Computational *de-novo* design of ester hydrolases

Florian Richter

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

David Baker, Chair

Ron Stenkamp

Phil Bradley

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Computational *de-novo* design of ester hydrolases

Florian Richter

Chair of the Supervisory Committee:
Prof. Dr. David Baker
Biochemistry

Computational protein design is a relatively new technique used to devise amino acid sequences to fold into proteins having novel structures or functions. Here, we first give an overview about the approaches and algorithms used in computational protein design, together with examples of recently reported successful designs of protein structure and function. Then we present in-detail a computational algorithm for one specific application of computational protein design, namely the *de-novo* design of catalytic activity. Next, we apply this algorithm to design catalysts for ester hydrolysis. Several low-activity catalysts were obtained, highlighting both the potential and the challenges inherent in computational design. Finally, we introduce another novel computational procedure for the *de-novo* design of protein backbones that make specific required interactions.

Table of Contents

Chapter 1

Introduction to Computational Protein Design

Abstract

The field of computational protein design (CPD) has made rapid progress over the last ten years. Methods and Algorithms to address a variety of protein design problems have been developed, and are beginning to be used for practical applications. In this chapter, first a brief overview of the potential impact of CPD on synthetic biology and of commonly used general methods in CPD will be given. Next, several recent highlights in the fields of designing protein-protein interactions and catalytic activity will be presented, together with an introduction to the specialized computational algorithms used to approach these problems. A brief look will be taken at thermostabilization and the design of novel protein structures through CPD, and finally, the relative advantages and disadvantages of CPD compared to other protein engineering approaches will be examined.

Introduction

The objective of (computational) Protein Design and Engineering is to create proteins with functions that are not available in natural proteomes. In many synthetic biology applications, just like in natural organisms, proteins are the workhorses that carry out the actual desired functions. But since natural proteins and the functionalities they exhibit evolved to facilitate the survival and maintenance of cells and organisms, the synthetic biologist's toolbox is limited to functions necessary for that purpose. And while the set of naturally available proteins is already very large and can be used to design cells and organisms with novel properties, synthetic biology would benefit tremendously from transcending this barrier and being able to design proteins with functional properties that so far have not naturally evolved. For example, if one wants to engineer a strain of *E. coli* that produces a certain small molecule of interest, one is dependent on the existence of an enzymatic synthesis pathway for said molecule. However, if the molecule of interest is artificial, such as a drug or biofuel candidate, it is unlikely that a natural synthesis pathway exists. In this case, a novel enzyme needs to be designed that catalyzes the desired reaction.

In a way, a synthetic biologist without the capability to create custom-tailored proteins is like an architect who is limited to using only naturally occuring materials like mud, wood, and stones to erect buildings. And while buildings with these materials are good enough for certain applications, the development of more advanced materials like brick, steel and glass immensely increased the size and type of possible buildings. Similarly, once proteins with new functions can be reliably engineered, synthetic biology will take a huge leap forward.

Computational Protein Design is by no means the only method available to engineer proteins. Other approaches such as Directed Evolution, which is discussed in another chapter in this book, have been employed to obtain impressive results. Computational design does however have some unique advantages that allow it to address problems not amenable to directed evolution, as we will demonstrate in this chapter. Conversely, if sufficient high-throughput assays for the function of interest exist, directed evolution is better suited to improve proteins starting from a threshold level of initial activity. Thus, computational design and directed evolution are perfectly complementary, and we anticipate that these two methods will often be used hand-in-hand when designing new proteins for real-life applications.

In this chapter, we will first delineate where we expect computational protein design to have the biggest potential impact on synthetic biology. Then, we will give an overview of the general models and algorithms most often used in computational design. Next, we will give an introduction to the design of novel and specificity-changed binding proteins and enzymes, together with a brief description of the specialized computational algorithms used for these problems. Then, we will give examples of computational thermostabilization of proteins, followed by a brief overview of the design of novel protein folds. Finally, we will compare the relative strengths and weaknesses of computational design vs. directed evolution, and finish with an outlook on where computational protein design could have the most imminent impact on synthetic biology.

The potential impact of computational protein design on synthetic biology

Computational protein design (CPD) is still a relatively young technique, and so far most synthetic applications rely on reusing and recombining existing natural proteins as building blocks. However, as we will show in this chapter, the successes achieved with CPD over the last decade forecast the types of synthetic biology applications and devices that CPD will help enable. We anticipate an impact of CPD in six ways:

1) The design of novel protein-protein and protein-small molecule interactions will allow for the manipulation of signaling cascades to modify gene expression in response to designed, unnatural stimuli
2) The design of protein-protein interactions will also allow for the creation of tailor-made proteins that bind to protein targets and can either inhibit or elicit responses not related to gene expression in a target organism
3) The design of novel catalytic activities will allow for novel biosynthetic pathways for compounds of interest and will also enable the creation of synthetic organisms that break down environmental pollutants or toxins
4) The design of protein-small molecule interactions will allow for the creation of novel biosensors for compounds of interest
5) The design of self-assembling proteins could lead to the creation of novel biomaterials, such as delivery containers for drugs or conductive fibers or sheets for bioenergy applications
6) CPD enables the thermostabilization of proteins with relative ease, and therefore could contribute to increased robustness of synthetic biology applications

For some of these applications, successes using computationally designed proteins have already been reported (Table 1).

Synthetic biology application	Protein design task	Inputs required	Representative examples
1) redesign of cell signaling	Protein-protein interface redesign	Crystal structure of complex	14
2) interfering with target proteins	Protein-protein interface redesign, de-novo protein interface design	Crystal structure of target for de-novo design, structure of complex for redesign	10, 11, 19, 24
3) new catalytic activity / metabolic pathways	Enzyme redesign or de-novo enzyme design	Crystal structure of enzyme with substrate for redesign, theozyme for de novo design	32, 37, 38, 47
4) small molecule binders and sensors	Same as for enzyme design, without catalytic machinery	Crystal structure of wild type receptor with ligand	none yet
5) material design	Protein-protein interface redesign or de-novo protein protein interface design	Crystal structure of scaffold protein	None yet
6) stability increase	Monomer design	Crystal structure of protein of interest	54, 55

For example, regarding point #2, CPD has been used to create proteins that inhibit viral infection or the build-up of amyloid fibrils. Regarding point 3, one case has been reported where a computationally designed enzyme could be used in a novel biosynthetic pathway, and another novel enzyme was designed to break down a

component representative of a class of pollutants. For point #6, several examples have been reported where proteins have been stabilized, leading to higher expression or increased half-lives. And for the types of applications where no examples have been reported yet, it is conceivable to achieve designs with the required functionality using computational algorithms very similar to the ones used to obtain the successful results. We are thus hopeful that CPD will have a broad impact on synthetic biology and will put applications within reach that could not be created otherwise.

Methods overview

The term Computational Protein Design (CPD) as used in this chapter describes the design of amino acid sequences based on computational structural modeling of the to-be-designed protein. While some results have been reported with design algorithms that are not based on an underlying structural model of the protein[1], these approaches will not be described in this chapter.

The computational methods and algorithms typically used in CPD can broadly be divided into three categories:

1) sidechain placement algorithms that, given a model of the protein backbone, select a set of amino-acid sidechain conformations compatible with that backbone. Since the amino acid identities of the selected sidechain set can be any of the 20 natural amino acids (or even unnatural ones), the sequence design happens at this stage.
2) backbone conformation generating algorithms, whose purpose is to generate models of backbone conformations according to the requirements of the specific design task. The backbone models generated by these algorithms are usually passed to sidechain placement algorithms for sequence design.
3) rigid-body placement algorithms, which are used to place two protein models or a protein and a small molecule model in a relative spatial orientation to each other. These algorithms are often the first step when designing binding or catalytic proteins, where one has to design a functional site on one of the proteins and thus needs to first design the spatial relation of the two interacting partners.

CPD algorithms from all three categories generally employ classical molecular mechanics[2] representations of the designed system. All atom models of the protein are used, where bond lengths and angles are usually held constant to increase computational speed. Similar to Molecular Dynamics force fields, CPD energy functions usually contain terms for van der Waals interactions, bond-dihedral potentials, hydrogen-bonding, simplified electrostatics, and implicit solvation[3]. A peculiarity of CPD energy functions is the additional inclusion of sequence composition terms, which are parameterized to make the distributions of amino acid identities in designed sequences similar to those in natural proteins. In this section, we will focus on describing the most often used sidechain placement algorithms (category 1), since these are broadly utilized in virtually all CPD calculations. Examples of category 2 and 3 algorithms will then be presented in subsequent sections, along with the specific design tasks these algorithms are meant to address.

The problem of computational protein design was first presented in 1983 as the so-called inverse protein folding problem.[4] Whereas the task in protein folding and structure prediction is to derive a protein's tertiary structure given the primary amino acid sequence, the objective of the inverse folding problem is, given a certain backbone tertiary structure, to find a sequence that will fold into this template. The inverse folding problem is an extension of the threading problem in homology modeling. In both cases, a sidechain placement algorithm is tasked with finding the optimal combination of sidechain conformations on the template backbone. In threading, only one amino acid, namely that from the wild-type sequence of the threaded protein is allowed at each residue position, whereas in design, all amino acids can be considered at each residue position. For example, for the full sequence design of a relatively small 100 residue long backbone, the sidechain placement algorithm is tasked with finding the optimal sequence out of the astronomically large number of $20^{100} \approx 10^{130}$ possible sequences.

Virtually all sidechain placement algorithms approach this problem by first discretizing sidechain conformational space into a library of so-called rotamers for each amino acid type, where each rotamer represents a frequently observed conformation for that amino acid. This approach can be justified by the observation that amino acid sidechains prefer a limited number of low-energy conformations in high-resolution crystal structures of natural proteins. The library of rotamers for each amino acid type can thus be derived from statistical analysis of protein crystal structures[5], and as a simple rule of thumb, the rotamer library for a sidechain with *n* chi angles will contain 3^n rotamers (with one rotamer in staggered and two rotamers in *gauche* conformation for each chi angle). For example, the rotamer library for valine, a residue with one chi angle contains three rotamers, whereas the rotamer library for lysine (four chi angles) contains 81 rotamers. The combined rotamer library for all 20 canonical amino acids contains 367 rotamers by this rule of thumb.

Using the rotamer concept, the sidechain placement algorithm's task of finding the optimal sequence for a given backbone can be formulated more specifically, namely as the task of finding the set of rotamers that give the lowest energy (as determined by the energy function used) when placed on the template backbone. This lowest energy conformation is often referred to as the GMEC (Global Minimum Energy Conformation). Thus, for the aforementioned 100 residue case, with 367 rotamers being allowed at every position, the sidechain placement algorithm needs to select a combination of rotamers out of $367^{100} \approx 10^{256}$ total possibilities. Even for a smaller problem, such as the redesign of a 20 residue binding site, there are $367^{20} \approx 10^{51}$ possible combinations.

The simplest imaginable side chain placement algorithm would be a brute-force approach that simply enumerates all possible rotamer combinations, scores each of them with the energy function and remembers the GMEC. However, such an enumerative algorithm is evidently impractical considering the large number of possible solutions for even small design problems. Assuming that assembling and scoring one conformation takes a millisecond on modern computer hardware, an enumerative algorithm would take 10^{48} seconds to find the GMEC for the above presented hypothetical 20 residue binding site design problem, which is roughly 31 orders of magnitude longer than the estimated age of the universe. The most important aspect of a viable sidechain placement algorithm is thus its ability to reduce the combinatorial complexity of the problem and select a low-energy rotamer combination within a short amount of time. An in-detail comparison of several algorithms developed for this purpose was done by Mayo *et al.*[6] Today, the most commonly employed ones are Monte-Carlo type algorithms[7] and the so-called FASTER algorithm[8].

Computational Design of Protein-Protein interactions

Protein-Protein interactions are involved in a large number of cellular processes from signal transduction to differentiation to apoptosis and others. Being able to create new or modify existing protein-protein complexes in a rational fashion would thus endow the synthetic biologist with the capability to change cellular behavior at will. Potential synthetic biology applications include the rewiring of signal-transduction pathways to turn on a reporter gene in response to an environmental stimulus, or the design of proteins that bind to functional sites on (and thus inhibit) target proteins.
In this section we will describe methods that are commonly used for the design of protein-protein interactions and introduce examples of several studies done in this regard so far. The goals in CPD of protein interactions can broadly be divided into two areas: redesigning existing interactions towards higher affinity or changed specificity, and the design of novel binding interactions. Depending on the specific problem, either the sequence of both binding partners in the complex may be modified by the design algorithm ("two-sided design"), or the algorithm is only allowed to design the sequence of one partner while leaving the other partner constant ("one-sided design").

Computational redesign of protein-protein interactions

There are usually two motivations to redesign an existing protein-protein interface: 1) increasing the affinity to make the interface more stable and 2) changing the specificity of the interface, meaning to redesign the interface in such a way that the affinity of one pair of desired binding partners is retained, while the affinity towards another, competing binding partner is reduced. In both cases, a structural model of the to-be-redesigned complex, preferably a crystal structure, needs to be available, since this serves as an input for the CPD calculations. When the goal is to increase the affinity, the computational workflow used is usually some iterative combination of rigid-body docking algorithms developed for virtual protein docking[9] and general sidechain placement algorithms, while also taking into account a set of empirical rules governing affinity. When the goal is to modify specificity, these docking and sidechain placement algorithms need to be augmented by specialized algorithms that take into account and penalize the competing states.
There are several representative examples of increasing affinity by computational design. Roberts *et al.*[10] presented a study where a peptide inhibitor of a PDZ domain involved in cystic fibrosis was redesigned for higher affinity towards its target. Starting from an NMR structure of the natural ligand – PDZ domain complex, three mutations were introduced into the sequence to obtain a hexameric inhibitor that had 170-fold increased activity compared to the natural ligand. Lippow *et al.*[11] applied computational design to the problem of antibody affinity maturation, which is a field of broad therapeutic significance. In their work, the authors increased the affinity of two antibodies, one, a lysozyme-binding model antibody, by 140-fold through mutation of four residues, the other, an epidermal growth factor receptor binding therapeutic antibody, by 10-fold through mutation of three residues. Computational design was also used by Haidar *et al.*[12] to introduce four mutations into a solubilized T-cell receptor, increasing the affinity toward its cognate peptide MHC complex by 100 fold. The redesigned receptor is potentially better suited than the wild type for diagnostics applications.

There has also been significant progress in the field of specificity redesign in the last several years. Perhaps the most challenging aspect of this problem from the standpoint

of computational design is the need to incorporate 'negative design' into the calculation, meaning to consider the unwanted, competing stages and design a sequence that disfavors these. Another side effect of this requirement is that the designed sequences might not have the highest possible affinity against the target of interest, since the identities of the interface positions are not just determined by how well they interact with the target state, but also how well they discriminate against the unwanted states. In virtually all CPD algorithms, the designed sequence is a product of the sidechain placement algorithm, but most of these algorithms, i.e. Monte-Carlo schemes and FASTER (see section above) have been developed to optimize the energy function for a single state, and extending them to consider multiple states is not trivial. Therefore, algorithms specifically developed for this purpose need to be employed.

An elegant and very general approach to this problem was presented by Leaver-Fay *et al.*[13]. In this approach, the input to the calculation is a set of backbone models together with an algebraic rule of how to sum the scores (as determined by the energy function) of a certain sequence on each backbone model into one value. During the calculation, the sequence search is done by a genetic algorithm, and the value that is being optimized is the user-specified sum over the scores of all considered states instead of the score of a single state. When calculating the sum, absolute as well as relative scores of states can be considered in positive (for the unwanted states) or negative (for the desired states) fashion. In their work, the authors used this general framework to design sequences that *in silico* have favorable interaction scores for a set of desired states while having unfavorable scores for another set of competing, unwanted states. As this approach is still very recent, no experimental data on sequences designed with it exists yet, but it does represent the most general and comprehensive theoretical approach to the problem.

Several illustrative demonstrations of successful specificity redesign have been reported in recent years, some of them with direct applications in synthetic biology contexts. In a textbook example of two-sided design, Kapp *et al.*[14] designed an orthogonal GTPase / GEF pair, which could both be used as a valuable tool to study cell signaling as well as serve as a component in a synthetic signaling pathway. Starting from crystal structures of the GTpase Cdc42 and its activator GEF, intersectin, the authors first identified positions in Cdc42 at which mutations would interfere with intersectin binding without disrupting any of the known interfaces with other binding partners or the active site. After identifying one position (Phe-56), a cognate position on intersectin was identified where a salt-bridge could be introduced if Phe-56 was concurrently mutated to Arg. The mutated GEF stimulated nucleotide exchange in the mutated GTPase but not in the wild-type, while the mutated GTPase could be activated by the mutated GEF but not the wild-type, demonstrating the orthogonality of the new pair. The new pair retained (albeit lower than wild-type) signaling activity *in vivo*.

Grigoryan *et al.*[15] introduced a computational framework that explicitly considers competing states. In this study, the authors set out to design proteins that interact with individual members of a class of transcription factors known as bZIPs. This class comprises about 20 families, which share extensive structural and sequence similarity, making the design of inhibitors specific for only a subset very challenging. To address this problem, the authors developed an algorithm that first designs a sequence with highest possible affinity for the target, and then in a second step modifies the found sequence to increase the gap between the target state and the nearest competing state. 46 designs were characterized, ten of which interacted more strongly with their intended target than with any competitor. However, to make the used algorithm computationally tractable, a scoring function specifically developed for this class of proteins had to be used, making this approach not easily extensible to other problems.

In another study, Yosef *et al.*[16] redesigned the calcium dependent second messenger protein calmodulin towards preferably binding only one of its two major interaction partners. One of the designs, containing six mutations, had significantly reduced affinity towards the undesired partner, but retained affinity for the desired partner, resulting in a 900-fold specificity switch. Several more examples of successful computational redesign of protein-protein interactions have been described in a recent review.[17]

Computational design of novel protein interactions

Complementary to redesigning existing protein interfaces, CPD can be used for the *de novo* design of protein interactions. Generally, the objective in a *de novo* interface design problem is to design a protein that binds a target protein of interest, starting from a structural model (preferably a crystal structure) of the target protein alone. The computational workflow used to tackle this problem can roughly be divided into three steps:
1) choosing a 'scaffold' protein, i.e. a protein that can serve as the backbone template that the binding sequence can be designed onto
2) finding a productive relative spatial orientation of the target protein and the scaffold
3) designing the amino acid sequence of the scaffold (and potentially the target) to stabilize this spatial orientation

The first two steps represent computational challenges unique to this problem. In step 1, the scaffold protein can either be taken from a library of existing proteins for which the crystal structure is known, or it can be designed *de-novo* using category 2 CPD algorithms. In step 2, after a scaffold has been decided upon, category 3 (rigid-body placement) algorithms then need to be used to place it in a spatial orientation towards the target, thus giving an initial model of the complex. Essentially, the 'global' shape of the complex and the location of the binding interfaces on the two partners are determined in step 2. This model is then passed to sidechain placement algorithms (or to the specialized algorithms developed for protein interface design as described above) to design the novel amino acid sequence.
Arguably, step 2 is the most critical stage in this workflow, because it is necessary to find a spatial orientation that is 'designable', i.e. where a sequence can be designed such that the resulting binding interface has sufficient size, shape complementarity, and interactions for the complex to be energetically favorable compared to the unbound state and thus lead to a high-affinity interaction. Moreover, many applications require that a particular region of the target-protein is part of the binding interface, meaning the rigid-body placement algorithm needs to be able to orient the scaffold to optimally interact with the desired surface patch on the target. In recent years several impressive examples of successful *de novo* protein interface design have been reported, using different approaches for steps 1 and 2. These different approaches can be divided into two groups: general approaches that can in principle be used to design a binder for any arbitrary target protein, and more specialized approaches that take advantage of certain structural properties of the target.
Perhaps the most general method for scaffold selection and placement is Fleishman *et al*'s so called "Hotspot-design" method[18], which was recently used to design proteins that bind to a conserved region of influenza hemagglutinin and inhibit this protein's ability to undergo conformational changes that underlie influenza infectiousness[19]. This method is based on the observation that in many natural protein-protein interfaces, the affinity is mostly mediated by a small subset of the residues making up the interface. When

making a series of single-point mutants of a binding protein, in which residues belonging to the interface have been mutated to alanine, and measuring the resulting binding affinities, usually only a subset of variants will have significantly reduced affinities. The residues that were replaced by alanine in these variants are thus the most important interface residues, and are often referred to as 'hotspot' residues[20]. From this observation, Fleishman *et al.* reasoned that a viable strategy to design novel interfaces would be to first place a small number of side chains in favorable locations on the target interface, and then use these proto-hotspot residues as anchors that the rest of the interface is designed around. In the Hotspot-design method, a small (usually on the order of 1-3) set of disembodied amino acids is placed in the vicinity of the target-protein surface patch in an energetically favorable fashion. For example, if there are exposed hydrophobics on the target surface, an aromatic amino acid might be chosen as the Hotspot residue, or in case unsatisfied hydrogen-bonding atoms are present, an amino-acid with complementary hydrogen-bonding functionality could be picked. The exact location for each hotspot residue can either be set explicitly or found with the help of ligand-docking approaches[21]. Once the hotspot residues have been placed, a candidate scaffold protein is placed such that the backbone atoms of the hotspot residues are superimposable onto the backbone atoms of any one residue of the candidate scaffold, without the scaffold backbone overlapping with the target protein. To achieve this, protein-protein docking algorithms are used, where the target protein is held constant and the candidate scaffold is docked. To guide sampling, coordinate restraints between the candidate scaffold and the hotspot residues are imposed during the docking calculation. After the scaffold has been properly placed, all residues in the vicinity of the target protein (except the hotspot residues) are redesigned with sidechain placement algorithms to obtain the final designed sequence. In practice, after the hotspot residues have been placed, the docking and design procedure is carried out millions of times, with the candidate scaffolds picked from a library of several hundred small, monomeric, globular proteins for which high-resolution crystal structures are available. The resulting designs are then ranked by energy, and a few dozen of the highest ranking designs are considered for expression.

The most impressive example of a successful execution of the hotspot-design strategy is represented by Fleishman *et al.*'s designed protein that binds to an evolutionary conserved surface region of influenza hemagglutinin, referenced above. Hemagglutinin is a protein on the flu surface that plays a vital role in the virus' capacity to infect cells, undergoing a conformational change when the virus binds to a host cell and gets endocytosed. After inspecting a co-crystal structure of hemagglutinin with a broadly neutralizing antibody, the authors placed hotspot residues in the vicinity of the antibody binding site, and were subsequently able to design two proteins that bind hemagglutinin, one with a K_D of 200nM and one more weakly. After affinity maturation, the binding affinity was increased to a K_D of 22nM and 38nM for the two designs, and the designed proteins inhibited the conformational change that hemagglutinin undergoes during infection. X-ray crystallographic analysis confirmed that the designed proteins bound in the intended location and orientation. It is well conceivable that such designed proteins can play a role as therapeutics in the future.

Another general approach for steps 1 and 2 is represented by Jha *et al.*'s[22] so-called DDMI method ("Docking, sequence Design, and gradient-based Minimization"). In this method, first protein-protein docking methods are used to place a candidate scaffold (taken from a library) and the desired target in close proximity to each other. During the docking phase, the energy function is augmented with restraints that direct the scaffold protein towards the desired binding site on the surface of the target protein. At the end of the docking phase, a rough score filter is used to determine whether the candidate

complex will be discarded or subjected to the next, sequence design phase of the protocol. During this phase, the docked complex model is then subjected to iterative rounds of sequence design and gradient-based minimization, and resulting sequences are selected for expression based on a combination of energy function scores and structural parameters. The authors used their method to generate six designs against PAK1, a human kinase, the best of which bound with an affinity of 100 μM to the desired target.

Several other approaches for steps 1 and 2 exist that rely on the target protein to have certain structural features. While none of them are as general as the hotspot-design or DDMI approach, they have been shown to yield high-affinity binders against targets having the required features.

Two cases of exploiting solvent-exposed edges of beta-sheets to form a binding site have been reported. This strategy makes use of the fact that such edges have accessible unsatisfied hydrogen-bonding functionalities, which can be bound relatively easy by any scaffold that also has a beta-sheet with an exposed edge, thus forming an intermolecular beta-sheet upon complex formation. The residues surrounding this beta-sheet anchored interface are then redesigned to further increase the affinity and modulate specificity beyond that provided by the sheet-contacts alone. Stranges *et al.*[23] used this approach to turn a monomeric protein into a symmetric homodimer. Starting from a library of proteins with exposed beta-strands, the authors generated initial models consisting of two copies of the monomer interacting through the strand, followed by, like in most design protocols, cycles of sequence design and minimization. Four designs were experimentally characterized, the best of which had a K_D of 1 μM, and a crystal structure showed that the interface was virtually super-imposable onto the design model. Sievers *et al*. used a very similar approach to design peptides that inhibit amyloid formation.[24]

The strong interactions that some amino acids can make with metals can also be harnessed to design protein interfaces. The same way in which a metal site can significantly stabilize a protein structure, it can also be used to stabilize an interface. The coordination spheres of metals usually contain between four and six ligand sites. Thus, if a site containing half of a metal coordination sphere can be designed on the surface of a protein, two copies of this protein could then dimerize and form the complete metal site. Der *et al.*[25] demonstrated the feasibility of this approach by designing a zinc-mediated homodimer from a monomeric scaffold protein. The design had a very tight interface (K_D < 30nM) in the presence of zinc, but in the absence thereof the K_D increased ~100 fold, indicating the importance of the metal. Interestingly, as demonstrated by Salgado *et al.*[26], this approach can also be used as a stepping stone to interfaces not dependent on metal. In their example, the authors first created a minimalist interface featuring a zinc-binding site, and then used CPD to create additional interactions across the interface, resulting in a design where binding was independent of metal presence.

Finally, another viable strategy to create novel protein binders against a given target is the so-called grafting of functional epitopes. This method relies on the availability of an already existing binder for the target of interest, and preferentially a crystal structure of the complex between the two. In this approach, the binding residues of the existing binder are transferred onto another scaffold, thus endowing this scaffold with binding affinity for the target of interest. This method is useful in cases where, for example, the original scaffold is difficult to express or problematic in other ways. Further, as demonstrated by Sia *et al.*[27], protein grafting can lead to higher binding affinities in cases where the binding epitope is a short, relatively unstructured peptide that is grafted onto a more rigid scaffold. Grafting strategies don't necessarily require computational modeling of the system, however, as shown by Azoitei *et al.*[28], computational algorithms allow for

the atomically exact modeling of the connection between the grafted epitope and the new scaffold. In their study, the authors took a discontinuous epitope (i.e. an epitope consisting of two backbone segments) from the HIV protein gp120 that a broadly neutralizing antibody, b12, had been raised against, and, after scanning a large scaffold library for complementary to both parts of the epitope, then transplanted both segments onto a different scaffold in the same relative orientation. After creating a library (also guided by computational design) around the grafted epitope, a design was obtained that bound b12 with similar affinity as the original viral protein. In theory, when challenging the immune system with these designs, they could raise antibodies with similar binding properties as b12, and thus this strategy could be used to design vaccines.
In summary, several methods have been developed recently for the *de novo* design of protein-protein interactions. Which method is most suited in a given situation must be decided on a case-by-case basis. If the target of interest has certain structural features that can be exploited to create a binding site, then approaches like beta-strand assembly or metal-mediated interfaces can be considered. If other binders against the target protein already exist, epitope grafting or the hotspot-design method are good strategies. If neither is available, DDMI can be used, or ligand-docking algorithms can be used to generate inputs for hotspot-design.

