rattus norvegicus*; (tp) *Treponema pallidum*. The length of the horizontal bar corresponds to 0.1 substitutions per site. This figure was reused and modified based on [Loedige et al., 2014] with permission from *Cold Spring Harbor Laboratory Press*.

3.1.2 Docking of putative light-inducible inhibitors to β-galactosidase

Isolating single enzyme molecules in femtoliter-sized reaction chambers is an exciting technique for basic research of enzyme kinetics [Liebherr and Gorris, 2014]. By combining this method with light-switchable inhibitors, the on- and off-states of single molecule reactions can be monitored. This project is part of a collaboration with Nadja Simeth and Karin Rustler from the group of Prof. Dr. Burkhard König as well as Matthias Mickert from the group of Prof. Dr. Hans-Heiner Gorris to create photo-switchable inhibitors of β-galactosidase (β-gal) for use in single molecule reactions. For photoswitching, azobenzene- and dithienylethene-derived photoswitches were considered [Brieke et al., 2012]: Azobenzene-derivatives share two phenyl rings linked by an N-N double bond; photoisomerization from the trans-isomer to the cis-isomer is triggered by ultraviolet light (see Fig. 3.2). Dithienylethene-derivatives have aromatic groups bonded to each end of a C-C double bond; the molecular structure can be switched between ring-open and ring-closed isomers by ultraviolet light (see Fig. 3.2). If an isomer has inhibitory activity, the conformational changes induced by switching to the other isomer may modulate this activity and the compound can act as a light-switchable inhibitor.

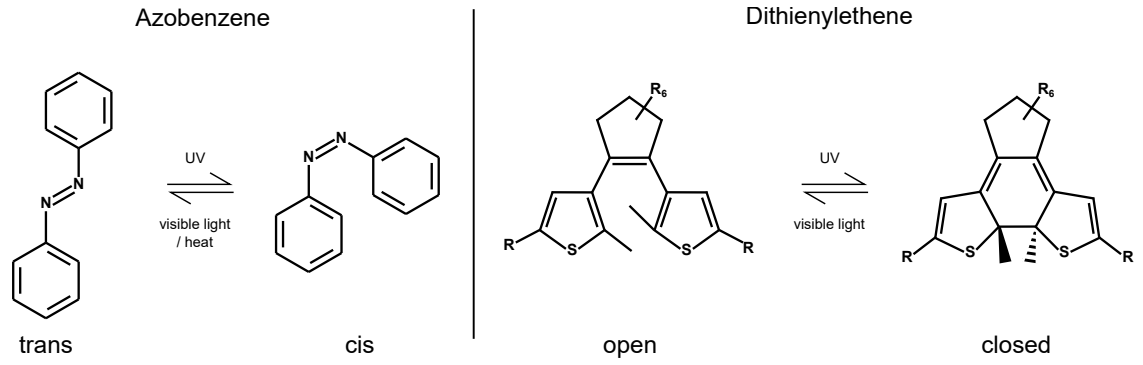


Fig. 3.2 **Photoswitching of azobenzenes and diarylethenes.** (**Azobenzene**) Photoisomerization of the trans-isomer is triggered by exposure with ultraviolet light; the cis-isomer can thermally or by exposure with blue light relax back to the trans-isomer. (**Dithienylethene**) The molecular structure can be switched from ring-open to ring-closed isomers by exposure with ultraviolet light , and vice versa by exposure with visible light.

The 3D structures of 121 ligands were provided for an analysis via docking that served as a preselection for later biochemical characterization. This dataset

contained substrates of β-gal, a known inhibitor for β-gal, and 118 inhibitor designs. Most inhibitor designs were based on a new class of potent β-gal inhibitors [Greul et al., 2001; Kleban et al., 2001], but also potentially poor candidates (decoys, non-photoswitchable) were considered to test the confidence of the docking algorithm. The small number of targets allowed docking with flexible binding site residues and the consideration of multiple β-gal structures: In total, 55 crystal structures of β-gal from *E. coli* were downloaded from the PDB and prepared for docking. Residues in the substrate binding site were held flexible and docking was performed using AUTODOCKVINA as described in Section 2.2.

After docking, the results for each single ligand were extracted based on the best hit sorted by binding energy from all 24 respective docking runs on all 55 crystal structures. Table 3.1 lists the three ligands with the highest and lowest binding energy after docking. For reference, AUTODOCKVINA defines the binding energy (given in $^{kcal}/_{mol}$, more negative means stronger binding) as the Gibbs-free energy by computing a conformation-dependent score of intra- and intermolecular energies; the standard error is 2.85 $^{kcal}/_{mol}$ as assessed on a benchmark dataset [Trott and Olson, 2010]. The binding energy of Resorufin-β-D-galactopyranoside (= positive control), which is a substrate of β-gal, was -11.23 $^{kcal}/_{mol}$. In comparison, decoys having no assumed inhibitory activity showed binding energies in a range of -6.434 $^{kcal}/_{mol}$ to -12.233 $^{kcal}/_{mol}$ (mean -8.473 $^{kcal}/_{mol}$). As -11.23 $^{kcal}/_{mol}$ lies still within the mean decoy $\pm$ standard error (-8.473-2.85 $^{kcal}/_{mol}$), the accuracy of docking is rather uncertain. The range of binding energies in the full dataset was -6.639 $^{kcal}/_{mol}$ to -13.067 $^{kcal}/_{mol}$ (mean -10.6493 $^{kcal}/_{mol}$), indicating that there are binders as well as non-binders for β-gal in this dataset of ligands, when compared to the above values.

Fig. 3.3 illustrates the docked poses of ligands with highest and lowest binding energies. ns_switch16_closed (Fig. 3.3 A), the closed-isomer of a dithienylethene-derivative, tightly binds to β-gal by form-complementarity and several hydrogen bonds. On the other hand, ns_cmpd2_ki8µM (Fig. 3.3 B), a decoy, also seems to be fairly stabilized at the correct binding site by hydrogen bonds. However, the form-complementarity is low and the desolvation penalties are high. In contrary to ns_switch16_closed, the poor binding energy is also reflected in the docking trajectory of ns_cmpd2_ki8µM, where diverse modes of binding occur at different locations in the binding site (data not shown).

Certainly, the total binding energy is no means for the actual switching capability. The perfect candidate for a putative switching inhibitor would possess an inhibitory isomer with a binding energy lower than the positive control and its non-inhibitory

Table 3.1 **The three ligands with the highest and lowest binding energy.** Row coloring serves visual guidance.

name	binding energy [$kcal/mol$]
ns_switch16_closed	-13.067
ns_switch5_closed	-12.996
ns_switch15_open	-12.95
...	...
ns_false1_trans	-7.958
ns_cmpd3_ki4-5µM	-6.847
ns_cmpd2_ki8µM	-6.639

isomer a preferably high binding energy. Hence, all variants were sorted by their switching capability (see Table 3.2 for the top three compounds), resulting in the ordered list of variants preselected for biochemical characterization. Although several variants exist with a predicted binding energy in the range of the positive control, switching to the respective isomer does not increase the binding energy above the level of the standard error (2.85 $kcal/mol$), see Table 3.2. In conclusion, the *in silico* analysis predicts that binders with inhibitory activity are available in this dataset but their switching capability is relatively uncertain and can only be determined experimentally, which is ongoing work at the time of writing. The preselection thus serves to differentiate between non-binders and putative inhibitors and increases the chance of finding a switchable one. An analysis that increased the accuracy was not feasible for this particular problem size. Methods that consider solvation, protein backbone flexibility and more sophisticated force fields rely on rescoring docked poses [Bienstock, 2015] and are computationally too expensive for a dataset of this size.

The software used for this project, AUTODOCKVINA, is one of the most frequently used state-of-the-art tools for fast protein-ligand docking. Computing time however was not negligible: With the given parameters for this docking experiment, one ligand required about 300 core hours when docked on all of the 55 crystal structures using one node[1] of the HPC cluster of Regensburg. All 121 ligands thus required around 36000 core hours, which testifies to a computing time of approximately four years for a modern workstation computer. Methods that increase the accuracy by

[1] 8 cores of an Intel Xeon Processor E5-2650 v2

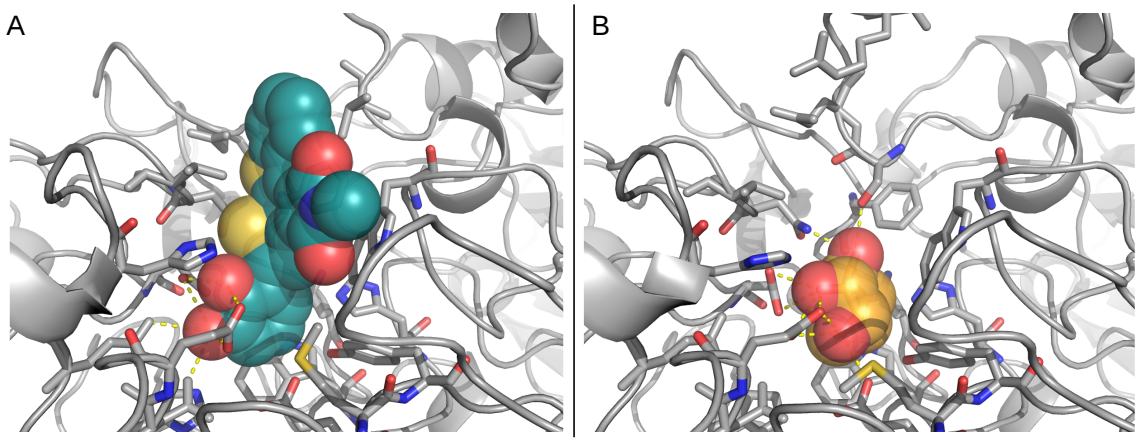


Fig. 3.3 **Protein-ligand complexes of the two ligands with the lowest and highest binding energy.** Conformation of ns_switch16_closed (**A**) and ns_cmpd2_ki8µM (**B**) docked into the binding pocket of an ensemble of 55 structures of β-galactosidases. The ligand conformer with the best binding energy after docking 24 times on an ensemble of 55 structures is shown. Hydrogen bonds are depicted as yellow dashed lines.

Table 3.2 **Top three dithienylethene- and azobenzene-derivatives.** Compounds were sorted by their highest difference in binding energy ($|\Delta|$=switching capability) of the two respective switching modes.

name	binding energy [$kcal/mol$]		
	closed	open	$\|\Delta\|$
ns_switch14	-12.668	-10.307	2.361
ns_switch5	-12.996	-11.003	1.993
dl_structure7	-12.019	-10.169	1.85
	trans	cis	$\|\Delta\|$
ns_switch11	-10.635	-9.026	1.609
kr_switch7	-11.338	-9.845	1.493
ns_switch24	-10.508	-9.248	1.26

considering explicit solvation and backbone flexibility require a computing time several times higher. In the next section, a method that uses a sophisticated force field and explicit solvation will be applied, e.g. to understand the binding modes of a single known light-controllable enzyme inhibitor (Subsection 3.2.1).

3.2 Molecular modeling considering structural dynamics

In contrast to the traditional rigid perception of visualized protein 3D models, proteins in solution exist as conformational ensembles [Wei et al., 2016]. The native state of a protein is consequently a distribution of conformations on the energy landscape determined by statistical thermodynamics. However, those conformational states are optimized by evolution to perform a protein's native biological functions. When modeling a protein, it is thus of utmost importance to capture the dynamic native state as close as possible.

To understand biochemical processes, methods for molecular modeling rely on experimentally determined structure databases such as the PDB. At the time of writing, 89% of all structures in the PDB are solved by X-ray crystallography. Structures determined this way provide static but high-resolution structural information based on artificially generated protein crystals. An alternative way to capture dynamic properties is to obtain 3D structures from proteins in solution: For this purpose, either NMR structure determination (9% of the PDB) or electron microscopy (1% of the PDB) is used; However, NMR is limited in the overall protein size while the resolution of EM structures - although improving - is still relatively low. For this reason, high-resolution structures are usually static snapshots from protein crystals while NMR and electron microscopy offer lower-resolution approximations of the native conformational ensemble.

Although the determination of a 3D model is an essential step to solve a problem in structural biology, a molecular model is much more than a simple representation of a static structure. If structure models are available, physics-based methods like MD simulations can be used to gather information about the time-dependent behavior of molecules on atomic level. However, the drawback of MD simulations is the computational cost: Depending on the system-size, atomically accurate simulations in explicit water require a day of computing time for few nanoseconds of simulation time on a single modern workstation. In this period of time, the observation of local

motions such as atomic fluctuations as well as side-chains and loop dynamics are possible. Extracting this information can be critical to solve biological questions in an intuitive way - by studying the molecular context; three examples are presented in this section: MD simulations were used to refine structure models (Subsection 3.2.1), increase the information value of a static model (Subsection 3.2.2) and observe structural changes upon modifying a model (Subsection 3.2.3).

3.2.1 Differences in binding modes of light-controllable enzyme inhibitors elucidated by molecular dynamics

As described in Subsection 3.1.2, light-mediated regulation of proteins is an exciting tool to understand and control biological processes. In publication **B**), several light-switchable inhibitors targeting the phosphoribosyl isomerase A from *M. tuberculosis* (mtPriA) were designed. mtPriA is part of the amino acid synthesis pathway and catalyzes two sugar isomerization reactions in tryptophane and histidine biosynthesis. In the latter pathway, the aminoaldose N′-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) is converted to the aminoketose N′-[(5′-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR). The fold of mtPriA is a $(\beta\alpha)_8$-barrel which is characterized by a twofold-symmetry. The substrate ProFAR binds to two phosphate binding sites that are located opposite each other.

In total, 13 C_2-symmetric photoswitches based on 1,2-dithienylethene (DTE) with terminal phosphate anchors were designed. Five out of 13 compounds acted as photoswitches as tested in UV light absorption experiments for several ring-closing/ring-opening cycles. Next, the inhibitory effect of those five compounds was analyzed in steady-state enzyme kinetics: All variants showed inhibitory effects comparable to the natural substrate ProFAR. The strongest difference in switching capability was found for *compound 6*; here the reaction rate could be enhanced by about threefold when switching from the open to the ring-closed isomer. However, the molecular basis of the different binding affinities for the open and closed state remained unclear. Thus, we structurally characterized the interaction of the strongest switching inhibitor with mtPriA via MD simulations.

First, the protein-inhibitor complexes for the open and closed conformer were modeled (see Section 2.4). Second, we performed three 10 *ns* simulations for each conformer and computed binding energies of the inhibitors for all snapshots of the MD trajectory. During MD simulation, the open and closed form of *compound 6* are

strongly fixed at the phosphate binding sites (Fig. 3.4 A,C). However, striking differences can be observed in their structural cores: The open isomer displays similar, C_2-symmetric conformers in three independent calculations (Fig. 3.4 B), while in the ring-closed isomer a terminal phenyl ring is twisted to facilitate phosphate group coordination (Fig 3.4 D). The experimentally determined difference in inhibition activity could be confirmed by the binding energies determined during simulation, which concordantly show that the open form is more favorable (Table 3.3). We concluded, that the increased flexibility of the open form allows a better structural adaption to the binding site which overcompensates the loss in conformational entropy upon binding.

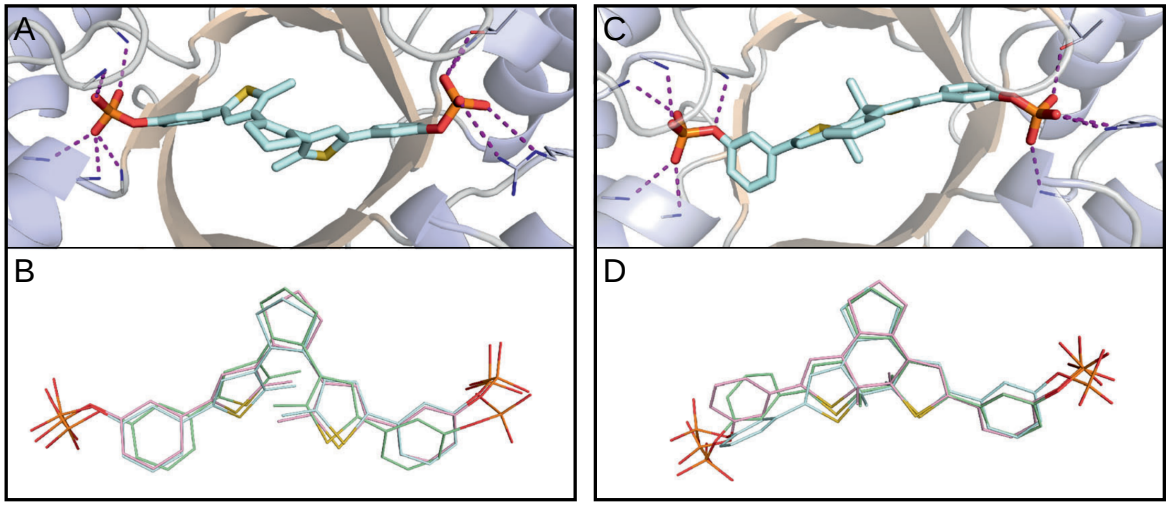


Fig. 3.4 **MD simulations of mtPriA and bound *meta*-phosphate 6.** For each isomer, three independent calculations were performed and representative enzyme structures for the open (**A**) and closed (**C**) form are shown. A superposition of the energetically most favorable conformer of each of the three simulations is depicted for the open conformation (**B**) and the closed conformation (**D**). This figure was reused and modified based on [Reisinger et al., 2014] with permission from *John Wiley and Sons*.

Table 3.3 Ligand binding energies derived from MD simulations of mtPriA and *compound 6* in its open and closed form

run	interaction energy [$^{kJ}/_{mol}$]	
	open form	closed form
1	-2171 ±55	-2062±64
2	-2162 ± 59	-1919±71
3	-2122 ± 77	-2049±60

3.2.2 Nucleosomal cores of human and plasmodial histones show similar binding affinities in MD simulations

Cellular DNA usually does not occur loosely but is condensed into a dense structure named chromatin [Cutter and Hayes, 2015]. From a functional point of view, this packaging allows a clear separation of chromosomes during mitosis, prevents DNA damage, and regulates gene expression as well as DNA replication. The basic elements of chromatin are nucleosomes, which are protein-DNA complexes that wrap DNA and are positioned on it like beads-on-a-string interconnected by sections of linker DNA. Approximately 146 base-pairs of DNA are wrapped around a single nucleosome, that consists of 2×4 subunits (Histone proteins H2A, H2B, H3, and H4). Besides a compact core that directly interacts with DNA, histones possess long tail domains with high intrinsic flexibility that are not suggested to contribute to complex stability [Luger et al., 1997]. Acting as general DNA-packers, nucleosomes are thought to bind DNA non-specifically. However, they are engineered by Nature to be meta-stable which facilitates assembly and disassembly [Workman and Kingston, 1992]. By affecting the accessibility of the genome, nucleosomes regulate the gene expression which is proposed to be governed by sequence preferences that control nucleosomes-positioning on the DNA.

The parasite of malaria, *P. falciparum*, has a complex life cycle involving transmission from a mosquito vector to a human host and several sexual and asexual development stages [Bousema et al., 2014]. Its chromatin structure is distinct from those of other eukaryotes, featuring high-accessibility and poorly positioned nucleosomes. This difference is suggested to be anchored in its genome: The genome-wide AT-content of *H. sapiens* is around 60% [Cohen et al., 2005], while that of *P. falciparum* is around 80% [Gardner et al., 2002] resulting in one of the most skewed eukaryotic

base pair compositions [Hamilton et al., 2016]. Given that also plasmodial nucleosome sequences are highly divergent from those of other eukaryotes, an adaption of nucleosomes to their AT-rich genome is assumed. In publication **C)**, human and plasmodial nucleosomes were analyzed and their binding to GC- and AT-rich DNA was studied via *in vitro* experiments: Interestingly, the sequence differences observed in plasmodial nucleosomes did not improve binding to AT-rich DNA. Moreover, for *P. falciparum*, reduced thermal and salt stabilities were measured, indicating that nucleosomes were also less stable on GC-rich DNA. In order to understand the structural differences of human and plasmodial nucleosomes, we performed a comparative analysis *in silico*.

To begin with, a homology model for the plasmodial nucleosome was built based on the human nucleosome structure (PDB ID 3afa). The fact that 84% of all amino acid residues in the human nucleosome are identical ensures a high quality 3D-model of the plasmodial nucleosome. Due to their flexibility, histone tails are structurally not determined in the crystal structure and hence, modeling was done without considering histone tails: Compared to the full-length sequence, the first 15 and last 13 residues of H2A and the first 27 of H2B, 42 of H3, and 23 residues of H4 were missing. For both the human and plasmodial model, 50 *ns* MD simulations were performed and DNA-protein interactions were scored for individual residues as well as for the full complex. For the residue-wise comparison, we assessed π-π stacking, cation-π stacking, *vdW* contacts, hydrophobic interactions, and hydrogen bond networks based on all snapshots of the MD trajectory.

The residue-wise comparison revealed that π-π stacking did not contribute noticeable to DNA-protein interactions and the comparison of all other residue-specific scores did not indicate striking differences (compare Fig. 3.5). Regarding the location of amino acid differences, there are no sequence differences in H3 and H4 at sites of direct histone-DNA interaction. On the other hand, H2A and H2B possess most amino acid differences in their N-terminal tails and in the H2A C-terminal tail, which are not covered by simulation data (see red squares in the "Tail regions" of Fig. 3.5).

For an overall energetic assessment, the histone- and species-specific differences in binding energy were determined by subtracting for each snapshot the score of the plasmodial DNA-histone interaction from the median score calculated for the human DNA-histone interaction by means of FOLDX (see Fig. 3.6). Only H2A#1 showed a slightly stronger DNA binding in human histones, which we did not consider significant for the following reasons: The reported accuracy of FOLDX

is 0.46 $^{kcal}/_{mol}$, which is the standard deviation of the difference between $\Delta\Delta G$ calculated by FOLDX and the experimental values. Human H2A sequences differ from plasmodial H2A sequences by 24 residues in the modeled core region, while 14 interacted with DNA in our analysis (compare Fig. 3.5). The median difference in H2A#1 energies was 2.03 $^{kcal}/_{mol}$. Consequently, the mean contribution of each mutation was $\frac{2.03\ ^{kcal}/_{mol}}{14}$=0.15 $^{kcal}/_{mol}$, which was below the reported accuracy. The mean contributions of each mutation were even smaller for the other histones and were therefore considered as a neutral effect. Based on the computational analysis, we suggested similar DNA binding for the core regions of all corresponding human and plasmodial histones, both in total and on a per residue basis.

Summing up, the results are controversial: The biochemical data showed clear differences in binding and nucleosome stability, while our *in silico* results proposed no difference. However, as an analysis of histone tails had not been possible, the computational analysis was blind for their contribution to binding energy. In addition, most sequence differences are found in tail regions of histones. Hence, this analysis suggested that mutations in the flexible histone tails and not in the histone cores of *P. falciparum* decrease nucleosome stability and DNA binding strength. Although histone tails had not been suggested to contribute to complex stability in the literature, this study suggested a crucial contribution which is a hypothesis to be tested in future studies.

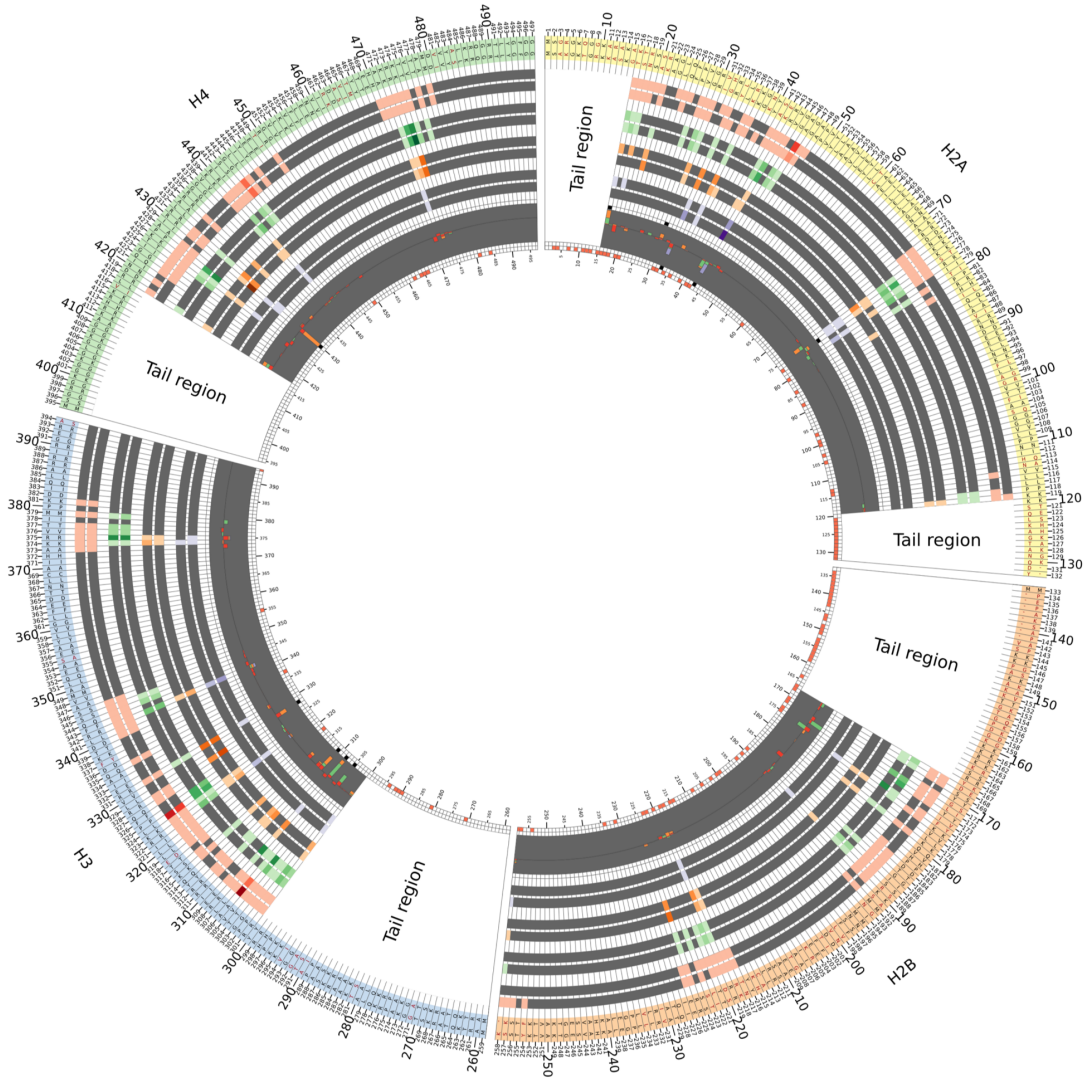


Fig. 3.5 **A scoring of interaction-differences for nucleosomal residues.** The outermost circle is an alignment of the residues from the four histones H2A, H2B, H3, and H4 from *H. sapiens* (outer sequence) and *P. falciparum* (inner sequence). Positions occupied by different residues are printed in red and additionally highlighted by a red rectangle in the innermost circle. The in-between circles consist of color-coded score values that indicate a higher score, if the color is dark (see Section 2.5). The order of the interactions is, if listed from the outer to the inner circles: hydrogen bond networks, hydrophobic interactions, *vdW* contacts, and cation-π stacking. Circle six summarizes the differences as stacked bars. Black boxes mark residues that possess at least one score belonging to the 10% most extreme values. No scores are given for the trimmed N- and C-termini (Tail regions) not considered during modeling. This figure was reused based on [Silberhorn et al., 2016] with permission according to the *Creative Commons Attribution license*.

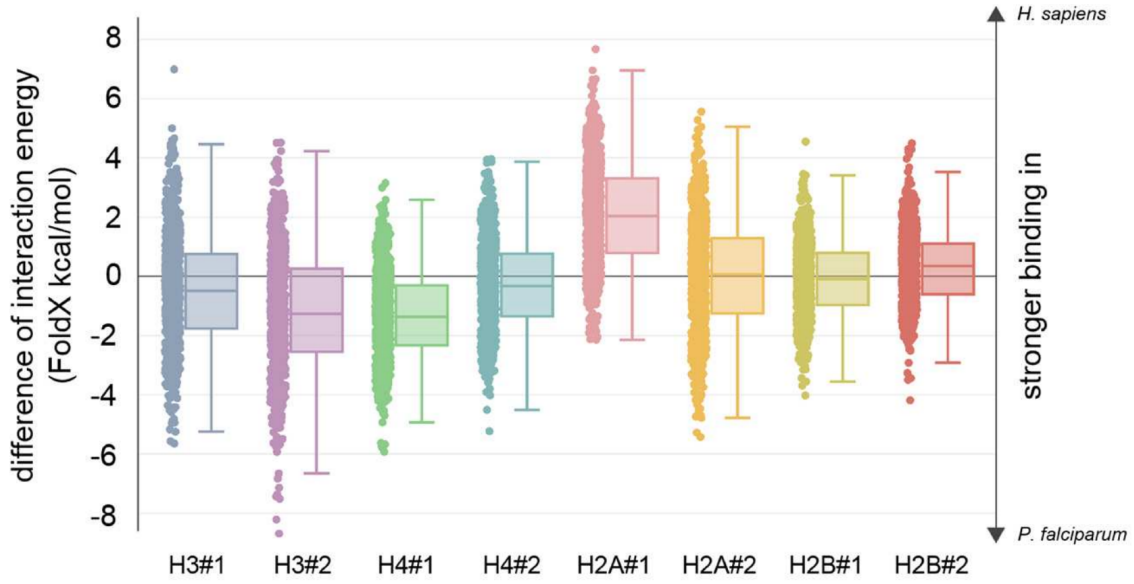


Fig. 3.6 **Differences in FoldX for the histone-DNA interaction.** For the *H. sapiens* histone cores, the available crystal structure (PDB ID 3afa), and for *P. falciparum* models of the histone cores were used; see Section 2.5. Energy differences were analyzed for each of the eight individual histones; individual histone copies are indicated by #1 or #2 respectively. Each dot of a scatter plot represents the difference between the median human interaction energy and a snapshot-specific plasmodial energy value. Interactions were further characterized by means of box plots. Whiskers indicate the lowest and the highest datum still within the 1.5 interquartile. This figure was reused based on [Silberhorn et al., 2016] with permission according to the *Creative Commons Attribution license*.

3.2.3 The substrate specificity of chorismate-utilizing enzymes correlates with a change in putative nucleophile channels

An organism's primary metabolism governs the processes that allow growth, development, and reproduction. In contrast, the secondary metabolism contributes to the fitness and the survival of the organism in its environment, for example by producing competitive weapons like antibiotics. It is still unclear, where the secondary metabolism evolved from, but the primary metabolism is debated to be a general predecessor.

In publication **D**), it was shown that the primary-metabolic enzyme anthranilate synthase (AS) from *S. typhimurium* (stAS, subunits TrpE:TrpG) can be converted into a bifunctional variant adopting the function of the secondary-metabolic isochorismate synthase by few mutations. The conversion from AS to ICS changes the nucleophile specificity from ammonia to water, which is transported via an intermolecular channel from the active site of subunit TrpG to the active site of subunit TrpE.

Due to the peculiarity of nucleophile-switching, the intermolecular channels were studied. First, stAS was analyzed using MOLE, a program to characterize channels, tunnels, and pores in molecular structures. MOLE found a 30 Å-long channel that connects the two active sites of stAS (see Fig. 3.7 A). This putative nucleophile channel near the chorismate ligand (CH) is lined by mainly three residues: Gln263, Met364, and Leu365. Based on data from a multiple sequence alignment of AS, ICS, and other homologous enzymes, 16 stAS variants were generated that carry mutations at these three channel-lining positions. Variants with the substitution Gln263Lys formed the product isochorismate (IC, originally produced by ICS) in the absence of an ammonia source when combined with a mutation at position 365. Interestingly, a striking increase in IC formation was observed for variations of the amino acid (Leu/Val/Ser/Ala) at this position.

The nucleophile-switching was suspected to be induced by structural rearrangements of the substrate channel. To gain further insight, we simulated all 16 variants via MD as described in Methods (Section 2.6); for each variant, 600 snapshots were generated over a total simulation time of 6 *ns*. Based on each MD snapshot and using MOLE, putative nucleophile trajectories (PNTs) were generated, where a PNT is defined by the centerline of the corresponding MOLE channel; for details, see Methods. Strikingly, a reduction of the size of the amino acid at position 365 corresponds to an increase in CH to IC conversion and is related to a shift in PNT trajectories which was quantified in a statistical analysis: First, all PNTs were classified as following

one of two paths (A- or L-path) traveling along its substrate tunnel (see equations (2.2),(2.3),(2.4) and Fig. 2.1). Next, the fraction of PNTs approaching CH along the A-path was computed and compared to the conversion rate of CH to IC (see Fig. 3.7 C). The localization of PNTs undergoes a notable shift when decreasing the size of residue position 365 (Leu>Val>Ser>Ala) from the L-Path to the A-Path while the percentage conversion of CH to IC changes in the same way. The PNTs confirming those findings were visualized in Fig. 3.8.

Intriguingly, few mutations suffice to establish secondary metabolic ICS activity based on the primary metabolic stAS. Moreover, all variants were bifunctional, as they were still able to form their natural product anthranilate. This conversion provides a cost effective evolutionary path from a primary to a secondary metabolic enzyme that does not require gene duplications by exploiting bifunctionality and thus supports the hypothesis that the secondary metabolism can evolve from the primary metabolism [Firn and Jones, 2000; Vining, 1992].

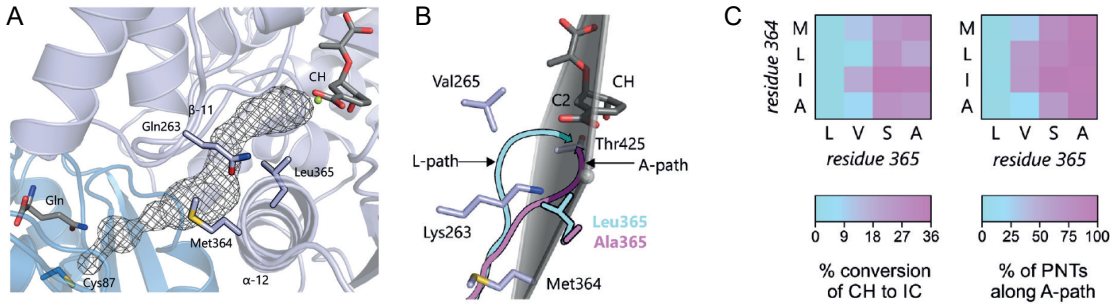


Fig. 3.7 **Visualization and quantitative analysis of PNTs in wild-type and mutant stTrpE variants.** (**A**) Nucleophile channel connecting the active sites of stTrpG (blue; represented by Cys87 and a glutamine ligand) and stTrpE (pale blue; represented by CH and a Mg^{2+} ion) in stAS (model based on PDB ID 1i7q). The channel-lining residues of stTrpE that had been mutated in experiments are shown as stick models. (**B**) The L-path (cyan) and A-path (magenta) reflect the majority of PNTs in variants with Leu365 and Ala365, respectively, and therefore show the boundaries of the PNT shift. The directions from which the two paths approach CH can be separated by a plane specified by CH-C2 and the Cα-atoms of Met364 and Thr425 (for details see Section 2.6 and Fig. 2.1). (**C**) Comparison of the average CH-to-IC conversion by all stTrpE variants with Lys263 and the fraction of PNTs proceeding along the A-path in these variants. This figure was reused and modified based on [Plach et al., 2015] with permission from *John Wiley and Sons*.

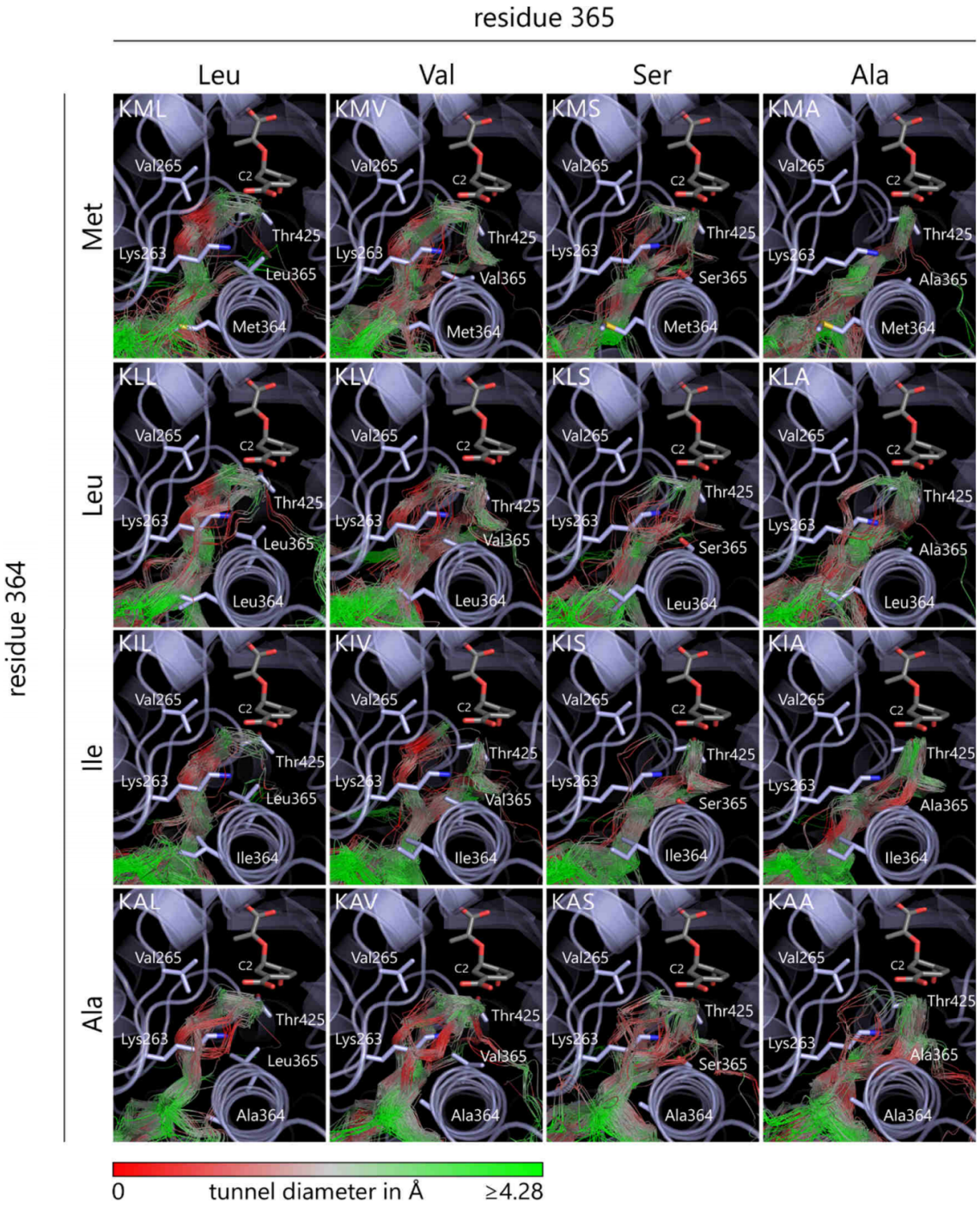


Fig. 3.8 **Comparison of PNTs in 16 stTrpE variants containing Lys263**. PNT localization is almost exclusively determined by residue 365. To a great extent, PNTs of Leu365-variants proceed along Val265, i.e. along the L-path (Fig. 2.1). In variants with Val365, approximately half of the PNTs proceed along the L-path; in the KAV variant this distribution is less pronounced. In variants with Ser365, PNTs further shift towards the A-path (Fig. 2.1). This effect is most pronounced in the KIS variant. In the Ala365-variants, nearly all PNTs follow the A-path. Diameters of the PNT associated channels can be deduced from the red-grey-green color scale. This figure was reused and modified based on [Plach et al., 2015] with permission from *John Wiley and Sons*.

3.3 ROSETTA:MSF: a modular framework for multi-state design

So far, computational protein design had mostly been carried out by optimizing sequences based on single conformations (i.e. design states) known as single-state design. However, as described in Section 3.2 a protein's structure is more accurately specified by a conformational ensemble. CPD on ensembles requires a multi-state design algorithm to simultaneously assess multiple conformations. Such an algorithm enables the protein designer to approach other challenging CPD objectives like multi-specificity design or the concurrent consideration of positive and negative design goals.

Rosetta [Leaver-Fay et al., 2011b] is a popular software suite to study and design proteins. Rosetta's protocols consist of specific procedures and a fine-tuned set of parameters to carry out a given task. However, most protocols were implemented for SSD. Thus, in the scope of this thesis, the framework MSF was developed that allows the user to apply existing Rosetta protocols in a multi-state environment. In the following, the implementation, architecture, and availability of MSF is described (Subsection 3.3.1). In Subsection 3.3.2, the performance of MSF is assessed on several benchmark datasets based on conformational ensembles. Finally, the enzyme design functionality of MSF is characterized and compared to the standard protocol (Subsection 3.3.3). The implementation of MSF together with the multi-state enzyme design of retro-aldolases described in Section 3.4 were submitted as publication **E)**.

3.3.1 Implementation and architecture

As part of MSF, we implemented a multi-state design logic based on Rosetta's genetic algorithm to explore the sequence space. As described in Subsection 2.7.2, the GA maintains a population of design sequences that are evolved for a number of generations. The fitness of an individual sequence results from the application-specific design protocol and a user-defined fitness function. For the initial implementation of MSF, two widely used Rosetta protocols were integrated: ENZDES provides ligand binding and enzyme design functionality by repacking and redesigning residues around the binding/active site and by optimizing catalytic contacts [Richter et al., 2011]. ANCHORED redesigns a protein-interface by using information from a known interaction at the same interface of one partner [Lewis and Kuhlman, 2011]. The resulting applications were validated and expose all options of the two protocols and

of the GA to the user. To accommodate the multi-state design approach, sequences are evaluated in the application specific design protocols for all given states. A combined score (=fitness) is calculated according to a so-called "dynamic aggregation function" (DAF, introduced by MPI_MSD [Leaver-Fay et al., 2011a]) that may weight different states individually and supports positive and negative design as well.

The framework MSF was integrated as an additional protocol into Rosetta and is purely written in C++98. It aims at significantly reducing the development effort of equipping arbitrary Rosetta protocols with multi-state design capability. Therefore, MSF bundles a number of classes responsible for distributing tasks to available computational resources, by establishing MPI [Gropp et al., 1996] based communication and task synchronization, as well as initialization and execution of Rosetta protocols. The software architecture is as follows for all protocols: One process is responsible for the logic of the GA and a user-defined number of additional processes are grafting and scoring, which guarantees high scalability. Every aspect of MSF was designed with the highest possible flexibility in mind. This allows the modification of superficial aspects of the algorithm as well as easy access to core elements due to a global allocation system managing the instantiation of polymorphic key classes. Simply put, most important functions of the multi-state design pipeline are exposed to the application and may be modified for each application individually without breaking compatibility to existing applications. Due to the modularity of the implementation, it is easier to extend single-state Rosetta protocols with multi-state design capability in contrast to already existing multi-state applications. In the following, ROSETTA:ENZDES (or for the sake of brevity ENZDES) and ROSETTA:MSF:GA:ENZDES (MSF:GA:ENZDES) are the names of the SSD and MSD implementations for enzyme design; ANCHORED and MSF:GA:ANCHORED are the SSD and MSD implementations for anchored protein-protein interface design.

Comparison to existing MSD approaches in ROSETTA

This is not the first implementation of multi-state design in Rosetta. MSD methodology extends the application spectrum and thus, Rosetta offers several multi-state applications; noteworthy are MPI_MSD [Leaver-Fay et al., 2011a] and RECON [Sevy et al., 2015]. MPI_MSD provides a generic multi-state design implementation based on a genetic algorithm that optimizes a single sequence on multiple states given a fitness function. RECON starts by individually optimizing one sequence for each state; subsequently the computation of a consensus sequence is promoted by incrementally increasing convergence restraints. However, the current implementations of

both methods are limited to certain design tasks and cannot make use of fine-tuned protocols, e.g. those required for enzyme design (ENZDES) or anchored design of protein-protein interfaces (ANCHOREDDESIGN). In order to overcome this limitation, MSF was developed and the integration of MSF into Rosetta enables the use of already proven single-state protocols in a MSD environment.

MSF relates to MPI_MSD as a progression, relying on the same GA protocol and DAF accounting for the different states. In MPI_MSD the maximal number of processes that can be used efficiently is limited by the number of design states, while MSF is highly scalable by being able to utilize up to states$\times$population processes. Also, MSF was written from scratch and designed with high maintainability and flexibility in mind, allowing any developer to extend existing Rosetta protocols with multi-state design capability. We thus present MSF as a third option for multi-state design in Rosetta which may especially be considered in cases where generic design algorithms cannot be applied or produce unsatisfactory results. MPI_MSD relies on the standard packing procedure which grafts sequences onto a pose in a generic way and does not support specialized tasks like enzyme or interface design out of the box. At the time of writing, it was also not possible to utilize RECON in the same manner, since it does not support the strict separation of processes required by the GA to synchronize the design processes during the transition of a generation.

Availability, installation, and command line options

The integration of MSF into Rosetta's master branch is current work of Samuel Schmitz at the time of writing and aims at providing MSF-support for upcoming ROSETTA releases. However, MSF is available via a git-branch based on version 2015.19.57819, which contains the final version of MSF used for benchmarking and retro-aldolase designs. The branch with the name SamuelSchmitz/msf_2015.19.57819 contains the two applications enzyme design (application msf_ga_enzdes) and anchored interface design (msf_ga_anchored) and can be accessed only from RosettaCommons developers by cloning https://github.com/RosettaCommons/main/tree/SamuelSchmitz/msf_2015.19.57819 via git [Torvalds and Hamano, 2010]. It is required to compile Rosetta with MPI support, e.g. ./scons.py mode=release extras=mpi msf_ga_enzdes msf_ga_anchored. The list of available command line options and an example command to run MSF are listed in Appendix A.

3.3.2 Comparing multi-state and single-state protein design performance via conformational ensembles

Compared to SSD, MSD approaches offer a broader functionality, allowing the integration of multi-specificity design, negative design, and mimicking conformational flexibility. A direct comparison between SSD and MSD approaches is thus not always possible. However, this thesis focuses on assessing the design performance of SSD and MSD methods based on conformational ensembles. This allows an easy comparison of the designed sequences by performing MSD on a conformational ensemble and SSD on all single rigid states of the ensemble. For SSD, the standard applications of Rosetta were applied that utilize Monte Carlo optimization; results of the computation were averaged over all conformations. For MSD, MSF:GA was utilized and results were computed ensemble-wise. Scores were computed by assessing the fitness according to Equation (2.10) based on the Rosetta total score (ts) averaged over all states of the ensemble.

MSD performs better than SSD in recapitulating a ligand binding site of an NMR ensemble

The most obvious usage of MSD is its application on an ensemble representing the native conformations of one protein. In solution, a protein's structure is varying and nuclear magnetic resonance (NMR) offers an experimentally determined estimation of its variability. Interestingly, in previous analyses SSD protocols performed better on crystal structures than on NMR templates [Allen et al., 2010; Schneider et al., 2009]. We speculated that this performance loss could be compensated, if MSD is applied to a whole ensemble and we decided to assess a ligand-binding design. Hence, for a first performance comparison of the SSD algorithm ENZDES, and the MSD algorithm MSF:GA:ENZDES, we chose an NMR ensemble of the human intestinal fatty acid binding protein (hIFABP) with bound ketorolac (PDB ID 2mji).

This ensemble, consisting of ten conformations, was prepared for ligand-binding design (see Subsection 2.7.1) and the design shell contained 21 residues in the vicinity of the ligand. Our protocol allowed Rosetta to find a low energy sequence by arbitrarily choosing any residues for all positions of the design shell, which is a predefined set of residues surrounding the ligand. 1000 randomly seeded $runs_l(i)$ of ENZDES (SSD) were started for each of the individual conformations $conf(l)$ and the design quality was monitored by computing for each number of runs i the score $ts_{SSD}^{hIFABP}(i)$. This is the mean total score deduced from corresponding conformations

(Equation (2.8)) and it is given in Rosetta Energy Units (*REU*). MSF:GA:ENZDES (MSD) was applied to the full ensemble and the GA was executed for 800 generations. Analogously, the mean total score $ts_{MSD}^{hIFAB}(j)$ was computed for each generation j (Equation (2.11)). As a second measure of design quality, we determined the native sequence similarity recovery (*nssr*). Commonly, the performance of design algorithms is assessed by means of the native sequence recovery (*nsr*) [Havranek et al., 2004; Hu and Kuhlman, 2006; Humphris and Kortemme, 2007], which is the fraction of identical residues at corresponding positions of the native and the designed sequence. The concept of *nsr* is blind for a more specific comparison of residues beyond identity, which may impede a detailed assessment. In contrast, for the computation of *nssr*, all residue pairs with a BLOSUM62 score > 0 are considered similar and contribute to the *nssr* value (equations (2.6) and (2.7)).

The plots shown in Fig. 3.9 indicate that the SSD and the MSD algorithm converged, both with respect to sequence recovery and the *ts* values of the chosen sequences. In summary, MSF:GA:ENZDES performed better than ENZDES; the mean *nssr* values after convergence were 46.66% and 41.90%, respectively. Moreover, only two of the ten ENZDES designs reached an *nssr* value (47.62% and 61.90%, respectively) that was higher than the mean *nssr* of MSF:GA:ENZDES. These findings suggest to prefer MSD, if sequences have to be designed for an ensemble. Regarding energies, the SSD solutions score on average better than those of the MSD solutions, with a difference of 7.11 *REU*. However, a comparison of *ts* scores is no proper means to compare SSD and MSD performance: In MSD, a sequence is a compromise that has to satisfy the constraints associated with all conformations in an acceptable manner. Contrariwise, in SSD the algorithm can select for each conformation a highly customized (low energy) sequence. Thus, it is no surprise that the mean *ts* values of SSD sequences are superior to those of the MSD results. On the other hand, due to these specific adaptations based on single, less-native conformations, the SSD sequences are receding from the native ones, which are considered as close to optimal [Kuhlman and Baker, 2000]. This undesired effect is less pronounced for MSD sequences computed on the whole native ensemble. Thus, the *nsr* and *nssr* scores are more suitable than *ts* values for a comparative benchmarking of SSD and MSD approaches.