Computational Design of catalytic activity

Of equal importance in biology as protein-protein interactions, the ability of enzymes to accelerate chemical transformations is involved in virtually every biological process. And just like being able to rationally engineer protein binding, being able to rationally design new or modify existing catalytic activity would tremendously expand the synthetic biologist's horizon. Potential synthetic biology applications are the redesign of metabolic pathway enzymes to yield novel biofuels or the *de novo* design of enzymes with catalytic activities not available in the repertoire of natural enzymes, which would allow for these non-biological transformations to be included in biological contexts and pathways.

In this section, we would like to highlight the advances made so far in the computational design of catalytic activity. Broadly speaking, there are three types of problems in the realm of computational enzyme design. These are, in order of increasing difficulty:

1) redesign of an existing enzyme active site to process a different substrate (so called "specificity redesign")
2) redesign of an existing enzyme active site to catalyze a different type of chemistry ("reactivity redesign")
3) *de novo* design of novel catalytic activity into a previously inert scaffold

Arguably, (re-)designing an enzyme is more challenging than designing a binding protein. For the latter, it is enough to present a rigid surface that is complementary in shape to the design target, and usually only one target needs to be recognized. An enzyme however at a minimum needs to be able to bind the substrate(s) with sufficient affinity, then stabilize the transition state with even higher affinity, but then not have too high affinity for the product. In addition to that, often residues in the enzyme active site show large pKa shifts or form covalent intermediates with the substrate(s), and thus the enzyme's active site environment needs to modulate active-site residue reactivity. This necessity to concurrently satisfy several, sometimes conflicting criteria has led to enzymes being referred to as 'masters of compromise'. Further, while the biophysical

principles governing binding are generally understood, and significant insight into the principles underlying enzymes' catalytic power has been obtained[29], there are still open questions about the relative importance of certain phenomena for catalysis, such as the trade-off between active-site dynamics and preorganization[30]. However, developing an algorithm to approach a certain CPD problem necessitates understanding of the biophysical factors playing a role in the problem, and if this understanding is incomplete, the resulting algorithms might be lacking critical features. Thus, while impressive early results have been achieved in computational enzyme design, the procedure is still less reliable than binding or thermostability design.
Another computational challenge unique to enzyme design is that to accurately model chemical reactivity, a quantum-level treatment of the system is required, for example to assess the effect of the active site electronic environment on the energy levels of the substrate's molecular orbitals. All CPD energy functions employ a classical physical representation of the system however, and since billions of different sequences are usually considered by the sidechain placement algorithm, using a more accurate quantum-level model would be prohibitively slow. Thus, the energy function is essentially blind with respect to the catalytic competence of the designed active site. The most-often used workaround for this problem is based on restraining the identity and allowed geometry for a handful of the active site residues[31].
In this section, we will introduce methods used and successful examples for each of the three cases.

Computational redesign of enzyme specificity

The objective of specificity redesign usually is to take an existing enzyme and change it such that it transforms a different substrate. Generally, the catalytic residues of the enzyme active site (i.e. the residues that mediate the chemical steps of the reaction), and consequentially the type of chemistry and mechanism that the enzyme performs, remain unchanged. Only those active site residues that play a role in binding the substrate are modified by the design algorithm. Thus, the new substrate needs to be similar to the enzyme's natural substrate, in that it must feature the same reactive moieties that are acted upon by the enzyme. The new substrate is only allowed to differ in the non-reacting parts of the molecule.
As for most other CPD projects, a structural model, optimally a crystal structure, of the to-be designed system is required as the input. For specificity redesign, the optimal starting point is a crystal structure of the enzyme of interest in complex with its native substrate, or with a substrate or transition state analog. Additionally, a basic understanding of the catalytic mechanism and knowledge of the most important active site residues is required, to prevent the design algorithm from mutating away these critical residues. The first step is to generate a model of the new desired substrate. If this substrate features rotatable bonds, an ensemble of possible conformations also needs to be generated. Usually, this ensemble has restricted diversity in the moiety of the target substrate that resembles the wild type substrate, but full diversity in the differing regions. The target substrate ensemble is then superimposed onto the wild type substrate, such that the shared moiety is in the same region of the active site and making identical contacts to the catalytic residues. Next, the sidechain placement algorithm is used to design a sequence that accommodates the new substrate. In this step, besides sampling the identity and conformational diversity of the active site residues, the substrate's conformational ensemble is also sampled. Usually, only the subset of residues that contact the differing moiety of the target substrate is allowed to mutate. After running the sidechain placement algorithm, the resulting structure is

usually refined through gradient-based minimization, and these two steps are iterated several times. If a stochastic sidechain placement algorithm is used, the calculation is carried out several hundred or thousands of times, leading to a large number of models. The best scoring models are selected, and depending on the throughput of the available experimental setup, anywhere between a handful to 100s of sequences are experimentally tested for the desired activity.

Several examples of successful computational redesign of enzyme specificity have been published in recent years, with applications in different fields. Chen *et al*. set out to redesign the specificity of the phenylalananine adenylation domain of the gramicidin S synthetase A (GrsA-PheA)[32]. This domain is part of the nonribosomal peptide synthetase (NRPS), which is a large multi-domain enzyme complex that assembles peptides in an assembly-line manner. Many of the product peptides have antimicrobial properties and are thus of pharmacological interest. Being able to redesign the specificity of individual NPRS domains could yield novel, unnatural peptides with potentially improved properties. In their work, Chen *et al.* succeeded in redesigning GrsA-PheA to accept several different substrates instead of the native substrate phenylalanine, namely leucine, arginine, aspartate, glutamate, and lysine. Their most successful variant, a redesign for leucine, had a 2168 fold increased preference for leucine over phenylalanine compared to the wild type enzyme, while maintaining about 1/6th of the catalytic proficiency of the native enzyme.

Another example of enzyme specificity redesign is Ashworth *et al.*'s redesign of the DNA cleavage site of the homing endonuclease I-MsoI[33]. Homing endonucleases are DNA-cutting enzymes that recognize target sites of ~15 nucleotides length, as opposed to the ~6nt cut-sites of restriction enzymes, and cut these with high specificity. Since the cleavage sites are fairly long, most homing endonucleases' cut-sites only occur once per genome, meaning that homing endonucleases are potentially valuable tools for genome engineering applications. Ashworth *et al*. succeeded in changing the specificity of I-MsoI for one base-pair in its recognition sequence, creating a variant that cleaves the new target site 10^4 more efficiently than the wild type enzyme, while having activity comparable to the wild type enzyme and with good discrimination against the original cleavage site. Ashworth *et al.*'s work represents an important first step towards the ultimate goal of being able to design an endonuclease for any cleavage site. Potential synthetic biology applications include the manipulation of specific genetic loci in living organisms, such as the introduction of new traits in plants[34] or the genomic engineering of entire mosquito populations.[35]

Murphy *et al*. used computational design to change the specificity of a human guanine deaminase[36], towards accepting ammelide instead of guanine as a substrate. The achieved specificity switch was $2.5*10^6$ fold, albeit the designed enzyme had significantly reduced activity. Ammelide is a structural intermediate between guanine and cytosine. There are no human cytosine deaminases known. Therefore, if the specificity switch could be extended towards cytosine, the resulting enzyme could find application as a prodrug-activating enzyme with low immunogenicity. To achieve their results, Murphy *et al.* used a more advanced computational algorithm, where, similar to the hotspot-design method, first a disembodied sidechain was placed in ideal relation to the new substrate and then a segment of nearby backbone was remodeled to support this desired side-chain placement.

Lippow *et al.* succeeded in turning a galactose 6-oxidase into a novel glucose 6-oxidase.[37] This designed enzyme could serve as the starting point in a designed efficient metabolic pathway for the synthesis of the value-added chemical D-glucaric acid in *E. coli*. Because no crystal structure of the wild type enzyme with substrate was available, the authors first had to create a model for galactose in the wild-type active site through *in*

sillico docking. Since a medium-throughput plate screening assay was available, the authors ran the design algorithm several thousand times to yield 2379 unique sequences, and then devised a strategy to create a library of size 10^4 that encompassed the sequence diversity generated by the computational algorithm. 10^4 clones were screened and 402 hits were found (3.8% hit rate), one of which had a 400 fold increased activity for glucose and, though still having better activity with galactose, the preference for galactose over glucose was decreased 13000 fold.

Computational redesign of enzyme reactivity

The redesign of enzyme reactivity, while still considered redesign, represents a more drastic intrusion into the redesigned enzyme than mere specificity redesign. In this approach, the active site of an existing enzyme is redesigned, or 'repurposed', to catalyze a different type of reaction. Arguably, the task becomes easier the more similar the two reaction types are. A class of enzymes inherently suited for this approach is metalloenzymes. In many metalloenzymes, the metal carries out the essential chemical step, while the surrounding active-site amino acids serve to bind and orient the metal and the other (usually organic) substrate(s) in the proper orientation to each other, while also activating relevant functional groups of the organic substrate. Depending on what type of organic functional group is bound proximal to the metal, the same metal center can catalyze different chemistries. For example, metals such as zinc have an inherent affinity for water molecules. An H_2O molecule has reduced pKa when bound to zinc, and the resulting ZnOH species is a more potent nucleophile than unactivated H_2O. Depending on the electrophile present, the zinc bound OH^- could then either undergo a hydrolytic or a nucleophilic addition reaction.
In the inaugural example of this approach, Khare *et al.* redesigned an adenosine deaminase into an organophosphate hydrolase.[38] The wild-type enzyme, which was part of a set of enzymes featuring mononuclear Zn-sites with at least one of the Zn-coordination sites not occupied by a sidechain, contains a Zn binding site that activates a water molecule for nucleophilic attack onto the amino group of adenosine. In their work, the authors first placed the design substrate, a model organophosphate featuring an activated leaving group, in the active site such that the Zn was coordinating to the phosphate's keto-oxygen, thus rendering the phosphorus atom more susceptible to nucleophilic attack by a water molecule. Next, the sidechain placement algorithm was used to redesign the surrounding active site residues to accommodate the new substrate. During this step, the Zn-coordinating residues were held constant. The resulting design featured eight mutations and hydrolyzed the model substrate with a k_{cat}/K_M of 4 M^{-1} s^{-1}, and after three rounds of directed evolution, a variant with a total of 13 mutations and a k_{cat}/K_M of ~10^4 M^{-1} s^- for the designed substrate was obtained. Further analysis of the importance of each of the mutations indicated that a minimal set of four mutations was absolutely required to confer hydrolytic activity for the target substrate. This suggests that obtaining a comparable result through directed evolution alone would necessitate screening of an enormous library containing all possible quadruple mutations of the protein.

Computational *de novo* design of enzyme activity

The third and most difficult type of problem in computational enzyme design is the design of catalytic activity from scratch. In this case, both a catalytic mechanism as well as a protein site to carry it out need to be devised. The currently most viable approach

consists of developing a so-called theozyme for the reaction of interest, and then trying to graft this theozyme into a scaffold protein. A theozyme, or "theoretical enzyme", is a three-dimensional model of a minimal active site necessary to catalyze the desired reaction[39]. Usually it consists of a model of the energetically highest transition state on the reaction pathway, together with a set of disembodied amino acids placed around it that are meant to stabilize this state or perform chemical transformations on it. For example, if negative charge develops on a certain substrate atom over the course of the reaction, a strategy to stabilize this buildup and thus accelerate the reaction would be to place a positively charged sidechain, such as an arginine or lysine, next to this substrate atom. If the substrate gets deprotonated, a protic amino acid, such as a glutamate, aspartate, or histidine could be placed next to the mobile proton in the theozyme. Several strategies to develop a theozyme for a given reaction of interest have been devised. The most comprehensive and most difficult one represents an approach using quantum-mechanical modeling as described by Zhang *et al.* [40] Other, more *ad hoc* approaches relying on chemical intuition or biological precedent will be described below.

Once a theozyme has been devised, the next step in the process is to graft it into a protein scaffold. In this step, called matching, a library of scaffold protein structures is searched for attachment sites for the theozyme. A theozyme can be attached in a scaffold if the theozyme ligand can be placed in a scaffold cavity and if simultaneously every theozyme sidechain can be grafted onto a scaffold backbone position, such that the desired relative geometric orientation between theozyme sidechains and theozyme ligand is maintained, with no clashes between the theozyme components and the scaffold backbone. Thus, the rigid body orientation between the substrate and the to-be designed enzyme is determined at this stage, and hence matching algorithms fall into category 3 of computational protein design algorithms.

Several matching algorithms have been proposed, such as a simple enumerative algorithm that places each side chain sequentially[41], an algorithm that places the ligand for each theozyme side chain in parallel and then employs six-dimensional geometric hashing to determine ligand positions that can make all desired theozyme contacts,[42] or an algorithm that scans through all pairwise combinations of scaffold residues to determine geometric overlap with theozyme sidechains.[43]

Independent of the matching algorithm used, the result of the matching stage is a set of so called matches, which are models featuring the desired minimal active site in a protein cavity. However, the theozyme usually only contains a handful of sidechains (2-4), while in the matches usually several dozen sidechains are within the shell contacting the placed ligand. This means that only a subset of the sidechains making up the new active site have been assigned their ideal identities at this stage. To determine sequences identities for the remaining active site residues, standard sidechain placement algorithms are run in the next stage of the design process. Usually several iterations of sequence design and gradient-based minimization of the designed sequence are carried out. During this stage, the original theozyme residues are not allowed to change their identities any more, and the energy function is augmented with terms that ensure that the theozyme residues stay in their desired theozyme geometry[31]. If stochastic side chain placement algorithms are used, this stage is usually carried out several dozen times for each match. Depending on the initial number of matches, 100 to 10000 design models are generated. The models are then usually filtered and ranked according to several criteria reflecting the degree of realization of the ideal theozyme geometry, the affinity of the designed site for the new ligand, and the structural integrity of the scaffold.

The current strategy of designing a novel catalyst by placing a theozyme into a known protein structure is rooted in two observations. First, many enzymological studies have

shown that the catalytic prowess of natural enzymes is usually caused by a small number of catalytic residues, while many other residues that are also found in the active site play a far less important role in catalysis. Knocking out the catalytic residues, i.e. replacing them with chemically inert alanine, will have a huge effect on efficiency of the enzyme, while changing other binding site residues to Ala often has no influence on the activity at all. One conclusion from this observation is that to design an active site, it is most important to properly position a handful of key catalytic residues.

Second, computational protein design as a technique is still far from perfect, and, despite Kuhlman *et al.*'s work, it is still unclear whether current algorithms and force fields are reliably capable of designing sequences that fold into a given protein architecture or fold. Further, as of yet there are no algorithms that would design a reasonable protein backbone conformation starting from a theozyme only. Placing a theozyme into an already known structure and then designing only the surrounding site, which usually means mutating less than 10% of the scaffold residues, circumvents both of these problems. Since the backbone structure is given by the scaffold there is no need to come up with a new one, and since the final designed sequences is highly homologous to the wild type scaffold sequence, one can reasonably assume that the designed protein will be expressed and folded.

Several examples of successful *de novo* computational enzyme design have been accomplished in recent years. These include two reports[44,45] of designed catalysts for the Kemp elimination, a well-studied model reaction, one set of proteins catalyzing a retro-aldol reaction through a similar mechanism as natural aldolases[46], and one study reporting the design of a “diels-alderase”[47] catalyzing the Diels-Alder reaction, which is a carbon-carbon bond forming reaction for which no natural enzymes exist.

The Kemp elimination is an isoxazole ring opening reaction that can be initiated by deprotonation of the carbon adjacent to the azole nitrogen, and the substrate used in both studies featured a benzyl ring adjacent to the isoxazole moiety. In the transition state, negative charge accumulates on the isoxazole oxygen. Thus, for the Kemp eliminases, in both studies the theozyme consisted of three elements: a base (either a Glu/Asp or His) to deprotonate the relevant carbon, a hydrogen-bond donor to stabilize the isoxazole oxygen, and an aromatic side chain (Phe/Tyr/Trp) to stack against the substrate’s phenyl group and thus aid with binding the substrate. In the work by Roethlisberer *et al.*[44], a scaffold library was then scanned for matches to this theozyme, and after designing the sequence and ranking the resulting matches, 59 designs were selected for expression. Eight of the designs showed detectable activity, and for each of the designs, replacing the catalytic base by a non-protic residue resulted in significant decrease of the catalytic activity.

In the second study, by Privett *et al.*[45], a slightly different approach was followed. Instead of searching through a number of scaffolds, the authors identified an aspartate in a hydrophobic cleft of a xylanase, and chose this residue to be the catalytic base. The substrate was placed in a position to interact with this residue, and the surrounding pocket was then designed to accommodate the substrate in this position. The resulting design had activity comparable to the designs reported by Roethlisberger *et al*.

Jiang *et al.*[46] succeeded in designing proteins catalyzing a retro-aldol reaction, which is a multistep transformation capable of breaking carbon-carbon bonds. Several natural enzymes exist that accelerate the complementary reverse reaction. The most critical active site element of these natural aldolases is a strategically placed lysine that carries out a nucleophilic attack on a ketone moiety of the substrate to form a Schiff-base covalent adduct and thus initiate the reaction. Several other protic residues act in concert with the lysine to carry out a number of proton shuffling steps necessary to complete the reaction. Consequently, Jiang *et al.* picked this lysine as the central

element of the theozyme used for the retroaldolase designs. The authors used three variations of the theozyme: one featuring the lysine as the only amino acid but with explicit water molecules to carry out the proton transfers, the other two more similar to the natural active sites, with different combinations of protic sidechains supporting the lysine. Out of a total of 72 designs that were experimentally characterized, 32 showed activity, albeit at far lower efficiencies than natural enzymes. Somewhat surprisingly, designs based on the first, simple theozyme showed a higher rate of success than those featuring the more complicated theozyme.

Siegel *et al.* presented the first example of an enzyme catalyzing a non-natural reaction, in designing the first biocatalyst for a Diels-Alder reaction[47], which is a cycloaddition reaction that simultaneously forms two carbon-carbon bonds and four stereocenters and is thus highly useful in organic synthesis. This study demonstrates that, in principle, the catalytic repertoire of proteins is not limited to the types of reactions observed in nature, and thus suggests new possibilities for the biotechnological application of enzymes, such as the incorporation of novel steps into biosynthesis routes. In their study, the authors employed quantum-mechanical methods to compute the relative orientation of the two reacting molecules in the transition state and to devise the placement for protein functional groups to lower the activation barrier of the cycloaddition. Out of a total of 82 designs, two showed low activity, and for one of them, the activity could be improved 100 fold through five point mutations. In addition, the resulting variant showed high selectivity for the designed stereo-configuration of the product.

In summary, a number of impressive breakthroughs have been achieved with computational enzyme design over the last few years, suggesting that this method might play an important role in future synthetic biology applications. What is important to understand however, is that the current computational methods still lack an accurate modeling of the quantum-mechanical aspects of reactivity, and thus computational design of catalytic activity requires an exquisite understanding of the reaction of interest. For a successful project, the user needs to determine which aspects of a designed active site to enforce and which parts to allow CPD to determine. A future avenue of research is the incorporation of more advanced quantum-mechanical modeling protocols into the currently used design codes. Another caveat is that for successful design of catalytic activity, i.e. discrimination between ground and transition states, placement of the active site functional groups with sub-angstrom precision is required, whereas in the design of a binding interface, somewhat more ‘wiggle room’ is allowed. For these two reasons, *de novo* enzyme design will remain a challenging endeavor in the foreseeable future.

Protein thermostabilization by Computational Design

The earliest application of CPD was the design of mutations to increase the stability of a protein of interest. As sidechain placement algorithms were developed, their ability to predict stabilizing mutations in naturally occurring proteins was used as the first experimental verification of the methodology[48]. It should be noted that, like for other CPD applications discussed so far, an experimental structure of the protein of interest needs to be available in order to stabilize it by CPD. If this is the case, CPD can be used as a valuable tool to create thermostable variants of the protein of interest for synthetic biology applications. There are three main reasons why the problem of protein stabilization is very amenable to CPD, and thus became CPD’s first application:

1) Protein stability is correlated with the number of intramolecular interactions occurring in a protein, and thus can be approximated relatively well by the energy function.

When comparing structures of related mesophilic and thermophilic proteins, the thermophilic variant will often have tighter packing in the hydrophobic core, as well as more hydrogen bonds and salt bridges on the surface. The simplest strategy to stabilize a protein is thus to introduce as many additional packing and hydrogen bonding interactions as possible. Fortunately, these types of interactions can be relatively well approximated by CPD energy functions through their van der Waals and hydrogen bonding terms.

2) Native proteins tend to be only marginally stable, and therefore there are usually many possible mutations that improve stability.

A surprising fact of protein biochemistry is that many natural proteins are only marginally stable, with a ΔG of folding between -5 and -20 kcal/mole. By comparison, a single hydrogen bond can contribute between -1 and -4 kcal/mole to stability[49], meaning that overall protein stability is often equivalent to only a few interactions. Explanations for this at-first paradoxical observation have been offered in detail elsewhere[50], but this fact suggests that the stability of most proteins can be increased.

3) Thermostable variants of proteins can be obtained that have virtually unchanged backbone structure compared to their mesophilic counterpart.

When comparing variants of a certain enzyme adopted for different temperatures, it is often found that the structures are virtually identical, despite often having only low sequence identity[51]. This perhaps surprising finding has favorable implications for the design of thermostable proteins: after all, if stability (and catalytic activity) at different temperatures can be realized with the same tertiary structure in many cases, then it should also be possible to create thermostable versions of other mesophilic proteins without significantly changing the backbone structure.

These three observations in combination lay out a straight-forward approach to stabilize a protein through computational design: starting from the structure of the to-be-stabilized mesophilic protein, use the sidechain placement algorithm to introduce as many favorable additional intramolecular interactions as possible. The backbone can stay fixed throughout the calculation. The presumption that the starting protein is only marginally stable implies that potentially a number of different mutations beneficial for stability can be made, and CPD energy functions can usually identify these. And indeed, there are many examples in the literature where this straight-forward approach succeeded in designing variants with improved stability, often with retained functional properties.

In an early example, Malakauskas *et al.*[52] were able to increase the stability of a model system protein by 4.3 kcal / mol and shift the T_M by more than 20ºC while maintaining (albeit reduced) affinity to a binding partner. In this example, seven mutations were introduced that mostly increased hydrophobic packing. In a comprehensive study by Dantas *et al.*[53], nine globular proteins were completely redesigned. On average, only 35% of the wild-type sequence was retained in the designed variants. Six of these had a T_M above those of the wild-type template. Computational thermostabilization of a functional protein was first described by Korkegian *et al.*[54], who identified a set of three mutations that increased the T_M of yeast cytosine deaminase, an enzyme with potential

applications in prodrug therapy, by 10ºC while retaining the wild-type catalytic efficiency. Recent work by Borgo *et al.*[55] introduced an improved computational protocol that explicitly focuses on packing defects in the hydrophobic core and identifies mutations that are most likely to fix these defects. Using this protocol, the authors were able to increase stability of one of the unsuccessful cases from Dantas *et al.*'s[53] test set by 2.3 kcal / mol.

Computational Design of (novel) protein folds

One of the first and longest standing objectives in CPD is that of full-sequence design for a given backbone conformation. Several breakthroughs in this area have been achieved. However, most of the studies done so far were only concerned with finding a sequence that folds into the target conformation, without regard to any specific function of interest. Therefore this application of CPD is only of limited interest for synthetic biology, and we will cover it only briefly.

The approach of designing a novel protein structure can be subdivided into two stages: 1) generating a three-dimensional model of the desired backbone conformation and 2) designing a sequence for that backbone conformation. During stage 1, the protein is often modeled as consisting of a poly-alanine or poly-valine chain, and during stage 2, the earlier mentioned standard sidechain-placement algorithms are used to find a sequence compatible with the new backbone.

Perhaps the earliest major breakthrough in CPD was Mayo's full sequence design for a 28-residue zinc finger[56]. In this study, the authors took the backbone from a previously solved crystal structure as the template for stage 2, and designed a sequence that was 21% identical to the template's original sequence. The *de novo* designed sequence was shown to fold into the desired structure by NMR. While the authors in this early study didn't design a novel fold and bypassed stage 1 by using a known backbone conformation, this study still represents an impressive early demonstration that sidechain-placement algorithms are capable of designing a sequence for a whole protein fold.

The first example of a completely *de novo* designed protein fold was presented by Kuhlman *et al.* in 2003[57]. In this study, the authors first devised a 106-residue protein fold topology not observed in nature. The novel α/β topology, which was devised as a back-of-the-envelope sketch, consisted of a five-strand antiparallel β-sheet packaged against two α-helices, with a hydrophobic core between these three secondary structure elements. Algorithms developed for protein folding[58] were then used to create three-dimensional models of the desired topology (with poly-Val as sequence). Next, an iterative protocol of sidechain placement algorithms and gradient-based minimization was used to design a sequence compatible with the backbone model. The expressed sequence was well-folded and exceptionally stable, and a crystal structure showed close agreement with the design model.

The approach developed by Kuhlman *et al.* represents a general algorithm to design a sequence that folds into a desired fold (assuming the fold is 'designable'), starting from nothing more than a very rough, back-of-the-envelope draft of the fold. It should be noted that other rational/computational approaches to design novel protein structures have been developed, but most of these are only suitable to design a certain class of protein fold and are not as general as Kuhlman *et al.*'s. The most prominent example is probably the work by Harbury, DeGrado and others[59], who succeeded in designing alpha-helix bundle assemblies not seen in nature. Their approach relies on exploiting a unique and well understood property of a protein-alpha helix: amphipathic alpha helices

can be designed by assigning alternating hydrophobic and hydrophilic amino acids to the seven positions in the heptad-repeat unit that helices are made up of. These amphiphatic helices can then be designed to assemble into helix bundles by choosing amino acids complementary to each other at the positions of the heptad that form the hydrophobic interface (knobs-into-holes approach). Most recently this technique was used to design an alpha helix bundle consisting of six helices (bundles containing this number of helices are not known in nature) that featured a water-filled pore on the inside[60].

A common property of *de novo* designed folds so far is that most of them have fairly compact, convex shapes, and therefore these folds don't have any cavities or pockets that could be used as binding or active sites to mediate a certain function. Designing folds with concave features and cavities that could harbor functional sites is arguably harder, as the fold needs to be stable enough to not collapse around and bury the cavity. Computational algorithms to generate backbone templates that feature cavities and are at the same time designable have yet to be developed.

Complementarity with directed evolution

Before the onset of the computational methods described in this chapter, the design and engineering of protein function was mainly carried out by directed evolution (DE). Many powerful high-throughput screening techniques (HTS) have been developed to address the various engineering challenges, such as a number of display strategies for evolving protein-protein interactions (e.g. phage-, yeast-, *E. coli*-, ribosome-, or mRNA-display) and growth selections, microfluidic / *in vitro* compartmentalization systems, or plate screens for the evolution of new catalysts. Computational approaches and DE are by no means mutually exclusive. Since each of these two methods offers its own unique advantages but also has shortcomings, they are perfectly complementary, and we anticipate that these methods will usually be applied hand in hand in future design efforts. This prognosis is supported by the fact that many of the designed proteins described in this chapter have been optimized by DE.

Perhaps the currently biggest shortcoming of CPD methods are the inaccuracies in the energy function, both in terms of estimating the precise energetics of a candidate designed sequence and in terms of translating this estimated value into a precise estimate for the functional parameter of interest, i.e. a reaction rate or a dissociation constant. However, energy functions are certainly accurate enough to identify the usually extremely small subset of sequence space that is considered compatible with a function of interest, and, as described earlier in this chapter, current algorithms are fast enough to search through sequence spaces of size 10^{130} and above.

The throughput of DE methods can approach 10^{12} variants for some of the *in vitro* display techniques, ~10^{8} for in-vivo display approaches and growth selections, and 10^{4} for plate screening methods, and is thus far below that of CPD. Considering that most active or binding sites are comprised of one or two dozen residues (i.e. a sequence space of ~10^{26} for a 20 residue site), no currently available HTS method can cover more than a tiny subset of possible sequences in a functional design problem. In addition, experimental work is usually much more laborious and resource-intensive than computation. However, DE allows for the selection of the best variant from the sequence pool based on the actual activity of interest instead of based on an (inaccurate) energy function value.

Thus, an ideal way to combine CPD and DE (assuming that an HTS assay is available) is to use computation to come up with initial designs, and then use those initial designs

that show measurable activity as a foothold in sequence space and starting point for successive rounds of DE. This strategy was used for several of the cases described here, and the reported successes could not have been achieved with either CPD or DE alone. In Fleishman *et al.*'s[19] influenza binder, the affinity of the initial computational designs was increased 10-fold. In Khare *et al.*'s[38] designed organophosphate hydrolase, the best variant after DE was four orders of magnitude more active than the original design, but to obtain activity by DE alone, a very large library containing all possible quadruple mutations of the scaffold's active site would have to be screened. In Azoitei *et al.*'s two-loop graft[28], the authors directly compared a computation-informed library to a naïve library, and were able to isolate much tighter binders from the former. Further, in the future, with both advanced DNA synthesis and deep sequencing technologies becoming more readily available, we expect that approaches like the one presented by Lippow *et al.*[37], where computation is used to design a library instead of a single sequence, will become more widely used.

Conclusion and Outlook

The recent advances in computational protein design highlighted in this chapter suggest the likely areas in which synthetic biology could be impacted by CPD almost immediately. Generally, CPD should have an advantage over directed evolution in cases where the sequence space related to the design problem is too large for library construction and/or there is no high-throughput assay. Of the different available techniques discussed here, the redesign of protein-protein interactions is currently the most robust, and thus most readily applicable. This puts within reach the redesign of cell-signaling pathways (as presented by Kapp *et al.*[14]) or improved protein therapeutics that bind their targets with picomolar affinity (as presented by Lippow *et al.*[11]).

De novo design of binders against a given epitope on a target of interest, while shown to work in several examples, is probably still a few years away from being routinely used for practical applications and arbitrary targets. One of the problems in this area that is not yet solved completely is how to identify productive 'anchoring' points for the to-be-designed binder in cases where no other binder is known. For example, if the hotspot-design strategy is used, it is not yet clear how best to identify potential hotspot residues if only a structure of the target in its unbound form is known. However, in case a crystal structure of the target in complex with an already existing binder is known, but that binder happens to have undesirable biochemical properties (i.e. size, immunogenicity), the existing *de novo* design algorithms can be used out of the box to try to design a novel binder starting from a more favorable scaffold.

Specificity redesign of enzymes, i.e. for bioremediation of pollutants or modification of existing metabolic pathways, should also be achievable with the currently available methods, with the latter endeavor being more demanding, since, in the pathway of interest, possibly all enzymes downstream of the first redesigned enzyme would also have to be redesigned to transform the novel metabolite. *De novo* design of enzymes, being perhaps the hardest problem in CPD, will likely require several more years of research effort before it can reliably be applied for practical applications, and should be deployed in combination with directed evolution to increase the (probably low) activity of the initial designs.

Chapter 2

De novo enzyme design using Rosetta3

Disclosure: This chapter has been published as

Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) De Novo Enzyme Design Using Rosetta3. PLoS ONE 6(5): e19230. doi:10.1371/journal.pone.0019230

My contributions consisted in devising the protocol presented herein, implementing the software, running calculations for the benchmark described, and writing the paper.

Abstract

The Rosetta de novo enzyme design protocol has been used to design enzyme catalysts for a variety of chemical reactions, and in principle can be applied to any arbitrary chemical reaction of interest, The process has four stages: 1) choice of a catalytic mechanism and corresponding minimal model active site, 2) identification of sites in a set of scaffold proteins where this minimal active site can be realized, 3) optimization of the identities of the surrounding residues for stabilizing interactions with the transition state and primary catalytic residues, and 4) evaluation and ranking the resulting designed sequences. Stages two through four of this process can be carried out with the Rosetta package, while stage one needs to be done externally. Here, we demonstrate how to carry out the Rosetta enzyme design protocol from start to end in detail using for illustration the triosephosphate isomerase reaction.

Introduction

There has been exciting recent progress in computational enzyme design. Active enzymes have been designed for a variety of reactions including the Diels-Alder reaction[1], the Kemp elimination[2], and the retro-aldol reaction[3]. This paper describes in detail the protocol used to design these catalysts to help researchers apply the method to new reactions, and discusses routes for further improvement of the methods.

Enzymes accelerate reactions through interactions that stabilize the transition state[4]. The Rosetta enzyme design protocol starts with minimalist active site descriptions consisting of transition state models surrounded by disembodied sidechain and backbone functional groups positioned optimally for catalysis, so called theozymes[5]. The Rosetta "matching" algorithm[6] is then used to identify constellations of backbone positions in a set of scaffold proteins where the minimalist active sites can be realized. For each such recapitulation of the minimalist active site in a protein scaffold, the Rosetta design methodology is used to optimize the surrounding residues for transition state binding affinity and catalysis. Here, the Rosetta enzyme design protocol, which has been recently re-implemented as part of the re-architecturing of the Rosetta molecular modeling program[7], is described in detail using as an example the isomerization of Dihydroxy-acetone-phosphate (DHAP) to Glyceraldehyde-3-phosphate (GAP)[8](Figure 1). Figure 2 shows an overview of the different stages of the design process.

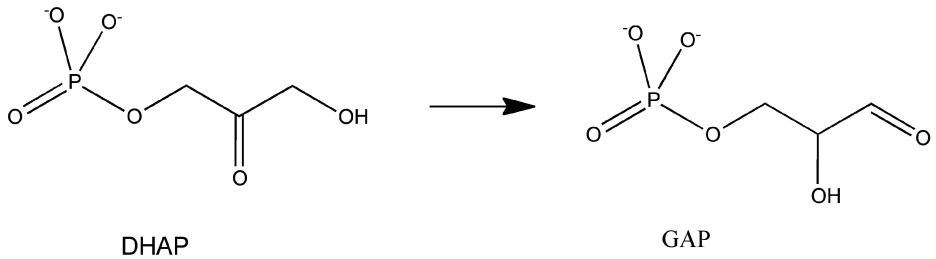


Figure 1 Isomerization of Dihydroxy-acetone-phosphate (DHAP) to Glyceraldehyde-3-phosphate (GAP)

Methods

1) Choice of catalytic mechanism and corresponding theozyme

The first step in the enzyme design process is to model the reaction transition state(s) and decide on a set of idealized active site descriptions consisting of disembodied side chain and backbone functional groups surrounding the transition state in positions optimal for catalysts. The latter include for example hydrogen bond acceptors and donors for proton abstraction and proton donation, positive charges to stabilize developing negative charge in the transition state, etc. These active site descriptions and geometries can be obtained from quantum chemistry saddle point calculations, as in the case of the theozymes pioneered by Ken Houk's group[9], from analogy to active sites of known enzymes, from chemical intuition, or in practice a combination of the above.

Once a theozyme is defined, it needs to be expressed in terms of a Rosetta geometric constraint file, a “cstfile.” The information in this cstfile is used both by the Rosetta matcher to try to graft the desired theozyme onto a scaffold structure, and by the enzyme_design code to restrain the theozyme residues to the desired theozyme geometry during sequence optimization and gradient-based minimization.

The cstfile consists of blocks; for each interaction between two residues, (i.e. for each theozyme interaction), there needs to be one block. The example below describes the interaction between a Glu or an Asp and a ligand abbreviated with name 1n1. In this cstfile, there needs to be a block of the following format for each catalytic interaction:

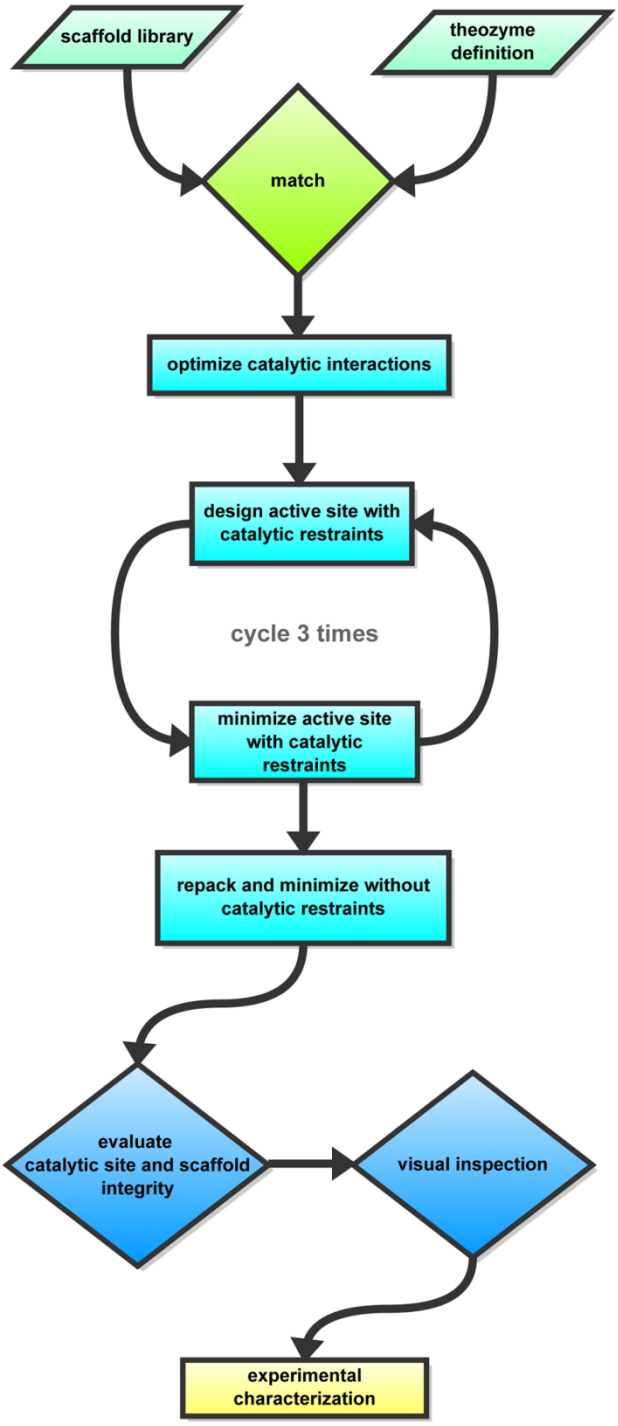


Figure 2 Flowchart of the enzyme design protocol, colored according to the different stages. Stage 1: light green; Stage 2: green; Stage 3: cyan; Stage 4: blue

```
CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: C1 C2 O2
  TEMPLATE::   ATOM_MAP: 1 residue3: 1n1

  TEMPLATE::   ATOM_MAP: 2 atom_type: OOC ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  ED

  CONSTRAINT:: distanceAB:   3.06  0.2  100.    0  0
  CONSTRAINT:: angle_A:     73.60 10.0  80.0  360. 1
  CONSTRAINT:: angle_B:    120.00 15.0  80.0  360. 1
  CONSTRAINT:: torsion_A: -101.20 15.0  60.0  360. 1
  CONSTRAINT:: torsion_AB: 180.00 90.0  0.00  360. 3
  CONSTRAINT:: torsion_B:  180.00 15.0  0.00  360. 1
CST::END
```

The information in this block defines constraints between three atoms on residue 1 and three atoms on residue 2. Up to six parameters can be specified, representing the ligand's six rigid-body degrees of freedom. These parameters are given as one distance, two angles, and three dihedrals.

The Records indicate the following:

'CST::BEGIN' and 'CST::END' indicate the beginning and end of the respective definition block for this catalytic interaction. The 'TEMPLATE:: ATOM_MAP:' records indicate what atoms are constrained and what type of residue they are in. The number in column 3 of these records indicates which catalytic residue the record relates to. It has to be either 1 or 2.
The 'atom_name' tag specifies exactly which 3 atoms of the residue are to be constrained. It has to be followed by the names of three atoms that are part of the catalytic residue or ligand. In the above example, for catalytic residue 1, the ligand, atom 1 is C1, atom 2 is C2, and atom3 is O2. The geometry specified is visualized in Figure 3, top left panel.

The 'atom_type' tag is an alternative to the 'atom_name' tag. It allows more flexible definition of the constrained atoms. It has to be followed by the Rosetta atom type of the residue's atom 1. In case this tag is used, Rosetta will set atom 2 as the ``base atom'' (the parent for an atom in the tree rooted at the backbone growing out along the side chain) of atom 1, and will set atom 3 as the base atom for atom 2. There are two advantages to using the 'atom_type' tag: first, it allows constraining different residue types with the same file, e.g. if a catalytic hydrogen bond is required, but either a SER or THR would do. Second, if a catalytic residue contains more than one atom of the same type (e.g. atoms OD1 and OD2 of Asp), but it doesn't matter which of these atoms mediates the constrained interaction, using this tag will cause Rosetta to evaluate the constraint for all of these atoms separately and pick the one with lowest score, i.e. the ambiguity of the constraint will automatically be resolved.

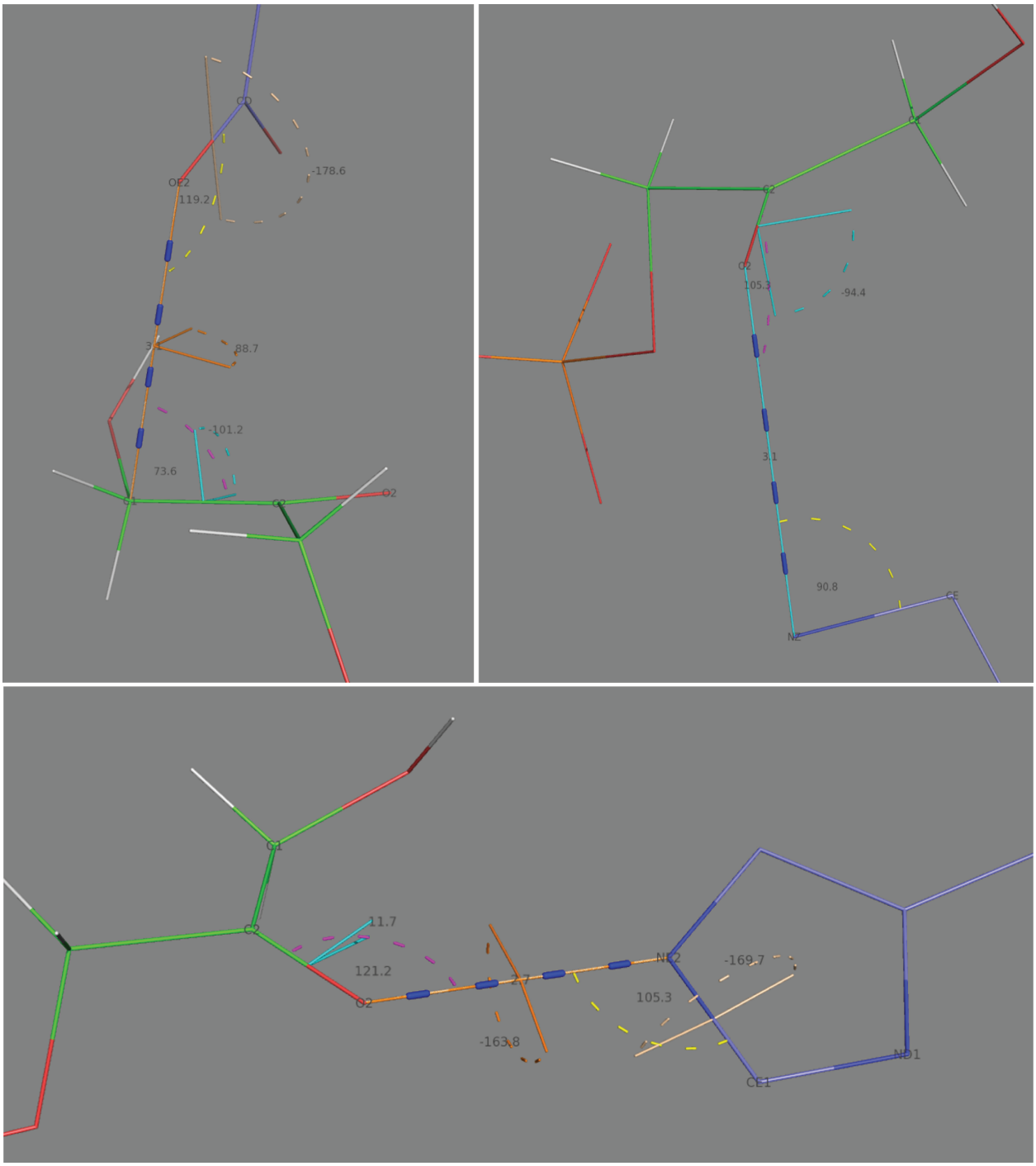


Figure 3 Theozyme geometries
Top left: Interaction 1; top right: Interaction 3; bottom: Interaction 2;
Color scheme: distanceAB: blue, angle_A: purple, angle_B: yellow, torsion_A: cyan, torsion_AB: orange, torsion_B: ligt brown

The 'residue1' or 'residue3' tag specifies what type of residue is constrained. 'residue3' needs to be followed by the name of the residue in 3-letter abbreviation. 'residue1' needs to be followed by the name of the residue in 1-letter abbreviation. As a convenience, if several similar residue types can fulfill the constraint (e.g. either Asp or Glu), the 'residue1' tag can be followed by a string of 1-letter codes of the allowed residues (e.g. “ED” for Asp and Glu, or “ST” for Ser and Thr).

The 'CONSTRAINT::' records specify the parameters and the strengths of the constraint applied between the two residues. Each of these records is followed by one string and four numbers. The string can have the following allowed values:
'distanceAB' means the distance Res1:Atom1 = Res2:Atom1, i.e. the distance between atom1 of residue 1 and atom1 of residue 2.
'angle_A' is the angle Res1:Atom2 - Res1:Atom1 - Res2:Atom1
'angle_B' is the angle Res1:Atom1 - Res2:Atom1 - Res2:Atom2
'torsion_A' is the dihedral Res1:Atom3 - Res1:Atom2 - Res1:Atom1 - Res2:Atom1
'torsion_AB' is the dihedral Res1:Atom2 - Res1:Atom1 - Res2:Atom1 - Res2:Atom2
'torsion_B' is the dihedral Res1:Atom1 - Res2:Atom1 - Res2:Atom2 - Res2:Atom3

Each of these strings is followed by 4 (optionally 5) columns of numbers: x0, xtol, k, covalent/periodicity, and number of samples. The 1st column, x0, specifies the optimum distance x0 for the respective value. The 2nd, xtol, column specifies the allowed tolerance xtol of the value. The 3rd column specifies the force constant k, or the strength of this particular parameter. If x is the value of the constrained parameter, the score penalty p(x) applied will be:

$$p(x) = \begin{cases} 0 & if \ \ |x - x0| < xtol \\ k * (|x - x0| - xtol) & otherwise \end{cases}$$

The 4th column has a special meaning in case of the distanceAB parameter. It specifies whether the constrained interaction is covalent or not. 1 means covalent, 0 means non-covalent. If the constraint is specified as covalent, Rosetta will not evaluate the vdW term between Res1:Atom1 and Res2:Atom1 and their [1,3] neighbors.
For the other 5 parameters, the 4th column specifies the periodicity per of the constraint. For example, if x0 is 120 and per is 360, the constraint function will have its minimum at 120 degrees. If x0 is 120 and per is 180, the constraint function will have two minima, one at 120 degrees and one at 300 degrees. If x0 is 120 and per is 120, the constraint function will have 3 minima, at 120, 240, and 360 degrees.
The 5th column is optional and specifies how many samples the matcher, if using the classic matching algorithm, will place between the x0 and x0 +- tol value. The matcher interprets the value in this column as the number of sampling points between x0 + xtol and x0 – xtol, i.e. in the above example, for angle_A, the matcher will sample values 63.6, 73.6, and 83.6 degrees. Generally, if the value in this column is n, the matcher will sample 2n+1 points for the respective parameter. Note that the number of samples is also influenced by the periodicity, since the matcher will sample around every x0.

When determining how many different values to sample for each parameter, it is important to remember that the number of different ligand placements attempted for every protein rotamer built is equal to the product of the samples for each of the six parameters. For example, in the above block there is one sample for distanceAB, three samples for angle_A, three samples for angle_B, three samples for torsion_A, three samples for torsion_B, and seven samples for torsion_AB, meaning that for every protein rotamer, the matcher will attempt to place the ligand in a total of 1*3*3*3*3*7 = 567 different conformations.

2) Matching: identifying sites in the scaffold library where the theozyme can be placed

In this stage, the hypothetical theozyme will be placed into an existing protein structure with the help of the RosettaMatch module. The inputs for this stage are the theozyme expressed in the cstfile format as described above, and a list of protein scaffolds. The RosettaMatch executable will be run once for each scaffold and, if the theozyme and the respective scaffold are compatible, will output a number of so-called matches. A match is defined as the theozyme grafted into a scaffold, i.e. the amino acid side chains of the theozyme have been placed on the scaffold backbone, the ligand has been placed into a cavity of the scaffold without clashing with the protein backbone or the theozyme side chains, and the geometric relation between the ligand and the theozyme side chains is as specified in the theozyme.

To run RosettaMatch, the user has to prepare each scaffold and decide which of two available algorithms to use for each side chain of the theozyme. Preparing the scaffold consists of deciding which residues of the scaffold should be considered when trying to place the theozyme.

Preparing the scaffold for matching
Usually only a subset of the scaffold residues are considered during the matching process, for two reasons: first, residues lining a concave pocket or cleft of the scaffold are more likely to form a binding site than residues that are buried in the hydrophobic core or residues that protrude into the solvent. Trying to design a binding site in the core of the protein is problematic because it will likely have a negative effect on the stability of the protein, while creating a binding site on the surface of the protein is difficult because there are possibly not enough side chains to contact the ligand from several sides and thus form a binding surface that is complementary in shape to the ligand. Second, the more residues that are considered by the matcher, the higher the computational requirements in terms of runtime and memory become, and for theozymes with many degrees of freedom the matching runtime can quickly become the bottleneck of the whole process.
Therefore, for each scaffold that is considered by the matcher, a subset of residues most likely to be part of the binding site need to be selected, such as residues lining a pocket or cleft. In case the scaffold was crystallized with a natural ligand, one could for example select all residues that are within a certain distance from that natural ligand.

Choosing a matching algorithm for each theozyme interaction

RosettaMatch can place the theozyme side chains through two algorithms: classic matching as introduced by Zanghellini et al.[6], and "secondary matching." Both algorithms can be used in the same matching run.

Classic matching has been described in detail before[6]. Briefly, for every interaction/side chain of the theozyme, the matcher builds rotamers at each of the scaffold active-site positions. For each rotamer built, the ligand is placed according to the geometry specified in the theozyme/cstfile. To use the classic matching algorithm, sample values must be given for all six parameters connecting the side chain to the transition state. Note that depending on the theozyme and the set of sample values specified in the cstfile, the number of different ligand placements for each side chain rotamer can grow to be quite large. Each ligand placement is checked for collisions with the scaffold's backbone, and, if there are none, the location and orientation of the ligand in space is recorded in a 6D coordinate (3 Euclidean coordinates and 3 Eulerian coordinates). The 6D coordinate for each collision free ligand placement along with the side chain rotamer used to build this coordinate, is called a "hit."

For a theozyme interaction/side chain to be placed with the secondary matching algorithm, the matcher proceeds as follows: First, just like in classic matching, rotamers at all scaffold active-site positions are built. Then, for each rotamer r, the geometry between r and each of the previously generated hits is evaluated. If the geometry of rotamer r and a particular hit h is compatible with the desired theozyme geometry, the 6D coordinate of the ligand is copied from h and stored with rotamer r as a new hit for this theozyme interaction.

After hits have been generated for all N side chains/interactions of the theozyme, all hits are then binned according to the 6D coordinate of the ligand and placed into a hash table. Then, the hash bins are checked for whether they contain at least one hit from each of the N theozyme side chains. For every bin that does, the ligand placements stored in it, together with the side chain rotamers they were built from, are considered a "match", i.e. a successful graft of the theozyme onto the scaffold.

Classic matching and secondary matching each have their own advantages and disadvantages. Which algorithm to choose for a given theozyme interaction depends on several factors.

There are two advantages to the secondary matching algorithm: first, since the ligand is not built from the side chain rotamer, but instead taken from a previously generated hit, not all six degrees of freedom must be specified. For example, if a theozyme interaction depends only the distance and the two angles, while the three dihedrals are unimportant, then the secondary matching algorithm can be given ranges for only the important parameters, while ignoring the unimportant parameters. If such an interaction were to be described to the classic matching algorithm, the unimportant dihedral parameters would have to be sampled over the whole 360 degree range, resulting in long running times. Sampling these dihedrals coarsely at 10 degree increments still requires building 36^3 ~ 46 thousand ligand conformations per assignment to the other three parameters per side chain rotamer.

Second, the secondary matching algorithm can also be used to find theozyme interactions between two side chains. Since both the ligand placement and the rotamer from which it stems are stored in a hit, the secondary matching algorithm can equally well evaluate the geometry between a rotamer built for one theozyme interaction and a rotamer for a previously generated hit.

The advantage of the classic matching algorithm is that it performs considerably faster than the secondary match algorithm in cases where a large number of ligand placements have been generated for preceding interactions of the theozyme. The speed of placing one theozyme interaction with the secondary matching algorithm decreases with an increasing number of previous ligand placements, since every one of them needs to be examined again. The classic match algorithm, on the other hand, always generates the ligand placements from the coordinates of the side chain rotamer and the information in the geometric constraint file, and is thus independent of previously generated ligand placements. It also is less sensitive to the lever-arm effect than secondary matching; secondary matching may not capture ligand interaction geometries very accurately forcing the user either to tolerate very broad ranges of values for a particular interaction or to sample ligand geometries very densely in other theozyme interactions.

As a rule of thumb, secondary matching should be used for theozyme interactions where not all six degrees of freedom are clearly defined and for side chain-side chain theozyme interactions, while classic matching should be used in other cases. Note that since secondary matching must rely on the hits generated prior to its execution, classic matching must be used for the first theozyme interaction.

Command line options affecting this stage:

Matching is carried out by the `match` executable available in the Rosetta 3.2 release and is sensitive to the following command line options:

```
-extra_res_fa <filename>                  path to rosetta parameter file of
ligand to be matched
-lig_name <string>                        name of the ligand to be matched
-geometric_constraint_file <filename>           path to constraint file /
theozyme
-s <filename>                             path to scaffold pdb
-scaffold_active_site_residues <filename> file containing what residues of
the scaffold to match at
-ex1, -ex2 <value>                        optional parameter governing size
of rotamer library
```

3) Designing the found sites

After matches have been found, optimal residue identities for other scaffold positions need to be determined to build an active site that is complementary in shape to the ligand while also stabilizing the catalytic side chains in their matched conformations. The basic Rosetta enzyme design protocol used for this consists of four steps, and is all carried out by the `enzyme_design` executable available in the Rosetta 3.2 package:

A. Determining which residues to design and which to repack
B. Optimizing the catalytic interactions
C. Cycles of sequence design/minimization (with catalytic constraints if specified)
D. Unrestrained fixed sequence rotamer pack /minimization

Determining which residues to design and which to repack
There are two ways of doing this: using a standard Rosetta resfile to exactly specify which residues are allowed at which position or automatic detection of the design region. In case there is only a small number of different starting structures, it is probably better to invest the time and use intuition to decide which positions in the protein to redesign or repack and which amino acids to allow.