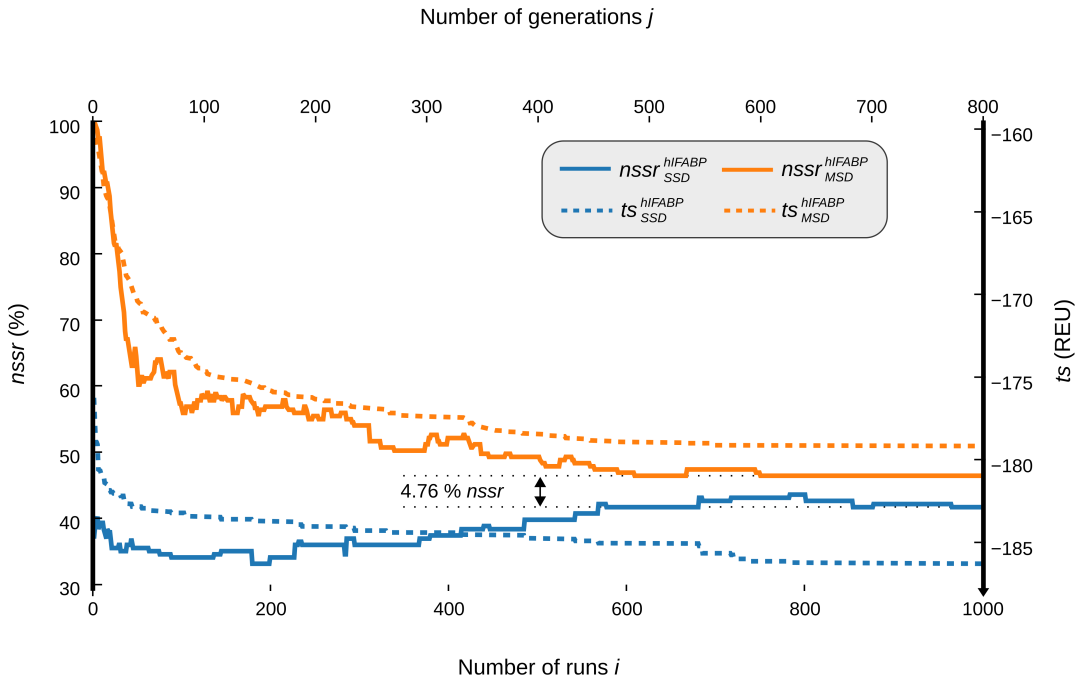


Fig. 3.9 **Performance of SSD and MSD on the NMR ensemble hIFABP.** ENZDES (results in blue lines) was executed for 1000 runs i for each of the ten conformations in the ensemble. For each number of runs i, the $ts_{SSD}^{hIFABP}(i)$ value (dotted line) is the mean of the ten lowest-energy sequences (Equation (2.8)). The corresponding $nssr_{SSD}^{hIFABP}(i)$ value (solid line) is the mean recovery value deduced from the same sequences (equations (2.7) and (2.9)). MSF:GA:ENZDES (results in orange lines) was carried out for 800 generations j on the whole ensemble using a population of 210 sequences. For each generation j, the $ts_{MSD}^{hIFABP}(j)$ value (dotted line) is the mean of the ten lowest-energy sequences of the corresponding population (Equation (2.11)). The corresponding $nssr_{MSD}^{hIFABP}(j)$ value (solid line) is the mean recovery value deduced from the same sequences (equations (2.7) and (2.12)).

A novel benchmark dataset for ligand-binding based on conformational sampling

A standard dataset for the assessment of ligand-binding and enzyme design is *Rosetta's scientific sequence recovery benchmark*. It consists of 51 representative proteins in which the ligand is bound with an affinity of 10 µM or lower [Nivón et al., 2014]. During benchmarking, it is the task of a given CPD algorithm to redesign residues of the design shell enclosing these ligands. The ability of the algorithm to recapitulate for the native backbone the native sequence (*nsr* and *nssr* values) is taken as performance measure. However, for an assessment of *de novo* design algorithms, this approach may be misleading, because the required remodeling of a chosen protein is more demanding than the recapitulation of its native binding pocket. This is why we created a novel dataset that is devoid of a perfect backbone and rotamer preorganization and is thus more suitable for the assessment of *de novo* design algorithms. For feasibility reasons, we randomly selected 16 $prot(k)$ of the above proteins. The corresponding ligands were removed and for each of the 16 apoproteins, an ensemble of 20 conformations was created by means of the BACKRUB server [Lauck et al., 2010], which is known to generate near-native conformational ensembles [Davis et al., 2006; Lauck et al., 2010]. Next, by superposition of each conformation with the corresponding crystal structure, the ligands were transferred to the binding pockets. Thus, the resulting dataset *BR_EnzBench* features for each of the 16 proteins 20 backbone conformations that are near to native but lack the implicit preorganization induced by a bound ligand in a crystal structure. A graphical overview of the benchmark setup is given in Fig. 3.10.

MSD performs better than SSD on a benchmark dataset mimicking *de novo* ligand-binding design applications

We used *BR_EnzBench* to compare the performance of SSD and MSD for *de novo* ligand-binding design. To begin with, all design shell residues were mutated to alanine and the conformations were energy-minimized to further increase the difficulty for CPD algorithms to recover the native sequence. To prevent a hydrophobic collapse of the alanine-only design shells, the energy minimization was performed with backbone constraints (see Appendix B.1). Thus, the CPD problem to be solved within the scope of this benchmark was to design a binding pocket by sequence optimization of the all-alanine design shells.

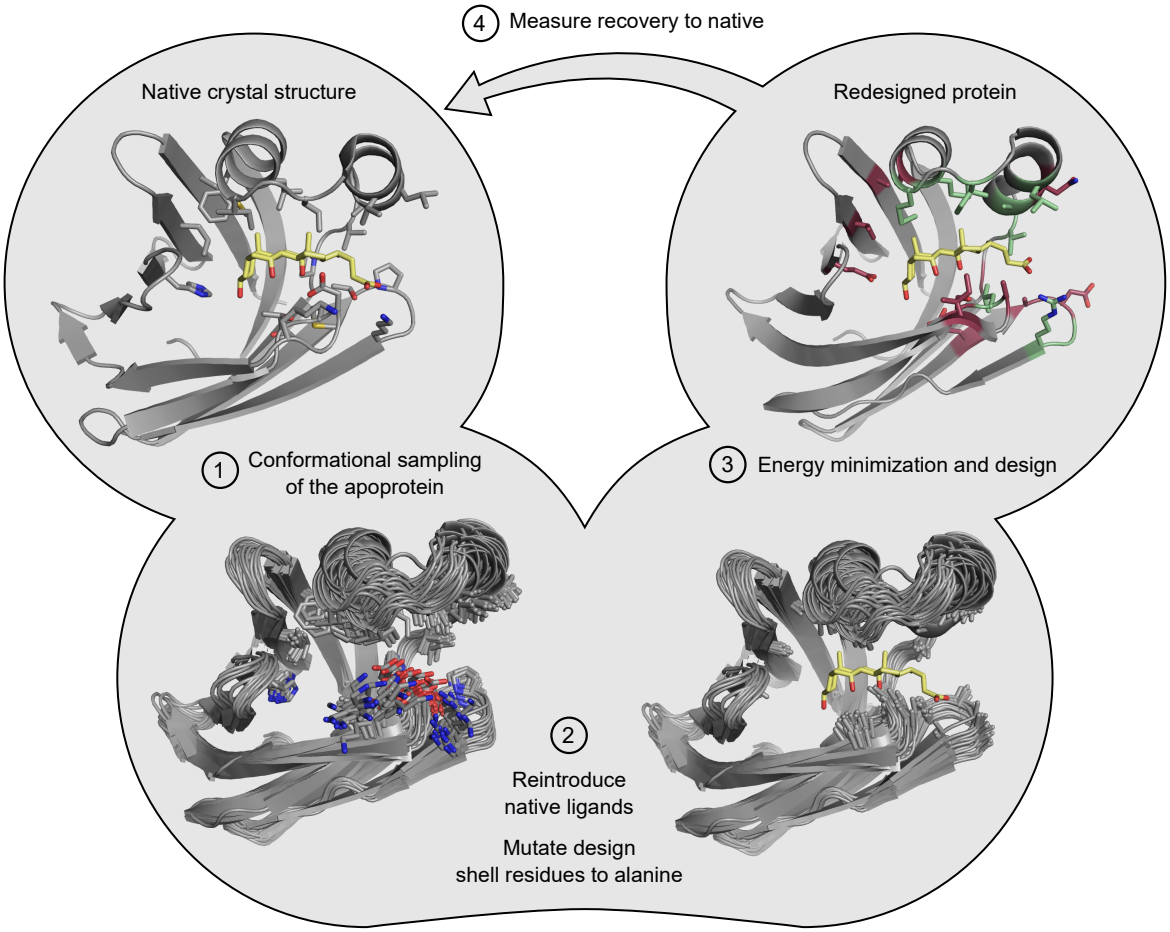


Fig. 3.10 **The compilation of novel benchmarks for *de novo* computational ligand-binding design.** (**1**) 16 PDB structures were taken from *Rosetta's scientific sequence recovery benchmark*. For each structure, the ligand was removed and the protein structure was conformationally sampled with either Backrub (*BR_EnzBench*) or molecular dynamics (*MD_EnzBench*). (**2**) For each protein conformation, the respective native ligands were reintroduced by means of PYMOL. To delete the native preorganization and increase the difficulty for the design algorithm, all design shell residues were mutated to alanine. (**3**) All conformations were relaxed with backbone constraints to adapt structures for Rosetta computations; then, all conformations were computationally designed. (**4**) For performance comparison the sequence recovery to native was measured.

For SSD with ENZDES, all conformations of each protein were considered independently for design. For each conformation, 1000 randomly seeded designs were performed and their quality was assessed by means of the three parameters *nsr*, *nssr*, and *ts*. The respective values were taken from best results after $i = 1000$ runs and averaged for each of the 16 $prot(k)$ (equations (2.8) and (2.9)) and are listed in Table 3.4. Additionally, the convergence of the design process was followed by monitoring the mean performance for each number i of design runs; these values are plotted in Fig. 3.11. To conduct multi-state design by means of MSF:GA:ENZDES, for each $prot(k)$ in the benchmark dataset, the 20 conformations were divided into $m = 4$ ensembles ens_m^k each containing five conformations. The GA was executed for 600 generations on a population consisting of 210 sequences. Analogously, *nsr*, *nssr*, and *ts* values (equations (2.11) and (2.12)) were determined for each MSD run and averaged for each of the 16 $prot(k)$. The results for $j = 600$ were added to Table 3.4. As above, the convergence of the GA was followed by monitoring the mean performance for each number j of generations; these values are also plotted in Fig. 3.11.

The protein-wise comparison (Table 3.4) indicates that in ten out of 16 cases, the *nsr* and in 13 out of all 16 cases, the *nssr* values of MSF:GA:ENZDES designs are superior to the corresponding values of ENZDES designs. As summarized in Fig. 3.11, MSF:GA:ENZDES recovers on average a higher percentage of native residues (Δnsr = 2.65%) and a higher percentage of similar residues ($\Delta nssr$ = 6.79%). Thus, with respect to the more adequate similarity measure *nssr*, MSD performs 15% better than SSD. In addition, MSD designs have slightly better energies (Δts = 2.51 *REU*), which is in contrast to the *hIFABP* results and is most likely due to the smaller ensemble size. Fig. 3.11 reflects the differences in convergence speed of both algorithms and indicates that the better performance has its price: The MC optimization utilized by ENZDES leads to acceptable design solutions even after a low number of runs. In contrast, the GA of MSF:GA:ENZDES is slower and more than one hundred generations are required to surpass the performance of the SSD algorithm.

MSD compared to SSD on a benchmark dataset for protein-protein interface design

As mentioned before, MSF was also equipped with MSD functionality for anchored protein-protein interface (PPI) design by integrating the *AnchoredDesign* protocol. This protocol [Lewis and Kuhlman, 2011] serves to redesign the PPI of one partner

Table 3.4 **Performance of SSD and MSD for individual proteins from *BR_EnzBench*.** *nsr*, *nssr* and *ts* values were determined for each of the 16 $prot(k)$ from *BR_EnzBench* after convergence of ENZDES and MSF:GA:ENZDES. For details, see Subsection 2.7.2.

PDB ID	***nsr* (%)**		***nssr* (%)**		***ts* (*REU*)**	
	ENZDES	MSF:GA: ENZDES	ENZDES	MSF:GA: ENZDES	ENZDES	MSF:GA: ENZDES
1fzq	53.25	37.75	58.25	48.75	-325.16	-328.55
1hsl	34.74	33.95	60.00	59.47	-448.58	-447.39
1j6z	29.81	34.81	41.11	51.30	-771.96	-774.62
1n4h	28.80	28.80	53.00	59.40	-484.49	-488.89
1nq7	30.89	32.32	51.79	57.68	-506.56	-511.51
1opb	24.77	35.68	45.00	52.27	-307.57	-307.50
1pot	12.11	17.89	41.84	43.68	-613.10	-613.72
1urg	16.05	32.63	26.05	42.63	-796.85	-799.61
2b3b	24.41	41.47	32.35	50.59	-831.19	-831.17
2dri	21.58	25.79	42.89	55.26	-611.75	-613.74
2ifb	24.77	30.23	41.82	49.09	-305.08	-305.86
2q2y	38.70	39.13	48.48	56.52	-609.27	-611.17
2qo4	45.91	40.68	56.82	62.27	-271.47	-277.49
2rct	27.27	20.45	49.32	47.27	-317.51	-320.33
2rde	14.50	19.00	25.50	37.75	-463.52	-471.90
2uyi	38.26	37.61	47.17	56.09	-640.19	-641.01
Average	29.11	31.76	45.09	51.88	-519.02	-521.53

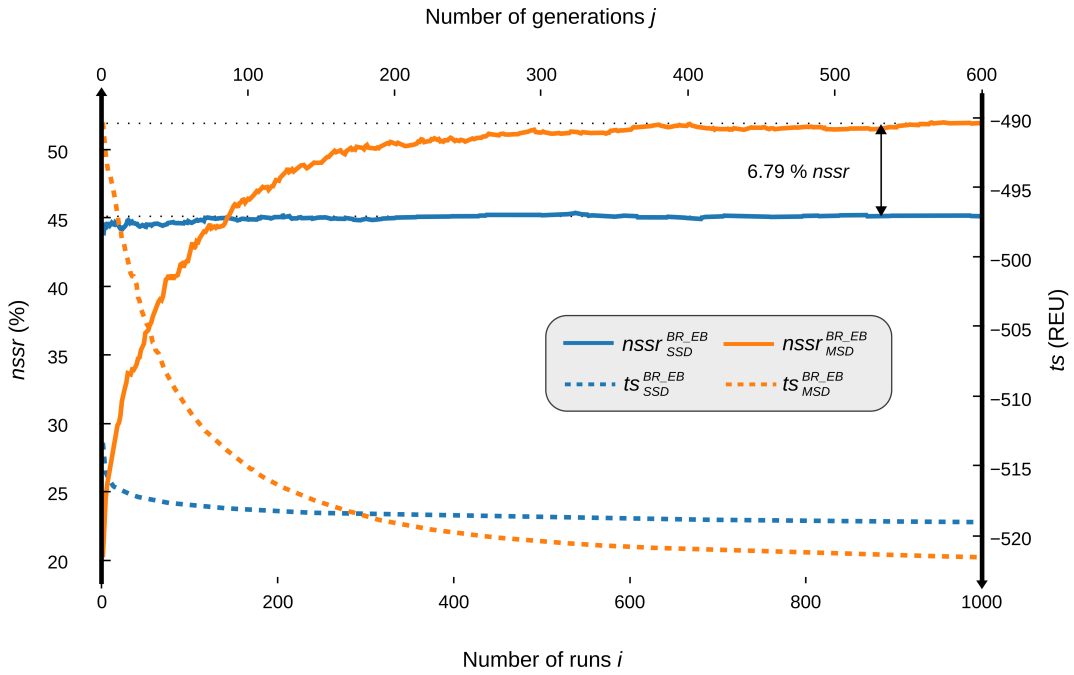


Fig. 3.11 **Convergence of SSD and MSD algorithms on the benchmark set *BR_EnzBench*.** ENZDES (results in blue lines) was executed for 1000 runs i on all 20 conformations of each $prot(k)$ from *BR_EnzBench*. For each number of runs i, the $ts_{SSD}^{BR\_EB}(i)$ value (dotted line) is the mean of the twenty lowest-energy sequences (Equation (2.8)). The corresponding $nssr_{SSD}^{BR\_EB}(i)$ value (solid line) is the mean recovery value deduced from the same sequences (equations (2.7) and (2.9)). MSF:GA:ENZDES (results in orange lines) was carried out for 600 generations j on all ensemble groups using a sequence population of 210. For each generation j, the $ts_{MSD}^{BR\_EB}(j)$ value (dotted line) is the mean of the five lowest-energy sequences of each of the four ensemble groups (Equation (2.11)). The corresponding $nssr_{MSD}^{BR\_EB}(j)$ value (solid line) is the mean recovery value deduced from the same sequences (equations (2.7) and (2.12)).

by exploiting already known interactions (the anchor) of the target protein with any other partner. In contrary to *de novo* interface design, anchored interface design simplifies the problem by providing the anchor as a known starting point for PPI design.

To assess the design performance of MSD and SSD for PPI design, a benchmark dataset was required. However, the original dataset that served to benchmark the ANCHOREDDESIGN protocol is a fixed sequence structure prediction benchmark [Lewis and Kuhlman, 2011]. It consists of $k = 16$ protein-protein complexes $complex(k)$ that possess interface loops (length 8-16) mediating binding; this benchmark aims at testing the performance of the algorithm to correctly predict the structure of those interface loops given their native sequence and an anchor residue part of the interface loop.

Although the fixed sequence structure prediction benchmark provides a quick verification of the algorithm, it is not meaningful for a comparison of the SSD and MSD approach. Hence, we were interested in the sequence design performance of this protocol for SSD and MSD. For this purpose, we compiled *BR_IfaceBench* in analogy to *BR_EnzBench* by applying the BACKRUB server to generate ensembles of the native protein-protein complexes as described in Subsection 2.7.1. Prior to design, all residues part of the interface loop except the anchor were mutated to alanine. Next, ANCHORED and MSF:GA:ANCHORED was given the task of designing the sequence and structure of the interface loop by sequence design and rigid-body docking using the anchor residue as a starting position. This is obviously much more difficult than just predicting the structure of the interface loop given the native sequence.

For SSD, eight randomly seeded runs of ANCHORED were applied on all $l = 20$ conformations $conf(l)$ of a $complex(k)$ as described in Subsection 2.7.2. For each k and each l, the *ts* , *nsr*, and *nssr* values were extracted and are listed in Table 3.5 (equations (2.8) and (2.9)). For MSD, the 20 conformations were divided into $m = 4$ ensembles ens_m^k each containing five conformations. Then, a two-step protocol was applied to each ensemble in order to speed up the computation (see Subsection 2.7.2 for details): MSF:GA:ANCHORED was executed for 1000 generations on a population of 50 sequences using a coarse-grained atom model. The energetically best 50 sequences were extracted and used to seed a refinement run for 500 generations on a population of 50 sequences. For each k and each m, the *ts*, *nsr*, and *nssr* values were extracted and added to Table 3.5 (equations (2.11) and (2.12)).

In contrast to the results obtained from the ligand-binding benchmarks, MSF:GA:ANCHORED does not show a clear performance advantage over ANCHOREDDESIGN in this benchmark dataset: Although the *nsr* value is 5.9% higher for MSD than for SSD, the *nssr* value of MSD is 3.68% lower in total. The results were not further analyzed, for the following reasons: Due to the much higher computational complexity of the ANCHOREDDESIGN protocol for sequence design purposes, it was not possible to run exhaustive calculations as done for the ENZDES protocol. For MSD, the number of generations in the refinement run and the population size were chosen relatively small. Also, the SSD protocol was executed only eight times for each conformation. This is much lower than the 1000 runs performed for ENZDES. For feasibility, the parameters controlling the number of loop design cycles were set to a relatively small number (see subsections B.3 and B.3), allowing only limited loop movements. For a typical production run, the parameters -AnchoredDesign::perturb_cycles and -AnchoredDesign::refine_cycles would be set to values over 1000, allowing extensive loop movement but requiring at least 20 times more computational ressources. It was thus not possible to deduce meaningful conclusions from the available data without further computations.

Table 3.5 **Performance of SSD and MSD for individual proteins from *BR_IfaceBench*.** *nsr*, *nssr* and *ts* values were determined for each of the 16 $prot(k)$ from *BR_IfaceBench* after convergence of ANCHORED and MSF:GA:ANCHORED. For details, see Subsection 2.7.2.

PDB ID	*nsr* (%)		*nssr* (%)		*ts* (*REU*)	
	ANCHORED	MSF:GA: ANCHORED	ANCHORED	MSF:GA: ANCHORED	ANCHORED	MSF:GA: ANCHORED
1dle	32.86	36.43	39.29	52.14	-732.80	-704.43
1fc4	31.00	29.00	63.00	36.00	-1355.7	-1291.40
1fec	15.00	16.43	21.43	22.86	-1426.38	-1418.50
1jtp	35.00	40.00	76.67	57.22	-371.86	-361.16
1qni	12.14	12.50	22.14	19.64	-1580.52	-1571.14
1u6e	25.00	31.67	47.78	52.78	-997.23	-985.55
1zr0	12.86	20.71	27.86	27.14	-306.34	-300.43
2bwn	45.00	52.31	59.62	62.31	-1253.90	-1229.77
2hp2	22.00	27.67	47.00	32.00	-1179.75	-1154.77
2i25	23.57	24.29	67.14	47.14	-633.55	-613.92
2obg	7.22	0.00	22.78	15.00	-634.44	-626.59
2qpv	21.43	32.86	47.86	40.71	-339.21	-326.45
2wya	18.89	28.89	59.44	48.33	-1363.26	-1346.46
3cgc	24.62	28.46	36.54	40.77	-1359.91	-1334.30
3dxv	21.82	28.64	51.82	36.82	-1320.67	-1307.15
3ean	12.14	16.43	16.43	24.29	-1463.59	-1354.14
average	20.59	26.49	42.96	39.28	-1193.84	-995.39

3.3.3 Characteristics of ENZDES

In the following subsection, the benchmark results of SSD and MSD for ligand-binding design were further analyzed. First, we show that the MSD concept in fact accounts for the performance advantage over SSD. Second, the different sequence preferences of MSD and SSD are studied.

The MSD concept is crucial for performance on *BR_EnzBench*

The sequence recovery results of the *hIFABP* benchmark and for *BR_EnzBench* strongly suggest that MSF:GA:ENZDES is superior to ENZDES in more complex design applications. However, it was unclear to us, whether the different concepts (single-state versus multi-state) or the different optimizers (MC versus GA) contributed most to the performance. Choosing a MSD approach increases computational cost which has to be substantiated by making plausible that the choice of the optimizer has less effect on the performance.

As described before, the performance of MSF:GA:ENZDES on *BR_EnzBench* was assessed ensemble-wise by determining for each ens_m^k the *nssr* scores, which were averaged (Equation (2.12)). Due to the stochastic approach of the Backrub algorithm, which was used to create the conformational ensembles (see Subsection 2.7.1), the conformations that are combined in each of the ensembles ens_m^k are unrelated. As these ensembles contain not more than five conformations each, the $nssr_{MSD}(ens_m^k)$ values (Equation (2.13)) vary due to the small sample size and one can sort for each $prot(k)$ the four ens_m^k on their $nssr_{MSD}(ens_m^k)$ value. The result is a ranking $ens_{rank=u}^k (1 \leq u \leq 4)$ of the four ensembles and we created the set ES_1 that contained the 16 ensembles (one for each $prot(k)$) with the lowest $nssr_{MSD}(ens_m^k)$ value. Analogously, we compiled the sets $ES_2 - ES_4$; consequently, ES_4 consisted of those 16 ensembles that had the highest $nssr_{MSD}(ens_m^k)$ value; for details see Subsection 2.7.3.

For these four sets ES_i, boxplots of the corresponding $nssr_{SSD}$ and $nssr_{MSD}$ values were determined; see Fig. 3.12. The boxplots characterizing the SSD results are nearly identical; this finding indicates that the conformations allocated to the four sets ES_1-ES_4 give rise to a similar SSD performance. Moreover, the boxplots representing the $nssr_{SSD}(ES_1)$ and $nssr_{MSD}(ES_1)$ values are nearly identical (median values 47.60% and 47.76%), which indicates that the optimizer GA is not generally superior to MC. Additionally the continuous increase observed for $nssr_{MSD}(ES_1) \rightarrow nssr_{MSD}(ES_4)$ (but not for $nssr_{SSD}(ES_1) \rightarrow nssr_{SSD}(ES_4)$ values) supports the notion that it is the combination of conformations that strongly affects MSD performance. Thus, we

concluded that the MSD approach (and not the optimizer) contributes most to the performance of MSF:GA:ENZDES.