In case there are a lot of input structures to be designed, it is also possible to automatically determine which residues to redesign.
Rosetta can divide the protein's residues into five groups of increasing distance from the ligand:
1) Residues that have their Cα within a distance cut1 Å of any ligand heavy atom will be set to designable
2) Residues that have their Cα within a distance cut2 of any ligand heavy atom and the Cβ closer to that ligand atom than the Cα will be set to designable. cut2 has to be larger than cut1
3) Residues that have their Cα within a certain distance cut3 of any ligand heavy atom will be set to repackable. cut3 has to be larger than cut2
4) Residues that have their Cα within a distance cut4 of any ligand heavy atom and the Cβ closer to that ligand atom will be set to repackable. cut4 has to be larger than cut3
5) All residues not in any of the above four groups are kept static.

Residues declared as catalytic in the input pdb will always be repackable (except if turned off by an option). At residue positions that are set to designable, every amino acid except cysteine will be allowed. Values (in Å) for the different cuts commonly used are: 6.0 (cut1), 8.0 (cut2), 10.0 (cut3), and 12.0 (cut4).

Command line options affecting this stage:

```
-resfile \<name of resfile\>       specifies the use of a resfile
-enzdes:detect_design_interface    invokes automatic detection of designable
region
-enzdes:cut1 \<value\>             value used for cut1 (in Å)
-enzdes:cut2 \<value\>             value used for cut2 (in Å)
-enzdes:cut3 \<value\>             value used for cut3 (in Å)
-enzdes:cut4 \<value\>             value used for cut4 (in Å)
```

```
-enzdes:fix_catalytic_aa          prevents catalytic residues from being
repacked or minimized
```

Further, these two ways of declaring the design and repack regions can be combined, i.e. a resfile and the detect_design_interface mechanism can be used concurrently. If the default behavior in the resfile is set to 'AUTO', the behavior of every residue which is not specifically declared in the resfile will be determined according to the -detect_design_interface logic.

Optimizing catalytic interactions

This stage consists of a gradient-based minimization of the input structure before design. During this minimization, all active site residues that are not catalytic (i.e. not part of the theozyme) are mutated to alanine (i.e. the active site is reduced to substrate and catalytic residues only), and a reduced energy function that does not contain vdW-attractive or solvation terms is used for the minimization. Restraints as specified in the cstfile are placed on the interactions between catalytic residues and the ligand. The purpose of this stage is to move the substrate to a position where the catalytic interactions are as ideal as possible.

Command line options affecting this stage

```
-enzdes:cst_opt   will invoke this stage
-enzdes:bb_min    optional but recommended. Allows the backbone to be
slightly flexible during the minimization
-enzdes:chi_min   optional but recommended. Allows the dihedrals of the
catalytic residues to move during the minimization
```

For backbone minimization, only the backbone phi/psi angles of residues in the designable/repackable region will be allowed to move. A special fold tree [10] is created to constrain backbone movement to the designed site, i.e. there will be no conformational changes in regions that are neither repackable nor designable. Further, to prevent the backbone of the active site from moving considerably during the gradient-based minimization, the Cα atoms are restrained to within 0.5 Å of their original positions.
An alternative to the gradient-based restraint optimization is running a short docking trajectory of the ligand with Monte Carlo rigid body sampling. It is invoked by

```
-enzdes:cst_predock       invokes this stage
-enzdes:trans_magnitude   largest allowed displacement of the ligand (in Å)
-enzdes:rot_magnitude     largest allowed rotation of the ligand (in degrees)
-enzdes:dock_trials       number of rigid body moves attempted
```

The trans_magnitude and rot_magnitude values are sampled with a Gaussian distribution of zero mean and one standard deviation. The rotation and translation center is taken as the centroid of the set of ligand atoms which have distance restraints to the protein. This ensures the most efficient sampling of the ligand with respect to the restraints. As in Bi, all designable residues are mutated to alanine to allow the ligand to sample the whole active site region prior to design.

Cycles of sequence design and minimization

This is where the actual sequence design happens. At the designable positions, the Rosetta standard Monte Carlo algorithm is employed to find a new lower energy sequence for the non-catalytic residues. The catalytic restraints are kept on through the entirety of this stage. After sequence design, the resulting structure is minimimized. These two steps are typically iterated a small number of times (2-4).

Command line options affecting this stage:

```
-enzdes:cst_design          will invoke this stage
-enzdes:design_min_cycles   how many iterations of design/minimization will be
done
-enzdes:lig_packer_weight   determines the relative importance of protein-
substrate interactions vs. protein-protein interactions in the sequence
selection calculation
-enzdes:cst_min             necessary to invoke minimization after sequence
design
-enzdes:bb_min              same as for stage B
-enzdes:chi_min             same as for stage B
-packing:ex1                optional but highly recommended. improved rotamer
sampling around the first dihedral for every amino acid
-packing:ex2                optional but highly recommended. improved rotamer
sampling around the second dihedral for every amino acid
-packing:use_input_sc       optional but highly recommended. include the input
rotamer of every side chain in the calculation
-packing:soft_rep_design    triggers use of the soft-repulsive force field in
design.
-packing:linmem_ig 10       optional but highly recommended. speeds up the
sequence design step while at the same time reducing memory requirements.
-packing:unboundrot         optional. pdb files that contain additional
rotamers to use in rotamer packing calculations.
```

Unrestrained fixed-sequence rotamer packing and minimization

After Rosetta has designed a new sequence, a final repack/minimization will be done without the catalytic restraints. This is to check whether the designed sequence actually holds the catalytic residues in their catalytically active conformations; in a good design, the catalytic residues should adapt their theozyme conformations without artificial restraints enforcing them.

Command line options affecting this stage:

```
-enzdes:no_unconstrained_repack   will prevent this stage from being invoked
-packing:ex1                      same as for stage C
-packing:ex2                      same as for stage C
-packing:use_input_sc             same as for stage C
-enzdes:cst_min                   same as for stage C
-enzdes:bb_min                    same as for stage B+C
-enzdes:chi_min                   same as for stage B+C
```

4) Evaluating and ranking the resulting designed sequences

In a typical enzyme design project, often hundreds or thousands of input structures will be designed. Typically when matches are found, they produce many structures which are very similar to each other (i.e. they derive from matches with very similar ligand placements). It is also recommended to redesign every starting structure a few times, since the stochastic Monte Carlo algorithm can lead to slightly different results every time. Together, this means that there are thousands of designs produced by stage 3. The problem then becomes analyzing, and ranking all of the produced structures to find the best handful worth experimentally characterizing. There is no perfect or ideal way to do this, and only one of many possibilities is described here.

Every model that is output by Rosetta has the scores broken down by residue and score type appended after the atom records. One can simply select the model that has the best overall score, or the best ligand score, or the best constraint score, etc.
However, the Rosetta scores don't necessarily capture all the important characteristics of a given design. The `enzyme_design` application is set up to evaluate each output structure with respect to the following additional properties and metrics:
-number of hydrogen bonds (in the whole protein and catalytic residues)
-number of buried unsatisfied polar atoms (whole protein/catalytic residues)
-non-local contacts (i.e. contacts between residues that are far away in sequence, for both whole protein/catalytic residues)
-score across the interface between protein/ligand
-packstat [13] of the designed structure with and without ligand present
if the option -out:file:o <filename> is active, a scorefile will be written that contains one line for every output structure. The column labels in the score file have the following meaning:

General syntax:
pm = Column labels ending in "_pm" are determined using a pose metric calculator
The catalytic/constrained residues are SR1, SR2, SRN for N residues. e.g. if there is one catalytic residue only SR1. Note that if the same catalytic residue is involved in multiple interactions (such as in a triad), it will appear multiple times.
The ligand is the SR(N+1) and it is the last SR, e.g. if there is one catalytic residue it is SR2.

```
total_score:       score (excluding the restraint score)
fa_rep:            full atom repulsive score
hbond_sc:          hbond side chain score
all_cst:           all restraint score
tot_pstat_pm:      pack statistics [13]
total_nlpstat_pm:  pack statistics without the ligand present
tot_burunsat_pm:   buried unsatisfied polar atoms
```

Command line options affecting this stage:

```
-out:file:o                    will trigger writing of the output file
```

```
-final_repack_without_ligand   will repack the design apo-structure and
evaluate RMSD
```

Results

To demonstrate the Rosetta enzyme design protocol, a full calculation for the triose phosphate isomerase (TIM) reaction (Figure 1) was carried out. This reaction is an essential component of glycolysis, and is one of the best-studied model reactions in enzymology[8]. High-resolution crystal structures of native enzymes have been obtained[11], and the fold they adopt was named after the reaction[12]. Numerous mechanistic, mutational and computational studies have been performed. Here, we try to design a TIM active site into a thermophilic scaffold of the same fold family as the native TIM.

1) Defining the theozyme

Defining the theozyme is easy in this case, since several crystal structures of native enzymes exist, and the most important residues in the active site have been determined previously[11]. Quantum mechanical calculations of the natural enzyme have rationalized the geometry observed between the catalytic residues and the substrate[8]. The theozyme is based on the crystal structure of *S. cerevisiae* TIM (PDB code 1ney, Figure 4) and defined according to a mechanism advocated by Guallar *et al.*[8], depicted in Figure 5. It contains three side chains: 1) a Glu/Asp (Glu165 in 1ney) to carry out two proton shuffling steps, 2) a His (His95 in 1ney) to polarize O1 and O2, the two substrate oxygen atoms that change hybridization during the reaction, and 3) a Lys (Lys12 in 1ney) to additionally polarize O2 and facilitate the initial enolate formation.

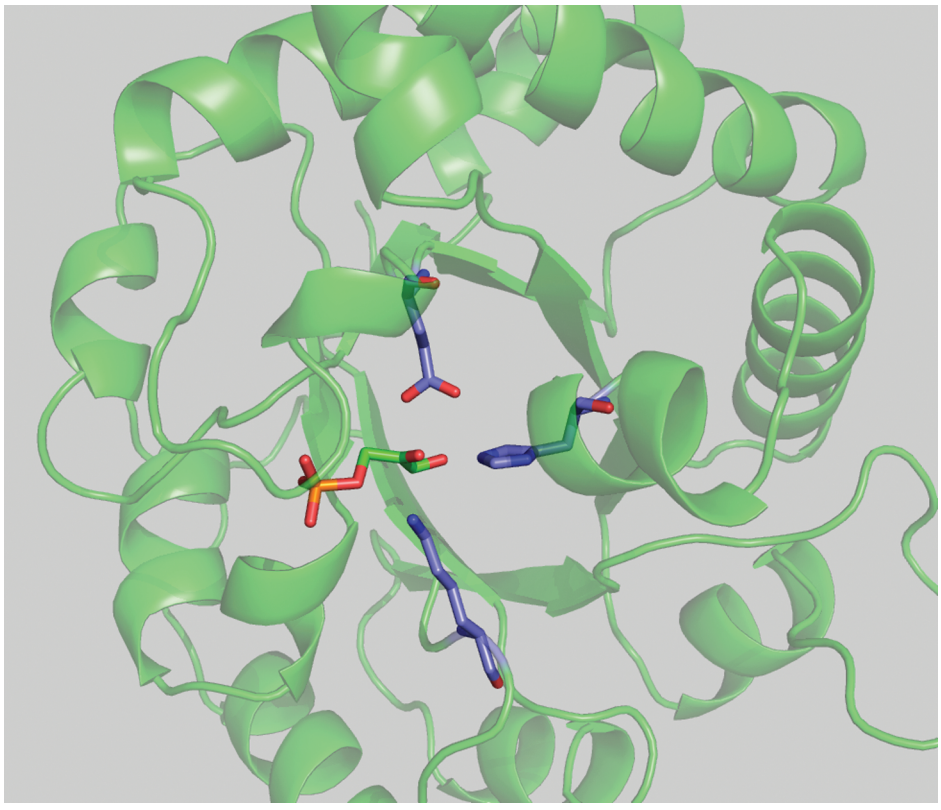


Figure 4 Crystal structure of *S. cerevisiae* TIM with substrate and the three most critical catalytic residues shown in stick representation

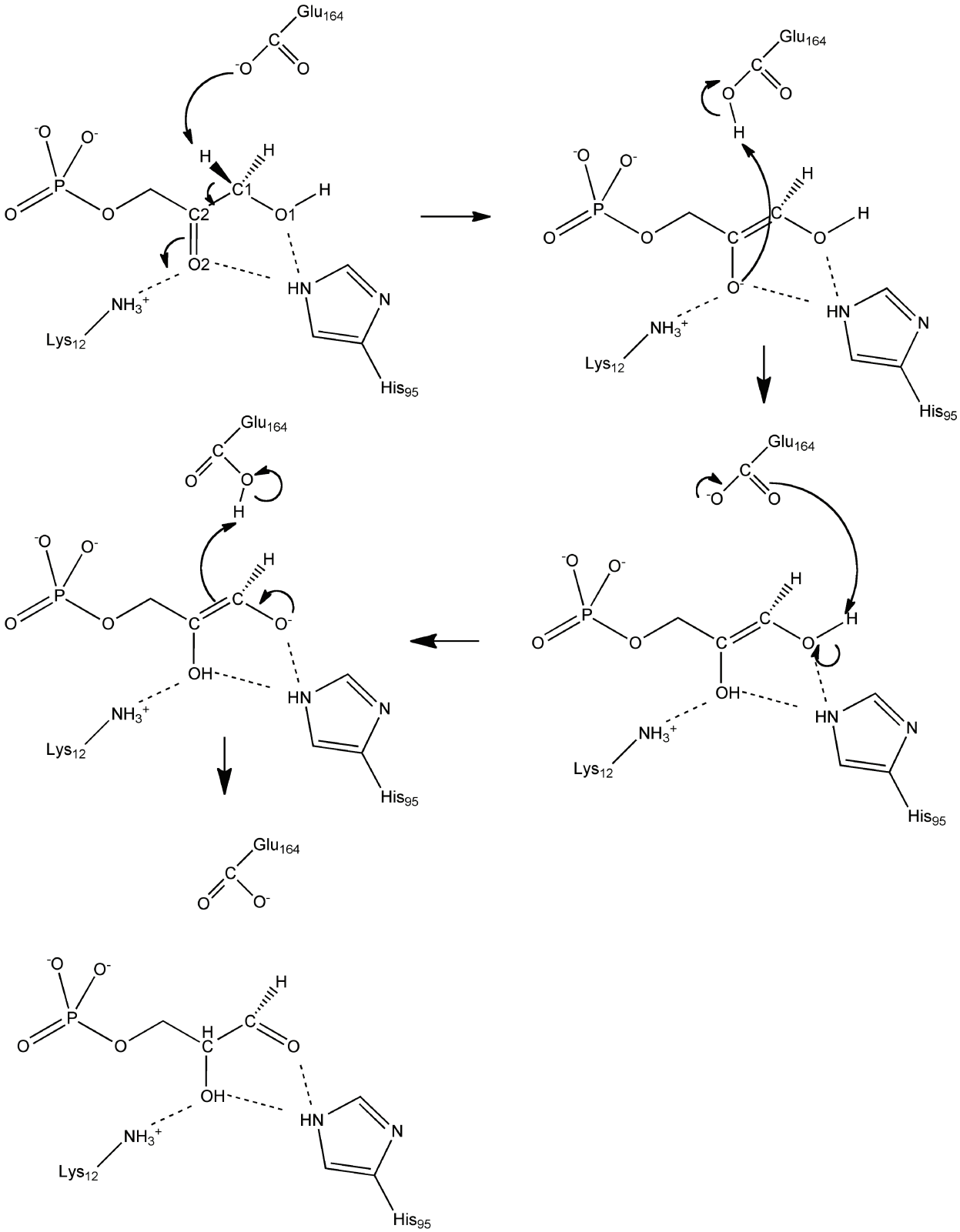


Figure 5 Proposed reaction mechanism of the DHAP to GAP isomerization as catalyzed by *S. cerevisiae* TIM.
In the top left panel, substrate atoms are labeled according to their label in the theozyme model.

The geometries used in matching, shown in Figure 3 and summarized in Table S1, are fairly constrained for the first two side chains and more relaxed for the third interaction. For Interaction 1, the carboxylate moiety of an Asp or Glu side chain needs to be in a position to abstract the *pro-R* proton from C1. This dictates the values for distance_AB, angle_A, and torsion_A to be within a fairly small range. Further, since the deprotonation is happening through the empty *anti* orbital of one carboxylate-O, angle_B and torsion_B are clearly defined. There is more freedom in the last parameter, torsion_AB, which governs the rotation around the DHAP:C1 – Asp:OD1/2 / Glu:OE1/2 vector, and accordingly the matcher is set to sample a wider range for this.
For the second side chain, a His, the protonated N of the imidazole ring needs to be between DHAP:O1 and DHAP:O2, within hydrogen bonding distance to both, and in the plane created by DHAP:C1, DHAP:C2, and DHAP:O2. This requirement clearly determines the values for distance_AB, angle_A, and torsion_A. The protonated N-sp2 orbital is pointing between the two oxygen atoms, and thus determining values for angle_B and torsion_B. There is a little more variability possible for torsion_AB, which governs the relative rotational orientation of the imidazole plane to the substrate plane.
The third interaction, between the Lys and the DHAP, is more loosely defined. The desired hydrogen bonding geometry between Lys:NZ and DHAP:O2, both sp3 hybridized after the initial reaction step, necessitates tight ranges for the values for distanceAB, angle_A, and angle_B. To restrict the Lys to the side of the DHAP plane opposite the theozyme Asp/Glu, torsion_A is also restricted. The values for torsion_AB and torsion_B, on the other hand, do not have much influence on the quality of the interaction, and therefore are allowed to have any arbitrary value.

2) Matching

In the example presented here, a TIM active site will be placed into a thermophilic scaffold with a TIM β/α barrel fold[12]. Six scaffolds of this fold from three thermophilic organisms, listed in Table 1, were selected from the PDB. None of these proteins has been annotated as catalyzing the TIM reaction. For each scaffold, residues lining the natural binding pocket were selected as match positions.

PDB ID	Natural function	Source organism	Number of matches	runtime / seconds
1tml	Endocellulase	*T. fusca*	9	16
1i4n	Indole-glycerol phosphate synthase	*T. maritima*	123	49
1thf	Imidazole-glycerol phosphate synthase	*T. maritima*	32	23
1igs	Indole-glycerol phosphate synthase	*S. solfataricus*	85	36
1dl3	Phosphorybosyl antranilate isomerase	*T. maritima*	22	17

1qo2	Ribonucleotide isomerase	*T. maritima*	101	42

Table 1 Matching results

For theozyme interactions 1 and 2, as described above, the catalytic geometry is fairly restricted, meaning that for all six degrees of freedom there are catalytically necessitated values. Therefore, the classic match algorithm was used to place these side chains. To increase the number of matches found, additional samples were done at small deviations from the ideal value. Theozyme interaction 3, featuring two degrees of freedom that can take on any value, is less clearly defined. Therefore the secondary matching algorithm was used to place the Lys, and during the matching process the torsion_B and torsion_AB parameters between candidate Lys rotamers and potential ligand placements were not evaluated. An overview over the matching setup is given in Table S1, and the results of the matching stage are in Table 1. In total, 372 unique matches were found.

3) Design

All unique matches were designed ten times with the design protocol as described in the methods section. Specifically, the design shell was defined as all residues within a cut1 of 4 Å and cut2 of 6 Å of the ligand, and the repack shell as all residues within a cut3 of 10 Å and a cut4 of 12 Å. The catalytic residues as found by the matcher were considered to be part of the repack shell. This led to design shells having on the order of 20 residues and repack shells on the order of 50 residues.
As a first step, all design shell residues were mutated to Ala, and a gradient-based optimization of the poly-Ala structure was done in order to optimize the catalytic interactions and ligand position (cst_opt stage).
The structures generated by the cst_opt stage were subjected to two rounds of design and minimization. During the design stage, interactions between the ligand and the protein were upweighted by a factor of 1.6, to favor the selection of sequences that make good protein-ligand interactions with the ligand over sequences that only make good protein-protein interactions. Further, to reduce the number of mutations that Rosetta introduces, at every design position the native residue identity was favored by a small bonus of 0.8 Rosetta energy units (REU). The purpose of this is to make sure that the native residue is kept at positions where there is no other residue that makes clearly better interactions, such as for exposed surface positions, hopefully avoiding inadvertent destabilization of the scaffold.
The structures thus generated were then repacked and minimized without the catalytic constraints, to ascertain the unbiased conformation of the designed sequence. Finally, to test whether the designed structures stay in the same conformation irrespective of the presence of the ligand, the ligand was removed and the structures repacked one last time. The complete protocol took on the order of 8 minutes per structure, and since a total of 3720 structures were generated, the complete CPU time for this stage was about 500 hours.
The resulting designs are then evaluated with respect to several factors as described in the methods section. The results are in Table 2. As a comparison, the values observed when repacking and minimizing the native 1ney enzyme are also listed.

Feature	av. value +- std.dev observed in 3270 designs	lowest/highest value observed in 3270 designs	av. value +- std.dev observed for repacking 1ney native 10 times	cutoff for selecting designs	# of designs passing cutoff
ligand binding and catalytic site measures					
total restraint score /REU	25.79 +- 31.95	0 / 303.21	3.50 +- 2.81	< 6.5	1255
restraint score catres1/REU	3.47 +- 6.01	0 / 149.45	0.76 +- 0.28	< 1.2	1413
restraint score catres2/REU	3.89 +- 8.53	0 / 108.17	0.33 +- 0.20	< 1.0	2136
restraint score catres 3/REU	5.53 +- 11.91	0 / 83.37	0.66 +- 1.57	< 2.3	2696
ligand binding score/REU	-2.35 +- 1.34	-7.03 / 2.62	-9.77 +- 0.22	< -8.5	0
active site repack rmsd w/o ligand /Å	0.33 +- 0.21	0 / 1.46	0.06 +- 0.02	< 0.5	3024
catres 1 repack rmsd w/o ligand Å	0.43 +- 0.76	0 / 4.43	0.12 +- 0.11	< 0.5	2737
catres 2 repack rmsd w/o ligand Å	0.73 +- 0.79	0/4.56	0.0 +- 0.0	< 0.5	1971
catres 3 repack rmsd w/o ligand Å	0.76 +- 0.94	0 / 5.28	0.06 +- 0.18	< 0.5	2121
scaffold integrity measures / differences to native					
total score /REU	-91.35 +- 30.24	-158.97 / 186.32	n/a	< 0.0	3705
# buried non H-bonded polar atoms	-2.10 +- 4.94	-19 / 19	n/a	< 5	3413
# non-local contacts	6.16 +- 3.83	-10 / 19	n/a	> -2	3615
packstat [13]	-0.02+- 0.04	-0.14 / 0.10	n/a	> -0.05	2565

Table 2 Design results

4) Ranking and Selection

After the design stage, the question becomes which of the resulting 3720 designs to visually inspect and eventually select for expression. The philosophy applied here for selecting designs

consists of two considerations: 1) for a design to be active, the ligand needs to have a good score (i.e. binding energy), the catalytic residues need to be in a competent conformation, and the active site needs to be preorganized; and 2) a designed protein must be folded, stable, soluble, and expressible in a standard *E. coli* production strain.

For consideration 1, since there is a natural enzyme known, this can be scored in Rosetta under the same conditions as the designs were generated and used as a benchmark. Designs are then selected only if they have comparable ligand-binding scores and comparable catalytic-constraint scores. In terms of preorganization, the RMSD of the catalytic residues between the final designed structure with a ligand and the structure that was repacked without the ligand has to be small. To enable selection of designs according to these criteria, the model of triose phosphate isomerase from *S. cerevisiae*[11], was repacked and minimized in Rosetta, with the same cstfile as used in the designs.

Regarding consideration 2, it is important to note that the absolute Rosetta score does not necessarily correlate with properties such as protein stability, solubility, or a clearly defined fold. Yet, a designed protein should feature each of these. To judge the qualities of a certain design in this regard, one approach is to compare the designed protein to the scaffold that it originated from. Since the original scaffold was a well behaved protein, one can reasonably assume that a design based on it will also be, provided that not too many of the scaffold interactions and features have been corrupted. Here, designs were required to have comparable scores, packing quality[13], contact order, and numbers of hydrogen bonds and buried polar atoms as the scaffolds they came from. Table 2 shows each feature that was used for selecting designs, together with the cutoffs used.

Of the 3720 designs, unfortunately none passed all the cutoffs, because none of the designs featured ligand binding scores comparable to the native. For the remaining selection parameters however, there were generally designs that had values comparable to the native active site or their respective scaffold. Not counting the ligand binding energy parameter, there were 43 designs that passed all cutoffs. The question then becomes if any of these 43 designs should be experimentally tested, or if matching in more scaffolds should be done to find matches that give rise to designs with better ligand binding scores. In the example design project presented here, only five starting scaffolds were used, whereas in most real-world design projects in our group, hundreds or thousands of scaffolds are considered. If one wanted to improve the binding score of the resulting designs, one possibility would be to include binding interactions with the ligand phosphate group (which is exquisitely bound in the 1ney native) in the theozyme, so that all matches would already feature better binding sites. Alternatively, the selected designs can be subjected to more thorough examination techniques such as MD simulations before experimental characterization is attempted[14].

Discussion

The Rosetta3 enzyme design protocol as presented here constitutes a general method to create suggestions for protein catalysts, for any arbitrary reaction of interest. Though it has been

shown to be capable of designing active enzymes in three cases [1,2,3], in each case the best designed proteins had only very modest activity, while many of the designs tested had no activity at all. Thus, while this protocol constitutes a powerful tool in the development of novel catalysts, success is by no means guaranteed. Several shortcomings and potentials for improvement exist, some of which have been showcased in this study. We consider three areas where we could improve the protocol.

First, to increase the quality of designs, it is beneficial to include as many interactions in the theozyme as possible, and concurrently run matching for as many scaffolds as possible. In the TIM example presented here, none of the designs showed sufficient ligand binding score, so for a new round of designs, it might be beneficial to include ligand binding interactions in the theozyme. However, the more complicated the theozyme becomes, the smaller the number of matches that are found; each additional geometric requirement made on the scaffold makes it less likely any particular scaffold will meet all the requirements. Incorporating backbone flexibility into the matching stage, possibly in a manner similar to the method developed by Havranek *et al.*[15], would likely increase the number of matches that can be found for complicated theozymes.

Second, the enzyme design protocol so far only considers one state of the reactant, or one snapshot of the reaction trajectory. This means that Rosetta will try to design a sequence that optimally stabilizes this state, while ignoring the other states that also occur along the reaction coordinate. For example, when designing a catalyst for a reaction featuring large spatial rearrangements of atoms, Rosetta might converge on a sequence that sterically clashes with one of the substrate or product conformations. Natural enzymes have evolved to exquisitely compromise between stabilization requirements for every stage of the reaction trajectory[16]. To design efficient enzymes, it may be that all states of the reaction need to be modeled simultaneously, so that the designed sequence stabilizes the transition state more than any other, without destabilizing any other state too much. Developing a sequence selection algorithm that simultaneously takes a complete reaction trajectory into account is perhaps the biggest remaining challenge in computational enzyme design.