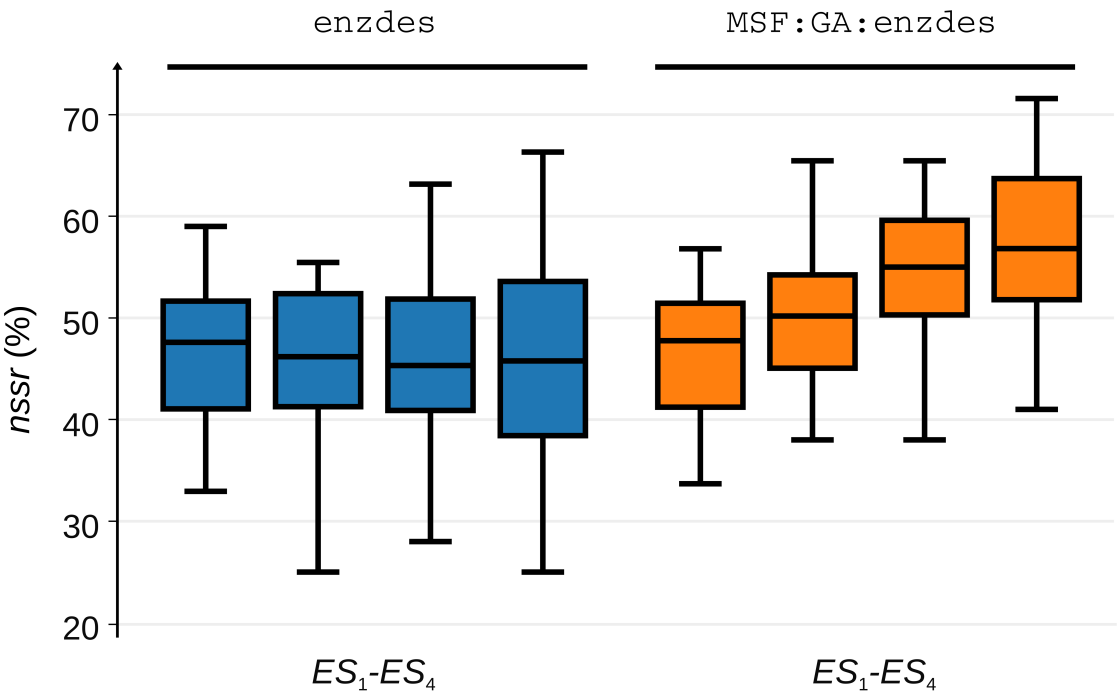


Fig. 3.12 **Performance of ENZDES and MSF:GA:ENZDES on a distinct grouping of conformations.** Each of the sets ES_1-ES_4 contains a quarter of the conformations from *BR_EnzBench*, which were grouped according to their $nssr_{MSD}$ values as described in Subsection 2.7.3. ES_1 contains all ensembles with the lowest and ES_4 those with the highest recovery values. For each set ES_i, the corresponding $nssr_{SSD}(ES_i)$ and $nssr_{MSD}(ES_i)$ values are represented by two boxplots. (**Left**) Performance of ENZDES (blue boxplots). (**Right**) Performance of MSF:GA:ENZDES (orange boxplots). Whiskers indicate the lowest and the highest datum still within the 1.5 interquartile range.

The residue preferences of ENZDES and MSF:GA:ENZDES differ

It is known that ENZDES has a certain bias in recapitulating native residues [Leaver-Fay et al., 2013]. Therefore it is reasonable to assess and compare the bias introduced by ENZDES and MSF:GA:ENZDES. For the assessment of the ENZDES outcome, we selected the 13440 sequences representing the best designs on *BR_EnzBench* and determined $nssr_{SSD}(aa_j)$ values. This distribution represents for all amino acids aa_j

the fraction of similar residues recovered at all design shell positions. Analogously, the distribution $nssr_{MSD}(aa_j)$ was computed that indicates the fraction of similar residues recovered by MSF:GA:ENZDES; for details see Subsection 2.7.3.

The two distributions that are plotted in Fig. 3.13 indicate that the recovery rates are similar and are below the optimal value of 100% for all residues. Generally, sequence recovery for large polar or charged residues is low, which contributes to the weakness of Rosetta to accurately design hydrogen bonds and electrostatics [Stranges and Kuhlman, 2013]. Interestingly, ENZDES is slightly better in recovering polar and charged residues (D, E, H, K, N, R, S), whereas MSF:GA:ENZDES clearly recovers a higher fraction of hydrophobic residues (A, F, I, L, P, V, W, Y).

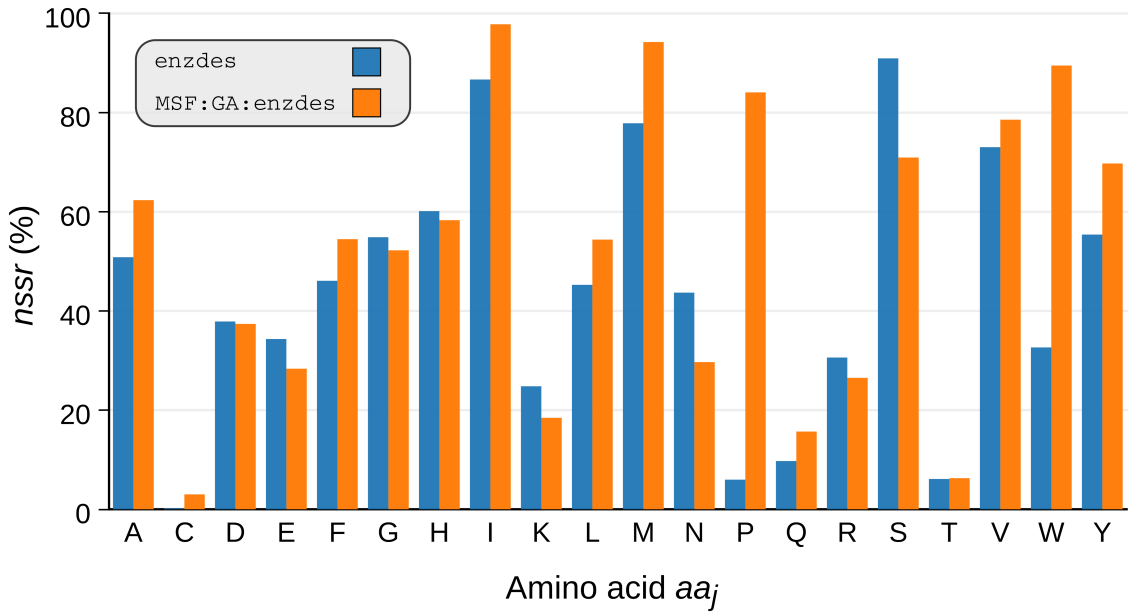


Fig. 3.13 **Recovery of design shell residues from *BR_EnzBench* by means of ENZDES and MSF:GA:ENZDES.** The distributions $nssr_{SSD}(aa_j)$ (blue bars) and $nssr_{MSD}(aa_j)$ (orange bars) represent for each amino acid aa_j the *nssr* value deduced from 13440 design sequences. These were created by ENZDES or MSF:GA:ENZDES for the benchmark proteins *BR_EnzBench*, respectively. *nssr* takes into account the recovery of all residues which are similar to the native aa_j. For details, see Subsection 2.7.3.

This general trend is most evident in the two benchmark proteins with the largest difference in $nssr_{SSD}$ and $nssr_{MSD}$ values: ARL3-GDP (PDB ID 1fzq) is a distinct GTP binding protein [Hillig et al., 2000] from *Mus musculus* and both the ligand and the native binding pocket are considerably polar. Fig. 3.14 A shows that ENZDES

correctly recovered the residues interacting with the guanine group (colored in teal) of GDP, while MSD was less successful. On the other hand, in the glucose binding protein (PDB ID 2b3b) from *Thermus thermophilus* four tryptophan residues provide tight binding to glucose by shape complementarity. Fig. 3.14 B shows that MSF:GA:ENZDES recovered three critical tryptophan residues (colored in teal) in most designs, whereas ENZDES preferred small polar residues that do not provide tight packing. It seems that the representation of a protein by means of an ensemble improves hydrophobic packing but not the formation of polar interaction networks. Their design is considerably more difficult than hydrophobic packing due to the partially covalent nature of a hydrogen bond and the geometric requirements for orientations and distances [Boyken et al., 2016; Leaver-Fay et al., 2013].

3.4 Proof of concept - designing *de novo* retro-aldolase activity

The ultimate proof of concept for any CPD algorithm is the design of functionally active proteins. A reaction that is frequently chosen for enzyme design is the amine-catalyzed retro-aldole cleavage of 4-hydroxy-4-(6-methoxy-2-naphtyl)-2-butanone (methodol) into 6-methoxy-2-naphthaldehyde and acetone [Tanaka et al., 2004]. This multi-step reaction comprises the attack of an active site lysine side-chain on the carbonyl group of the substrate to form a carbinolamine intermediate that subsequently is dehydrated to a protonated Schiff base. The latter is then converted to the reaction products by acid/base chemistry [Fullerton et al., 2006; Heine et al., 2001]. The most active *de novo* retro-aldolase designs have been established on a jelly roll and several $(\beta\alpha)_8$-barrel proteins [Althoff et al., 2012; Bjelic et al., 2014; Jiang et al., 2008]. We chose this reaction as a proof of concept for MSF and selected a previously used thermostable $(\beta\alpha)_8$-barrel scaffold for comparison purposes, namely the indole-3-glycerolphosphate synthase from *S. solfataricus* (ssIGPS).

To implement the MSD approach, we first tested MD simulations as an alternative to the BACKRUB server to generate ensembles that feature higher conformational variability. Next, the multi-state enzyme design procedure is described: We used MD simulations to create a conformational ensemble based on ssIGPS; afterwards, MSF:GA:ENZDES was applied to introduce retro-aldolase activity into the ssIGPS scaffold by considering its ensemble during MSD; finally, an *in silico* stabilization method was utilized to improve the production of stable and soluble protein.

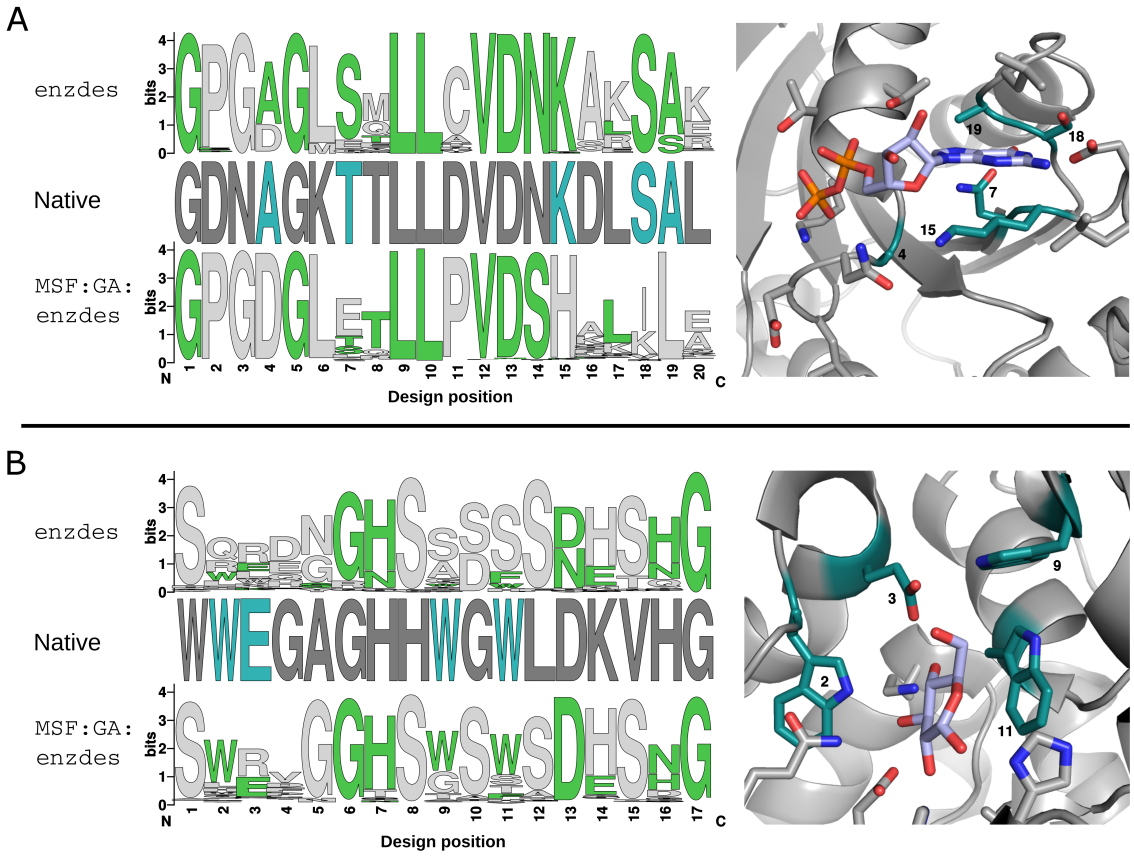


Fig. 3.14 **Recovery of two striking binding pockets by means of ENZDES and MSF:GA:ENZDES.** (**A**) The 3D structure of the binding pocket of ARL3-GDP is shown on the right, the ligand GDP is colored light blue. The residues of the corresponding design positions are shown on the left (labeled "Native"). The sequence logos labeled "ENZDES" and "MSF:GA:ENZDES" represent for each design position the distribution of residues as generated by the corresponding protocols. Residues that are similar to the native ones are colored in green. In the native sequence, residues are colored in teal, if the outcome of the two protocols differs drastically. (**B**) The 3D structure of the binding pocket of the glucose binding protein is shown on the right; the bound glucose is colored light blue. Native residues and sequence logos are shown on the left and were prepared and colored as described for panel (**A**).

3.4.1 Molecular dynamics simulation is well-suited to compute ensembles with higher structural variability

MD simulation is a well-established and reliable method for modeling conformational changes linked to the function of proteins [Klepeis et al., 2009]. Therefore, MD provides an alternative to the Backrub approach for the generation of ensembles to be utilized in MSD. We were interested in assessing the designability of conformations resulting from unconstrained MD simulations of length 10 *ns*. Thus, in analogy to *BR_EnzBench*, we compiled the dataset *MD_EnzBench* consisting of 1000 conformations generated for each of the 16 benchmark apoproteins by means of YASARA as described in Subsection 2.7.1. Note that according to our benchmark protocol, all design shell residues were replaced with alanine prior to design.

To assess the structural variability of *MD_EnzBench* conformations, C_α-RMSD values of design shell residues were determined in a protein-specific all-against-all comparison and then averaged. Analogously, the structural variability of *BR_EnzBench* conformations was determined. Interestingly, the structural variety of the binding pockets generated by the MD simulations is much larger than that generated by the BACKRUB server: The mean RMSD of *MD_EnzBench* is 0.62 Å and that of *BR_EnzBench* is 0.17 Å, which indicates that a 10 *ns* MD simulation generates an ensemble with higher structural diversity than the BACKRUB server. As a control of design performance, the 16×20 $nssr_{BR}$ values of (single) ENZDES designs generated for 20 protein-specific conformations from *BR_EnzBench* were summarized in a boxplot, which had a mean value of 43.88%. To assess the designability of the *MD_EnzBench* conformations, for each of the 1000 protein-specific conformations, one sequence was designed by means of ENZDES and the resulting $nssr_{MD}$ values were averaged protein-wise. Fig. 3.15 shows 100 boxplots each representing 16×10 $nssr_{MD}$ values resulting from ten conformations generated by the MD simulation in a 100 *ps* interval for each of the 16 $prot(k)$. The mean of these $nssr_{MD}$ values is 42.53%, which testifies to a satisfying design performance compared to $nssr_{BR}$ and given that only one sequence was designed per each MD conformation. Moreover, the boxplots indicate that performance did not decrease for conformations generated at later phases of the MD simulation: The median $nssr_{MD}$, and the first and third quartile of the most left and the most right boxplots are 42.10% [35.40%, 45.89%] and 42.24% [34.78%, 50.00%], respectively. In summary, these findings suggest for *de novo* design to consider MD simulations for the generation of ensembles that combine high structural variability and appropriate designability.

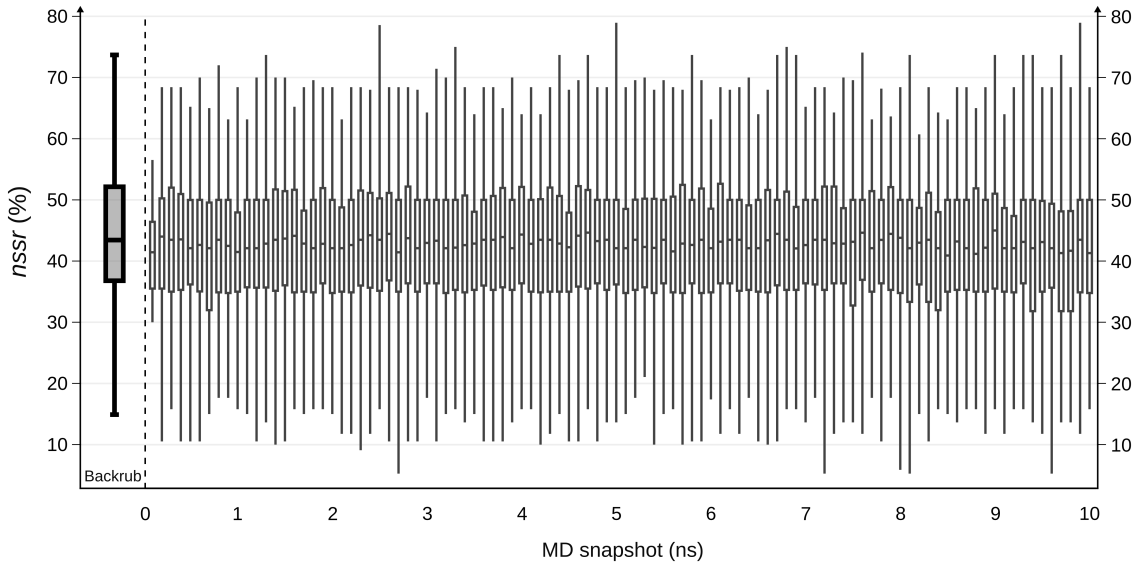


Fig. 3.15 **Performance of ENZDES and MSF:GA:ENZDES on a distinct grouping of conformations.** Each of the 100 boxplots on the right represents 16×10 *nssr* values resulting from ten conformations generated by the MD simulation in a 100 *ps* interval for each of the 16 $prot(k)$. As a control, the 16×20 $nssr_{SSD}^{BR\_EB}$ values of (single) ENZDES designs generated for 20 protein-specific conformations from *BR_EnzBench* were summarized in a boxplot shown on the left (label Backrub). Whiskers indicate the lowest and the highest values of the 1.5 interquartile.

3.4.2 Multi-state design of retro-aldolases

As part of this thesis, a novel design procedure for multi-state enzyme design was developed; a graphical overview with brief explanations is given in Fig. 3.16. In the following, the procedure is described in more detail. The information given here is complemented by additional information in Appendix C.

The structure of the scaffold protein chosen for enzyme design was ssIGPS complexed with its product indole-3-glycerol-phosphate (PDB ID 1a53). In order to create a conformational ensemble of the pure scaffold protein for MSD, the preorganization introduced by ligand-binding should be completely disbanded. The BACKRUB server provides a convenient and well-tested way to create structural ensembles. However, as shown in the previous subsection, the generated structural variability is lower than that of MD simulations. Hence we relied on MD simulations in order to create a conformational ensemble free of a ligand-binding bias.

To begin with, the native ligand was removed from the dataset and the apoprotein was subjected to conformational sampling. Using the protocol validated with *MD_EnzBench*, three individual MD simulations were performed for 10 *ns*. A clustering of MD snapshots based on RMSD values helps to choose near native conformations [Zhang and Skolnick, 2004b]. Thus, we used DURANDAL [Berenger et al., 2012] to cluster the snapshots (conformations) generated with each MD run and picked four conformations from the largest cluster. These 3×4 conformations and the crystal structure of the apoprotein constituted the structural ensemble for the subsequent enzyme design.

Enzyme design usually starts with generating a theozyme, which is a model for the proposed active site that is based upon the geometric constraints dictated by the expected transition state. To design retro-aldolase catalysis, we used a previously designed theozyme containing the carbinolamine reaction intermediate as transition state surrogate covalently bound to the catalytic lysine. In addition, the theozyme contained a glutamate or an aspartate residue to function as general acid/base as well as a serine or a threonine residue to provide additional hydrogen-bonding interactions [Bjelic et al., 2014]. In the next step, we used ROSETTA:MATCH to create in all conformations several thousand matched transition states (*mTS*) with catalytic triads $Lys_i[Asp, Glu]_j[Ser, Thr]_k$ located at markedly different residue positions. A critical step of MSD is the compilation of the ensembles that are used concurrently as states. For enzyme design, ensembles ens_{mTS} of *mTS* are needed and we compiled them the following way: First, *mTS* judged as binding the transition state only weakly were discarded. Second, *mTS* derived from different conformations were

added to the same ens_{mTS}, if identical catalytic triads were located at matching residue positions. Thus, each ens_{mTS} contained a certain number of conformations accommodating the same catalytic triad. Third, the consistency of each ens_{mTS} was assessed by superposing the transition states and by comparing the corresponding conformations. We chose 23 ens_{mTS} consisting of four to 13 conformations (states) and their design and repack shells were defined by merging the output created by ENZDES:AUTODETECT for all conformations.

MSF:GA:ENZDES was executed with each ensemble until convergence; see Section C.1 for details of the protocol. In brief, to assess the designs we compared active-site geometry as well as total and interaction energies and the best 100 variants were subjected to MD simulations of 10 *ns* length. For each variant, we analyzed in detail catalytic site geometries of 100 snapshots (see Section C.2) and nine variants named RA_MSD1 to RA_MSD9 were chosen for biochemical characterization; see Table 3.6. Previous experience had shown that not only the catalytic efficiency but also the conformational stability of initial designs is often poor [Khersonsky et al., 2012]. This is why the sequences are generally further optimized with the help of FOLDIT or other software tools to revert unnecessary mutations back to the native sequence of the scaffold [Bjelic et al., 2014] or, alternatively, by means of directed evolution [Althoff et al., 2012]. We, however, initially did not introduce subsequent stabilizing mutations into the sequences of RA_MSD1 to RA_MSD9 prior to a first experimental characterization. In doing so, we wanted to demonstrate the potential and also the limitations of multi-state designs.

For a comparison of these novel designs with previous ones, we compiled a list of 42 retro-aldolases RA* from the literature (see Subsection C.4) that were also created in the ssIGPS scaffold by means of SSD in Rosetta. These RA* sequences differ on average at 15 positions from the native ssIGPS sequence; in contrast, the nine RA_MSD* (see Section C.2) sequences contain on average 21 amino acid substitutions. Moreover, RA* sequences deviate on average from RA_MSD* sequences at 24 positions and 18 substitutions distinguish the most similar pairs of variants (RA41 versus RA_MSD9 and RA90 versus RA_MSD8). Even the two designs (RA114 versus RA_MSD1) that share the same catalytic residues K210 and S110 differ at 25 positions. Although we utilized the same transition state and the same scaffold that was used for the design of RA114 - RA120 [Bjelic et al., 2014], our MSD approach has generated a set of entirely novel catalytic sites located in the same shell as used for previous designs, compare Fig. 3.17.

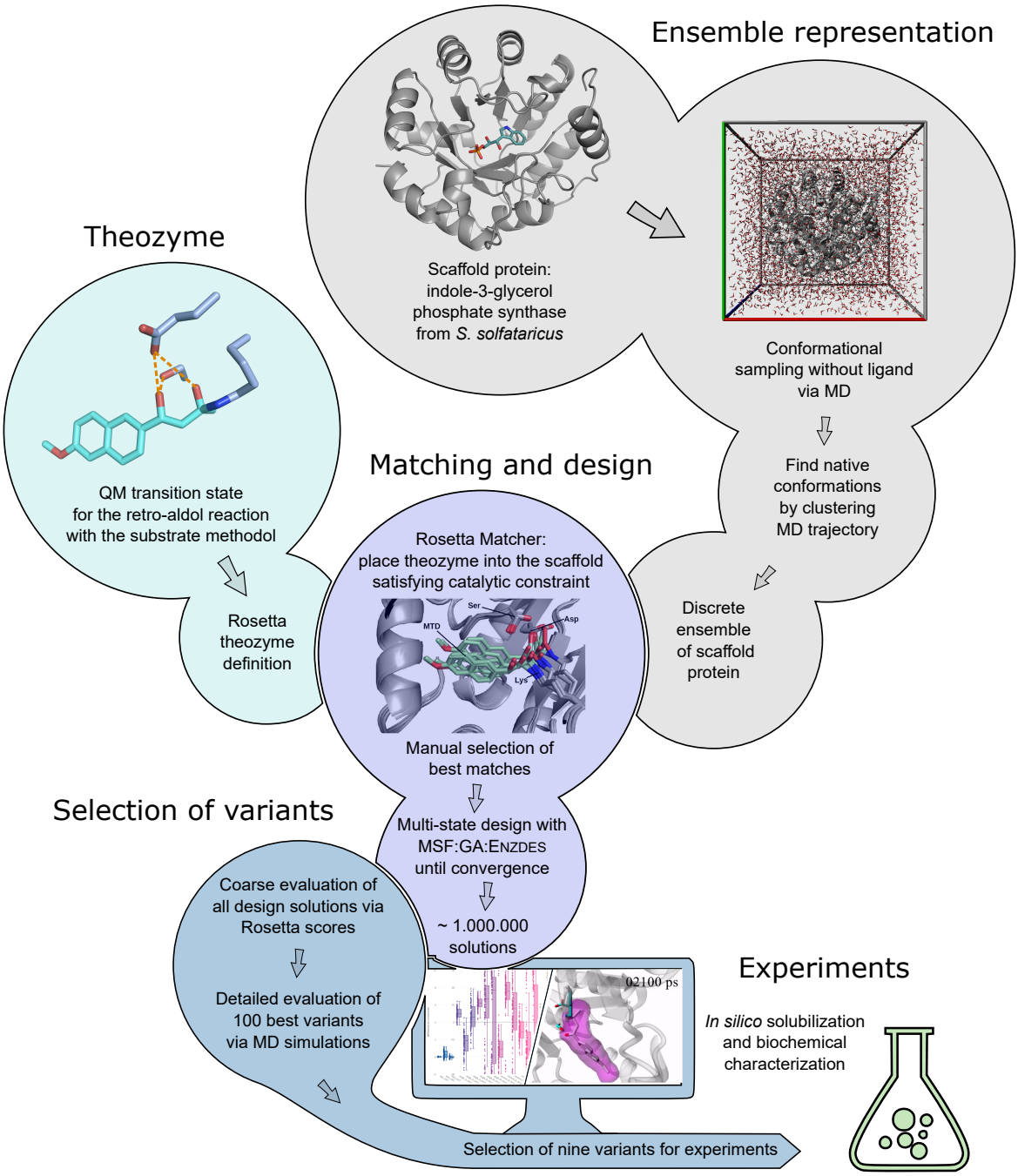


Fig. 3.16 **Overview of the retro-aldolase design process.** (**Ensemble representation**) The scaffold protein was sampled via MD; the MD trajectory was clustered and snapshots were extracted to generate an ensemble representation. (**Theozyme**) The theozyme for the retro-aldol reaction was defined based on previous work [Bjelic et al., 2014]. (**Matching and design**) ROSETTA:MATCH was utilized to place the theozyme into the scaffold ensemble. The best matches were filtered and manually selected; multi-state design was performed using MSF:GA:ENZDES. (**Selection of variants**) A coarse evaluation based on Rosetta scores and a more detailed evaluation via MD were performed to choose variants for biochemical experiments. (**Experiments**) Retro-aldolase activity was verified by *in vitro* experiments. The solubility of one variant was further increased using the PROSS server.