Third, ranking and selection of designs could be improved by the development of faster more thorough computational examination methods. Often, the catalytic side chains in the final designs will deviate from the idealized theozyme geometry. Further, the electrostatic potential created by the designed side chains that were not part of the theozyme also has an effect on the reaction's energy barrier, and thus the hypothetical stabilization achieved in a raw theozyme might be much less, depending on what scaffold this theozyme was placed in. Combined QM/MM hybrid approaches might be used to address this problem, but so far these methods are far too slow to be routinely employed for screening the hundreds or thousands of designs that Rosetta can suggest. Another factor that must be taken into account is the structural integrity of the designs. Even though the Rosetta designed sequence represents an energy minimum for the scaffold conformation, this does not mean that this sequence cannot fold into a different conformation. Local rearrangements of the backbone are not unlikely and have been reported for designed proteins. A possible method to examine designs for structural integrity is to run MD simulations[14], although this can quickly become prohibitively slow for large numbers of designs.

Despite the limitations listed above, from a purely practical standpoint, the Rosetta3 enzyme design protocol is still very useful. What distinguishes Rosetta's computational approach is that it is capable of generating catalytic activity from an inert scaffold, whereas most experimental methods, such as directed evolution approaches, rely on an existing catalytic activity as a starting point. Rosetta designed low-activity enzymes have been evolved to respectable catalysts[17], which shows that in combination with a high-throughput screening or selection strategy, Rosetta can facilitate the *de novo* creation of reasonable active enzymes.

Chapter 3

Computational design of catalytic dyads and oxyanion holes for ester hydrolysis

Disclosure: This chapter has been published as

Richter F, Blomberg R, Khare SD, Kiss G, Kuzin AP, Smith AJ, Gallaher JL, Pianowski Z, Helgeson RC, Grjasnow A, Xiao R, Seetharaman J, Su M, Vorobiev S, Lew S, Forouhar F, Kornhaber GJ, Hunt JF, Montelione GT, Tong L, Houk KN, Hilvert D, Baker D. "Computational design of catalytic dyads and oxyanion holes for ester hydrolysis." J Am Chem Soc. 2012 (epub ahead of print) Aug 7

My contributions consisted in running the computational design procedure, selecting designs to express, making the initial activity measurements, devising several of the optimization mutants and writing parts of the paper.

Abstract

Nucleophilic catalysis is a general strategy for accelerating ester and amide hydrolysis. In natural active sites, nucleophilic elements such as catalytic dyads and triads are usually paired with oxyanion holes for substrate activation, but it is difficult to parse out the independent contributions of these elements or to understand how they emerged in the course of evolution. Here we explore the minimal requirements for esterase activity by computationally designing artificial catalysts using catalytic dyads and oxyanion holes. We found much higher success rates using designed oxyanion holes formed by backbone NH groups rather than by sidechains or bridging water molecules and obtained four active designs in different scaffolds by combining this motif with a Cys-His dyad. Following active site optimization, the most active of the variants exhibited a catalytic efficiency (k_{cat}/K_M) of 400 $M^{-1}s^{-1}$ for the cleavage of a *p*-nitrophenyl ester. Kinetic experiments indicate that the active site cysteines are rapidly acylated as programmed by design, but the subsequent slow hydrolysis of the acyl-enzyme intermediate limits overall catalytic efficiency. Moreover, the Cys-His dyads are not properly formed in crystal structures of the designed enzymes. These results highlight the challenges that computational design must overcome to achieve high levels of activity.

Introduction

Hydrolytic enzymes play an important role in numerous physiological and pathological processes such as inflammation,[1] angiogenesis,[2] cancer,[3,4] and diabetes.[5] These enzymes are also established tools for the industrial synthesis of fine chemicals. For example, lipases and hydrolases are often used for the production of optically pure molecules[6] and the modification of complex natural products such as antibiotics,[7] steroids[8] and the anti-cancer drug taxol.[9] Hydrolytic enzymes have been mimicked by cyclodextrins,[10-12] cyclophanes[13] and other synthetic molecules.[14] These organic compounds provide hydrophobic binding sites for their substrates, and, like natural lipases and serine proteases, employ alcohols as nucleophiles to effect ester cleavage.[12] However, in these systems only the first step of ester hydrolysis is usually accelerated, namely the nucleophilic attack of a hydroxyl group on the ester substrates to give a covalent acyl intermediate. True turnover catalysis has been achieved by catalytic antibodies.[13] Structural studies show that these antibodies usually operate by stabilizing the negatively charged intermediate of the hydrolysis reaction rather than by nucleophilic catalysis.[14] Phosphonate haptens generally fail to program for more elaborate arrays of catalytic functionality, although a nucleophilic histidine was elicited in at least one case.[15,16] A nucleophilic histidine was also used in a designed thioredoxin with hydrolytic activity.[17]

Natural hydrolytic enzymes often utilize a serine or cysteine as a nucleophile, which is deprotonated by a hydrogen bonded histidine.[18] Precisely positioned hydrogen bond donors, so called “oxyanion-holes”,[19] stabilize the oxyanion intermediate. The importance of these elements has been demonstrated by mutagenesis experiments in which the removal of any of these functional groups leads to drastic losses in activity. However, to our knowledge there have to date been no complementary efforts to build esterase catalysts in a bottom up approach based on this catalytic machinery. With the approaches enumerated in the previous paragraph it is challenging to program an appropriately positioned nucleophile and oxyanion stabilization in the same catalyst, and hence difficult to systematically assess the extent to which functional esterase active sites can be built using combinations of these catalytic elements.

Here, we explore the fundamentals of esterase catalytic machinery using computational enzyme design.[20-22] The Rosetta3[23] *de novo* enzyme design protocol[24] is used to embed catalytic dyads and appropriately positioned oxyanion holes into catalytically inert protein scaffolds. We find that a minimalist catalytic schema consisting of a cysteine nucleophile, a nearby histidine together with a backbone NH group to stabilize the oxyanion intermediate is sufficient to generate primitive esterases. The relatively high success rate in generating esterases with this strategy suggests that a similar mechanism could have been employed by the nascent ancestors of modern enzymes.

Results

Computational design

We set out to explore the extent to which active esterases could be generated using the catalytic elements found in natural hydrolytic enzymes. Esterases often employ serines[25] and cysteines[26] as nucleophiles. We focused on cysteine as the reactive group as it is more nucleophilic than serine, has a lower pK_a and is hence a better leaving group. In natural cysteine hydrolases, a histidine residue, usually oriented and activated by another hydrogen bond acceptor such as Asp/Glu or backbone oxygen, acts as a general acid/base to deprotonate the nucleophilic cysteine in the first step and the water in the second, and to protonate the leaving group of the tetrahedral intermediates.[27] Two or three hydrogen bond donors stabilize the oxyanion reaction intermediate. A backbone amide group usually forms at least one of these oxyanion contacts.

The general hydrolysis mechanism catalyzed by this active site arrangement is depicted in Figure 1A.

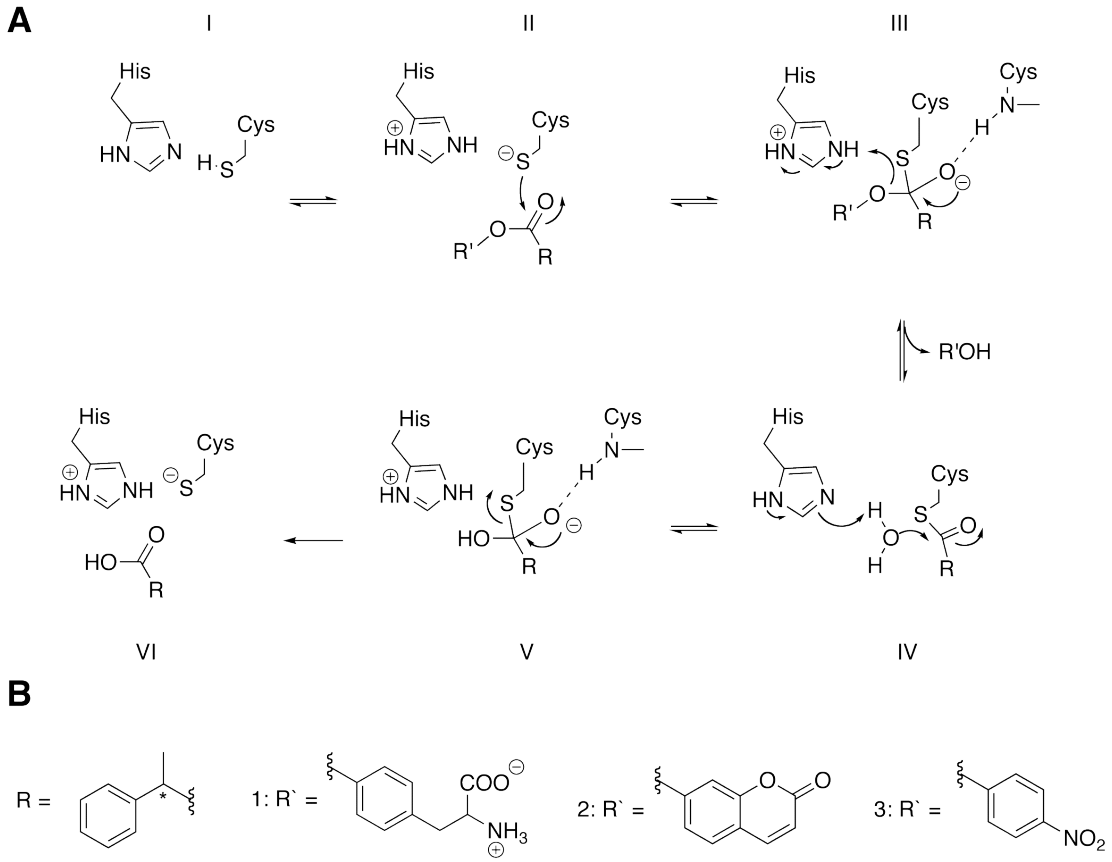


Figure 1 Schematic representation of the reaction catalyzed by the *de novo* designed esterases and of the employed substrates. A) Representation of the programmed reaction scheme and B) representation of the substrates: tyrosyl ester 1 is the computationally designed substrate, whereas the fluorogenic coumarin ester 2 and the chromogenic *p*-nitrophenyl ester 3 are utilized to facilitate the activity screens.

Three esters (**1**, **2**, and **3** in Figure 1) were chosen as model substrates. These three compounds have identical acyl-groups, but differ in the degree of activation of their aromatic leaving groups. Tyrosyl ester **1** was used for the computational design process. Although it is the least activated substrate, cleavage of this compound would lead to the production of tyrosine, and thus allow for the development of a high-throughput growth selection assay based on complementation of an auxotrophic bacterial strain unable to biosynthesize this essential amino acid.[28] Hydrolysis of umbelliferyl ester **2**, which exhibits intermediate reactivity, can be monitored by a sensitive fluorescence assay. The *p*-nitrophenyl leaving group of ester **3**, the most activated substrate, is identical to that of *p*-nitrophenyl acetate, a model substrate often used in other studies of ester hydrolysis.

In a set of 214 scaffold proteins,[21] we used RosettaMatch[29] to search for constellations of protein backbones that could accommodate these functional groups (Cys, His, Asn/Gln, and two

backbone-NH oxyanion hole contacts) in cysteine hydrolase-like geometries. Initial calculations showed that no placements could be found for this five-residue arrangement (data not shown).

Mutation of the catalytic triad asparagine/glutamine to alanine does not abolish catalytic activity in natural cysteine hydrolases,[30] suggesting that a Cys-His dyad should be sufficient for activity. Furthermore, the oxyanion intermediate can, in principle, be stabilized by any appropriately positioned hydrogen bond donors including water molecules (which presumably perform this function in the uncatalyzed reaction). Therefore, we generated theozymes[31,32] as described in the methods section consisting of the central cysteine residue programmed to carry out nucleophilic attack, a histidine residue to assist with the various proton shuffling steps occurring during the reaction, and three possible oxyanion-stabilization schemes (Figure S10). In the first set of designs, a backbone-NH group serves as the oxyanion stabilizer, and in the second and third theozymes, explicit water molecules or sidechain functional groups were used. An example of the first theozyme is shown in Figure 2A. RosettaMatch[29] was then used to scan for matches to these theozymes in the set of 214 protein scaffolds. For the three-residue theozyme I (featuring one backbone-NH), a total of 207 unique matches was identified in 81 distinct scaffolds. In 178 of the matches, the amide backbone of the cysteine itself provided a hydrogen bonding contact to the oxyanion of the tetrahedral intermediate, as often observed in natural cysteine and serine hydrolases. Every match was designed 100 times and the resulting designs were filtered and ranked as described in the methods section. We selected 31 theozyme I designs, 12 theozyme II designs, and 12 theozyme III designs for experimental testing (Table 1S).

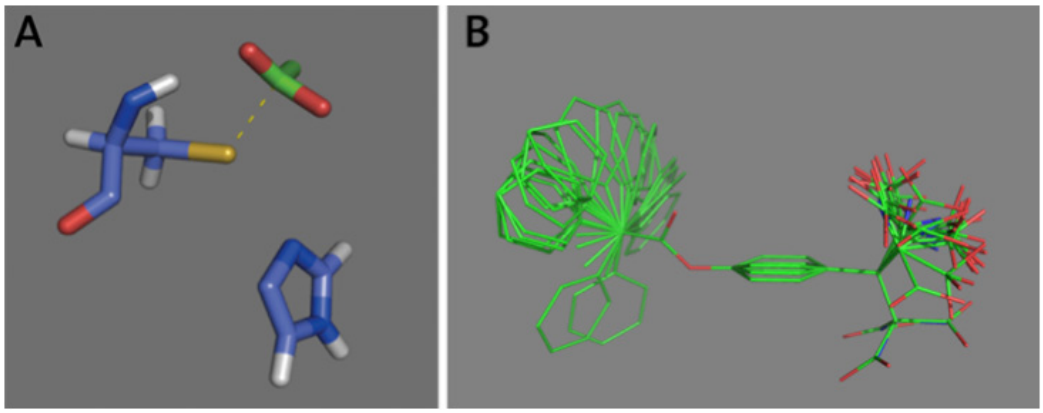


Figure 2 Snapshots of the computation design process. A) Representation of the calculated theozyme of the ester substrate framed by the catalytic dyad (Cys-His) and the backbone-NH – oxyanion contact. Note that in this case, the backbone-NH contact is made by the cysteine itself. B) Image of the conformer ensemble of the tyrosyl ester as created by the software Omega (OpenEye).

Initial activity screen

Genes encoding the 55 designs were cloned into the pET29b+ vector (Novagen) and the proteins were expressed and purified by Ni-NTA affinity chromatography. Purified soluble protein was obtained for 19 theozyme I designs, five theozyme II designs, and eight theozyme III designs; the remaining designs did not yield soluble protein. Hydrolytic activity was evaluated by measuring the increase in fluorescence due to the hydrolysis of coumarin ester **2**. Four of the theozyme I designs showed measurable activity, but none of the theozyme II or theozyme III designs were active, suggesting that backbone-NH groups might make more robust interactions

with the tetrahedral oxyanion than polar side chains or localized water molecules. For each of the four active designs, substitution of the catalytic cysteine and histidine residues with alanine, as well as the single knockout of the catalytic cysteine, either abolished activity completely (FR29, ECH13, ECH19) or decreased it considerably (ECH14), indicating that the source of the observed activity is in fact the designed active site (Figure 3).

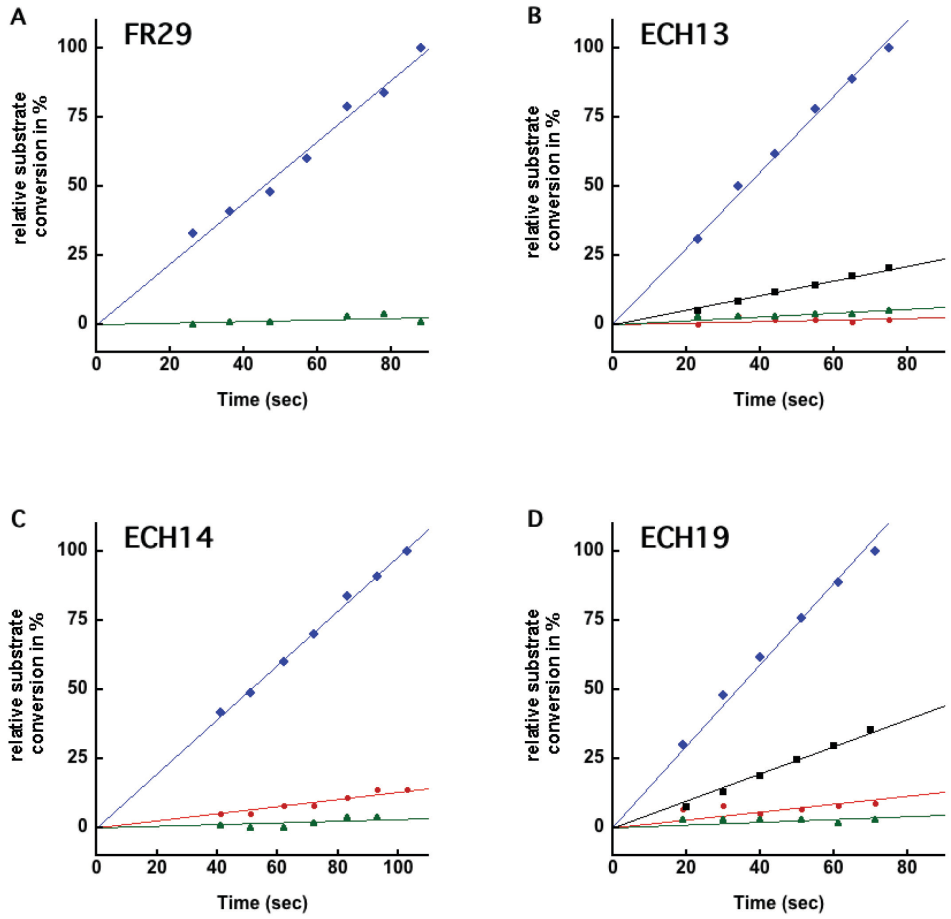


Figure 3. Experimental characterization of the active designs and their respective knockout variants. The progress curves of the parental designs are depicted in blue, the traces of the single knockout variants (cysteine) are shown red, traces of the double knockout variants (cysteine and histidine) are illustrated in green, and the histidine knockout variants for ECH13 and ECH19 are shown in black: A) FR29; B) ECH13; C) ECH14; D) ECH19. The enzymes (5 μM) were tested with the coumarin ester 1 (FR29 design: 20 μM; ECH designs: 50 μM) and the reaction progress was monitored by measuring the appearance of the fluorescent coumarin product (excitation wavelength: 340 nm; emission wavelength: 452 nm). For each graph, the amount of substrate that was converted by the active designs after 70-100 sec was set to 100 % and then used to normalize the entire data set. The background was subtracted in all cases and the linear fits were extrapolated to zero substrate conversion.

Optimization of the active designs

Although the computational design process afforded novel active sites capable of cleaving the activated coumarin ester, the rate accelerations over background provided by the esterases are modest. Guided by visual inspection and by evaluation of the active site dynamics, we explored the effects of mutations on catalytic activity for each of the designs. The optimized variants were

generated as described in the SI. For FR29 we constructed and analyzed 46 single mutants and for the ECH designs we screened 5-8 mutants each. Mutations increasing activity more than 1.5 fold were combined into second-generation variants and the hydrolytic activity was re-tested. For design FR29 this step was carried out twice to generate third-generation variants, which contained up to seven mutations compared to the parental design (SI, Figures 1S and 2S). A set of three mutations (A44S/T112L/V151L) was found to increase the overall catalytic efficiency 17-fold relative to the original design. ECH13 showed a low tolerance to additional mutations, and most variants exhibited a dramatic decrease in expression yield (SI, Figure 1S A). ECH14 was more tolerant of substitutions, but since mutation of the active site cysteine does not completely eliminate activity we did not explore these further.

ECH19 is based on a periplasmic binding protein (PBP)[33] scaffold that contains two large domains connected by a flexible hinge region, allowing the two domains to open and close around a cavity. The crystal structure of the closed form was used in the design, but most PBPs are in the open conformation in the absence of ligand. MD simulations[34] of ECH19 show an irreversible transition from closed to open conformation, both with and without substrate (Fig S7). Single-mutant DDG calculations[35] on the open and closed forms of the ECH19 scaffold suggested that incorporation of a proline at position 354, which is located in the hinge region, would increase the stability of the closed form. A variant with two mutations, K354P and P364W, enhanced the esterase activity 4-fold compared to the original design (SI, Figure 1S C). The P364W mutation was intended to enhance the binding of the esters by optimizing packing around the acyl moiety of the substrate.

Biophysical and kinetic characterization of the ester hydrolases

For more detailed biophysical and biochemical characterization, the designed hydrolases were purified by an additional anion exchange step following the standard Ni-NTA affinity chromatography. Variant FR29 was further purified by either gel filtration or GST-affinity chromatography. The specific activity was found to be independent of the purification method, arguing against contamination by endogenous esterases. Consistent with this conclusion, active site alanine mutants of the designed esterases, which were purified in an identical manner to the active biocatalysts, did not convert the ester substrates above the buffer background rate. Mass spectrometric analysis confirmed the identity of the individual variants (SI, Table 2S), and circular dichroism (CD) measurements verified that they were folded and exhibited similar stabilities to the parental designs (SI, Figure 3S, Table 3S).

For all four designs, cleavage of ester **2** exhibits a biphasic time course with an initial fast phase followed by a second slow phase (Figure 4). The slow phase is roughly 2-3 fold above the spontaneous hydrolysis reaction, whereas the fast phase varies considerably depending on the analyzed variant. Similar behavior was observed for the cleavage of ester **3**. The burst phase was studied as a function of substrate concentration to determine steady-state parameters (Figure 4, Table 1). The computationally designed ester hydrolases exhibit Michaelis-Menten kinetics. Apparent bimolecular rate constants (k_{cat}/K_M) for the acylation step range between 10 $M^{-1}s^{-1}$ (FR29) and 80 $M^{-1}s^{-1}$ (ECH13) for the conversion of coumarin ester **2**. Para-nitrophenyl ester **3**, which is also a much better mimic of tyrosyl ester **1**, towards which the computational designs were generated, is hydrolyzed with up to 4-fold higher catalytic efficiency. In this case, the values for k_{cat}/K_M range between 30 $M^{-1}s^{-1}$ (FR29) and 320 $M^{-1}s^{-1}$ (ECH13). The slightly higher turnover numbers can be attributed to the greater reactivity of substrate **3** compared to ester **2**. However, cleavage of ester **3** is also characterized by a two to three-fold lower Michaelis constant, consistent with a better fit to the active site.

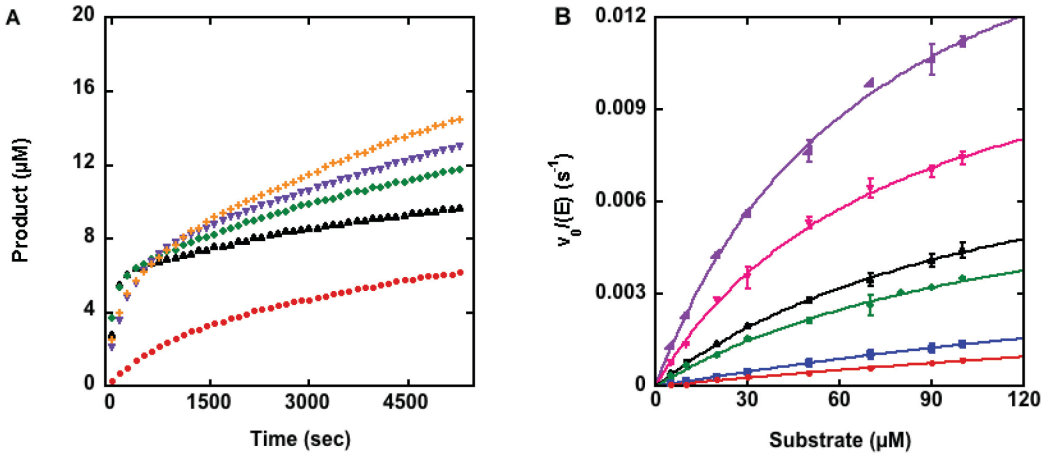


Figure 4 Kinetic analysis of the designed esterases: A) Two-phase progress curves of selected *de novo* designed ester hydrolases. The conversion of 100 μM coumarin ester 1 by 10 μM of FR29 (red), ECH13 (black), ECH19 (green), FR29 T112L (purple) and FR29 A44S/V151L (orange) consists of a initial fast phase followed by a second, slow phase. B) Michaelis-Menten plots of the hydrolysis of coumarin ester 2 and by the *in silico* designed ester hydrolases and their best evolved variants (red: FR29; blue: ECH14; green: ECH19; black: ECH13; pink: FR29 A44S T112L V151L: purple: ECH19 K354P P364W). Only the slopes of the fast phases were considered for the determination of k_2 and K_m.

As summarized in Table 1, the catalytic efficiency of the acylation step could be successfully improved up to 17-fold by introducing point mutations into the parental designs. The best third-generation ester hydrolase variant, FR29 A44S/T112L/V151L, exhibits k_{cat}/K_M values of approximately 400 $M^{-1}s^{-1}$. This value is within the range achieved by typical catalytic antibodies. However, the best hydrolytic antibody has been found to display a 1000-fold higher catalytic efficiency[15,16] and the natural cysteine protease papain cleaves *p*-nitrophenyl hippurate with a k_{cat}/K_M of 1.8×10^5 $M^{-1}s^{-1}$.[36]

To assay the cleavage of tyrosyl ester **1** – which is the designed substrate, but less reactive than esters **2** and **3**, we monitored formation of the product tyrosine by HPLC. Although no cleavage of this substrate was detected in the presence of the designed proteins, this compound inhibits the hydrolysis of ester **2** indicating binding to the active site. IC_{50} values were determined for the most active hydrolases FR29 A44S/T112L/V151L and ECH19 K354P/P364W by recording esterase activity in the presence of increasing concentrations of inhibitor. Tyrosyl ester **1** inhibits the FR29 and ECH19 variants with an IC_{50} of 37 μM and 45 μM corresponding to K_i values of 23 μM and 27 μM, respectively (Figure 5).[37]

Enzyme	Substrate	$k_2 * 10^3$ (s^{-1})	K_M (μM)	k_2/K_M ($M^{-1}s^{-1}$)
FR29	**2**	5.1 ± 0.7	500 ± 180	10
	3	4.1 ± 0.6	120 ± 20	34
A44S/T112L/V151L	**2**	13.3 ± 0.6	78 ± 7	170
	3	15.4 ± 1.1	38 ± 5	405
ECH13	**2**	9.6 ± 0.4	120 ± 10	80
	3	17.6 ± 1.9	57 ± 10	309
ECH14	**2**	6.3 ± 0.9	360 ± 60	17
	3	8.2 ± 1.9	130 ± 40	63
ECH19	**2**	7.7 ± 0.7	125 ± 20	62
	3	10.0 ± 0.7	44 ± 5	227
K354P/P364W	**2**	19.5 ± 0.7	73 ± 5	267
	3	n.d.	n.d.	n.d.

Table 1 Kinetic parameters of the *in silico* designed ester hydrolases and their evolved variants. Measurements were done in buffer (25mM HEPES, 100mM NaCl, 5% acetonitrile) at pH 7.5 and 29°C. Only the rates of the initial phase were considered for the determination of k_2 and K_m. n.d. = not determined

Burst kinetics is expected for a 2-step reaction mechanism in which the substrate rapidly reacts with the enzyme to form stoichiometric amounts of a stable enzyme-bound intermediate that subsequently breaks down slowly (eq 1). The observation of the burst phase is thus evidence that the four novel hydrolases function through the designed mechanism; the fact that biphasic kinetics is not observed for the corresponding cysteine-knockout variants supports this conclusion. For the ECH designs and the optimized FR29 variants the second phase is much slower than the first phase ($k_2 << k_1$) and the observed bursts approximately correspond to the concentration of enzyme employed, and hence to a single turnover. This correlation provides additional evidence that the observed activities are due to the designs and not to a high activity esterase contaminant.

$$E + S \underset{P_1}{\overset{k_1}{\rightleftharpoons}} EI \xrightarrow{k_2} E + P_2 \qquad (1)$$

Typically the acyl-enzyme intermediate that is formed in serine and cysteine proteases undergoes hydrolysis, but can also undergo aminolysis when in the presence of amine nucleophiles.[38] Since aminolysis of thioesters is known to be significantly faster than for esters,[39] we attempted to facilitate liberation of the acyl-enzyme complex by adding amine nucleophiles spanning a range of reactivity to the assay mixtures (SI, Figure 6S) but no increase in activity was observed.

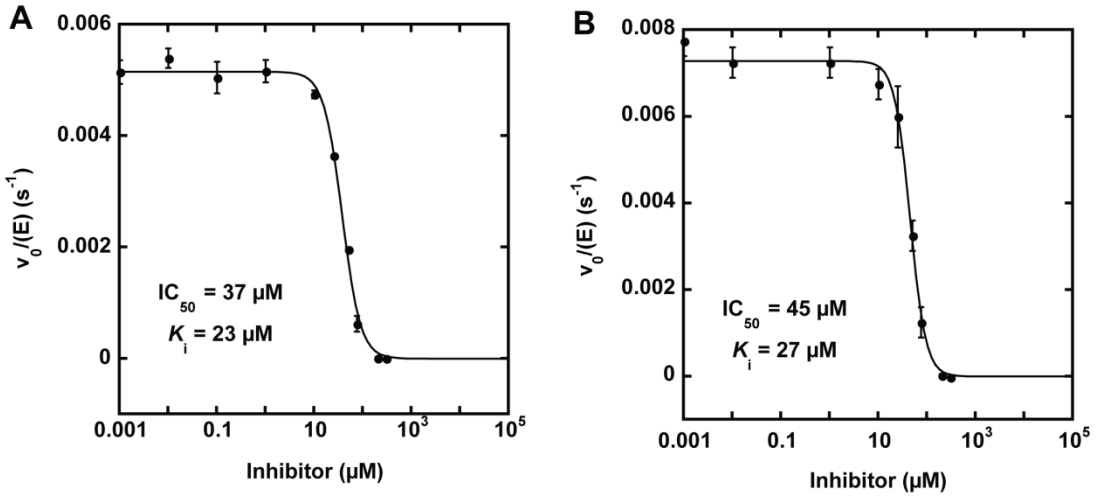


Figure 5 Inhibition of a FR29 and a ECH19 variant by tyrosyl ester 1. The IC_{50} of tyrosyl ester 1 was determined by incubating A) FR29 A44S/T112L/V151L ([E] = 2 μM) and B) ECH19 K354P/P364W ([E] = 2 μM) with increasing amounts of the inhibitor ([c] = 0.001 – 300 μM) before 50 μM of coumarin ester 2 was added and the release of coumarin was monitored. The IC_{50} was determined by curve fitting and converted into the corresponding K_i using the Cheng-Prusoff equation.[37]

Characterization of the acyl-enzyme intermediate by mass spectrometry

In analogy to natural cysteine hydrolases, the most likely intermediate formed during reaction of the designed proteins with substrate is an acyl-enzyme adduct. To provide support for this hypothesis, we attempted to isolate this species and characterize it by mass spectrometry. For FR29, ECH13, and ECH19, a shift of 133 Da was observed for the major mass peak of the protein after incubation with substrate **2**. This increase corresponds to the acyl fragment of the substrate, as expected for acylated proteins. Consistent with this interpretation, the identically treated variants of FR29 and ECH19, in which the active site cysteine and histidine were mutated to alanine, did not show an increase in mass following incubation with substrate. In the case of ECH14, the protein was modified up to seven times and mutation of the active site cysteine did not alter the acylation pattern (Figure 5S, Table 4S). In contrast to the results obtained with activated ester **2**, no acyl-enzyme intermediate was observed with tyrosyl derivative **1**, which inhibits the enzymes but is not cleaved and thus does not covalently modify the proteins (Table 5S).

To reveal the location of the modification, the acylated proteins were analyzed by matrix-assisted laser desorption/ionization tandem mass spectrometry (MALDI-MS/MS). For variants

FR29 A44S/V151L/M133Y (a precursor of the most active FR29 variant) and ECH13 the active site cysteine was found to be acylated. In the case of ECH13, an additional modification site was identified, namely a cysteine at the C-terminus of the protein. Unfortunately, low peptide intensities precluded conclusive MALDI-MS/MS analysis of ECH14 and ECH19.

In summary, the MS data show that the designed cysteine functions as the catalytic nucleophile in three of the four designs (ECH13, ECH19, FR29), while in ECH14 the results are less clear. These findings fit well with the previously reported observation that ECH13 and ECH19, but not the cysteine-knockout variants, efficiently react with cysteine-hydrolase specific probes.[41]

Structural characterization

Crystal structures of the apo forms of ECH13 (at 1.6 Å resolution, PDB code 3u13), ECH14 (at 3.2 Å, 3uak), ECH19 P364W (at 2.55 Å, 3u1o) and FR29 A44S/T112L/V151L (at 2.8 Å, 3u1v) were determined (Figure 6, table 10S). The overall backbone structures of the designed proteins were similar to the design models. However, in each case either the designed histidine residue or the nucleophilic cysteine adopt a conformation different from the design model, and the dyad is not formed as desired; this was also observed in MD simulations (Figures S7, S8, S9). Furthermore, in each case, either the histidine or cysteine is in regions of the protein with relatively high flexibility. For two designs that were based on the ligand-bound, holo conformation of scaffolds which undergo global conformational changes upon ligand binding (ECH19, based on a periplasmic binding protein[34] and FR29, based on a tryptophanyl-tRNA synthetase[40]), crystal conformations resemble the unliganded, apo conformation of the scaffold. MD simulations also suggest that the unbound conformation is thermodynamically more favored. Incorporation of backbone flexibility and avoidance of alternative, non-catalytic, states can consequently be envisioned as avenues for improvement of the design methodology.

Even though dyads are not formed in the apo-crystal structures of ECH13 and ECH19, substituting the catalytic histidine with alanine leads to diminished catalytic activity (Fig 3). This observation indicates that the histidine does take part in the reaction and that the energy gap between the designed (holo) and observed (apo) conformations may be overcome upon ligand binding. Although preliminary experiments to crystallize the enzymes with compound **1** and **2** have been unsuccessful, structures with bound ligand will be important for verifying this hypothesis. The catalytic activities would presumably be significantly higher if the desired theozyme geometry had been fully realized in the designs.

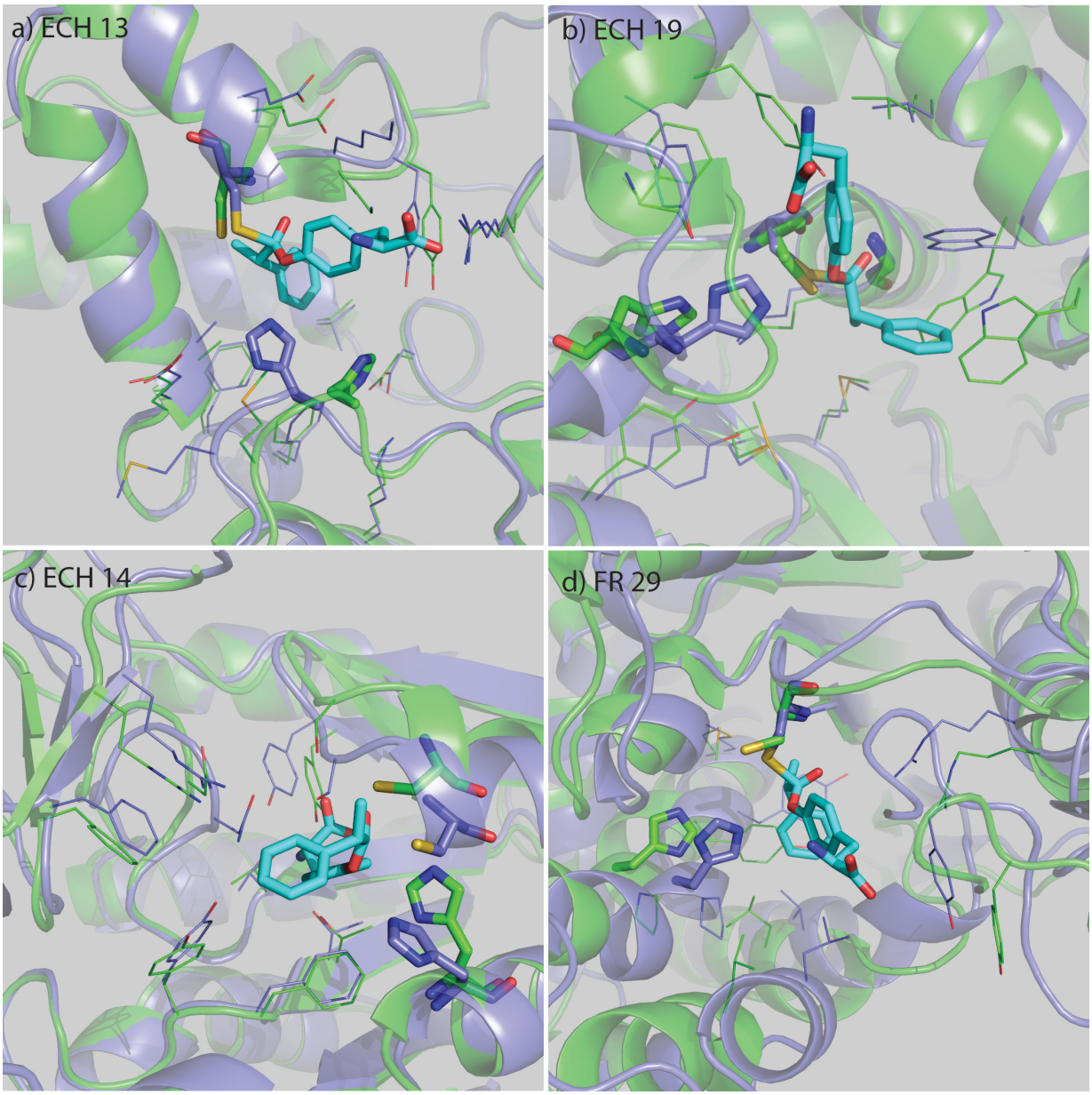


Figure 6 Crystal structures of the four active designs. In each case, the design model is shown in purple (ligand in cyan) and the crystal structure in green. The theozyme residues and the ligand are shown in stick representation, and selected other active site residues in line representation. A) ECH13: The Cα RMSD between design model and crystal structure is 0.97 Å over the 15 active site residues. The catalytic histidine, His100, is in a rotameric conformation different from the design model, and instead of pointing towards the ligand and Cys45, it makes a hydrogen bond with Asp10. This alternative conformation is facilitated by a small backbone shift between residues Pro99 and Phe103, and the observed close interaction between His100 and Asp10 would not be possible with the scaffold backbone conformation that served as the template for the design.
B) ECH19 P364W: The design was based on the closed conformation of a periplasmic binding protein, but the apo protein crystallized in the open form, with an RMSD of 4.1 Å to the design model but an RMSD of only 1.6 Å to the open form of the scaffold protein (PDB 2uvg). The designed active site is mostly located in one of the scaffold's two domains, close to the interdomain cleft. When superimposing design model and crystal structure based solely on the

active-site containing domain, the resulting RMSD is 1.5 Å. However, the catalytic His226 does not interact with Cys161 as designed, but adopts a different rotameric conformation to interact with the sidechain-hydroxyl of Tyr250 and the backbone oxygen of Phe221. The stretch from Tyr218 to Lys230 that contains His226 has high relative B-factors, suggesting that it is fairly flexible.
C) ECH14: the crystal structure has an RMSD of 1.4 Å to the design. The catalytic dyad is not formed, as the Cys132 containing loop-helix stretch between residues 127 and 140 moves upward away from the active site and His104 reorients around chi2. This unexpected movement may result from the W130S mutation, since W130 stacks against the PLP cofactor of the wild type scaffold and thus locks this backbone segment into the conformation used as the design template.
D) FR29 A44S/T112L/V151L: The apo-structure of FR29 is more similar to the unliganded, more open conformation of the scaffold (PDB 1D2R, 0.86 Å RMSD) than to the ligand-bound structure (PDB 1mau, 2.7 Å RMSD) which was used as the template for the design. The catalytic dyad is not formed, since in the apo form of the scaffold the helix-turn-helix motif between residues 106 and 132 that contains the catalytic His125 moves outward relative to the catalytic Cys9, leading to a shift in the His125 Cα – Cys9 Cα distance from 10.9 Å to 12.4 Å. As the backbone of most of the designed active site residues shifts between the liganded and the apo structure, the active site is generally more open than in the design model.

Discussion

We have described the construction of primitive esterases by combining Cys-His dyads with appropriately positioned oxyanion holes. Comparison of the structures and activities of the designs to those of native hydrolases and catalytic antibodies (Table 2) provides insight into the contributions of these functional elements to catalysis and where improvements can be made in the design process.

Enzyme	$k_{cat} \times 10^{-3}$ [s^{-1}]	K_M [μM]	k_{cat}/K_M [$M^{-1}s^{-1}$]
FR29 (best variant)	15.4 ± 1.1	38 ± 5	405
CNJ206[52,a]	7 ± 0.8	80 ± 10	87
48G7[53,b]	35	113	310
17E8[54,c]	817 ± 28	215 ± 33	3.7×10^3
43C9[55,d]	460 ± 50	470 ± 160	979

Table 2 Kinetic parameters for ester cleavage catalyzed by artificial biocatalysts. [a] **30mM TBS, pH 8.0**
[b] **10mM Tris-HCl, 50mM NaCl, pH 8.2, 37°C**
[c] **50mM Tris-HCl, 150mM NaCl, pH 8.7**
[d] **100mM ACES, 50 mM Tris-HCl, 50mM CAPS, pH 8.5, 25°C**

The first insight relates to the relative ease with which the developing negative charge on the carbonyl oxygen can be stabilized by backbone NH groups. We experimented with different strategies for oxyanion stabilization using oxyanion holes formed by backbone NH groups, sidechain NH groups, or water molecules. Notably, all of the active designs utilized a backbone NH group for oxyanion stabilization. Four out of 19 (~20%) soluble designs with backbone NH based oxyanion holes had activity, while none of the designs using sidechain or discrete water-mediated oxyanion stabilization were active. Although direct evidence for an oxyanion binding site is lacking, and there may be other reasons why the theozyme II and III designs did not work, this anecdotal correlation is suggestive. In all active designs except ECH19, the backbone amide of the active site cysteine residue itself provides oxyanion stabilization, and in all active designs except FR29, the catalytic Cys residue and the oxyanion donor are at the N-terminus of an alpha-helix which may provide additional oxyanion stabilization due to its helix dipole, and this might represent an intrinsic advantage of theozyme I. Backbone oxyanion holes are found in almost all proteases and many esterases and have evolved independently multiple times during evolution. Because the protein backbone is on average more rigid than sidechains, it is possible that the preferential use of backbone amides for oxyanion hole stabilization in designed (and natural) hydrolases reflects the advantages of pre-organized active sites that are poised to stabilize the oxyanion intermediate, thereby aiding nucleophilic attack. In natural hydrolases, the full oxyanion hole is often formed by two backbone-NH groups, and the removal of only one of these can reduce catalytic efficiency 10^2 to 10^3-fold.[46] Adding a second interaction to a partial oxyanion hole of the computational designs might therefore improve catalytic efficacy. This suggestion is supported by comparison with the antibody esterases, which have more completely formed oxyanion holes and increased activity even without a catalytic dyad or triad.[14]

The second insight concerns the relationship between the reactivity of the catalytic cysteine and the effectiveness of the active site for catalysis. The active site cysteines in ECH13 and ECH19 react as strongly with cysteine protease-specific probes as the nucleophiles in natural cysteine proteases.[40] Nevertheless, the acylation efficiency of the designed esterases is more than three orders of magnitude lower than that observed for the natural hydrolase papain.[43] Thus, activating the cysteine is not difficult for even primitive designs to accomplish. The contrast between high nucleophilicity of the active site cysteine yet low catalytic activity may be due to relatively poor oxyanion stabilization, or as suggested in studies with thiosubtilisin,[44] a failure to protonate the substrate leaving group. Nevertheless, simply placing a cysteine in a protein binding pocket is not sufficient to generate an active hydrolase, as evidenced by the lack of activity observed for the other soluble designs.

The third insight concerns the effectiveness of the combination of the catalytic dyad and minimalist oxyanion holes utilized in the designs for acylation versus deacylation. The most proficient of the designs accelerates acylation, the first step of the reaction sequence, with an apparent bimolecular rate constant of about 400 $M^{-1}s^{-1}$ and a rate acceleration of approximately 3.7×10^3 -fold over background. However, in contrast to naturally occurring enzymes and antibody catalysts, the computationally designed proteins are not efficient multiple turnover catalysts. Kinetic and mass spectrometric studies establish that the nucleophilic cysteine attacks the substrate and becomes acylated as programmed by design, but the resulting acyl-enzyme intermediate is hydrolyzed only slowly. Deacylation evidently places more demands on the catalytic machinery than the acylation step. This is due, at least in part, to the fact that the designs work on activated esters. It may also reflect the fact that the theozyme does not capture the full complexity of this multi-step transformation. While theozyme I may be a good representation for the initial transthioesterification step, it does not explicitly model general base

delivery of water to the acyl-enzyme intermediate or possible contributions of conformational dynamics and preorganization towards catalysis.

The fourth insight relates the contribution of the His of the engineered Cys-His dyad to the actual geometry of the active site. The crystal structures and the MD simulations show that the engineered Cys-His dyad is usually not in the designed conformation, yet mutation of the histidine to alanine reduces the rate of catalysis. Evidently histidine can promote weak ester hydrolysis by a nearby cysteine in the absence of a stable hydrogen bond between the two, which suggests that the familiar catalytic triad could have evolved in a stepwise fashion, with a histidine in the vicinity of a cysteine promoting catalysis and only later fixed in place by the third amino acid in the triad. By analogy, increasing the catalytic efficacy of the designed esterases will likely require expansion of the dyad motif into a full catalytic triad to stabilize the histidine in the desired conformation and promote its alternating roles as general base and acid in the overall catalytic cycle. The hydrogen bonds between thiol donors and nitrogen acceptors are not as strong as those between oxygen and nitrogen donors,[45] and hence the interaction energy between the cysteine and the histidine might not suffice to hold the histidine in position. Backing up the histidine with a hydrogen-bond acceptor would also favor the correct tautomeric state necessary for the histidine to deprotonate the cysteine in the acylation step and water in the deacylation step.[46] The extent to which such a designed triad is exposed to bulk-solvent is also likely to be important. Mutation of naturally occurring enzymes illustrates the importance of these effects; for example the turnover number (k_{cat}) of papain drops by two orders of magnitude when the catalytic asparagine of the Cys-His-Asn triad is replaced with an alanine residue[28] and larger effects have been observed in other hydrolases.[47]

There are a number of avenues for improving the primitive esterases described in this manuscript. Foremost among these are converting the poorly formed catalytic dyad into a full catalytic triad to hold the histidine in place for cysteine activation, leaving group protonation, and water activation in the deacylation step. Achievement of this goal would be expected to speed up the rate-limiting deacylation step as well as acylation of the enzymes by less activated ester and amide substrates. Supplementing the single backbone NH group in the primitive oxyanion holes with additional backbone or sidechain hydrogen bond donors for better tetrahedral intermediate stabilization is also likely to be important. Creating these more sophisticated active sites will require going beyond the fixed-backbone approach utilized in this work since, as found in the RosettaMatch calculations described in the results section, the five required functional groups cannot be placed in the proper relative orientations without modification of the backbone of the scaffolds.

Conclusions

Our “inside out” *de novo* design results complement extensive “top down” studies on the effects of removing catalytic triad residues from naturally occurring proteases and esterases. In nature, this special constellation of amino acids possibly evolved by the stepwise addition of the individual catalytic groups to an ancestral binding pocket. Our results suggest that nascent catalysts could have utilized a cysteine as a nucleophile with a histidine nearby and a backbone hydrogen-bond to stabilize the oxyanion. How subsequent evolution matured such primitive sites into the extremely proficient modern day catalysts is unclear; but improved design methods and laboratory evolution should not only afford progressively more active variants but also further insights into the origins of the strong inter-residue synergies that distinguish highly evolved hydrolytic enzymes.

Methods

Computational design

Quantum-mechanical methods were used to perform theozyme calculations. The optimal arrangement was computed for a model catalytic triad and oxyanion hole contacts along the reaction path of ester hydrolysis as reported by Smith et al.[18] Reactant, intermediates, and transition states were obtained by systematically stepping along the reaction coordinate, followed by optimizations towards the respective stationary point. Of particular interest for the purpose of enzyme design in general are the transition state (TS) geometries along the reaction steps. Here we focused on the first TS of the acylation step in which the ester substrate undergoes nucleophilic attack by a reactive thiol. The QM theozyme consists of a Cys-His-Glu/Asp triad and two oxyanion hole contacts and was computed at the B3LYP/6-31G(d) level of theory using Gaussian 03 (SI references). Currently, the computational expense of matching more than three groups into a protein scaffold presents a bottleneck in the design protocol, as a consequence of which a stripped down version of the theozyme was utilized (Figure 2A). The conformations of the catalytic Cys/His dyad and of an oxyanion hole are similar to the active site of human cathepsin K which was investigated in a previous QM/MM study.[27] An ensemble of ligand conformers was then generated employing OpenEye's Omega software[48] (Figure 2 B). The final theozyme thus consisted of a conformer library of the substrate in the transition state, the Cys-His catalytic dyad, plus one of the three oxyanion hole possibilities shown in Fig S10. Attachment sites for the theozyme in the scaffold set were found with an improved version[24] of the RosettaMatch algorithm.[28] For theozyme I, the 207 unique matches that were identified were subjected to 100 iterations of the standard Rosetta3 enzyme design protocol.[24] Briefly, all scaffold residues (except the matched catalytic residues) containing either a Cα within 6 Å of a matched ligand atom or both Cα and a Cβ atom within 8 Å of a matched ligand atom were considered design shell residues and mutated to alanine. The resulting structure was then subjected to a gradient-based minimization to optimize the three catalytic interactions. During this minimization, restraints were added to the energy function to enforce the desired theozyme geometry. Three rounds of sequence design with subsequent gradient-based minimization were carried out for the shell residues. Scaffold residues with a Cα within 13 Å of a ligand atom but not part of the design shell were allowed to change their rotameric state. Finally, the designed structures were repacked without the catalytic restraints. From the resulting 20700 design models, 1071 were selected based on the following criteria: (1) a ligand binding score less than -10.0 Rosetta energy units (REU), (2) no more than two unsatisfied, buried polar ligand atoms, (3) more than 66% of ligand surface area buried by the protein, (4) fewer than three buried unsatisfied polar atoms on the catalytic histidine, (5) fewer than two overall unsatisfied polar atoms compared to the respective wild type scaffold, (6) packing statistics[49] comparable to the wild type scaffold, and (7) greater than -2 nonlocal contacts compared to the wild type scaffold (residues with an interaction score <-1.0 REU and separated in sequence by at least eight residues were considered to be a nonlocal contact). The cutoff of -10.0 REU for the ligand binding score was determined according to the MASC method[50] by docking substrate **1** to a set of 68 random proteins with the Rosetta ligand docking protocol.[51]

Protein production, initial activity screening and biochemical characterization

All design constructs were cloned into pET29b and expressed in *E. coli* BL21 cells. Proteins providing crystal structures included NESG targets OR49 (ECH19 P364W), OR51 (ECH13),

OR52 (FR29 A44S/T112L/V151L), and OR54 (ECH14). pET expression vectors for these proteins and the single site mutants listed in Table 1 have been deposited in the PSI Materials Repository (http://psimr.asu.edu/). Purification was carried out as described in the SI. Initial activity determination, CD spectroscopy, melting curves and mass spectrometric analysis were carried out as described in the SI.

Kinetic measurements

The substrates were synthesized as described in the SI. Reactions were initiated by adding different amounts of coumarin ester **2** (5 μM to 100 μM final concentration) or *p*-nitrophenyl ester **3** (5 μM to 50 μM final concentration) in acetonitrile to 2 μM of protein (or no protein for the background reaction) in 25 mM HEPES buffer (pH 7.5), containing 100 mM NaCl and 5% acetonitrile. Initial reaction rates were determined as described in the SI. The initial rates divided by the catalyst concentration were plotted against substrate concentration, and k_{cat} and K_M were determined by fitting the data to the Michaelis-Menten equation $v/[\text{catalyst}]=k_2[\text{substrate}]/(K_M+[\text{substrate}])$ using Kaleidagraph software (Synergy Software). The aminolysis and KI determination measurements were performed as described in the SI.

Mass spectrometry

For standard mass determination, the protein samples were desalted using Illustra Nap-5^{TM} columns (GE Healthcare, Glattbrugg, Switzerland) and measured in 0.1% acetic acid (pH 2.0) by ESI-MS on a Daltonics maXis ESI-Q-TOF mass spectrometer (Bruker). The mass spectra of the proteins were deconvoluted using MaxEnt1 software. Protein samples were prepared as described in the SI. Then the samples were acidified, desalted using C_{18} ZipTips and analyzed by MALDI-MS in the positive-ion mode. Acylated peptides were identified by comparison of the mass spectra of treated proteins with the mass spectra of the corresponding negative controls. The identified peptides were additionally fragmented and the obtained fragments were again compared to those of the corresponding non-modified peptides. Additional details can be found in the SI.

Structure Determination and Molecular Dynamics Simulations

The structures were determined as a collaborative project of the Community Target Nomination program of the NIH PSI Northeast Structural Genomics Consortium (www.nesg.org). Details of the experimental procedure, as well as the Molecular Dynamics setup, can be found in the SI.

Chapter 4

A new method for Interaction-driven flexible backbone design

Abstract

In the current standard fixed backbone design protocol as described in Chapter 2, the backbone of the protein scaffold stays completely fixed in the scaffold conformation during the matching stage, and only small deviations are allowed during stage two. Experience has shown that with this fixed-backbone approach, a significant number of active sites will be found only for theozymes containing up to four amino acids. There are reactions however, for which more amino acids are needed to stabilize the transition state. Here, a generalized method, termed “Inverse Rotamer Remodelling” is described to vary a protein scaffold’s backbone conformation during the matching stage such that sites for theozymes with an arbitrary number of amino acids can be found. The method was benchmarked against a set of eight naturally occurring enzymes. The method was then used to generate variant models of ECH19 in which the catalytic histidine is making an additional hydrogen-bonding contact with either an amide (Asn or Gln) or carboxylate (Asp or Glu) side-chain.