Table 3.6 **MSD proteins and their retro-aldolase activity.** The catalytic triad designed for nine proteins (RA_MSD1 - RA_MSD9) is specified in the second column. The third column gives the number of residue exchanges compared to the native sequence of ssIGPS. The third column lists the conversion rates (rate of product formation divided by the enzyme concentration) in the presence of 500 µM S-methodol, and the last column the catalytic efficiency k_{cat}/K_M as determined from the linear part of the substrate saturation curves. ND: not determined.

Name	Catalytic triad	Number of exchanges	Conversion rate (s^{-1})	k_{cat}/K_M ($M^{-1}s^{-1}$)
RA_MSD1	K210/D131/S110	21	8.08×10^{-7}	ND
RA_MSD2	K210/D131/S110	22	3.14×10^{-7}	ND
RA_MSD2.4	K210/D131/S110	26	1.23×10^{-6}	ND
RA_MSD2.5	K210/D131/S110	29	1.49×10^{-6}	ND
RA_MSD3	K210/D131/S110	22	2.60×10^{-6}	ND
RA_MSD4	K51/E53/S83	20	3.03×10^{-6}	ND
RA_MSD5	K51/E53/S83	21	1.69×10^{-5}	3.47×10^{-2}
RA_MSD6	K231/E53/S83	25	2.82×10^{-6}	ND
RA_MSD7	K231/E131/T159	18	8.33×10^{-6}	1.41×10^{-2}
RA_MSD8	K231/E131/T159	18	5.61×10^{-6}	ND
RA_MSD9	K231/E53/T83	19	7.55×10^{-7}	ND

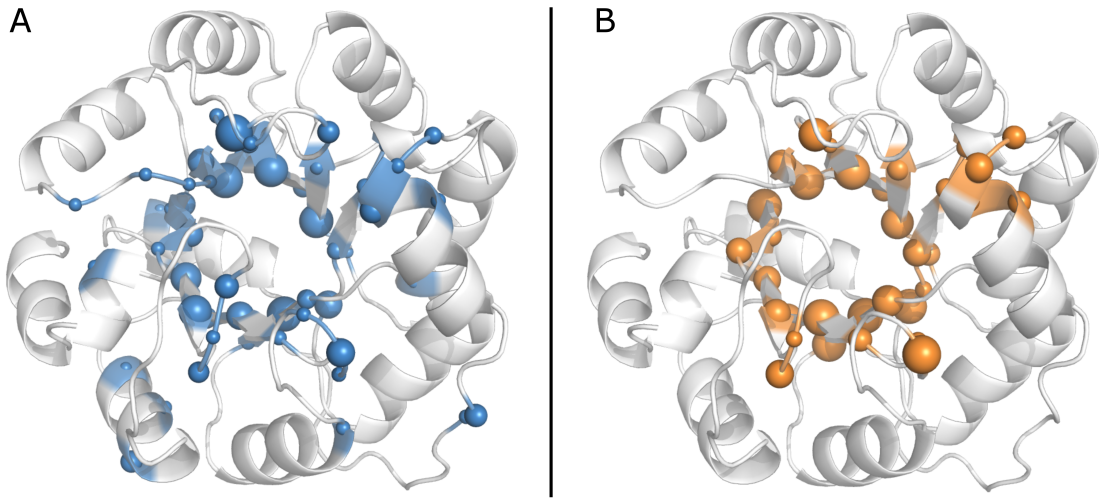


Fig. 3.17 **Mutations introduced into the ssIGPS scaffold to design retro-aldolase activity.** (**A**) An overview of all mutations introduced in 42 previous designs subsumed in the set RA* (sequences listed in Subsection C.4). Blue spheres indicate residue positions and sphere diameters are proportional to the frequency of the mutations in comparison to the native ssIGPS sequence. (**B**) Ditto, for nine RA_MSD* designs, mutations are visualized by means of orange spheres.

3.4.3 All initial multi-state designs possess activity but need further processing to improve solubility

The genes for RA_MSD1-RA_MSD9 were synthesized and expressed in *E. coli* as fusion constructs with the gene for the maltose binding protein (MBP). The fusion proteins were purified with metal chelate affinity chromatography via their N-terminal hexa-histidine tags, resulting in high yields (50-150 $mg\ protein/l\ expression\ culture$). RA_MSD5 could be produced in soluble form also without MBP, whereas the other designs precipitated in the absence of the solubility enhancer. All designs showed modest catalytic activity with low substrate affinity, leading to conversion rates for 500 µM S-methodol ranging from only 3×10^{-7} to 1.7×10^{-5} s^{-1} (see Table 3.6). For the best designs, RA_MSD5 and RA_MSD7, the linear part of the substrate saturation curve was used to determine k_{cat}/K_M values of 3.47×10^{-2} and 1.41×10^{-2} $M_{-1}s_{-1}$, which are similar to the values obtained for RA114-RA120 [Bjelic et al., 2014]. Moreover, the RA_MSD5 designs with and without MBP displayed virtually the same k_{cat}/K_M values, excluding an influence of the solubility enhancer on activity. All experimental work was done by the biochemists in the lab of Prof. Dr. Reinhard Sterner.

Due to the intentionally omitted step of secondary protein stabilization following the initial design process, eight of our nine designs were insoluble without MBP. We wanted to test whether protein stabilization would result in higher activity. Accordingly, we attempted to improve the stability of RA_MSD2, which has the lowest activity of all designs (see Table 3.6), by using the fully automated *in silico* method offered by the PROSS webserver [Goldenzweig et al., 2016]. The six conformations constituting the ensemble ens_{mTS} used to design RA_MSD2 were individually submitted to PROSS and the corresponding output sets that contained 6 to 21 stabilizing mutations were merged to five consensus sequences (see Section C.3). Variants RA_MSD2.4 and RA_MSD2.5 that contained the highest number of stabilizing mutations, could be produced in soluble form without MBP and were purified with high yield (about 25 *mg protein/l expression culture*).

Software protocols for *in silico* protein stabilization are known to include false positives when choosing stabilizing mutations [Magliery, 2015]. Thus, a single destabilizing mutation could undo the stabilizing effects of other mutations. This is why in previous studies, the number of mutations simultaneously introduced into the scaffold had generally been kept low and different types of stabilizing mutations had been tested. For the stabilization of RA_MSD2 via PROSS, we pursued a similar strategy of varying the number of stabilizing mutations and their location (see Fig. 3.18 and Fig. C.3). Interestingly, the variants with the highest number of introduced mutations (RA_MSD2.4 and RA_MSD2.5) showed the highest effect of stabilization, while those with the lowest number of stabilizing mutations (RA_MSD2.1 and RA_MSD2.2) were still completely insoluble without MBP. Although the solubility of variants RA_MSD2.4 and RA_MSD2.5 was strongly increased, their melting temperatures were unaffected by this stabilization; however, melting temperatures were already at a very high level due to selecting a thermostable scaffold (T_m around 80°C [Andreotti et al., 1997]). Also, activity measurements showed that mutations did not drastically improve the conversion rate of RA_MSD2 (see Table 3.6). We therefore can confirm that the PROSS server is able to predict stabilizing mutations at a low false-positive rate. In addition, we could not find drastic changes in activity, indicating that those mutations did not introduce dramatic structural changes.

In summary, our results showed that MSD (based on a structural ensemble) is comparably successful as SSD (based on a single structure) for establishing retro-aldolase activity on a thermostable $(\beta\alpha)_8$-barrel scaffold, indicating that this particular reaction requires only a limited degree of conformational flexibility. However, catalysis is often linked to conformational transitions, which can only be captured by

Table 3.7 **Stabilizing mutations predicted by PROSS that were introduced intro RA_MSD2 variants.** From RA_MSD2.1 to RA_MSD2.5, stabilizing mutations are marked in bold when appearing for the first time. Colored circles correspond to the sphere colors of Fig. 3.18 indicating the location in the scaffold protein.

Name		Mutations
RA_MSD2.1	●	**N34K S102A S117Y N164E N190D N204D N228H L248E** (8)
RA_MSD2.2	●	**S70A N161H C178G D180N I189V S234E** (6)
RA_MSD2.3		N34K S70A S102A S117Y N161H N164E I189V N190D N204D N228H S234E L248E (12)
RA_MSD2.4	●	**P8W M9L** N34K S70A **Y89F** S102A **S108L** S117Y N161H N164E C178G N190D N204D N228H S234E L248E (16)
RA_MSD2.5	●	P8W M9L **Q14E L15I S25E** N34K S70A Y89F S102A S108L S117Y N161H N164E C178G D180N I189V N190D N204D N228H S234E L248E (21)

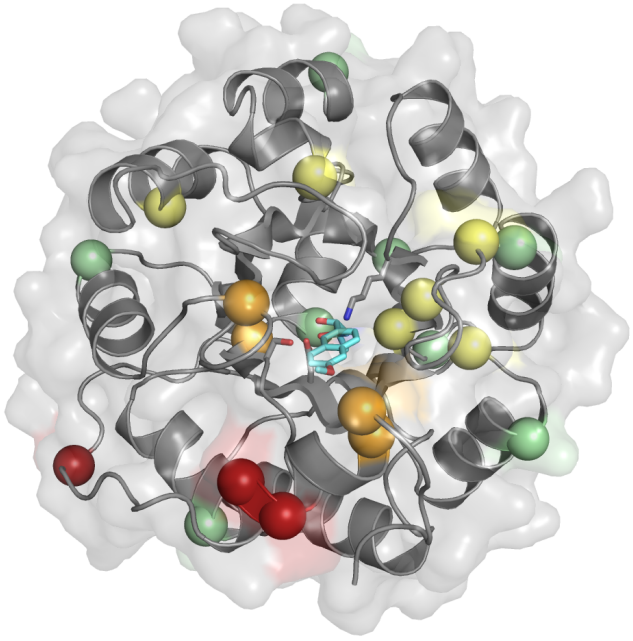


Fig. 3.18 **Location of PROSS mutations used to stabilize RA_MSD2.** Spheres mark positions proposed by PROSS and sphere colors correspond to the colored circles in Table 3.7. Stabilizing mutations for RA_MSD2.1 were selected on the protein surface (green); those for RA_MSD2.2 in the protein core but outside of the catalytic site (yellow). RA_MSD2.3 contained the union of the mutations from RA_MSD2.1 and RA_MSD2.2. For RA_MSD2.4 (orange) and RA_MSD2.5 (red), more stabilizing mutations were added irrespective of their distance from the active site.

MSD approaches. Moreover, in contrast to SSD, MSD offers a broader functionality and is also suited for more challenging tasks like negative design. The combination of positive and negative design offers the stabilization of desired conformations while simultaneously impairing unwanted conformations. This is particularly interesting for designing protein-ligand or protein-protein interaction specificities such as different oligomerization states [Davey and Chica, 2012].

Chapter 4

Outlook

The advent of ensemble methods has found its way into many scientific fields, e.g. weather prediction: Instead of generating a single prediction for the available input data, supercomputers now generate hundreds of predictions using different models to address multiple sources of forecast uncertainty [Bauer et al., 2015; Gneiting and Raftery, 2005]. In analogy, our framework ROSETTA:MSF allows a protein designer to address the structural uncertainty of a protein scaffold by considering an ensemble of structures instead of a single rigid model. In Subsection 3.3.2 we could show that this improves design accuracy. Unfortunately, due to the computational intensity of our algorithm, ensembles with only less than a dozen of models could be considered. On the other hand, generating structural data by simulation is relatively fast and cheap compared to CPD in terms of computing time. It would thus be desirable to consider a much bigger ensemble, which requires a computational speedup.

Relying on a genetic algorithm as its optimization routine is the weakest point of ROSETTA:MSF, leading to slow convergence times and a high consumption of computational resources. However, our implementation is modular and designed to easily integrate other optimization routines, if those are made available in Rosetta. An optimizer of preference supporting MSD is particle swarm optimization [Kennedy, 2011], which converges faster than a genetic algorithm thanks to the particle movement in the direction of the gradient. Although ROSETTA:MSF is an improvement compared to the default application, the accuracy of *de novo* ligand-binding design reached in our benchmark dataset is still rather low due to the inevitable simplification and lack of an accurate physical model in CPD. Fortunately, our recent efforts of integrating ROSETTA:MSF into the codebase of Rosetta, which is maintained by the community, will allow the framework to take part in all general improvements implemented in the future.

We applied ROSETTA:MSF to computationally design an enzyme. To our knowledge, this is the first *de novo* enzyme design approach that uses an ensemble of structures via multi-state methods. With our approach, we could not observe a higher catalytic activity than that described in previous works for single-state design methods. However, there may be a good reason: In a recently published work [Obexer et al., 2016], the authors evolved retro-aldolase activity on the same scaffold as in this work by fluorescence-activated droplet sorting, a technique allowing high-throughput screening. Their best evolved variants reached catalytic activity comparable to natural enzymes. Although CPD and experimental screening methods have different scenarios of use, the results offered an interesting fact: The crystal structure of the best evolved variant complexed with an inhibitor similar to the substrate was barely different to the apo structure of the evolved protein, suggesting negligible conformational flexibility in the binding site. Apparently, establishing native-like retro-aldolase activity does not require conformational dynamics. However, this can by no means be generalized and future enzyme design approaches for more complex reaction types will require the consideration of protein dynamics [Mukherjee and Gupta, 2015].

Still, computational enzyme design remains one of the most complex *de novo* design problems to be solved and most enzymes obtained this way are quite slow catalysts compared to native enzymes. Nevertheless, the era of *de novo* design has come. As mentioned in Section 1.1.2, the number of folds explored by evolution is very small and protein evolution has most probably sampled only a tiny fraction of the gigantic sequence space available to proteins [Huang et al., 2016]. Protein designers have already begun to explore this unknown space, by generating proteins with no sequence similarity to natural ones [Harbury et al., 1998; Walsh et al., 1999] as well as protein structures unseen in Nature [Brunette et al., 2015; Kuhlman et al., 2003]. This is not a coincidence: The fundamentals of protein design have been understood better and better in the last years. Recently, the atomically-accurate design of hydrogen bonds [Boyken et al., 2016], idealized folds [Lin et al., 2015], self-assembling oligomers with precisely defined PPIs [Gonen et al., 2015; King et al., 2012], repeat proteins [Doyle et al., 2015] as well as the recombination of pieces of existing proteins [Jacobs et al., 2016] have been described. Further understanding the fundamentals of protein folding, dynamics, and biophysics will enable to design from ground up a world of customized proteins. Such a powerful tool could help facing important challenges in the future, such as clean energy, customized drugs and materials as well as biosensors.

Abbreviations

Acronyms

AS	Anthranilate synthase
CAA	Canonical amino acid
CH	Chorismate
CPD	Computational protein design
DAF	Dynamic aggregate function
DTE	1,2-dithienylethene
EHM	Equilibrated homology model
GA	Genetic algorithm
HM	Homology model
ICS	Isochorismate synthase
IC	Isochorismate
MC	Monte Carlo method
MBP	Maltose-binding protein
MD	Molecular dynamics
MSA	Multiple sequence alignment
MSD	Multi-state design
MSF	Multi-state framework
MM	Molecular mechanics
NMR	Nuclear magnetic resonance
PLD	Protein-ligand docking
PNT	Putative nucleophile trajectory
PPI	Protein-protein interface
PRFAR	N′-[(5′-phosphoribulosyl)formimino] -5-aminoimidazole-4-carboxamide ribonucleotide

ProFAR	N′-[(5′-phosphoribosyl)formimino] -5-aminoimidazole-4-carboxamide ribonucleotide
pum	Pumilio protein
RMSD	Root mean square deviation
rotamer	Amino acid rotational isomer
SSD	Single-state design
TS	Transition state
mTS	Matched transition state
vdW	Van der Waals

Nomenclature

dmBrat	Brain tumor NHL-domain from *Drosophila melanogaster*
hIFABP	Human intestinal fatty acid binding protein
mtPknD	Serine/threonine protein kinase PknD from *Mycobacterium tuberculosis*
mtPriA	Phosphorybosyl isomerase A from *Mycobacterium tuberculosis*
rnPAL	Peptidyl-α-hydroxyglycine α-amidating lyase from *Rattus norvegicus*
ssIGPS	Indole-3-glycerolphosphate synthase from *Sulfolobus solfataricus*
stAS	Anthranilate synthase from *Salmonella typhimurium*
stTrpE	TrpE subunit from *Salmonella typhimurium*
stTrpG	TrpG subunit from *Salmonella typhimurium*

Units

fs	Femtosecond
K	Kelvin
k_{cat}	Turnover number
$kcal$	Kilocalorie
kJ	Kilojoule
K_M	Michaelis-Menten constant

ns	Nanosecond
nsr	Native sequence recovery
nssr	Native sequence similarity recovery
ps	Picosecond
REU	Rosetta energy unit
ts	Rosetta total score

Bibliography

Agarwal, P. K. (2006). Enzymes: An integrated view of structure, dynamics and function. *Microb Cell Fact*, 5(1):2.

Allen, B. D., Nisthal, A., and Mayo, S. L. (2010). Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci U S A*, 107(46):19838–19843.

Allert, M., Rizk, S. S., Looger, L. L., and Hellinga, H. W. (2004). Computational design of receptors for an organophosphate surrogate of the nerve agent Soman. *Proc Natl Acad Sci U S A*, 101(21):7907–7912.

Althoff, E. A., Wang, L., Jiang, L., Giger, L., Lassila, J. K., Wang, Z., Smith, M., Hari, S., Kast, P., Herschlag, D., et al. (2012). Robust design and optimization of retroaldol enzymes. *Protein Sci*, 21(5):717–726.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Andreotti, G., Cubellis, M. V., Michela, D., Fessas, D., Sannia, G., and Marino, G. (1997). Stability of a thermophilic TIM-barrel enzyme: Indole-3-glycerol phosphate synthase from the thermophilic archaeon *Sulfolobus solfataricus*. *Biochem J*, 323(1):259–264.

Ansari, A., Berendzen, J., Bowne, S. F., Frauenfelder, H., Iben, I., Sauke, T. B., Shyamsunder, E., and Young, R. D. (1985). Protein states and proteinquakes. *Proc Natl Acad Sci U S A*, 82(15):5000–5004.

Arama, E., Dickman, D., Kimchie, Z., Shearn, A., and Lev, Z. (2000). Mutations in the β-propeller domain of the *Drosophila* brain tumor (Brat) protein induce neoplasm in the larval brain. *Oncogene*, 19(33):3706.

Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J., Stoddard, B. L., and Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

Berenger, F., Shrestha, R., Zhou, Y., Simoncini, D., and Zhang, K. Y. (2012). Durandal: Fast exact clustering of protein decoys. *J Comput Chem*, 33(4):471–474.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1):235–242.

Bienstock, R. J. (2015). Solvation methods for protein–ligand docking. *Methods Mol Biol*, 1289:3–12.

Bjelic, S., Kipnis, Y., Wang, L., Pianowski, Z., Vorobiev, S., Su, M., Seetharaman, J., Xiao, R., Kornhaber, G., Hunt, J. F., et al. (2014). Exploration of alternate catalytic mechanisms and optimization strategies for retroaldolase design. *J Mol Biol*, 426(1):256–271.

Boas, F. E. and Harbury, P. B. (2007). Potential energy functions for protein design. *Curr Opin Struct Biol*, 17(2):199–204.

Bousema, T., Okell, L., Felger, I., and Drakeley, C. (2014). Asymptomatic malaria infections: Detectability, transmissibility and public health relevance. *Nat Rev Microbiol*, 12(12):833–840.

Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170.

Boyken, S. E., Chen, Z., Groves, B., Langan, R. A., Oberdorfer, G., Ford, A., Gilmore, J. M., Xu, C., DiMaio, F., Pereira, J. H., et al. (2016). *De novo* design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science*, 352(6286):680–687.

Brieke, C., Rohrbach, F., Gottschalk, A., Mayer, G., and Heckel, A. (2012). Light-controlled tools. *Angew Chem Int Ed Engl*, 51(34):8446–8476.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*, 4(2):187–217.

Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. (2015). Exploring the repeat protein universe through computational protein design. *Nature*, 528(7583):580–584.

Burgen, A. (1981). Conformational changes and drug action. *Fed Proc*, 40(13):2723.

Canutescu, A. A. and Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci*, 12(5):963–972.

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014.

Cho, A. E., Guallar, V., Berne, B. J., and Friesner, R. (2005). Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J Comput Chem*, 26(9):915–931.

Cohen, N., Dagan, T., Stone, L., and Graur, D. (2005). GC composition of the human genome: In search of isochores. *Mol Biol Evol*, 22(5):1260–1272.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Res*, 14(6):1188–1190.

Cutter, A. R. and Hayes, J. J. (2015). A brief review of nucleosome structure. *FEBS lett*, 589(20PartA):2914–2922.

D'Abramo, M., Rabal, O., Oyarzabal, J., and Gervasio, F. L. (2012). Conformational selection versus induced fit in kinases: The case of PI3K-γ. *Angew Chem Int Ed Engl*, 51(3):642–646.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J Mol Biol*, 332(2):449–460.

Davey, J. A. and Chica, R. A. (2012). Multistate approaches in computational protein design. *Protein Sci*, 21(9):1241–1252.

Davis, I. W., Arendall, W. B., Richardson, D. C., and Richardson, J. S. (2006). The Back-rub motion: How protein backbone shrugs when a sidechain dances. *Structure*, 14(2):265–274.

Davis, I. W. and Baker, D. (2009). RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol*, 385(2):381–392.

DeLano, W. L. (2008). The PyMOL molecular graphics system. *Schrödinger: PyMOL Schrödinger Inc*.

Desmet, J., De Maeyer, M., et al. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539.

Dewar, M. J., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. (1985). Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. *J Am Chem Soc*, 107(13):3902–3909.

Dill, K. A., Chan, H. S., et al. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol*, 4(1):10–19.

Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L., and Bradley, P. (2015). Rational design of α-helical tandem repeat proteins with closed architectures. *Nature*, 528(7583):585–588.

Edwards, T. A., Wilkinson, B. D., Wharton, R. P., and Aggarwal, A. K. (2003). Model of the brain tumor – Pumilio translation repressor complex. *Genes Dev*, 17(20):2508–2513.

Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 277:396–404.

Epstein, C. J., Goldberger, R. F., and Anfinsen, C. B. (1963). The genetic control of tertiary protein structure: Studies with model systems. *Cold Spring Harb Symp Quant Biol*, 28:439–449.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J Chem Phys*, 103(19):8577–8593.

Fenwick, R. B., van den Bedem, H., Fraser, J. S., and Wright, P. E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc Natl Acad Sci U S A*, 111(4):E445–E454.

Ferreiro, D. U., Komives, E. A., and Wolynes, P. G. (2014). Frustration in biomolecules. *Q Rev Biophys*, 47(04):285–363.

Firn, R. D. and Jones, C. G. (2000). The evolution of secondary metabolism – a unifying model. *Mol Microbiol*, 37(5):989–994.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., Wilson, I. A., and Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). SCOPe: Structural classification of protein – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*, 42(D1):D304–D309.

Fullerton, S. W., Griffiths, J. S., Merkel, A. B., Cheriyan, M., Wymer, N. J., Hutchins, M. J., Fierke, C. A., Toone, E. J., and Naismith, J. H. (2006). Mechanism of the Class I KDPG aldolase. *Bioorg Med Chem*, 14(9):3002–3010.

Gainza, P., Roberts, K. E., Georgiev, I., Lilien, R. H., Keedy, D. A., Chen, C.-Y., Reza, F., Anderson, A. C., Richardson, D. C., Richardson, J. S., et al. (2013). OSPREY: Protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol*, 523:87.

Garcia-Viloca, M., Gao, J., Karplus, M., and Truhlar, D. G. (2004). How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303(5655):186–195.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906).

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., et al. (2016). Automated structure-and sequence-based design of proteins for high bacterial expression and stability. *Mol Cell*, 63(2):337–346.

Gonen, S., DiMaio, F., Gonen, T., and Baker, D. (2015). Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science*, 348(6241):1365–1368.

Govindarajan, S., Recabarren, R., and Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins*, 35(4):408–414.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J Mol Biol*, 153(4):1027–1042.

Greul, J. N., Kleban, M., Schneider, B., Picasso, S., and Jäger, V. (2001). Amino (hydroxymethyl) cyclopentanetriols, an emerging class of potent glycosidase inhibitors – Part II: Synthesis, evaluation, and optimization of β-D-galactopyranoside analogues. *ChemBioChem*, 2(5):368–370.

Gropp, W., Lusk, E., Doss, N., and Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Comput*, 22(6):789–828.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol*, 320(2):369–387.

Hamilton, W. L., Claessens, A., Otto, T. D., Kekre, M., Fairhurst, R. M., Rayner, J. C., and Kwiatkowski, D. (2016). Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Res*, 45(4):1889–1901.