4.1. The limits of fixed – backbone design

While there are no accurate ways of estimating how many unique sites can be found for a given theozyme definition and scaffold, it is evident that the more complex a theozyme is, fewer unique sites can be found by the matching process. The number of matches that can be found in a given scaffold depends on several factors: the size and shape of the ligand, the number of pockets in the scaffold and the number of residues bordering and pointing into these pockets, the flexibility of the desired active site amino acids, the desired precision in the geometric constraints, and the number of catalytic contacts. For example, searching for an active site containing two flexible side chains will yield far more matches than searching for a site containing three short, less flexible side chains.
In practice, experience has shown that searching for theozymes with two to three catalytic interactions usually yields on the order of 10-100 unique matches in the current scaffold set (containing 214 scaffolds), while theozymes comprised of more than four catalytic interactions almost never yield any matches. Theozymes featuring interactions mediated by Arg or Lys can be an exception, due to the large rotameric diversity for these long sidechains. The experience with esterase design as described in Chapter 3 highlights this. The active sites of natural serine and cysteine hydrolases usually contain five catalytically critical interactions, as depicted in Fig 4.1a: the Ser/Cys nucleophile, the His proton shuttle, a residue to back up the histidine (usually N/Q/D/E), and two oxyanion hole contacts, which are usually provided by backbone N-H groups. However, a significant number of matches could be found only for the rudimentary theozyme (Fig 4.1b) consisting of the nucleophile, the histidine, and one oxyanion hole contact.

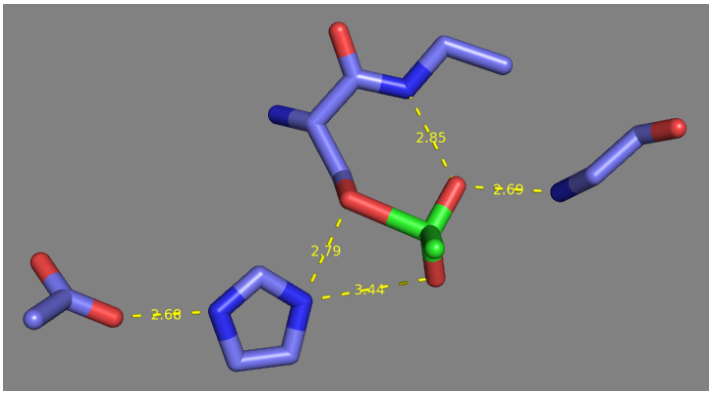


Figure 4.1a A typical natural esterase active site

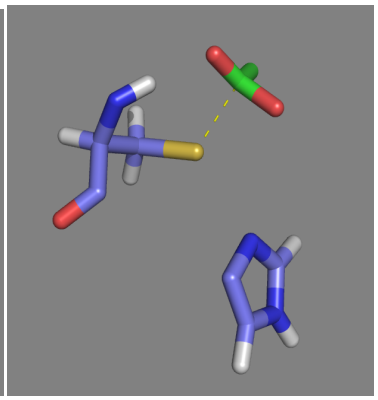


Figure 4.1b The minimal theozyme used in Chapter 3

The actual data for this example are shown below in Table 4.1: three searches for a cysteine protease-like active site were done. In the first, only the core three interactions (Cys-His dyad and one backbone N-H oxyanion-hole) were required, as schematically shown in Fig 4.1b. In the second, the core active site plus either an Asn or Gln to interact with the histidine at the other nitrogen was added to the active site (as observed in catalytic triads), while in the third search, a second oxyanion-hole contact was searched for. In the fourth search, the core active site and both the histidine-orienting Asn or Gln and the additional oxyanion contact, i.e. the full five element theozyme, were searched. While searching for the core active site yielded more than 100 unique sites, only a single match was found for the full, naturally observed active site.

Table 4.1 Number of matches for different active site definitions

Theozyme	Core (Fig 4.1b)	Core + 2nd histidine contact with either Asn or Gln	Core + 2nd backbone-NH oxyanion contact	Core + 2nd oxyanion contact + 2nd histidine contact
#unique matches	178	29	3	1

In summary, it is usually not possible to find large numbers of matches for active sites containing more than four predefined contacts. This constitutes a bottleneck in the current *de novo* enzyme design methodology for two reasons:

1) it limits the choice of target reactions to those where only 3-4 residues are involved in the actual chemical steps
2) it reduces possible active site definitions to 1st shell protein-ligand interactions, though often it would be desirable to include predefined 2nd shell contacts between the active site residues and other protein residues. For example, if, to ensure preorganization[1] of an active site, every 1st shell residue should make an additional contact to another protein residue, then a two residue active site definition becomes a four residue active site definition.

It would thus be highly desirable to improve the matching step of the design process such that matches for more complex active site definitions are routinely found.

The reason why fewer matches are found with an increasing number of catalytic interactions is simple: to find a match, the structure of the scaffold must provide peptide Cα-Cβ vectors where all the catalytic residues can be attached such that they are in the desired spatial arrangement with respect to each other and the substrate. The more catalytic residues there are, the less likely it is that the scaffold features a combination of Cα-Cβ vectors that supports this arrangement.
Thus, one strategy to overcome this problem is to not limit oneself to the Cα-Cβ vector configuration of the native scaffold, i.e. to allow for flexibility of the scaffold peptide backbone. Sampling different conformations of the protein backbone during the matching process could significantly increase the number of matches.

4.1.1 The challenge of placing additional interactions in ECH19

As introduced in the last chapter, the reason for the low catalytic activity of the computationally designed esterase ECH19 is possibly that the catalytic histidine is predominantly in an unproductive conformation, and stabilizing the histidine to be in the conformation as desired in the theozyme would possible lead to increased catalytic activity. A promising strategy to stabilize the histidine would be to extend the active site to include “backing up” interactions for the histidine, i.e. place side-chains that would hydrogen-bond with the histidine’s imidazole nitrogen atom that is pointing away from the substrate and not satisfied in the theozyme. If the histidine was making an additional hydrogen bond, it presumably would predominantly populate the desired conformation. The role of the Asp and Glu side chains in the catalytic triads of many naturally occurring hydrolases, besides activating the imidazole ring for deprotonation of the catalytic Ser, is also probably to orient and lock the histidine in the catalytically required conformation. However, assuming that an Asn/Asp or Gln/Glu would provide the backing up conformation, such an interaction is unfortunately not possible in ECH19. This is shown in

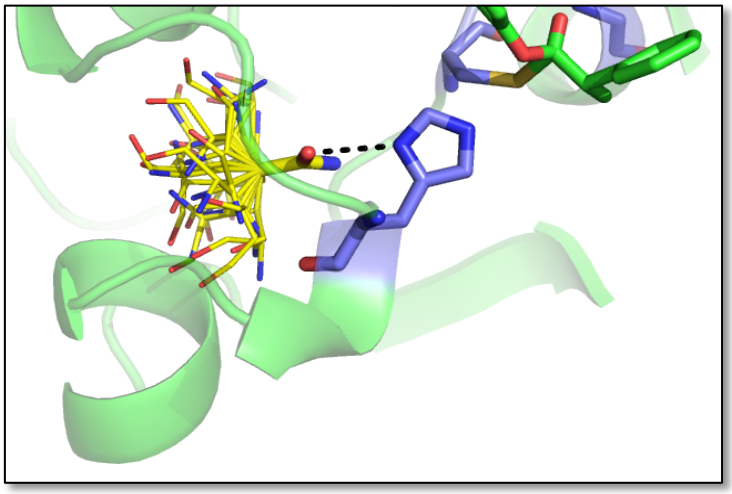


Figure 4.2 Asn rotamers positioned to back-up the histidine do not overlap with the ECH19 ABL backbone

Figure 4.2: if a library of Asn rotamers is placed such that they interact ideally with the histidine in its desired conformation, none of the Asn rotamers' backbones overlaps with the ECH19 scaffold backbone, indicating that it is not possible to introduce the desired interaction into ECH19 without altering the backbone. Therefore, to design an active site in ECH19 where the histidine makes the desired backing-up interaction, the backbone of ECH19 needs to be redesigned.

4.2 Inverse Rotamer Remodeling

While including backbone flexibility in the matching stage is a promising approach to generate a large number of new matches, there are two big challenges that need to be addressed. First, the sampling of alternative backbone conformations needs to be directed towards regions of conformational space where rotamers of the match residues can be attached such that they form the desired interactions with the ligand. Second, the generated conformations need to be "designable', i.e. there needs to exist an amino acid sequence (to be found in the second stage of the enzyme design process) that will fold into this conformation.
There exist a plethora of methods in the field of computational protein modeling to sample conformational space, a number of which have been implemented in the Rosetta package. The focus of this project will lie on modifying these existing methods of backbone sampling to preferentially sample regions of conformational space compatible with the two requirements mentioned above.

4.2.1 Means of backbone sampling

Several methods are available for sampling backbone conformational space in Rosetta3. These methods are all similar in that they combine random, stochastic backbone perturbations with a Metropolis-Monte Carlo algorithm that determines whether a certain perturbation will be accepted or not. This means that the conformations before and after the random backbone perturbation are evaluated with the scoring function, and the difference in score between the two conformations is the number subjected to the Metropolis criterion. Thus, if the random perturbation leads to a conformation with a better score, it is always accepted. If it leads to a higher score, it is accepted with a probability inversely proportional to the difference in score. In this way, low energy regions of backbone conformational space are explored. What the different backbone sampling methods differ in are the types of perturbations they make. The most frequently used method is so-called fragment insertion[2], described in more detail below. Briefly, the protein databank (PDB) is queried for backbone conformations of protein fragments (usually trimers and nonamers) that have similar sequence to the protein chain being modeled, and good scoring models are then assembled from these fragments. Starting from an extended conformation, compact, globular models can be obtained in this way, and this method has been successfully applied in structure prediction problems[3]. Fragment insertion can either be used to predict the structures of a whole protein or of isolated segments, such as in loop modeling. Another example of a backbone sampling approach is the so-called 'backrub' algorithm, which makes small rotations around Cα-Cα axis of two pivot residues close in sequence that lead to a slight difference in phi/psi and Cα-Cβ vector orientation of the residues in between these pivot residues[4]. This type of move is inspired by the observation of slightly different conformations in high-resolution crystal structures that can be interconverted through a backrub move[5]. A third method to generate new backbone conformations, called Kinematic Loop Closure (KIC), is to sample random phi/psi values from a Ramachandran distribution and then use equations developed for robotics to calculate values for three residues in the chain to ensure a closed conformation[6].

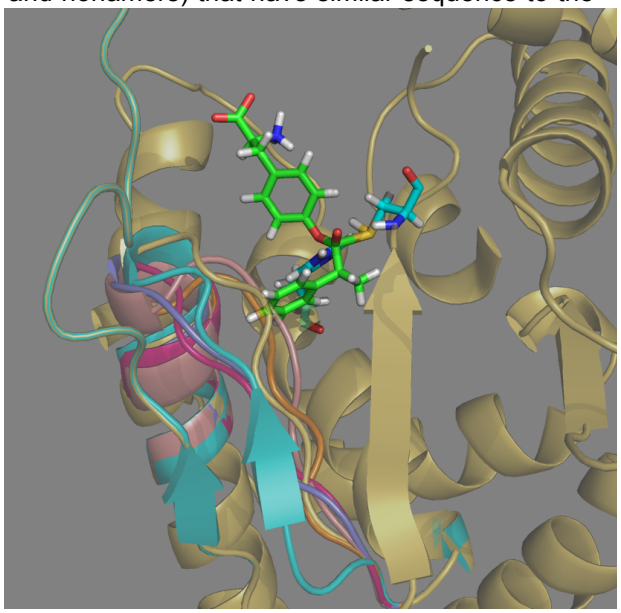


Figure 4.3 Example of using fragment insertion to sample a protein backbone: several alternative conformations for an active site loop are shown (native: cyan)

However, KIC is mostly suited for small loop-modeling problems. Even more methods of backbone perturbation are described in the literature[7]. Figure 4.3 shows an example of the conformational diversity observed when sampling a loop region in a protein with fragment insertion.

As described above, the challenge when including backbone flexibility at the matching stage is to find alternative backbone conformations where rotamers of the catalytic amino acids can be attached such that they form the desired interactions with the ligand. The new backbone conformation is thus required to have at least one amino acid with a Cα-Cβ vector that supports this orientation. Because in most theozymes the catalytic geometries are within very narrow, sub-angstrom, ranges, the part of conformational space where this requirement is met is very small, and thus acceptable alternative backbone conformations are very rare. Further, the standard Rosetta score function has no score terms that would favor these rare conformations, especially since a low resolution, centroid representation of the molecule is used during the backbone sampling stage. This means that the part of conformational space favorable in score and the part of conformational space where new catalytic contacts can be supported don't necessarily overlap, and therefore Monte Carlo sampling of backbone conformations with the above described methods will not always converge on a new catalytic backbone conformation. If one wants to bias sampling towards catalytic backbone conformations, one thus needs to modify the scoring function with a term that favors these conformations.
This problem is illustrated by the ECH19 case. When placing the backing up rotamers in the ideal conformation, the backbone atoms of several of them are positioned in spatial proximity to the active site loop following the catalytic histidine, and formed by residues 227-238 (from hereon referred to as ABL, for "Active site Backup Loop"). It is therefore conceivable that the ABL could be redesigned to have the backbone atoms of one of its residues be superimposable onto the backbone atoms of one of the inverse rotamers. However, when running the standard Rosetta flexible backbone protocol to generate new conformations for the ABL, conformations that would support a back-up side-chain are never found. As shown in Fig. 4.4a, the generated alternative conformations preferably sample a region of conformational space far away from the desired region where the backup is possible.

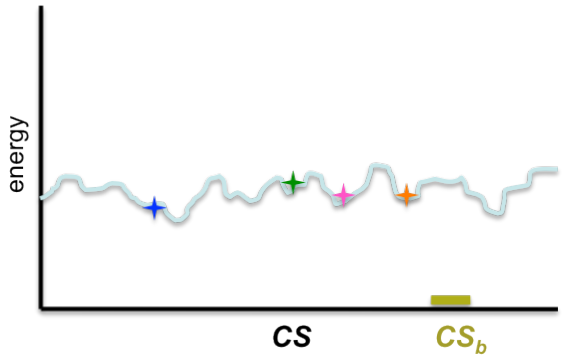


Figure 4.4a The energy landscape during the backbone remodeling stage is flat and frustrated. CSb, the region of conformational space CS where the backing up interaction is possible is not preferred

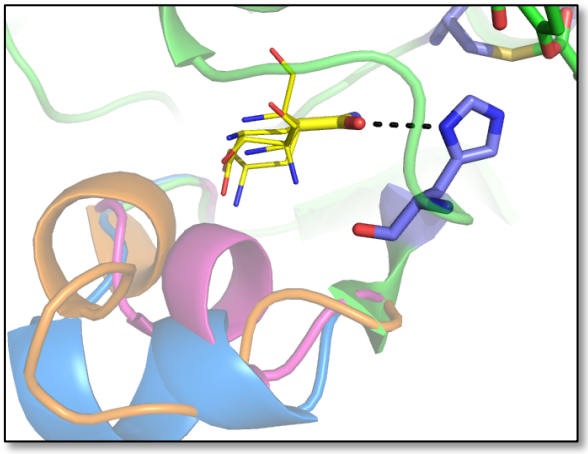


Figure 4.4b Alternative conformations for the ECH19 ABL generated with the standard protocol do not allow the desired back-up interaction

Thus, to design new ABL conformations that can support back-up interactions, the score-function needs to be modified to favor regions of conformational space where the back-up can be maintained. The strategy described here will combine inverse side chain rotamers with ambiguous harmonic restraints.

4.2.2 The concept of inverse rotamers

The term 'rotamer' describes a conformation of an amino acid side chain that is frequently observed. Analyzing 1000s of crystal structures in the PDB has shown that for every of the 20 amino acids, there is a discrete set of conformations, usually with the side chain dihedrals in staggered orientations. A complete set of these conformations of one amino acid is called a rotamer library[8].
Usually, rotamer libraries are used to determine the packing pattern of protein side chains. When used in this way, all members of the library are superimposed on the amino acid's backbone atoms (i.e. N – Cα – C), and the conformational diversity is expressed in the side chain atoms. However, the library rotamers can of course also be superimposed on other atoms, e.g. those on the tip of the side chain, and in this case the conformational diversity in the library is expressed in the backbone atoms. However, rotamers can also be superimposed on their side chain atoms, in which case they are commonly referred to as inverse rotamers[12]. Fig 4.5 shows a serine rotamer library superimposed both ways.

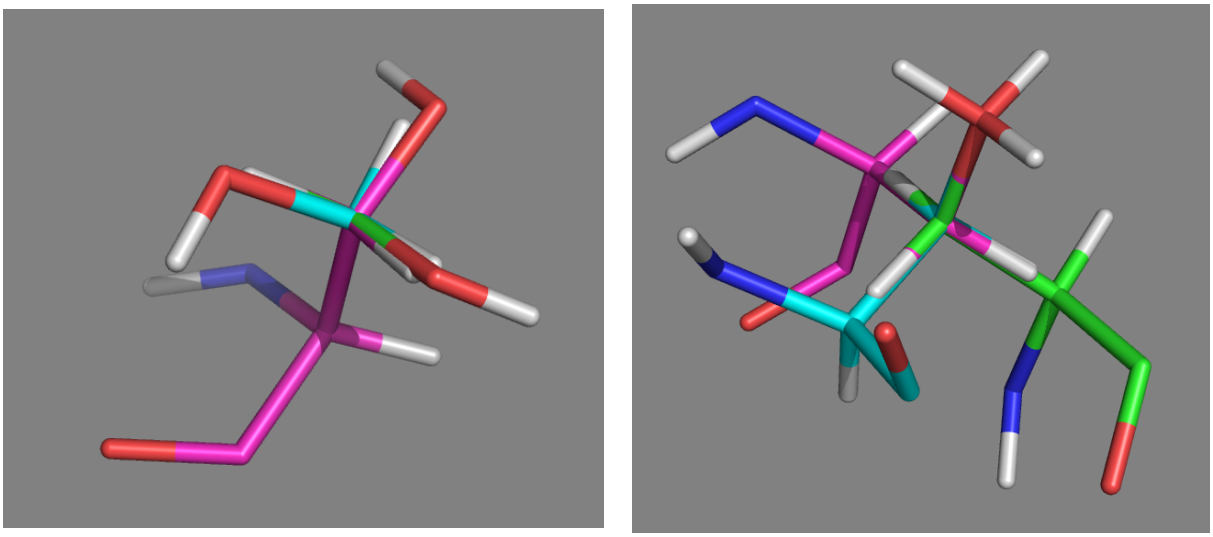


Figure 4.5

a) Ser-rotamers superimposed on peptide backbone b) Ser rotamers superimposed on the sidechain atoms

Inverse rotamers can be used to direct backbone sampling towards catalytic backbone conformations. If the catalytic geometry between a substrate and a side chain is well defined, and if the coordinates of the side chain in space are known, coordinates for the substrate can be unambiguously obtained (this is what happens in matching). Conversely, if the coordinates of the substrate are known, the side chain can be placed unambiguously in space. Given the coordinates of the placed side chain, other rotamers of the side chain can be superimposed on the catalytically relevant atoms (usually the functional groups on the tip of the side chain), and thus one obtains a rotamer library where each member makes a catalytic interaction with the substrate. The backbone atoms of the inverse rotamers then occupy regions in conformational space where catalytic backbone conformations are possible. When exploring alternative backbone conformations, sampling has to be directed such that the backbone atoms of one residue in the flexible region are superimposed onto the backbone atoms of one of the inverse rotamers. An example of using inverse rotamers to direct conformational sampling is the recently reported Hotspot approach by Fleishman *et al.*, described in the introduction, where inverse rotamers are used to find a productive rigid-body orientation between two proteins such that a binding interface can be designed between the two.

4.2.3 Directing sampling towards inverse rotamers

After inverse rotamers of the catalytic residue have been placed in the proper position, the question is how to best incorporate the information about their backbone atom locations into the conformational sampling algorithm. This problem has recently begun to be addressed in two papers from the Baker lab.

In Murphy *et al*'s method[8], the problem is approached in a divide-and-conquer fashion. After the inverse rotamers have been placed, one of them is randomly selected to be incorporated into the new backbone. Next, the section of backbone that is closest to the Cα of this rotamer is identified and removed. The two residues upstream and downstream of the removed section and the inverse rotamer are kept fixed, and two new backbone sections are rebuilt: one between the upstream residue and the inverse rotamer, the other one between the fixed rotamer and the downstream residue. Notable features of this approach are that it will always find a new catalytic backbone conformation, and that backbone conformations of differing lengths can be considered. On the downside, the generated conformations are prone to differ significantly from the starting conformation, which means that more aggressive sequence changes will have to be introduced in the design stage.

In Havranek *et al*'s approach[9], the closest overlap between a residue in the existing backbone conformation and any of the inverse rotamers is determined. A strong restraint between the backbone atoms of the closest existing residue and the corresponding inverse rotamer is then added to the scoring function. The backbone is then subjected to backrub[4] sampling, and ideally the restraint will guide the selected residue towards the inverse rotamer position. Notable features of this algorithm are that it is very fast and potentially solves the problem with minimal changes to the protein backbone, i.e. the subsequent design stage is less difficult. Disadvantages are that this algorithm is not always guaranteed to find a solution, for example in cases where the backbone of the inverse rotamer is far from any original residue.

Here, a new approach is described that is complementary to the two approaches above and overcomes some of their shortcomings. It is fairly simple, yet should have several advantages. This approach, called remodeling with ambiguous restraints, combines Rosetta's fragment library based backbone sampling with a score term called ambiguous restraint that guides sampling towards the inverse rotamer backbones, and should lead to the design of backbone structures that can support the additional catalytic interaction while also having an energetically favorable, native-like conformation.

The fragment-based backbone sampling method pioneered by Rosetta was originally developed for protein structure prediction[41]. Briefly, if a protein of length N is to be sampled, the sequence is subdivided into segments of three (trimers) and nine (nonamers) residues in a sliding-window fashion. For each of these segments, the PDB is searched for segments of the same sequence, and their backbone conformation is noted. One such conformation is termed a fragment in this context. Once a list of possible fragments for each of the segments that make up the protein is obtained, the backbone angles of a randomly picked segment are set to those in one of the fragments observed in the PDB for that segment, the resulting conformation is evaluated with the scoring function, and the Metropolis criterion is used to decide whether to accept or reject the fragment insertion. This is repeated several thousand times, and in the end a low energy conformation for the protein in question should be obtained. If only part of the protein is simulated, i.e. the coordinates of the beginning and the end of the flexible segment are fixed, then the chain is cut in the middle and a term penalizing the chainbreak is added to the scoring function to ensure that only closed loop conformations are sampled[10]. Fragment sampling for design problems was recently reimplemented in Rosetta3 under the name RosettaRemodel, as described by Huang, Ban, Richter, *et al*[11].

RosettaRemodel harvests fragments directly from a culled set of torsion angles from non-redundant x-ray structures and assembles them on the scaffold structure according to their secondary structure type. An advantage of harvesting fragments directly is that one can collect a new set of fragments during a protocol and not be limited to the pre-defined set provided at the start of the simulation, resulting in significantly expanded sampling diversity. The default number of fragments used is 200 segments of nine-, three-, and single amino acids for each position, as is commonly used for other Rosetta fold prediction projects[7,12]. Additionally, the protocol allows harvesting fragments that match

the entire length of a remodeled region, potentially with improved fragment qualities similar to those previously reported[13]. Since the objective is to design new structures and new sequences, all energy function terms that involve specific sequence information are explicitly turned off. Only van der Waals, radius of gyration, and Ramachandran probability terms are used to specifically address clashing, packing, and chain geometry, respectively. Residues in the backbone building stage are centroids of valines or alanines, and thus the Ramachandran term is the same for all moving positions, but is significantly scaled down to 1/10 its normal weight to avoid areas of very low probability.

Backbone modeling on internal loops is performed with random cut sites within the loops preceding fragment building. The internal chain breaks are subsequently reconnected using closure algorithms such as Cyclic Coordinate Descent (CCD)[14]. Only models with properly closed chains after the fragment assembly stages are passed along to the design stage. There are often cases where successful closure is rare; in such cases it may be that too few residues are being used or that the residues at the ends of the loop being modeled are in orientations not suitable for proper closure and should be allowed to move.

Energy function terms that enforce backbone geometry can be applied and adjusted to suit particular design problems. Backbone-specific terms, namely strand pairing and hydrogen bonding energies on helices and sheets, can be selectively applied for different types of designs; conversely, the terms used by default can be scaled down or turned off for purposes such as turning off minimization of the radius of gyration when building a polar surface loop. RosettaRemodel usually uses centroids of alanines in the fragment assembly stage as generic space fillers until the design takes place at the full-atom level. Although contacts between sidechains are evaluated and are part of the Monte Carlo simulation when sampling backbone conformations, evaluation of proper chain closure supersedes all other criteria

There is one caveat when using the fragment insertion strategy to find backbone conformations before sequence design takes place: the amino acid sequence is unknown at the time the new backbone conformation is generated. This means that the fragments to be used cannot be selected based on sequence similarity. What is commonly done in this case is that the fragments are selected based on secondary structure, i.e. every amino acid in the segment in question is assigned to roughly have a sheet, helix, or loop backbone conformation[15]. Fragments for this segment are then selected from fragments observed in the PDB that have the same secondary structure makeup.

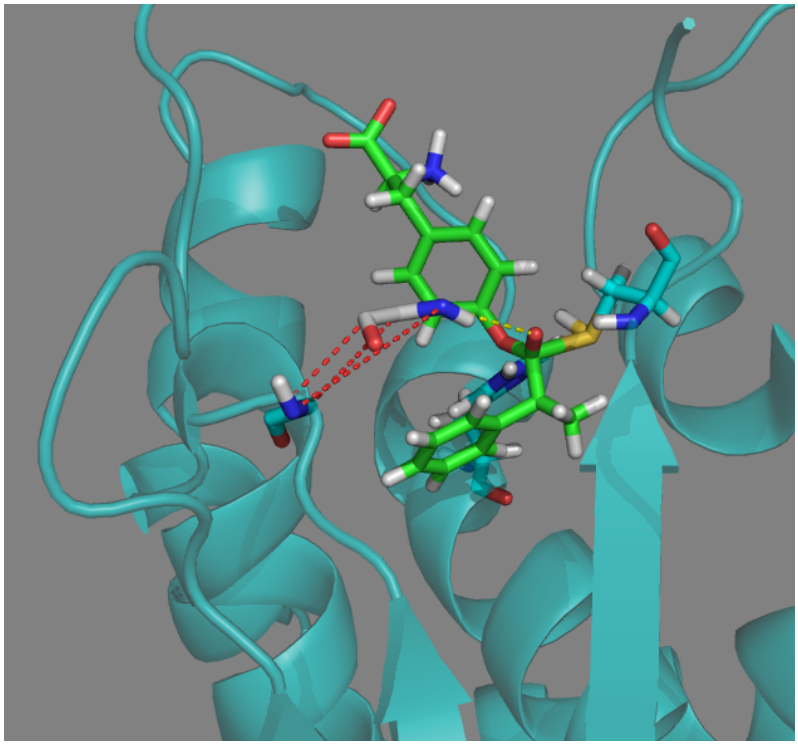


Figure 4.6: An inverse rotamer (white) has been placed such that it mediates a catalytic contact (yellow dashes) to the substrate (green). Distance restraints (red dashes) are added between backbone atoms in the inverse rotamer and in a residue (cyan) in a nearby segment to be sampled.