Hammes, G. G., Chang, Y.-C., and Oas, T. G. (2009). Conformational selection or induced fit: A flux description of reaction mechanism. *Proc Natl Acad Sci U S A*, 106(33):13737–13741.

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., and Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, 282(5393):1462–1467.

Hasegawa, H. and Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol*, 19(3):341–348.

Havranek, J. J., Duarte, C. M., and Baker, D. (2004). A simple physical model for the prediction and design of protein–DNA interactions. *J Mol Biol*, 344(1):59–70.

Heine, A., DeSantis, G., Luz, J. G., Mitchell, M., Wong, C.-H., and Wilson, I. A. (2001). Observation of covalent intermediates in an enzyme mechanism at atomic resolution. *Science*, 294(5541):369–374.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*, 4(3):435–447.

Hill, R. B., Raleigh, D. P., Lombardi, A., and DeGrado, W. F. (2000). *De novo* design of helical bundles as models for understanding protein folding and function. *Acc Chem Res*, 33(11):745–754.

Hillig, R. C., Hanzal-Bayer, M., Linari, M., Becker, J., Wittinghofer, A., and Renault, L. (2000). Structural and biochemical properties show ARL3-GDP as a distinct GTP binding protein. *Structure*, 8(12):1239–1245.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138.

Hu, X. and Kuhlman, B. (2006). Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins*, 62(3):739–748.

Huang, P.-S., Boyken, S. E., and Baker, D. (2016). The coming of age of *de novo* protein design. *Nature*, 537(7620):320–327.

Huang, S.-Y. and Zou, X. (2010). Advances and challenges in protein-ligand docking. *Int J Mol Sci*, 11(8):3016–3034.

Hubbard, R. E. and Kamran Haider, M. (2010). Hydrogen bonds in proteins: Role and strength. *eLS*.

Humphris, E. L. and Kortemme, T. (2007). Design of multi-specificity in protein interfaces. *PLoS Comput Biol*, 3(8):e164.

Huson, D. H. and Bryant, D. (2008). Estimating phylogenetic trees and networks using SplitsTree 4. *Software and information available at http://www.splitstree.org*.

Jacobs, T., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J., Szyperski, T., and Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. *Science*, 352(6286):687–690.

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem*, 23(16):1623–1641.

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., et al. (2008). *De novo* computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A*, 32(5):922–923.

Kafri, M., Metzl-Raz, E., Jona, G., and Barkai, N. (2016). The cost of protein production. *Cell Rep*, 14(1):22–31.

Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, 9(9):646–652.

Katoh, K. and Frith, M. C. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, 28(23):3144–3146.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780.

Kennedy, J. (2011). Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. New York: Springer.

Khersonsky, O., Kiss, G., Röthlisberger, D., Dym, O., Albeck, S., Houk, K. N., Baker, D., and Tawfik, D. S. (2012). Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci U S A*, 109(26):10358–10363.

King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., André, I., Gonen, T., Yeates, T. O., and Baker, D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science*, 336(6085):1171–1174.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov*, 3(11):935–949.

Kleban, M., Hilgers, P., Greul, J. N., Kugler, R. D., Li, J., Picasso, S., Vogel, P., and Jäger, V. (2001). Amino (hydroxymethyl) cyclopentanetriols, an emerging class of potent glycosidase inhibitors – Part I: Synthesis and evaluation of β-D-pyranoside analogues in the manno, gluco, galacto, and GlcNAc series. *ChemBioChem*, 2(5):365–368.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol*, 19(2):120–127.

Koshland, D. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104.

Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A., and Vriend, G. (2004). Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins*, 57(4):678–683.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins*, 77(S9):114–122.

Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins*, 47(3):393–402.

Krieger, E., Nielsen, J. E., Spronk, C. A., and Vriend, G. (2006). Fast empirical pKa prediction by Ewald summation. *J Mol Graph Model*, 25(4):481–486.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res*, 19(9):1639–1645.

Kubelka, J., Hofrichter, J., and Eaton, W. A. (2004). The protein folding "speed limit". *Curr Opin Struct Biol*, 14(1):76–88.

Kufareva, I. and Abagyan, R. (2012). Methods of protein structure comparison. *Methods Mol Biol*, 857:231–257.

Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A*, 97(19):10383–10388.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368.

Lane, T. J., Shukla, D., Beauchamp, K. A., and Pande, V. S. (2013). To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr Opin Struct Biol*, 23(1):58–65.

Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., and Kortemme, T. (2010). RosettaBackrub – a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res*, 38(suppl 2):W569–W575.

Leaver-Fay, A., Jacak, R., Stranges, P. B., and Kuhlman, B. (2011a). A generic program for multistate protein design. *PLoS One*, 6(7):e20937.

Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., et al. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol*, 523:109.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., et al. (2011b). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, 487:545.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226(2):507–533.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput Phys Commun*, 91(1-3):215–231.

Lewis, S. M. and Kuhlman, B. A. (2011). Anchored design of protein-protein interfaces. *PLoS One*, 6(6):e20872.

Liebherr, R. B. and Gorris, H. H. (2014). Enzyme molecules in solitary confinement. *Molecules*, 19(9):14417–14445.

Lin, Y.-R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., and Baker, D. (2015). Control over overall shape and size in *de novo* designed proteins. *Proc Natl Acad Sci U S A*, 112(40):E5478–E5485.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular cell biology, 4th edition*. New York: W.H. Freeman and Co.

Loedige, I., Stotz, M., Qamar, S., Kramer, K., Hennig, J., Schubert, T., Löffler, P., Längst, G., Merkl, R., Urlaub, H., et al. (2014). The NHL domain of Brat is an RNA-binding domain that directly contacts the hunchback mRNA for regulation. *Genes Dev*, 28(7):749–764.

Luger, K., Rechsteiner, T. J., Flaus, A. J., Waye, M. M., and Richmond, T. J. (1997). Characterization of nucleosome core particles containing histone proteins made in bacteria. *J Mol Biol*, 272(3):301–311.

Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol*, 33:161–168.

Malakauskas, S. M. and Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Mol Biol*, 5(6):470–475.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nat Biotechnol*, 30(11):1072–1080.

Marvin, J. S. and Hellinga, H. W. (2001). Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc Natl Acad Sci U S A*, 98(9):4955–4960.

Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., and Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, 7(1):1.

Morollo, A. A. and Eck, M. J. (2001). Structure of the cooperative allosteric anthranilate synthase from *Salmonella typhimurium*. *Nat Struct Mol Biol*, 8(3):243–247.

Mukherjee, J. and Gupta, M. N. (2015). Increasing importance of protein flexibility in designing biocatalytic processes. *Biotechnol Rep*, 6:119–123.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

Newton, I. (1999). *The Principia: mathematical principles of natural philosophy*. Oakland: Univ of California Press.

Nivón, L. G., Bjelic, S., King, C., and Baker, D. (2014). Automating human intuition for protein design. *Proteins*, 82(5):858–866.

Obexer, R., Godina, A., Garrabou, X., Mittl, P. R., Baker, D., Griffiths, A. D., and Hilvert, D. (2016). Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat Chem*.

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N. V., and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*, 4:e09248.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298.

Patil, R., Laguerre, A., Wielens, J., Headey, S. J., Williams, M. L., Hughes, M. L., Mohanty, B., Porter, C. J., and Scanlon, M. J. (2014). Characterization of two distinct modes of drug binding to human intestinal fatty acid binding protein. *ACS Chem Biol*, 9(11):2526–2534.

Pauling, L. (1946). Molecular architecture and biological reactions. *Chem Eng News*, 24(10):1375–1377.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612.

Plach, M. G., Löffler, P., Merkl, R., and Sterner, R. (2015). Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes. *Angew Chem Int Ed Engl*, 54(38):11270–11274.

Procko, E., Berguig, G. Y., Shen, B. W., Song, Y., Frayo, S., Convertine, A. J., Margineantu, D., Booth, G., Correia, B. E., Cheng, Y., et al. (2014). A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell*, 157(7):1644–1656.

Reisinger, B., Kuzmanovic, N., Löffler, P., Merkl, R., König, B., and Sterner, R. (2014). Exploiting protein symmetry to design light-controllable enzyme inhibitors. *Angew Chem Int Ed Engl*, 53(2):595–598.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The european molecular biology open software suite. *Trends Genet*, 16(6):276–277.

Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. (2011). *De novo* enzyme design using Rosetta3. *PLoS One*, 6(5):e19230.

Roberts, K. E., Gainza, P., Hallen, M. A., and Donald, B. R. (2015). Fast gap-free enumeration of conformations and sequences for protein design. *Proteins*, 83(10):1859–1877.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol*, 383:66–93.

Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195.

Sadowski, M. I. and Taylor, W. R. (2009). Protein structures, folds and fold spaces. *J Phys Condens Matter*, 22(3):033103.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

Schneider, M., Fu, X., and Keating, A. E. (2009). X-ray vs. NMR structures as templates for computational protein design. *Proteins*, 77(1):97–110.

Sehnal, D., Vařeková, R. S., Berka, K., Pravda, L., Navrátilová, V., Banáš, P., Ionescu, C.-M., Otyepka, M., and Koča, J. (2013). MOLE 2.0: Advanced approach for analysis of biomacromolecular channels. *J Cheminform*, 5(1):1.

Sevy, A. M., Jacobs, T. M., Crowe Jr, J. E., and Meiler, J. (2015). Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences. *PLoS Comput Biol*, 11(7):e1004300.

Shah, P. S., Hom, G. K., Ross, S. A., Lassila, J. K., Crowhurst, K. A., and Mayo, S. L. (2007). Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol*, 372(1):1–6.

Shapovalov, M. V. and Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858.

Shen, M.-Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–2524.

Shifman, J. M. and Mayo, S. L. (2003). Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci U S A*, 100(23):13274–13279.

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747.

Silberhorn, E., Schwartz, U., Löffler, P., Schmitz, S., Symelka, A., de Koning-Ward, T., Merkl, R., and Längst, G. (2016). *Plasmodium falciparum* nucleosomes exhibit reduced stability and lost sequence dependent nucleosome positioning. *PLoS Pathog*, 12(12):e1006080.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., et al. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res*, 43(Database issue):D376–D381.

Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, 33(Web Server issue):W244–W248.

Solomon, E., Berg, L., and Martin, D. (2002). *Biology, 6th edition*. Belmont: Brooks/-Cole Thomson Learning.

Sonoda, J. and Wharton, R. P. (2001). *Drosophila* brain tumor is a translational repressor. *Genes Dev*, 15(6):762–773.

Sousa, S. F., Fernandes, P. A., and Ramos, M. J. (2006). Protein–ligand docking: Current status and future challenges. *Proteins*, 65(1):15–26.

Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M.-C., Yanofsky, C., and Mills, S. E. (2001). The structures of anthranilate synthase of *Serratia marcescens* crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product, glutamate, and (ii) its end-product inhibitor, L-tryptophan. *Proc Natl Acad Sci U S A*, 98(11):6021–6026.

Stranges, P. B. and Kuhlman, B. (2013). A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci*, 22(1):74–82.

Tanaka, F., Fuller, R., Shim, H., Lerner, R. A., and Barbas, C. F. (2004). Evolution of aldolase antibodies in vitro: Correlation of catalytic activity and reaction-based selection. *J Mol Biol*, 335(4):1007–1018.

Torvalds, L. and Hamano, J. (2010). Git: Fast version control system. *Software available at http://git-scm.com*.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461.

Verlet, L. (1967). Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys Rev*, 159(1):98.

Vining, L. (1992). Secondary metabolism, inventive evolution and biochemical diversity – a review. *Gene*, 115(1-2):135–140.

Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H., and DeGrado, W. F. (1999). Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc Natl Acad Sci U S A*, 96(10):5486–5491.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 421(6921):397.

Wei, G., Xi, W., Nussinov, R., and Ma, B. (2016). Protein ensembles: How does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chem Rev*, 116(11):6516–6551.

Wernisch, L., Hery, S., and Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol*, 301(3):713–736.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70(3):697–701.

Workman, J. L. and Kingston, R. E. (1992). Nucleosome core displacement *in vitro* via a metastable transcription factor-nucleosome complex. *Science*, 258(5089):1780–1785.

Wu, S. and Zhang, Y. (2008). MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins*, 72(2):547–556.

Xu, D. and Zhang, Y. (2012). *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, 80(7):1715–1735.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nat Methods*, 12(1):7–8.

Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D., and Baker, D. (2006). New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci*, 15(12):2785–2794.

Zhang, Y. (2008). I-TASSER server for protein 3d structure prediction. *BMC Bioinformatics*, 9(1):1.

Zhang, Y. and Skolnick, J. (2004a). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710.

Zhang, Y. and Skolnick, J. (2004b). SPICKER: A clustering approach to identify near-native protein folds. *J Comput Chem*, 25(6):865–871.

Zhang, Y. and Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33(7):2302–2309.

Appendix A

List of command line options for MSF

Table A.1 **Options required for multi-state design.** Current MSF applications require the setup of an entity_resfile and fitness_file, which replicate the functionality of MPI_MSD's counterparts. See the documentation of MPI_MSD for these functions: https://www.rosettacommons.org/docs/latest/application_documentation/design/mpi-msd

Option (MSF namespace)	Description	Default
entity_resfile	Defines the size of the design shell	
fitness_file	Path to the DAF file which defines how to calculate an ensemble score	
read_states_on_demand	Load state structures on demand from disk; saves memory when a large number of states are used	false

A typical call of msf_ga_enzdes may look like the following: mpirun –np num_cpu $ROSETTA_ROOT/main/source/bin/msf_ga_enzdes.mpi.linuxgccrelease @flags; the number of processes should be chosen wisely to match half the size of the genetic algorithm's population size with the number of available processing cores. Please refer to the options table below to fill the flags file. It is required to specify most of the options. Example flag files are listed in the next chapter.

Table A.2 **Command line options of the genetic algorithm.** The following command line options define the parameters of the GA.

Option (MSF namespace)	Description	Default
checkpoint_write_interval	Write checkpoint every x generations	1
checkpoint_prefix	Location to the store path of the checkpoint files	
darwin_resume	Resume from checkpoint files	false
fill_gen1_from_seed_sequences	Initialize GA from distinct sequences	false
fraction_by_recombination	Fraction of the population that is recombined every generation	0.05
generations	Number of generations to be evolved	0
pop_size	Size of the population	0
resfile_tmpdir	Path to directory that allows saving temporary resfiles	
seed_sequences	List of sequences separated by ","	
seed_sequences_from_input_pdb	Fill the population from the sequence of the input structure	
seed_sequence_using_correspondence_file	Extract sequence from design shell	false

Appendix B

Details of benchmark datasets / protocols for their compilation

The datasets used for benchmarking are deposited on the website of our department: https://bioinf.ur.de/downl/MSF_bench.tar.gz; The following text lists the protocol of the energy-minimization performed via relax, the composition of the design and repack shell and the parameters used for benchmarking

B.1 Relax protocol

Structures were energy-minimized using a fast-relax protocol with backbone restraints, defined by the following flags:

```
-ignore_unrecognized_res
-relax:constrain_relax_to_start_coords
-relax:ramp_constraints false
-ex1
-ex2
-use_input_sc
-correct
-restore_pre_talaris_2013_behavior
-no_his_his_pairE
-no_optH false
-flip_HNQ
-nstruct 1
-relax:fast
-extra_res_fa ligand.params
```

B.2 Design and repack shell composition

The following table lists the design and repack shells used for recapitulation calculations on NMR ensemble *hIFABP* and those for *BR_EnzBench* and *MD_EnzBench*. Note that all numbers are based on structures, where missing residues (in the crystal structure) were loop-modeled with YASARA and residues were renumbered to start with 1.

Composition of design and repack shells for ligand-binding design

Table B.1 List of proteins, design shell and repack shell residues for benchmark datasets *hIFABP*, *BR_EnzBench* and *MD_EnzBench*.

PDB ID	design shell	repack shell
2mji	14 17 21 23 24 27 28 31 58 70 72 75 76 78 91 93 102 104 117 119 124	8 11 13 15 16 18 20 22 25 26 30 32 33 36 38 49 51 53 59 60 62 68 69 71 73 74 77 79 80 82 89 90 92 94 95 96 100 101 103 105 106 115 116 118 122 123 126
1fzq	23 25 26 27 28 29 30 31 32 33 66 90 92 125 126 128 129 158 159 160	21 22 24 34 35 51 53 64 65 67 88 91 93 94 95 102 123 124 127 130 131 132 156 157 161 162 163 164 165 177
1hsl	11 14 52 69 70 71 72 90 117 119 120 121 122 123 143 160 161 192 195	10 12 13 15 17 18 30 51 53 55 56 67 68 73 74 76 77 83 87 88 89 91 92 116 118 124 125 140 141 142 146 158 159 162 163 164 165 190 191 193 194 196 239
1j6z	8 10 11 12 13 15 30 71 151 153 154 155 178 179 180 182 183 207 210 211 214 298 299 300 302 303 333	9 14 24 27 28 29 66 68 69 70 103 105 106 134 138 152 156 157 158 181 184 185 186 203 204 206 208 209 212 213 215 216 254 255 258 297 301 304 306 332 334 335 336 337 338 370
1n4h	20 21 22 55 58 59 62 65 96 97 99 100 102 103 104 111 112 113 123 128 131 132 135 212 216	16 19 23 24 26 27 30 51 52 53 54 56 57 60 61 63 64 66 68 69 93 94 95 98 101 110 114 115 116 117 118 119 120 122 126 129 130 133 134 136 137 138 139 149 152 156 213 215 219 220 239 245

1nq7	20 21 22 55 58 59 62 65 93 96 97 99 100 102 103 111 112 113 123 126 128 131 132 135 136 212 213 216	16 17 18 19 23 24 26 27 43 51 52 53 54 56 57 60 61 63 64 66 68 69 92 94 95 98 101 104 110 114 115 117 118 119 120 121 122 124 129 130 133 134 138 139 149 152 156 209 210 211 214 215 217 219 220 239 243 245
1opb	20 25 29 33 36 38 40 42 51 53 55 57 58 59 60 62 76 77 106 108 117 119	4 8 10 13 16 17 19 21 22 23 30 31 32 34 35 37 39 41 44 49 50 52 54 56 61 63 64 72 74 75 78 84 86 93 94 95 97 104 105 107 109 110 115 116 118 120 121 126 128 130 134
1pot	9 10 11 12 37 58 107 143 145 146 186 187 204 206 207 230 232 268 302	8 13 33 34 35 36 38 39 40 56 57 59 60 61 104 105 106 108 109 141 142 144 147 149 150 185 188 189 190 191 202 203 205 210 228 229 231 233 266 301 303 323
1urg	43 44 45 64 66 67 112 145 148 149 150 151 206 224 226 257 259 327 337	8 9 10 38 42 47 62 63 65 68 70 71 97 109 110 111 113 114 144 147 153 205 222 223 225 227 254 255 256 258 260 295 296 328 329 333 334 336 338 340 374
2b3b	8 9 13 41 42 43 66 119 224 242 244 276 278 312 347 348 349	7 10 12 14 15 16 17 37 39 40 44 45 46 47 64 65 67 70 116 117 118 120 121 165 167 168 170 171 223 225 228 241 243 245 247 248 275 277 279 308 313 314 345 346 350 357 393
2dri	13 15 16 89 90 102 103 105 131 132 135 137 164 189 190 192 214 215 235	8 9 10 12 14 18 19 41 64 65 66 67 68 87 88 91 104 136 138 139 140 141 163 165 166 188 191 193 194 195 213 216 217 219 232 233 234 236 237 240 263 265 272
2ifb	17 18 23 27 30 31 38 49 51 60 62 70 72 73 74 78 82 91 93 102 104 117	2 6 11 14 15 19 20 21 22 24 26 28 29 32 34 36 39 40 47 48 50 53 54 55 56 58 59 61 63 68 69 71 75 76 77 79 80 81 89 90 92 94 95 103 105 106 115 116 118 119 122 124 125 126 132
2q2y	98 99 100 101 102 110 113 115 116 117 119 120 143 154 155 194 197 198 200 201 203 204 222	61 95 96 97 103 108 109 111 112 114 118 121 122 123 124 141 142 144 145 193 195 196 199 202 205 207 208 215 220 221 223 224 246 248 331
2qo4	17 18 21 23 27 30 31 34 36 51 53 54 55 56 72 73 74 75 76 98 111 118	8 11 13 14 15 16 20 22 24 28 29 32 33 35 37 38 49 52 57 58 60 70 71 77 78 91 93 95 96 97 99 100 109 110 112 113 116 119 120 122 126

2rct	20 24 29 33 37 40 42 44 55 57 59 61 62 63 64 66 80 81 110 112 121 123	8 12 14 17 21 23 25 26 27 28 30 34 35 36 38 39 41 43 45 46 48 53 54 56 58 60 65 67 68 76 78 79 82 88 90 97 98 99 101 108 109 111 113 119 120 122 124 125 130 132 134 142
2rde	73 75 79 112 113 114 115 117 139 140 141 143 144 145 146 184 185 196 197 198	17 18 20 71 72 74 76 77 78 80 81 110 111 116 118 119 120 137 138 142 147 148 158 179 181 182 183 186 187 194 195 199 204 207 211 225
2uyi	100 101 102 103 104 112 115 117 118 119 121 122 145 156 157 196 199 200 202 203 205 206 224	63 97 98 99 105 110 111 113 114 116 120 123 124 125 126 143 144 147 154 163 195 197 198 201 204 207 209 210 217 222 223 225 226 248 250 332

Composition of design shells for anchored design of protein-protein interfaces

Table B.2 List of proteins and design shell residues for *BR_IfaceBench*. All other residues of the proteins are part of the repack shell.

PDB ID	chain	anchor residue	design-shell residues
1dle	B	32	27 28 29 30 31 33 34
1fc4	B	71	66 67 68 69 70 72 73 74 75 76
1fec	B	458	455 456 457 459 460 461 462
1jtp	B	104	99 100 101 102 103 105 106 107 108
1qni	B	399	388 389 390 391 392 393 394 395 396 397 398 400 401 402
1u6e	B	96	90 91 92 93 94 95 97 98 99
1zr0	B	19	14 15 16 17 18 20 21
2bwn	B	85	77 78 79 80 81 82 83 84 86 87 88 89 90
2hp2	B	300	288 289 290 291 292 293 294 295 296 297 298 299 301 302 303
2i25	O	90	85 86 87 88 89 91 92
2obg	B	80	77 78 79 81 82 83 84 85 86
2qpv	B	55	50 51 52 53 54 56 57

2wya	C	89	83 84 85 86 87 88 90 91 92
3cgc	B	429	421 422 423 424 425 426 427 428 430 431 432 433 434
3dxv	B	289	286 287 288 290 291 292 293 294 295 296 297
3ean	B	464	458 459 460 461 462 463 465

B.3 Parameters for design

In this section, the ROSETTA parameters are listed that allow the reproduction of all design computations used for benchmarking.

Parameter set *ps_enzdes* for single-state ligand-binding design

Example flags are given for running ENZDES on the first benchmark protein 1fzq of *BR_EnzBench*. Using the information of the above shell composition, the benchmark computations can be easily reproduced for other benchmark proteins.

1fzq.flags:

```
-in:file:l ./lists/1fzq_all # a list with paths to all conformations
   to design
-resfile ./resfiles/1fzq.resfile # using above defined design shell
-no_his_his_pairE
-correct
-restore_pre_talaris_2013_behavior
-extra_res_fa ./params/1fzq.params
# ENZDES flags
-enzdes::cst_design
-enzdes::design_min_cycles 2
-out::nstruct 1000
-enzdes::cst_min
-enzdes::chi_min
-enzdes::bb_min
-enzdes::lig_packer_weight 1.8
-enzdes::final_repack_without_ligand
-ex1
-ex2
```

```
-ex1aro
-ex2aro
-extrachi_cutoff 1
-soft_rep_design
-flip_HNQ
-linmem_ig 10
-docking::ligand::old_estat
-out:file:o ./output/1fzq_energy
-out:prefix ./output/1fzq_design
-run:constant_seed
-run:jran 11111111
```

./resfiles/1fzq.resfile:

```
NATRO
start
...
21 A NATAA  EX 1 EX 2 EX ARO 1 EX ARO 2  # repack shell
22 A NATAA  EX 1 EX 2 EX ARO 1 EX ARO 2
23 A ALLAA  EX 1 EX 2 EX ARO 1 EX ARO 2  # design shell
...
```

Parameter set *ps_msf_enzdes* for multi-state ligand-binding design

In the following, example flags are given for running MSF:GA:ENZDES on the first conformation group of benchmark protein 1fzq. Using the information of the above shell composition and modifying few lines, the benchmark computations can be easily reproduced. The way how an MSF run is set up follows the way implemented in MPI_MSD and thus, the definition of *.resfile files, *.daf files, *.state files, *.corr files, *.2resfile files for MSF is exactly the same. A more detailed documentation on how to prepare and customize a multi-state run is provided here: https://www.rosettacommons.org/docs/latest/application_documentation/design/mpi-msd

1fzq.flags:

```
-entity_resfile ./resfiles/1fzq.resfile
-msf::fitness_file 1fzq.daf
-msf::pop_size 210
-msf::generations 600 # 800 for hIFABP
-msf::fraction_by_recombination 0.05
-msf::seed_sequences AAAAAAAAAAAAAAAAAAAA # length of DS - ALA seed
-msf::resfile_tmpdir ./tmp_resfiles/1fzq/ # temporary resfiles
-msf::checkpoint_write_interval 1
```

```
-msf::checkpoint_prefix ./checkpoints/1fzq/checkpoint
-no_his_his_pairE
-correct
-restore_pre_talaris_2013_behavior
-extra_res_fa ./params/1fzq.params
# ENZDES flags
-enzdes::cst_design
-enzdes::design_min_cycles 2
-enzdes::cst_min
-enzdes::chi_min
-enzdes::bb_min
-enzdes::lig_packer_weight 1.8
-enzdes::final_repack_without_ligand
-ex1
-ex2
-ex1aro
-ex2aro
-extrachi_cutoff 1
-soft_rep_design
-flip_HNQ
-linmem_ig 10
-docking::ligand::old_estat
-out:file:o msd_output/1fzq_energies
-out:prefix msd_output/1fzq_design
-run:constant_seed
-run:jran 11111111
```

./resfiles/1fzq.resfile:

```
20
ALLAA EX 1 EX 2 EX ARO 1 EX ARO 2
Start
```

1fzq.daf:

```
STATE_VECTOR state1 ./states/1fzq/state1
STATE_VECTOR state2 ./states/1fzq/state2
STATE_VECTOR state3 ./states/1fzq/state3
STATE_VECTOR state4 ./states/1fzq/state4
STATE_VECTOR state5 ./states/1fzq/state5
SCALAR_EXPRESSION best_state1 = vmin( state1 )
SCALAR_EXPRESSION best_state2 = vmin( state2 )
SCALAR_EXPRESSION best_state3 = vmin( state3 )
SCALAR_EXPRESSION best_state4 = vmin( state4 )
SCALAR_EXPRESSION best_state5 = vmin( state5 )
```

```
SCALAR_EXPRESSION best_sum = best_state1 + best_state2 + best_state3
   + best_state4 + best_state5
FITNESS best_sum
```

./states/1fzq/state1:

```
./1fzq/input/1_backrub_input.pdb ./corr/1fzq.corr ./resfiles/1fzq.2
   resfile
```

./corr/1fzq.corr:

```
1 23 A
2 25 A
3 26 A
4 27 A
5 28 A
6 29 A
7 30 A
8 31 A
9 32 A
10 33 A
11 66 A
12 90 A
13 92 A
14 125 A
15 126 A
16 128 A
17 129 A
18 158 A
19 159 A
20 160 A
```

./resfiles/1fzq.2resfile:

```
NATRO
start
21 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
22 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
23 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
24 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
25 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
26 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
27 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
28 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
29 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
```

```
 30 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 31 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 32 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 33 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 34 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 35 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 51 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 53 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 64 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 65 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 66 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 67 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 88 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 90 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 91 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 92 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 93 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 94 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
 95 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
102 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
123 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
124 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
125 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
126 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
127 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
128 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
129 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
130 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
131 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
132 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
156 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
157 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
158 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
159 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
160 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
161 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
162 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
163 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
164 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
165 A NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
177 B NATAA EX 1 EX 2 EX ARO 1 EX ARO 2
```

Parameter set *ps_anchored* for single-state protein-protein interface design

Example flags are given for running ANCHOREDDESIGN on the first benchmark protein 1dle of *BR_IfaceBench*. Using the information of the above shell composition, the benchmark computations can be easily reproduced for other benchmark proteins.

```
# protocol for coarse optimization
-unmute protocols.loops.CcdLoopClosureMover
#repeating options for safety
-run::version
-options::user
#packing options - these are about as high as they can go
-ex1
-ex2
-use_input_sc
-extrachi_cutoff 8
-linmem_ig 42
#minimization options
-run::min_type dfpmin_armijo
-nblist_autoupdate

#loops options
-loops::vicinity_sampling true
-loops::loop_file ../../loopsfile

#AnchoredDesign options
-AnchoredDesign
        -anchor ../../anchor
        -allow_anchor_repack false
        -vary_cutpoints true
        -debug false
        -show_extended false
        -refine_only false
        -perturb_show false
        -perturb_temp 0.8
        -refine_temp 0.8
        -refine_repack_cycles 50
        -rmsd false
        -unbound_mode false
        -no_frags false
        -perturb_CCD_off false
        -perturb_KIC_off false
        -refine_CCD_off false
```

```
        -refine_KIC_off false
        -chainbreak_weight 2.0
        -testing::VDW_weight 2

#sample-size command
-AnchoredDesign::perturb_cycles 50
-AnchoredDesign::refine_cycles 100
-nstruct 8
```

Parameter sets for multi-state protein-protein interface design

In the following, example flags are given for running MSF on the first conformation group of benchmark protein 1dle. Using the information of the shell composition described above and by modifying few lines, the benchmark computations for MSF:GA:ANCHORED can be easily reproduced. Two parameter sets are listed: *ps_msf_anchored_coarse* was run in centroid mode, providing sequence optimization while coarsely sampling loop conformations.

All optimized sequences from this first run were fed into a run with parameter set *ps_msf_anchored_refine* using `-msdesign::seed_sequences` and performing in both centroid and refinement mode by setting the following flagsnumber of `-AnchoredDesign::refine_cycles` in the below parameter set to `100` and running for `-msf::generations 1000` generations.

***ps_msf_anchored_perturb*:**

```
-unmute protocols.loops.CcdLoopClosureMover
#repeating options for safety
-run::version
-options::user
#packing options - these are about as high as they can go
-ex1
-ex2
-use_input_sc
-extrachi_cutoff 8
-linmem_ig 42
#minimization options
-run::min_type dfpmin_armijo
-nblist_autoupdate

#loops options
```

```
-loops::vicinity_sampling true
-loops::loop_file ../../loopsfile

#AnchoredDesign options
-AnchoredDesign
        -anchor ../../anchor
        -allow_anchor_repack false
        -vary_cutpoints true
        -debug false
        -show_extended false
        -refine_only false
        -perturb_show false
        -perturb_temp 0.8
        -refine_temp 0.8
        -refine_repack_cycles 50
        -rmsd false
        -unbound_mode false
        -no_frags false
        -perturb_CCD_off false
        -perturb_KIC_off false
        -refine_CCD_off false
        -refine_KIC_off false
        -chainbreak_weight 2.0
        -testing::VDW_weight 2

#sample-size command
-AnchoredDesign::perturb_cycles 50
-AnchoredDesign::refine_cycles 0
-nstruct 1

#msf flags
-msf:fitness_file ../daf1
-msf:entity_resfile ../entity_resfile
-msf:resfile_tmpdir ../tmp_resfiles1
-msf:checkpoint_prefix ../checkpoints1/checkpoint
-msf::pop_size 50
-msf::generations 500
-msf::fraction_by_recombination 0.05
-msf::checkpoint_write_interval 1
-ignore_unrecognized_res
```

Appendix C

Multistate approach to design retro-aldolases

This appendix complements the main text with more figures and information about the *in silico* design and evaluation of retro-aldolases. The sequences of initial variants chosen for expression as well as stabilized variants are listed in Section C.2. The RA* dataset used for comparison with RA_MSD* variants is listed in Section C.4.

C.1 Multi-state design

Multi-state design was performed on the scaffold protein indole-3-glycerolphosphate synthase from *S. solfataricus* (ssIGPS). First, the ligand was removed and conformations were sampled by means of MD simulations. To obtain representative conformations, the snapshots of the trajectory were clustered and in total 12 conformations were picked from different clusters. ROSETTA:MATCH was used to graft the transition state in those conformations and in the original crystal structure. Matched transition states (*mTS*) with putative weak binding were discarded and 23 ensembles of matched transition states (ens_{mTS}) were chosen for design. Here, each ens_{mTS} has a specific catalytic triad (unique positions and amino acid identities). Finally, ens_{mTS} were designed with MSF:GA:ENZDES.

MD simulations of the scaffold protein

Three 10 *ns* MD simulations were applied to the scaffold ssIGPS without ligand, generating a snapshot every *ps* (10,000 snapshots per trajectory). MD simulations were performed with YASARA (version 14.7.17) employing the YAMBER3 force

field. Simulations were run at 298 K under periodic boundary conditions and with explicit water, using a multiple time step of 1 fs for intramolecular and 2 fs for intermolecular forces. To perform three individual simulations, independent calculations were seeded by slightly changing the simulation temperature ($\pm 0.01K$) which reassigns the initial atom velocities. Lennard Jones forces and long-range electrostatic interactions were treated with a 7.86 Å cutoff, the latter were calculated using the Particle Mesh Ewald method. Temperature was adjusted using a Berendsen thermostat based on the time-averaged temperature and simulations were carried out at constant pressure. MD simulations require the definition of a simulation cell that should be adequately sized to prevent self-interaction through periodic boundaries. Simulation cells were thus defined as 5 Å larger than the protein along each axis. Cells were filled with water to a density of 0.997 g/ml, and counterions were added to a final concentration of 0.9% NaCl. Next, the protonation states of all molecules were assigned according. Before capturing production runs, usually an equilibration run is performed to remove conformational stress. Here, we performed an energy minimization, which was done as follows: After removing conformational stress by a steepest descent minimization, the procedure continued by simulated annealing (time step 2 fs, atom velocities scaled down by 0.9 every 10th step) until convergence was reached, i.e. the energy improved by less than 0.05 kJ/mol per atom during 200 steps.

Selection of representative conformations and matching

All snapshots (3 x 10,000 in total) were clustered with Durandal, using smart-mode enabled and semi-auto [0.03 .. 0.20]. Four structures were picked from each largest cluster of the three trajectories. In addition to those 12 conformations, the crystal structure of ssIGPS was used. Next, the theozyme was grafted onto the 13 conformations with ROSETTA:MATCH. The matcher proposes a number of ligand positions within a scaffold's cavity with regard to the given theozyme. The cavities of all conformations were detected by means of the tool Rosetta:gen_apo_grids as described in ROSETTA'S documentation. The theozyme definition was derived from previous work [Bjelic et al., 2014] and is given by:

```
CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: C5 C4 C3
TEMPLATE:: ATOM_MAP: 1 residue3: MTD
TEMPLATE:: ATOM_MAP: 2 atom_type: Nlys ,
TEMPLATE:: ATOM_MAP: 2 residue1: K
```

```
CONSTRAINT:: distanceAB:    1.51 0.2 50.0   1  0
CONSTRAINT::    angle_A:  110.   5.0  0.0  60. 0
CONSTRAINT::    angle_B:  110.  10.0  0.0  60. 1
CONSTRAINT::  torsion_A: -120.  20.0  0.05 60. 0
CONSTRAINT:: torsion_AB:    0. 180.0  0.05 60. 3
CONSTRAINT::  torsion_B:    0. 180.0  0.00 60. 3
CST::END

CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: O2 C5 C6
TEMPLATE:: ATOM_MAP: 1 residue3: MTD
TEMPLATE:: ATOM_MAP: 2 atom_type: OOC ,
TEMPLATE:: ATOM_MAP: 2 residue1: DE
CONSTRAINT:: distanceAB:   3.0   0.3 10.0   0  1
CONSTRAINT::    angle_A: 125.0  20.0  0.0  60. 0
CONSTRAINT::    angle_B: 125.0  25.0  0.05 60. 0
CONSTRAINT::  torsion_A: -60.0  20.0  0.01 60. 0
CONSTRAINT:: torsion_AB:   0.0 180.0  0.0  60. 3
CONSTRAINT::  torsion_B: 180.0 180.0  0.0  60. 3
CST::END

CST::BEGIN
TEMPLATE:: ATOM_MAP: 1 atom_name: O2 C5 C6
TEMPLATE:: ATOM_MAP: 1 residue3: MTD
TEMPLATE:: ATOM_MAP: 2 atom_type: OH ,
TEMPLATE:: ATOM_MAP: 2 residue1: ST
CONSTRAINT:: distanceAB:   3.0   0.30 10.    0  0
CONSTRAINT::    angle_A: 125.0  20.0   0.0  60. 0
CONSTRAINT::    angle_B: 125.0  25.0   0.05 60. 0
CONSTRAINT::  torsion_A:  60.0  20.0   0.01 60. 0
CONSTRAINT:: torsion_AB:   0.0 180.0   0.0  60. 3
CONSTRAINT::  torsion_B: 180.0 180.0   0.0  60. 3
CST::END
```

ROSETTA:MATCH was executed seven times with different seeds. For each uniquely specified catalytic triad (amino acid identities and positions), the resulting matched transition states (mTS) were collected, as described in methods. In total, 23 ensembles ens_{mTS} containing four up to 13 conformations were chosen. In Fig. C.1, the ens_{mTS} of design RA_MSD2 is shown, consisting of six conformations (states for MSD).

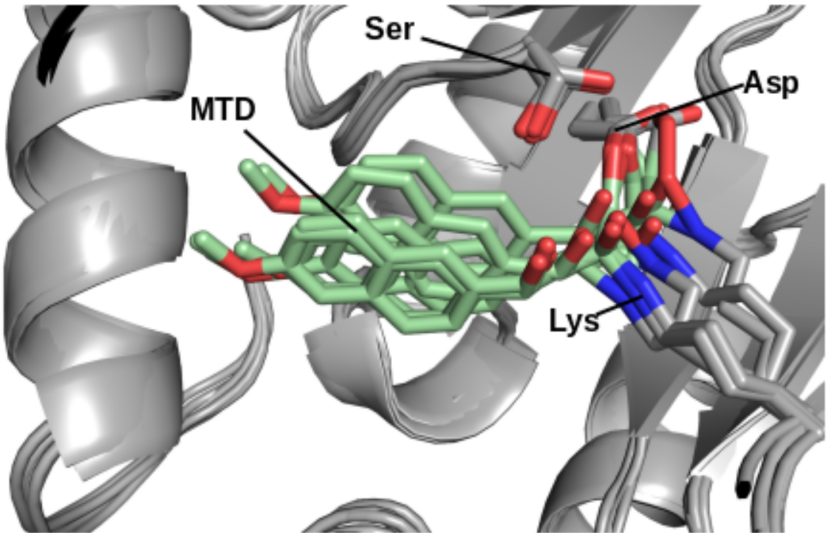


Fig. C.1 ens_{mTS} **from design of RA_MSD2.** Transition state of MTD (green) and the three catalytic residues (Lys/Asp/Ser) matched at positions 210, 131 and 110, respectively.

Multi-state design with MSF:GA:enzdes

All ens_{mTS} were designed with MSF:GA:ENZDES using the following flags, exemplarily for the design of RA_MSD2:

```
-entity_resfile ./corr_resfiles/RA_MSD2/entity_resfile
-msf::fitness_file ./daf/RA_MSD2.daf
-msf::pop_size 210
-msf::generations 2000
-msf::fraction_by_recombination 0.05
-msf::seed_sequence_using_correspondence_file ./corr_resfiles/RA_MSD2
   .corr
-msf::resfile_tmpdir tmp_resfiles/RA_MSD2/
-msf::checkpoint_write_interval 1
-msf::checkpoint_prefix checkpoints/RA_MSD2/checkpoint
-no_his_his_pairE
-correct
-restore_pre_talaris_2013_behavior
-extra_res_fa ./params/MTD_sb.params
-enzdes::cst_design
-enzdes::design_min_cycles 2
-enzdes::cst_min
-enzdes::chi_min
-enzdes::bb_min
-enzdes::favor_native_res 1.5
```

```
-ex1
-ex2
-ex1aro
-ex2aro
-use_input_sc
-extrachi_cutoff 1
-soft_rep_design
-flip_HNQ
-linmem_ig 10
-enzdes::lig_packer_weight 1.8
-docking::ligand::old_estat
-out:file:o ./energies/RA_MSD2/RA_MSD2_energies
-out:prefix ./output/RA_MSD2/
-enzdes::final_repack_without_ligand
-enzdes::cst_opt
-enzdes::cstfile ./cst/MTD.cst
-run:constant_seed
-run:jran 11111111
```

Here, the $fitness\_file$ defines the fitness as the sum of total energies over all states (conformations) in ens_{mTS}. Designs were then run for 97 to 710 generations and Fig C.2 shows the convergence for RA_MSD2:

C.2 Evaluation of designs

The structures of the best designed sequences of each design run were visually inspected, followed by an evaluation in terms of Rosetta. Because scores of design runs originating from different conformations are difficult to compare, we also compared the active-site geometries. Next, the best 100 variants were further analyzed via MD simulations.

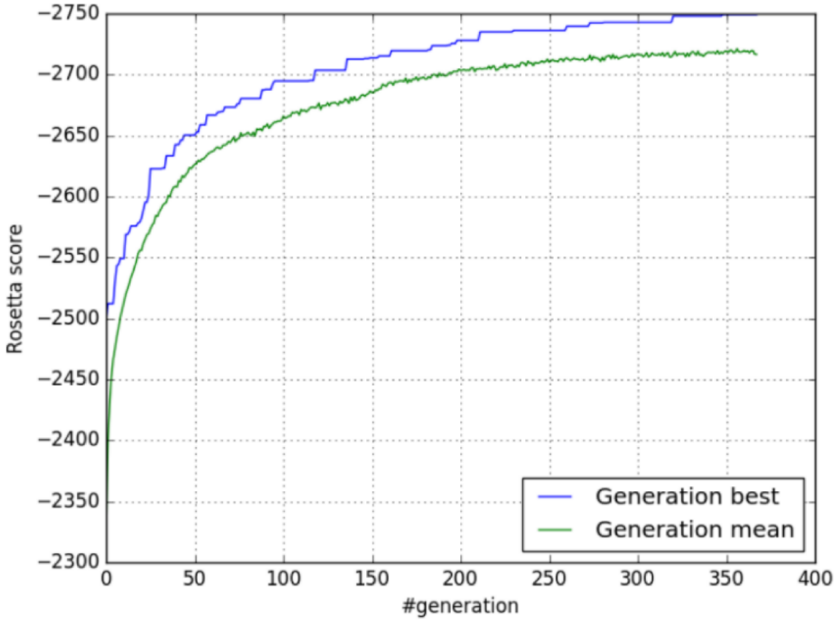


Fig. C.2 **Energetic convergence of design run RA_MSD2.** Generations of designed sequences plotted against their fitness (sum of Rosetta scores for all states). Lines show the fitness of each generation's best sequence (blue) and the mean fitness of the generation (green).

MD evaluation

As described in methods, the best 100 designs were selected for 10 *ns* MD simulations in water and 100 snapshots were generated per simulation. Simulations were performed for **i**) the enzyme/TS complex and **ii**) the enzyme/substrate complex. For each trajectory, the catalytic distances and angles of each snapshot were plotted as boxplot (Fig. C.3), in difference to the optimum as described in the theozyme definition C.1.

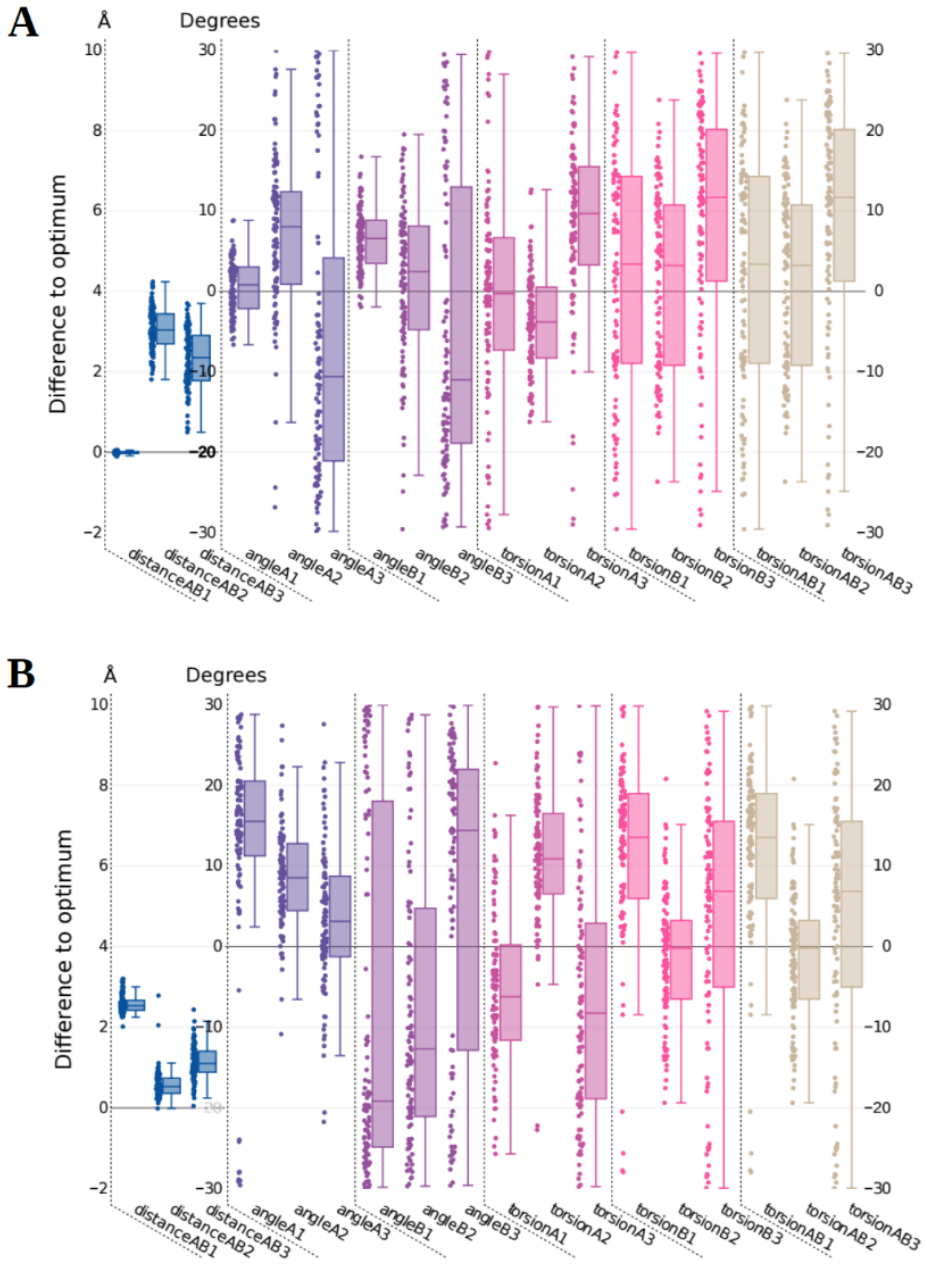


Fig. C.3 **MD evaluation of RA_MSD2.** (**A**) Catalytic distances and angles of all snapshots deviating from the optimum value of the theozyme definition after running the design of RA_MSD2 in MD simulations. Dots show raw data plotted as boxplots from each trajectory consisting of 100 snapshots. Deviations from the theozyme for the enzyme/*TS* complex. (**B**) Deviations from the theozyme for the enzyme/substrate complex.

Sequences of variants chosen for biochemical characterization

After filtering *in silico* for the best designs, the following sequences were chosen for biochemical characterization:

>RA_MSD1
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYSKFMERYAVGLTIVTFEWYYNGSYETLRKIASSVSIPISMSDIIV
KESQIDDAYNLGADTVDIIVKILTERELESLLEYARSYGMEPRIEINDENDLDIALR
IGARFIVIDSRDPETLEINKENQRKLISMIPSNVVKVAKSGISERNEIEELRKLGVN
AFDIGSSLMRNPEKIKEFIL
>RA_MSD2
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYSKFMERYAVGLTIVTFEKYYNGSYETLRKIASSVSIPISMSDVIV
KESQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDINDENDLDIALR
IGARFICIDSRDPETLEINKENQRKLISMIPSNVVKVAKSGISERNEIEELRKLGVN
AFDIGSSLMRNPEKIKEFIL
>RA_MSD3
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYSKFMERYAVGLTIVTFERYYNGSYETLRKIASSVSIPIAMSDIIV
KESQIDDAYNLGADTVDIIVKILTERELESLLEYARSYGMEPRICINDENDLDIALR
IGARFIAIDSRDPETLEINKENQRKLISMIPSNVVKVAKSGISERNEIEELRKLGVN
AFDIGSSLMRNPEKIKEFIL
>RA_MSD4
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAKYERKHP
SGLDVERDPIEYSKFMERYAVGLTISTLEKYFNGSYETLRKIASSVSIPIEMFDIIV
KESQIDDAYNLGADTVVLVVKLLTERELESLLEYARSYGMEPLIIITDENDLDIALR
IGARFIGIWSRDGETLEINKENQRKLISMIPSNVVKVADGGISERNEIEELRKLGVN
AFAIGESLMRNPEKIKEFIL
>RA_MSD5
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAKYERKHP
SGLDVERDPIEYSKFMERYAVGLMISTEEKYHNGSYETLRKIASSVSIPIAMFDIIV
KESQIDDAYNLGADTVVLVVGLLTERELESLLEYARSYGMEPLIIITDENDLDIALR
IGARFIGIWSRDGETLEINKENQRKLISMIPSNVVKVAIGGISERNEIEELRKLGVN
AFAIGESLMRNPEKIKEFIL
>RA_MSD6
MPRYLKGWVKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYERKSL
SGLDVERDPIEYSKFMERYAVGLFISTEEKYHNGSYETLRKIASSVSIPIGMVDGIV

```
KESQIDDAYNLGADTVVLVVRLLTERELESLLEYARSYGMEPLIVIKDENDLDIALR
IGARFIAIDSQDWETLEINKENQRKLISMIPSNVVKVAVNGISERNEIEELRKLGVN
AFKISASLMRNPEKIKEFIL
>RA_MSD7
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIALYGRKSP
SGLDVERDPIEYSKFMERYAVGLAIFTEEKYHNGSYETLRKIASSVSIPICMTDFIV
KESQIDDAYNLGADTVELYVKILTERELESLLEYARSYGMEPIITINDENDLDIALR
IGARFIGILSRDLETLEINKENQRKLISMIPSNVVKAAAEGISERNEIEELRKLGVN
AFKIWESLMRNPEKIKEFIL
>RA_MSD8
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAIYGRKSP
SGLDVERDPIEYSKFMERYAVGLQIFTEEKYHNGSYETLRKIASSVSIPICMSDFIV
KESQIDDAYNLGADTVELWVKILTERELESLLEYARSYGMEPIITINDENDLDIALR
IGARFIGILSRDLETLEINKENQRKLISMIPSNVVKAASEGISERNEIEELRKLGVN
AFKIWESLMRNPEKIKEFIL
>RA_MSD9
MPRYLKGWMKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAAYERKSP
SGLDVERDPIEYSKFMERYAVGLSITTEEKYGNGSYETLRKIASSVSIPIDMTDVIV
KESQIDDAYNLGADTVTLVVRILTERELESLLEYARSYGMEPLIVISDENDLDIALR
IGARFICIDSRDWETLEINKENQRKLISMIPSNVVKVAANGISERNEIEELRKLGVN
AFKIGSSLMRNPEKIKEFIL
```

C.3 *In silico* stabilization

Variant RA_MSD2 had the lowest activity of all designs and was insoluble on expression without MBP. To assess the relationship between activity and solubility, we predicted stabilizing mutations utilizing the PROSS server as described in Methods (Section 2.8.3). Each sequence logo shown below is computed from an MSA that was generated by predicting seven stabilized variants for each conformational state of RA_MSD2 and merging the predicted mutations for each degree of stabilization.

Fig. C.4 **Sequence logos 1-7 from PROSS predictions with an increasing number of putatively stabilizing mutations.** Sequence logos were determined using WEBLOGO [Crooks et al., 2004]. Black letters correspond to the native sequence and green letters indicate stabilizing mutations. The height of the letter represents the consensus from all designed conformational states.

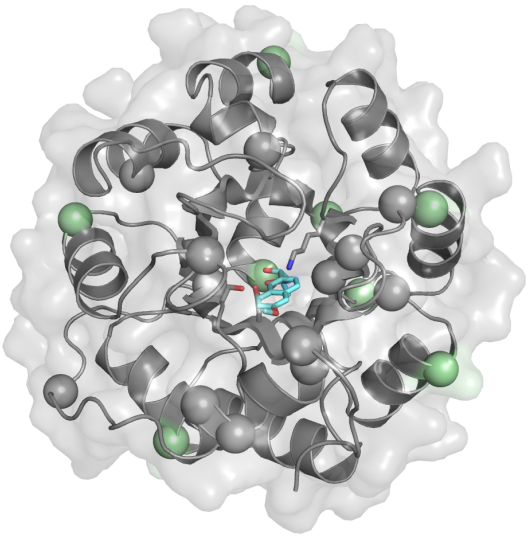
Sequences of variants chosen for stabilization

For the sake of completeness, this is the list of stabilized sequences chosen for expression:

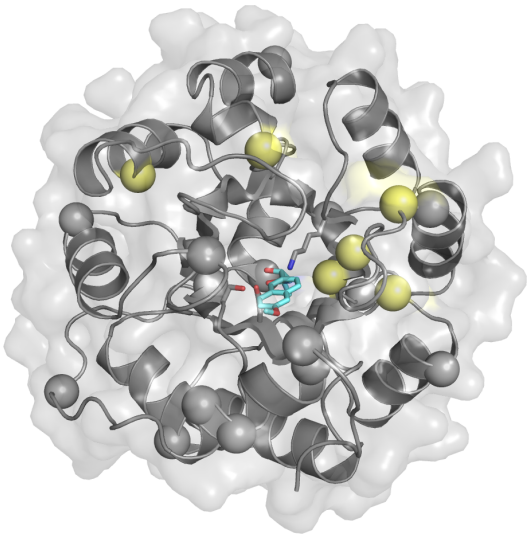
>RA_MSD2.1
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLKERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYSKFMERYAVGLTIVTFEKYYNGSYETLRKIASAVSIPISMSDVIV
KEYQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDINDEEDLDIALR
IGARFICIDSRDPETLEIDKENQRKLISMIPSDVVKVAKSGISERNEIEELRKLGVH
AFDIGSSLMRNPEKIKEFIE
>RA_MSD2.2
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYAKFMERYAVGLTIVTFEKYYNGSYETLRKIASSVSIPISMSDVIV
KESQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDIHDENDLDIALR
IGARFIGINSRDPETLEVNKENQRKLISMIPSNVVKVAKSGISERNEIEELRKLGVN
AFDIGESLMRNPEKIKEFIL
>RA_MSD2.3
MPRYLKGPMKRAVQLSLRRPSFRASRQRPIISLKERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYAKFMERYAVGLTIVTFEKYYNGSYETLRKIASAVSIPISMSDVIV
KEYQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDIHDEEDLDIALR
IGARFICIDSRDPETLEVDKENQRKLISMIPSDVVKVAKSGISERNEIEELRKLGVH
AFDIGESLMRNPEKIKEFIE
>RA_MSD2.4
MPRYLKGWLKRAVQLSLRRPSFRASRQRPIISLKERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYAKFMERYAVGLTIVTFEKYFNGSYETLRKIASAVSIPILMSDVIV
KEYQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDIHDEEDLDIALR
IGARFIGIDSRDPETLEIDKENQRKLISMIPSDVVKVAKSGISERNEIEELRKLGVH
AFDIGESLMRNPEKIKEFIE
>RA_MSD2.5
MPRYLKGWLKRAVEISLRRPSFRAERQRPIISLKERILEFNKRNITAIIAEYKRKEP
SGLDVERDPIEYAKFMERYAVGLTIVTFEKYFNGSYETLRKIASAVSIPILMSDVIV
KEYQIDDAYNLGADTVDLIVKILTERELESLLEYARSYGMEPRIDIHDEEDLDIALR
IGARFIGINSRDPETLEVDKENQRKLISMIPSDVVKVAKSGISERNEIEELRKLGVH
AFDIGESLMRNPEKIKEFIE

Locations of PROSS stabilizing mutations

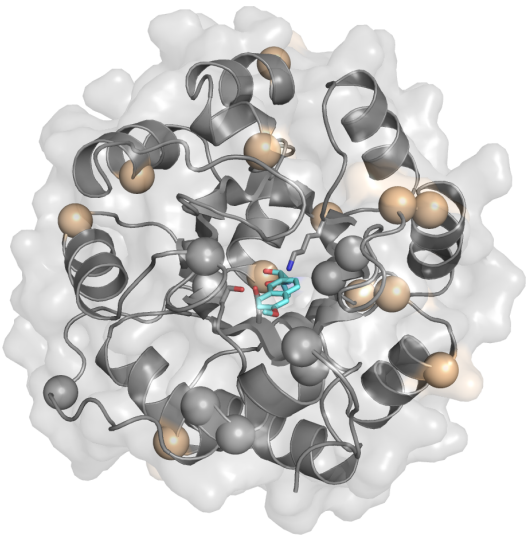
In analogy to Fig. 3.18, the following figures show the locations of stabilizing mutations for variants RA_MSD2.1-RA_MSD2.5



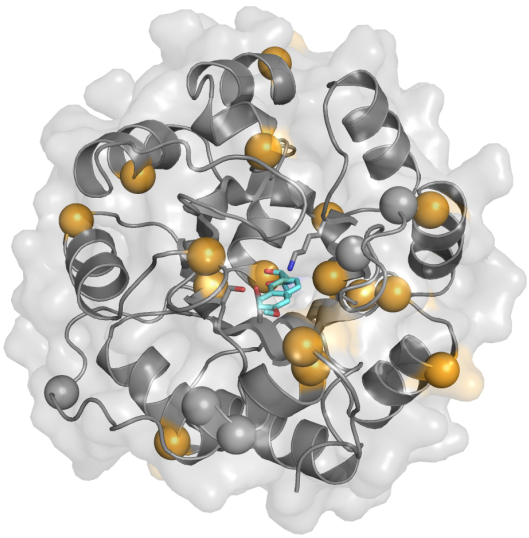
RA_MSD2.1



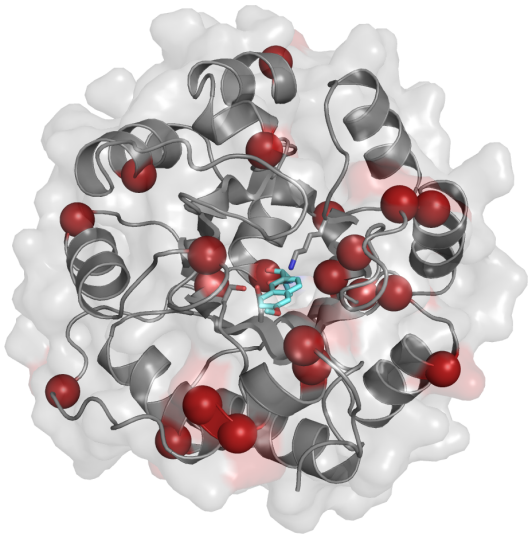
RA_MSD2.2



RA_MSD2.3



RA_MSD2.4



RA_MSD2.5

C.4 List of retro-aldolase sequences (RA*) used for comparison with multi-state variants (RA_MSD*)

For an easier comparison, termini were defined from `N-MPRYL...` until `...PEKIKE-C`. The complete sequence of RA variants contains – depending on the variants - more amino acid at the C-termini and a his-tag (For example `...PEKIKELIEGSLEHHHHHH-C` for RA114.3.

```
>RA41
MPRYLKGWAKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYSRKSP
SGLDVERDPIEYSKFMERYAVGLTIYTEEKYWNGSYETLRKIASSVSIPILMADLIV
KESQIDDAYNLGADTVVLIVPILTERELESLLEYARSYGMEPLIVIVDENDLDIALR
IGARFIKIKSRDWETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFIIGSSLMRNPEKIKE
>RA115
MPRYLKGTLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIASYTRKEP
SGLDVERDPIEYAKFMERYAVGLSILTEEKWSNGSYETLRKIASSVSIPILMKDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIEINDENDLDIALR
IGARFIGINSRDRETWEINKENQRKLISMIPSNVVKVAEKGISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE
>RA114
```

MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKDP
SGLDVERDPIEYAKFMERYAVGLSILTEEKYFNGSYETLRKIASSVSIPILMSDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIIINDENDLDIALR
IGARFIGIASRDWETGEINKENQRKLISMIPSNVVKVAKEGISERNEIEELRKLGVN
AFEIGSSLMRNPEKIKE
>RA116
MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAGYSRKSP
SGLDVERDPIEYAKFMERYAVGLSIMTEEKYFNGSYETLRKIASSVSIPIMMLDFIV
KESQIDDAYNLGADTVLLLVKILTERELESLLEYARSYGMEPLIAINDENDLDIALR
IGARFIGIYSRDPETLEINKENQRKLISMIPSNVVKVAIGGISERNEIEELRKLGVN
AFKIESSLNRNPEKIKE
>RA117
MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKSP
SGLDVERDPIEYAKFMERYAVGLKILTEEKYFNGSYETLRKIASSVSIPIAMSDVIV
KESQIDDAYNLGADTVVLIVKLLTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIGISSRDWETLEINKENQRKLISMIPSNVVKVAISGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA118
MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAGYHRKDP
SGLDVERDPIEYAKFMERYAVGLAIATEEKYANGSYETLRKIASSVSIPIEMWDFIV
KESQIDDAYNLGADTVCLIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSRDFETLEINKENQRKLISMIPSNVVKVASFGISERNEIEELRKLGVN
AFSILSSLMRNPEKIKE
>RA119
MPRYLKGWLEDAVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIALYMRKMD
AGLDVERDPIEYAKFMERYAVGLSILTSEKNHNGSYETLRKIASSVSIPILMWDMIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGISSEDSETLEINKENQRKLISMIPSNVVKVAQSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA120
MPRYLKGWLEDAVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKSP
SGLDVERDPIEYAKFMERYAVGLSILTSEKYFNGSYETLRKIASSVSIPIMMKDMIV
KESQIDDAYNLGADTVKLTVKILTERELESLLEYARSYGMEPLIEINDENDLDIALR
IGARFIGINSRDSETLEINKENQRKLISMIPSNVVKVAQSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA95.5-8

MPRYLKGWLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAIIAYYLRKSP
SGLDVERDPIEYAKYMERYAVGLSIKTEEKYFNGSYEMLRKIASSVSIPILMNDFIV
KESQIDDAYNLGADTVLLIVNILTERELESLLEYARSYGMEPLILINDENDLDIALR
IGARFIVIFSMNFETGEINKENQRKLISMIPSNVVKVAHLDISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE

>RA95.5-8F

MPRYLKGWLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAIIAYYLRKSP
SGLDVERDPIEYAKYMEPYAVGLSIKTEEKYFDGSYEMLRKIASSVSIPILMNDFIV
KESQIDDAYNLGADTVLLIVEILTERELESLLEYARGYGMEPLILINDENDLDIALR
IGARFITIYSMNFETGEINKENQRKLISMIPSNVVKVPLLDFFEPNEIEELRKLGVN
AFMISSSLMRNPEKIKE

>RA114.3

MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKDP
SGLDVERDPIEYAKFMERYAVGLFISTEEKYFNGSYETLRKIASSVSIPILMYDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIIINDENDLDIALR
IGARFIGIAARDWETGEINKENQRKLISMIPSNVVKVAKEGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE

>RA95.0

MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAVYERKSP
SGLDVERDPIEYAKFMERYAVGLSITTEEKYFNGSYETLRKIASSVSIPILMSDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLILINDENDLDIALR
IGARFIGIMSRDFETGEINKENQRKLISMIPSNVVKVAKLGISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE

>RA95.5-5

MPRYLKGWLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKSNITAIIAYYTRKSP
SGLDVERDPIEYAKFMERYAVGLSIKTEEKYFNGSYEMLRKIASSVSIPILMNDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLILINDENDLDIALR
IGARFISIFSMNFETGEINKENQRKLISMIPSNVVKVAKLGISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE

>RA95.5

MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAYYSRKSP
SGLDVERDPIEYAKFMERYAVGLSIKTEEKYFNGSYETLRKIASSVSIPILMSDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLILINDENDLDIALR
IGARFIGIFSMNFETGEINKENQRKLISMIPSNVVKVAKLGISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE

>RA117.1

MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKSP
SGLDVERDPIEYAKFMERYAVGLKILTEEKYFNGSYETLRKIASSVSIPIAMSDAIV
KESQIDDAYNLGADTVVLIVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIGIESRDWETLEINKENQRKLISMIPSNVVKVAIAGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA114.4
MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAEYKRKDP
SGLDVERDPIEYAKFMERYAVGLFISTEEKYFNGSYETLRKIASSVSIPILMYDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIIINDENDLDIALR
IGARFIGIAARDWETGEINKENQRKLISMIPSNVVKVAKEGISERNEIEELRKLGVN
AFVTASGSLMRNPEKIKE
>RA22
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYDRKSP
SGLDVERDPIEYSKFMERYAVGLSITTEEKYFNGSYETLRKIASSVSIPILMADFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSADWETLEINKENQRKLISMIPSNVVKVAAFGISERNEIEELRKLGVN
AFSIHSSLMRNPEKIKE
>RA34.6
MPRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIATYMRKSP
WGLDVERDPIEYAKFMERYAVGLSICTEEKYANGSYETLRKIASSVSIPILMADFIV
KESQIDDAYNLGADTVPLIVKILTERELESLLEYARSYGMEPIIKINDENDLDIALR
IGARFIGICSRDWETLEINKENQRKLISMIPSNVVKVASTGISERNEIEELRKLGVN
AFSIISSLMRNPEKIKE
>RA67
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAKYKRKHP
SGLDVERDPIEYSKFMERYAVGLSIWTEEKYFNGSYETLRKIASSVSIPILMSDFIV
KESQIDDAYNLGADTVVLYVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIEIVSRDLETGEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFSIGSSLMRNPEKIKE
>RA90
MPRYLKGSLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYSRKSP
WGLDVERDPIEYSKFMERYAVGLTILTEEKYFNGSYETLRKIASSVSIPILMSDVIV
KESQIDDAYNLGADTVKLIVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIGILSRDLETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA92

MPRYLKGSLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYTRKHP
SGLDVERDPIEYSKFMERYAVGLSILTEEKYLNGSYETLRKIASSVSIPILMVDLIV
KESQIDDAYNLGADTVVLIVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIGIKSRDFETLEINKENQRKLISMIPSNVVKVALSGISERNEIEELRKLGVN
AFLITSSLMRNPEKIKE

>RA98

MPRYLKGSLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAKYLRKSP
WGLDVERDPIEYSKFMERYAVGLSILTEEKYTNGSYETLRKIASSVSIPILMVDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIEINDENDLDIALR
IGARFILINSRDHETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE

>RA68

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAKYKRKSP
TGLDVERDPIEYSKFMERYAVGLSISTEEKYHNGSYETLRKIASSVSIPILMMDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLILINDENDLDIALR
IGARFIGINSRDYETGETNKENQRKLISMIPSNVVKVAIYGISERNEIEELRKLGVN
AFLISSSLMRNPEKIKE

>RA53

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYSRKHP
SGLDVERDPIEYSKFMERYAVGLSIYTEEKYTNGSYETLRKIASSVSIPILMVDVIV
KESQIDDAYNLGADTVVLIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIVISSYDWETLEINKENQRKLISMIPSNVVKVASGGISERNEIEELRKLGVN
AFSIGSSLMRNPEKIKE

>RA43

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYSRKSP
SGLDVERDPIEYSKFMERYAVGLLIWTGEKYGNGSYETLRKIASSVSIPILMVDWIV
KESQIDDAYNLGADTVLLVVKILTERELESLLEYARSYGMEPLISIYDENDLDIALR
IGARFIKIASRDPETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFVIGSSLMRNPEKIKE

>RA40

MPRYLKGWVKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYMRKSP
SGLDVERDPIEYSKFMERYAVGLTIYTEEKYFNGSYETLRKIASSVSIPILMVDFIV
KESQIDDAYNLGADTVVLFVPILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIKILSSDVETLEINKENQRKLISMIPSNVVKVASHGISERNEIEELRKLGVN
AFSIGSSLMRNPEKIKE

>RA26

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYSRKSP
WGLDVERDPIEYSKFMERYAVGLLILTEEKYFNGSYETLRKIASSVSIPILMHDFIV
KESQIDDAYNLGADTVKLIVKILTERELESLLEYARSYGMEPLIAIHDENDLDIALR
IGARFIGISSRDPETLEINKENQRKLISMIPSNVVKVALSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA63
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIALYMRKSP
WGLDVERDPIEYSKFMERYAVGLSILTEEKYFNGSYETLRKIASSVSIPILMHDFIV
KESQIDDAYNLGADTVKLSVYILTERELESLLEYARSYGMEPLISINDENDLDIALR
IGARFIGIVSRDPETLEINKENQRKLISMIPSNVVKVAISGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE
>RA57
MFRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYMRKSP
SGLDVERDPIEYSKFMERYAVGLSIWTEEKYSNGSYETLRKIASSVSIPILMLDFIV
KESQIDDAYNLGADTVVLIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSRDWETLEINKENQRKLISMIPSNVVKVASHGISERNEIEELRKLGVN
AFTIYSSLMRNPEKIKE
>RA56
MPRYLKGRLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYIRKHP
SGLDVERDPIEYSKFMERYAVGLAIYTEEKYTNGSYETLRKIASSVSIPILMIDFIV
KESQIDDAYNLGADTVVLIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIHSRDWETFEINKENQRKLISMIPSNVVKVATSGISERNEIEELRKLGVN
AFSIYSSLMRNPEKIKE
>RA55
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAYYTRKSP
WGLDVERDPIEYSKFMERYAVGLSILTEEKYFNGSYETLRKIASSVSIPILMTDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSRDWETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFSIVISLMRNPEKIKE
>RA49
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAMYSRKSP
WGLDVERDPIEYSKFMERYAVGLVILTGEKYANGSYETLRKIASSVSIPILMVDWIV
KESQIDDAYNLGADTVVLHVKILTERELESLLEYARSYGMEPLITINDENDLDIALR
IGARFIKISSRDHETLEINKENQRKLISMIPSNVVKVAALGISERNEIEELRKLGVN
AFIIGSSLMRNPEKIKE
>RA48

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAMYSRKSP
LGLDVERDPIEYSKFMERYAVGLAIFTEEKYWNGSYETLRKIASSVSIPILMLDFIV
KESQIDDAYNLGADTVKLSVLILTERELESLLEYARSYGMEPLISIYDENDLDIALR
IGARFILIVSRDPETLEINKENQRKLISMIPSNVVKVALSGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE

>RA46

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYSRKSP
SGLDVERDPIEYSKFMERYAVGLSIYTEEKYWNGSYETLRKIASSVSIPILMVDLIV
KESQIDDAYNLGADTVVLIVSILTERELESLLEYARSYGMEPVIVINDENDLDIALR
IGARFILIKSRDLETLEINKENQRKLISMIPSNVVKVASWGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE

>RA45

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIALYSRKHP
SGLDVERDPIEYSKFMERYAVGLSIWTEEKYVNGSYETLRKIASSVSIPILMVDFIV
KESQIDDAYNLGADTVLLFVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIGIKSRDWETLEINKENQRKLISMIPSNVVKVAMSGISERNEIEELRKLGVN
AFLITYSLMRNPEKIKE

>RA42

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIALYSRKSP
WGLDVERDPIEYSKFMERYAVGLVIATEEKYTNGSYETLRKIASSVSIPILMWDFIV
KESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIKISSMDYETLEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFVIYSSLMRNPEKIKE

>RA6

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAMYSRKSP
WGLDVERDPIEYSKFMERYAVGLVILTEEKYANGSYETLRKIASSVSIPILMVDWIV
KESQIDDAYNLGADTVVLVVKILTERELESLLEYARSYGMEPLIVINDENDLDIALR
IGARFIKISSEDLETLEINKENQRKLISMIPSNVVKVAAHGISERNEIEELRKLGVN
AFLIGSSLMRNPEKIKE

>RA47

MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYMRKSP
WGLDVERDPIEYSKFMERYAVGLAITTEEKYANGSYETLRKIASSVSIPILMADFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSRDWETLEINKENQRKLISMIPSNVVKVASYGISERNEIEELRKLGVN
AFSIYSSLMRNPEKIKE

>RA39

```
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYSRKSP
TGLDVERDPIEYSKFMERYAVGLSILTEEKYFNGSYETLRKIASSVSIPILMTDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIVITDENDLDIALR
IGARFIKILSRDWETGEINKENQRKLISMIPSNVVKVASSGISERNEIEELRKLGVN
AFSIYSSLMRNPEKIKE
>RA36
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYVRKGP
WGLDVERDPIEYSKFMERYAVGLAIATEEKYWNGSYETLRKIASSVSIPILMTDFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSADWETLEINKENQRKLISMIPSNVVKVASFGISERNEIEELRKLGVN
AFAIYSSLMRNPEKIKE
>RA35
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAGYIRKSP
SGLDVERDPIEYSKFMERYAVGLAITTEEKYGNGSYETLRKIASSVSIPILMADFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIISRDWETLEINKENQRKLISMIPSNVVKVASYGISERNEIEELRKLGVN
AFSIYSSLMRNPEKIKE
>RA34
MPRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIALYMRKSP
WGLDVERDPIEYSKFMERYAVGLSITTEEKYANGSYETLRKIASSVSIPILMADFIV
KESQIDDAYNLGADTVALIVKILTERELESLLEYARSYGMEPLIKINDENDLDIALR
IGARFIGIVSRDWETLEINKENQRKLISMIPSNVVKVASYGISERNEIEELRKLGVN
AFSIGSSLMRNPEKIKE
```

List of figures

List of tables

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Dr. Rainer Merkl, for the constant support throughout this thesis and all related research, for his patience, motivation, guidance and the freedom to pursue various projects without objection. He has been an excellent supervisor, being always helpful, honest and encouraging when I faced problems in the course of my research.

I am very grateful to Prof. Dr. Reinhard Sterner who provided the creative, friendly and well-functioning work environment. He always had an open ear, was very helpful, and gave valuable support and advice throughout my work.

I would like to thank Prof. Dr. Wolfram Gronwald for interesting discussion at the Schleusenwärterhaus seminar and for reviewing this dissertation.

Also, I am grateful to Prof. Dr. Jens Meiler, who mentored this thesis in the context of the International Graduate School of Life Science, visited our group several times and gave extremely helpful advice.

A truly outstanding quality of this institute is its terrific staff. I would like to thank all current and former members of the Sterner group and especially the technical and administrational assistants who were always friendly, communicative and easy-going.

Many thanks go to my former student Samuel Schmitz. He is an excellent programmer, an extremely funny human being and I have deeply enjoyed his way of thinking.

I cordially thank all of my collaboration partners for their valuable contributions to this work. Special thanks go to Enrico Hupfeld for believing in retro-aldolase ac-

tivity, and to Samuel Schmitz for countless hours of collective coding and debugging.

My heartfelt thanks go to all current and former members of the Merkl group: I truly enjoyed the working atmosphere, open minded discussions about work and very unrelated topics, the darts/kicker/nerf gun battles and of course the time of solving problems together. Special thanks go to Jan-Oliver Janda for informative discussions, to Julian Nazet who will replace me in the best way possible, to Kristina Heyn without whom we would have missed many schedules, to Leonhard Heizinger the most tech-savvy person in the world who doesn't own a smartphone, and to Maximilian Plach, the biochemist, who has become an integral part of our computational team. Also, many thanks go to Michael Bernhard, Martin Winter, and Philipp Bittner for interesting technical discussions and valuable help.

I would like to thank my brothers, my sister, my parents-in-law, and most importantly my parents for their constant love and unconditional support. Without you, I would not have come that far.

Finally, but by no means least, I would like to thank my own little family. Vicky and Nick, you are the most important people in my life and I deeply appreciate your unlimited support.