With the inverse rotamers as targets for the new backbone conformation and the fragment insertion algorithm as a method of generating new backbones, one now needs a way of biasing the fragment insertion sampling such that fragments are selected that make the backbone partly overlap

with the inverse rotamers. The most straightforward way to do so is to modify the scoring function by adding a distance restraint between the backbone atoms of a residue in the flexible backbone segment and an inverse rotamer, as shown in Figure 4.6. The closer the backbone residue approaches the position of the inverse rotamer, the lower the penalty generated by the restraint will be, and thus the Monte Carlo search should converge on a backbone conformation where one residue superimposes onto the inverse rotamer as desired. This is what is done in Havranek *et al.*'s method. The difficulty however lies in deciding which residue in the flexible segment and which inverse rotamer to restrain. For example, if a backbone stretch of length n is treated as flexible, and there are m inverse rotamers, there are n*m possibilities of defining the restraint. Havranek *et al.* simply pick the closest residue/inverse rotamer pair. This might be viable in cases where one of the inverse rotamers is already very close to the backbone, but not in situations where no backbone residue is close to any inverse rotamer or where several residue/inverse rotamer pairs exist that have comparable distances. Moreover, in contrast to Havranek *et al.*'s method, the fragment insertion algorithm starts from an extended backbone conformation and not the original backbone conformation, and in this case the closest residue / inverse rotamer pair before the simulation is unlikely to be the pair best suited to form the desired contact in the final, new low energy conformation.

4.2.4 Deciding which residue-inverse rotamer pair to restrain

The choice of which residue – inverse rotamer pair to restrain is critical because it has a huge effect on the trajectory of the fragment insertion backbone sampling. Adding an artificial restraint to the score function can direct sampling of conformational space into a region where the other, empirical, terms present in the score function (van-der-Waals, backbone-torsions, etc) would not have directed sampling themselves. While this is desired to a certain degree (to find new catalytic conformations), the artificial restraint might force the flexible backbone segment into a conformation that, although overlapping with the inverse rotamer and thus being catalytically competent, might otherwise be unfavorable, i.e. the segment will have steric strain, or approach the surrounding protein too closely, or might be too far from the surrounding protein to make good interactions to be stabilized by it. In other words, the newly generated backbone conformation might be so-called 'undesignable', which means that no amino acid sequence exists that would actually take on this conformation.

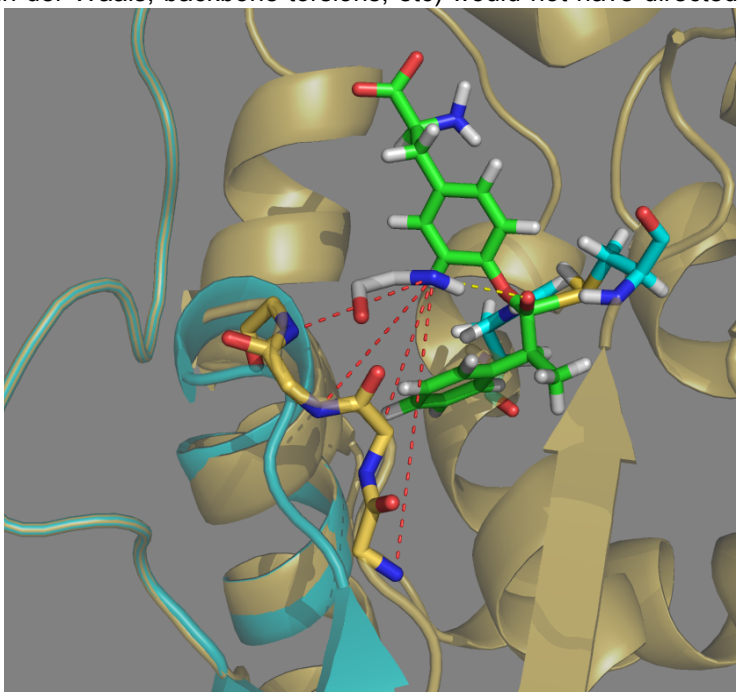


Figure 4.7: Sampling with ambiguous restraints. Between the inverse rotamer (white) and several residues in the loop being sampled (yellow, native in cyan), distance restraints have been added (red dashes). The penalty applied will be that of the restraint between the two pairs most likely to make the interaction.

Ideally, one would like to find a conformation that is both favorable with respect to the empirical, unmodified score function and can support the desired catalytic contact. This in turn necessitates that, ideally, the effect of the artificial restraint should be as small as possible and as large as necessary. Practically, in terms of making a decision how the

restraint should actually look like, i.e. which residue – inverse rotamer pair to restrain, this means that one has to know before running fragment insertion which of the n*m possible pairs is the one most likely to be superimposable if the restraint didn't exist. Making this prediction is a non-trivial task, because the optimal residue – inverse rotamer pair is governed not just by the position of the inverse rotamers and the fragments available for the flexible protein segment, but also by the positions of the residues upstream and downstream of the flexible segment, the position of the ligand, and the shape of the surrounding protein.

4.2.5 Using ambiguous restraints to guide sampling towards the optimal pair

A simple solution to this problem is to run the fragment insertion sampling n*m times, once for each possible pair restraint, but this is not computationally tractable, as n*m can easily be in the hundreds. However, because it is so difficult to predict the optimal pair, somehow one has to examine all n*m possible restraints to determine the best one. There exists a way to resolve this quandary: using ambiguous restraints.
An ambiguous restraint is a type of smart restraint that consists of a list of arbitrary other restraints. When this restraint is evaluated, each of the listed restraints gets evaluated, but only the score of the lowest scoring restraint counts as the score of the ambiguous restraint. This is conceptually shown in Fig 4.7
When using ambiguous restraints to find new catalytic backbone conformations, the m*n possible residue – inverse rotamer pair restraints could be treated as the member restraints of one ambiguous restraint. In this way, at every step of the fragment insertion sampling, the ambiguous restraint will only generate the lowest penalty possible. This means the effect of the restraint will be, as desired, as small as possible but as big as necessary. Sampling should be influenced such that the residue – inverse rotamer pair that is most likely to be superimposable given all factors will actually be superimposed in the final, converged structure at the end of fragment insertion sampling.

4.2.6 Advantages of this approach over previously developed ones

The advantages of this approach compared to Havranek *et al.* are obvious: larger backbone changes are possible, for example insertions and deletions of amino acids in the flexible segment. While this approach is more similar to Murphy *et al.*, it also features advantages compared to it. Most notably, in Murphy *et al.*, the decision of which inverse rotamer to use in the sampling and of how long the flexible segments upstream and downstream of the inverse rotamer should be is made before fragment insertion sampling. Compared to the approach proposed here, this is synonymous to choosing the residue – inverse rotamer pair beforehand, and so the Murphy *et al.* approach is more prone to force the backbone into a conformation that is unfavorable or even undesignable. Also, since Murphy *et al.* split the flexible segment into two, it is computationally slower, because two loop closure problems have to be solved.

4.3. Detailed protocol implementation

In this section, the detailed algorithm implementation of inverse rotamer remodeling with ambiguous restraints will be described. The protocol can roughly be divided into three stages: 1) generating a new backbone conformation that can make the desired interactions, 2) designing a new sequence for the new backbone conformation, and 3) structure prediction of the designed sequence to assess whether the desired backbone conformation is the lowest energy conformation of the designed sequence.

4.3.1 Backbone-generating stage

A schematic overview of the first stage is shown in Figure 4.8.

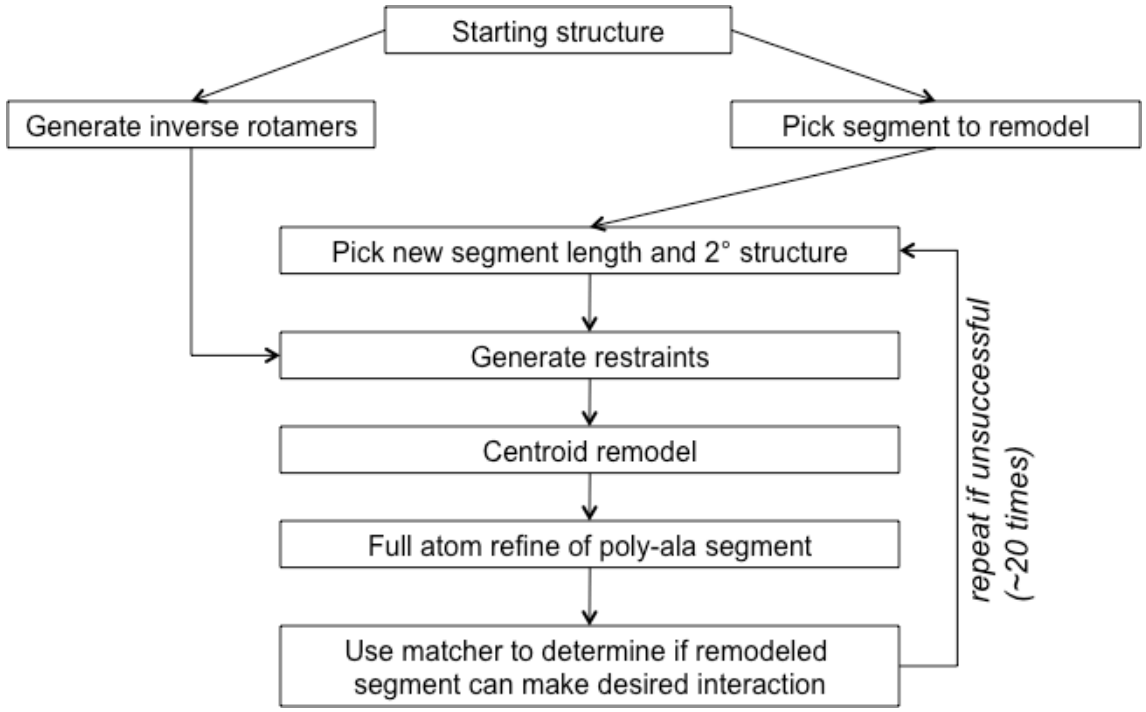


Figure 4.8 A schematic representation of the 1st (backbone generation) stage of Inverse Rotamer Remodel

At the outset of the calculation, three things need to be provided by the user: 1) a starting structure, 2) a definition of the desired geometry for the new interaction and 3) a definition of which segment of the starting structure to remodel. The starting structure needs to be provided in regular PDB format, while the geometry description needs to be in the Rosetta3 enzyme design .cst file format as described in Chapter 2. To specify the segment, three parameters need to be specified in a separate input file: the first and last residues of the loop to be remodeled and the desired new secondary structure. The secondary structure can either be specified directly as a string comprised of secondary structure characters H, L, or E, respectively for helix, loop, or strand specification. For example 'HHHHLLLEEE' would specify a 10 residue long helix-turn-strand motif. An arbitrary number of secondary structure strings can be specified. If more than one string is specified, a randomly picked one is used to specify

the new loop structure at the beginning of a run. Alternatively, the secondary structure can also be specified in a blueprint format specifying an approximate backbone conformation. This blueprint format (which is different from the one described for RosettaRemodel[11]) consists of a succession of pairs of secondary structure characters and associated length ranges. For example, the string 'H(4-8)L(3-5)E(4)' signifies a helix consisting of between four and eight residues, followed by between three and five residues in loop conformation, followed by four residues in strand conformation. Internally, blueprint strings get translated to a list of every explicit string that the blueprint specifies. For the above example, the list would contain 15 explicit strings (5 length possibilities for the helix * 3 possibilities for the loop * 1 possibility for the strand), ranging in length from 11 to 17. If the blueprint format is used, a random one from the list of explicit strings is picked at the beginning of every run. Optionally, values for the maximum and minimum allowed lengths of the new segment can be specified in the input file, and explicit strings generated from the blueprint string that fall out of this desired length range are not considered. Thus, the loop specifying input file contains blocks of the following make-up:

```
LOOP_BEGIN

start 20
stop 26
max_length 9
min_length 6
#ss_string

ss_blueprint E(0-3)L(5-10)H(0-2)
LOOP_END
```

In this example, the seven-residue segment between residue 20 and 26 will be remodeled using new backbone conformations ranging in length between six and nine, some having only a loop, others having a short strand followed by a loop, etc.

With both inverse rotamers and a new segment length (derived from the randomly picked secondary structure string) determined, the restraints to guide backbone sampling towards the inverse rotamers can be generated. Figure 4.9 illustrates the atoms used in the various constraints.

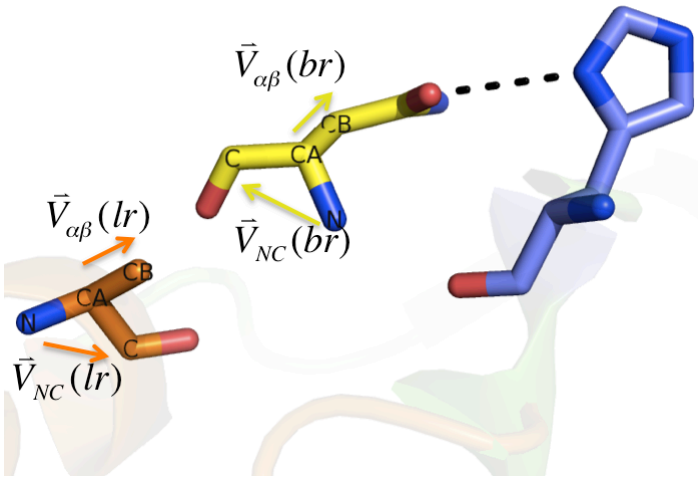


Figure 4.9 Atoms and angles used in the definition of the backbone stub constraint

Assuming there are *n* inverse rotamers and *m* mobile residues in the segment, restraints for a total of *m*n* inverse-rotamer backbone pairs have to be generated. For each pair *i,j*, several restraints are generated: for each of the backbone atoms N, CA, C and CB, a harmonic restraint between the

coordinates of that atom in the inverse rotamer *I* and the atom in the mobile residue *j* is created, where C represents the conformation of the system.
Besides the coordinate constraints, a second type of constraints is used, the so-called backbone stub constraint, which serves to ensure the proper orientation of mobile residue *Ir* onto the backbone of inverse rotamer residue *br*. This constraint is also used in hotspot-method for protein interface design[16] described in Chapter 1 and has the form

$$E_{bk} = \min(0,[(B + kD_{\beta}{}^{2})\cos(\theta_{NC})\cos(\theta_{\alpha\beta})])$$

Where:
B = superposition bonus
k = force constant
$D_{\beta}{}^{2}$ = squared distance C_{β}(br) – C_{β}(Ir)
$\theta_{\alpha\beta}$ = angle between $V_{\alpha\beta}$(br) and $V_{\alpha\beta}$(Ir)
θ_{NC} = angle between V_{NC}(br) and V_{NC}(Ir)

The atoms and angles used are also shown in Fig 4.9 for clarity. The backbone stub constraint and the four coordinate constraints are grouped together, and the total restraint score of the pair is

$$E_{i,j}(C) = E_{cst\_N}(C) + E_{cst\_C}(C) + E_{cst\_Ca}(C) + E_{cst\_Cb}(C) + E_{bstub}(C)$$

All *m*n* restraints are then grouped together, and at every evaluation of the scoring function during the remodeling trajectory, each pair restraint is evaluated. However, the penalty assigned by the scoring function to the conformation under consideration is that of the lowest pair restraint, i.e. the applied restraint score of conformation C is

$$E_{invrot}(C) = \min\{E_{i=1,j=1}(C),...,E_{i=n,j=m}(C)\}$$

Thus, as described earlier in this chapter, at every step in the remodel trajectory, the penalty added to the score of conformation C is determined by the inverse-rotamer backbone pair *I,j* that is most likely to be superimposable, i.e. the added penalty is as small as possible but as large as necessary. The artificial inverse rotamer restraint penalty should thus have the minimum necessary effect on the trajectory to guide sampling towards a conformation that can make the desired interaction while at the same time still having a decent score in the other, empirical scoreterms.

After a novel backbone has been generated, in this case with the score function augmented by the term, it still needs to be determined whether the desired interaction can indeed be placed on the novel backbone. After all, while adding constraints favors backup-compatible backbone conformations, it cannot guarantee that they will be found. For example, in some cases, the employed fragments might simply not be capable of forming a backbone superimposable onto an inverse rotamer, or the tried segment length and secondary structure might not be long enough to simultaneously overlap with an inverse rotamer while forming a closed chain where every backbone torsion is in allowed regions of the Ramachandran plot. To determine whether the generated backbones can indeed accommodate the new interaction, every candidate conformation is used as an input scaffold for the matching algorithm described in Chapter 2. In case the matcher can then place the new interaction, a new backbone conformation allowing for an active site that couldn't be placed on the original scaffold was successfully generated.
Finally, to design the full sequence for the new conformation, the regular phase 3 sequence design protocol as described in Chapter 2 is used.

4.4 Performance on a benchmark of eight natural enzymes and the ECH19 ABL

4.4.1 Benchmark of natural enzymes

To assess the performance of the inverse rotamer remodeling protocol, a benchmark consisting of eight natural enzymes was compiled. For each of these enzymes, a high-resolution crystal structure with a natural ligand was available. In each case, an 11-residue stretch of the active site that contained one or two residues making a contact to the ligand was chosen for remodeling. For the ligand contacting residues, inverse rotamers were built around the position of that side chain observed in the crystal structure. The 11 residue stretch was chosen such that the Inverse Rotamer (IR) residues were in the middle of the segment. An overview of the eight selected structures is given in Table 4.2 below:

Case (PDB ID)	**Remodel stretch**	**Catres 1**	**Catres 2**	**2° structure**
1c2t	103 - 113	Asn 106	His 108	LEEEEELLLLL
1dqx	88-108	Asp 91	Lys 93	EEELLLELLLH
1ecm	35-45	Lys 39	-	HHHHHHHHLLL
1jcl	197-207	Lys 202	-	LLLEEELELLL
1ney	6-16	Lys 11	-	EEEELLLLLH
1p6o	55-65	His 60	Glu 62	LLLLLLHHHHH
3h3j	171-181	His 179	-	LEEELLLLLLE
6cpa	64-74	His 69	Glu 72	EEELLLLLLHH

Table 4.2 Structures and backbone segments used for the benchmark

In each case, the residue indicated in red in the secondary structure column is the residue for which inverse rotamers were built and that had to be recovered by the matcher after the remodeling protocol was run. For the stretch to be remodeled, only the wild-type secondary structure and wild-type length was allowed. The protocol setup was as follows: the remodel stretch was remodeled 100 times with the E_{invrot} term and 100 times without it. After each run, the matcher was run to see if the wild-type interaction can be placed on the remodeled stretch. The results were analyzed with respect to four properties:

1) Percentage of models where the catalytic residue was recovered (success)
2) Percentage where it was recovered at the original position (exact success)
3) RMSD (Å) of the remodeled segments to the wild-type
4) In the exact successes, RMSD (Å) of the catalytic residue to the wild-type

Table 4.3 below shows the results for points 1 and 2, (results with E_{invrot} in black, without in red):

case	% success	av rmsd	% exact success	av rmsd
1c2t	97 38	1.05 +- 0.75 2.04 +- 2.27	94 38	0.97 +- 0.67 0.95 +- 0.42
1dqx	100 61	0.92 +- 0.39 1.28 +- 0.54	99 61	0.91 +- 0.39 0.98 +- 0.34
1ecm	100 100	0.38 +- 0.06 0.54 +- 0.07	100 100	0.38 +- 0.06 0.54 +- 0.07
1jcl	100 95	2.97 +- 1.66 2.65 +- 1.56	36 41	1.14 +- 0.41 1.26 +- 0.65
1ney	100 100	1.48 +- 0.32 1.59 +- 0.37	91 38	1.47 +- 0.30 1.47 +- 0.37
1p6o	100 98	1.63 +- 0.70 1.51 +- 0.73	100 97	1.63 +- 0.70 1.49 +- 0.63
3h3j	79 33	2.43 +- 1.34 3.39 +- 2.65	51 22	1.49 +- 1.00 1.43 +- 1.07
6cpa	95 38	1.47 +- 0.86 2.12 +- 0.77	59 24	0.88 +- 0.52 1.23 +- 0.37

Table 4.3 Benchmark results 1: Frequency of recovery of the interaction

Three conclusions can be drawn from this benchmark:
First, E_{invrot} clearly helps with recovering catalytic interactions. In all cases, the success rate, i.e. the rate of recovery of the catalytic interactions is higher (respectively equal in two cases) with E_{invrot} than without. Further, the RMSDs of the remodeled backbones are also usually closer to the wild-type than when using the unmodified score function.

Second, while E_{invrot} helps, it is not always necessary. There are four cases (1ecm, 1jcl, 1ney, 1p6o) in the benchmark where high recovery is achieved by the unmodified protocol. Interestingly, in these cases, the catalytic residue is either on a secondary structure element (1ecm, 1jcl, 1p6o) or is a long, flexible residue with four chi angles (1ney). This can be rationalized that in cases where the desired interaction can be made from a secondary structure element that is connected to the non-remodeled part of the scaffold, the score function does not need to be artificially modified to find a productive backbone conformation. In this case, the regular score function (which favors secondary structure contacts through the backbone hydrogen bonding and ramachandran terms), in combination with fragments of the right secondary structure, is sufficient to build backbones that can support the desired interaction. Further, in case the catalytic residue is long and flexible, more variation in the backbone conformation can likely be tolerated, as the increased conformational diversity of the side-chain can compensate for less accurate recovery of the backbone conformation. This is illustrated by the 1ney case: the catalytic interaction, mediated by a lysine, can be realized in 100% of runs both with and without E_{invrot}. However, without E_{invrot}, only in 38% of the cases is the catalytic lysine placed on the same sequence position, while this number jumps to 91% when using E_{invrot}.
Third, besides generally increasing success rates, E_{invrot} also helps with the accuracy with which the native backbone conformation is recovered, even in cases where the interaction is recovered without E_{invrot}. This is also exemplified in Table 4.4, showing the RMSDs of the catalytic residue backbone in cases where the catalytic residue was placed at the proper position. In all cases, E_{invrot} had lower RMSD to the wild-type backbone than the regular scorefunction.

case	Catres 1 rmsd	Catres 2 rmsd
1c2t	0.17 +- 0.08 0.37 +- 0.13	0.21 +- 0.16 0.40 +- 0.19
1dqx	0.30 +- 0.10 0.52 +- 0.17	1.05 +- 0.68 1.73 +- 0.65
1ecm	0.19 +- 0.05 0.73 +- 0.14	-
1jcl	0.48 +- 0.30 0.70 +- 0.34	-
1ney	1.85 +- 0.89 1.47 +- 0.37	-
1p6o	0.25 +- 0.15 0.49 +- 0.18	0.16 +- 0.04 0.32 +- 0.08
3h3j	0.69 +- 0.77 1.02 +- 1.04	-
6cpa	0.47 +- 0.32 1.27 +- 0.55	0.19 +- 0.14 0.68 +- 0.27

Table 4.4 RMSDs of the catalytic residues in cases where they were recovered

4.4.2. Performance on the ECH19 ABL

After having shown that the E_{invrot} score term helps with recovery of wild type catalytic interactions, its performance on the ECH19 ABL test case was examined. Earlier in this chapter in section 4.2.1 it was shown that the standard protocol is unable to find backbone conformations, which separates this test case from the eight cases presented for the natural enzyme benchmark, where for every case, the sought-for interaction was found at least some of the time. This suggests that the ECH19 ABL case is a much harder backbone design problem than the benchmark cases. Nevertheless, when using the E_{invrot} protocol, conformations of the ABL where the desired Histidine backup is possible are found in roughly 50% of the time (21 successful out of 40 runs). A few exemplary structures are shown in Fig 4.10.

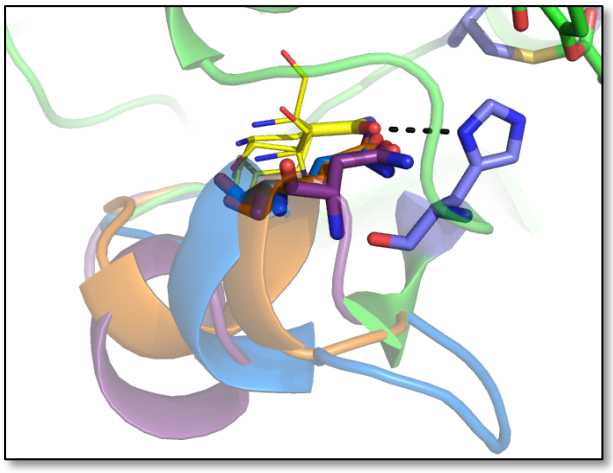


Figure 4.10 ECH19 ABL conformations obtained with the E_{invrot} protocol

4.5 Designability of the generated alternative backbones of the ECH19 ABL

As described in the previous sections, the E_{invrot} protocol allows for the design of novel backbone conformations where the catalytic histidine of ECH19 is backed up by a supporting residue in the ABL. However, the sequence used during the backbone design protocol is usually poly-alanine, and thus the redesigned backbone conformations do not yet have a designed sequence. The only side-chain whose identity has been assigned is the backing up residue (in this case either Asn/Gln/Asp/Glu), but all the remaining side chains on the ABL as well as the surrounding regions are still alanines. Designing a sequence for these is then done by the regular sequence design stage as described in Chapter 2. However, it is not clear whether the novel designed backbone conformations are 'designable', i.e. if a sequence can be designed that will actually fold into the desired conformation.

This question is relevant for two reasons: first, not every backbone conformation is designable. While it is certainly possible to design a sequence for any given backbone, it is not clear whether the lowest energy conformation of this sequence will then be the designed conformation or whether the sequence will preferably fold into an alternative, different

conformation. Second, since the protocol used to generate the novel backbone conformations for the ECH19 ABL features a score function modified by an artificial score term (the E_{invrot} term), the conformations generated with it might be physically unrealistic and thus not designable.

Two approaches were used to address this problem. First, the Rosetta score of the designed new structure/sequence pairs was compared to the score of the ECH19 starting model. Second, for those designs with Rosetta scores comparable to the starting model, loop modeling[17,18] calculations, where the structure of a certain part of the protein is predicted, were done, to see if the predicted conformation was similar to the designed conformation. Ideally loop modeling should be done for every design, however, since loop modeling calculations are computationally very intensive, this is not realistically possible. To only do the full structure prediction for designs that have scores comparable to or better than the starting structure was done because there is presumed to be a correlation between the score and the likelihood of the designed sequence being at its low energy conformation.

In total, several 10000 designs were done for the ECH19 ABL. Of these, 176, or about 0.1%. had scores comparable to the starting structure. When doing structure prediction on these 176, four general cases are observed:

1) the designed sequence prefers to take on an alternative minimum,
2) the designed sequence does not have a well-defined minimum, thus being likely unstructured,
3) the designed sequence takes on the designed backbone conformation, but the backing up side chain is not in the desired conformation,
4) the designed sequence takes on the designed backbone and sidechain conformation.

Table 4.5 below shows how many of the 176 sequences fall into each category

category	1	2	3	4
#designs	47	71	38	20

Table 4.5 Structure prediction results for the ECH19 ABL designs

Thus, 10% of the designs for which structure prediction was done do fold back to the designed conformation. This represents a success rate of 0.01% of all designs. Two conclusions can be drawn from this: first, most of the backbone conformations designed by this algorithm are likely not designable. While this rate might vary depending on the exact structural problem (the structures generated for the wild-type benchmark set are most likely all designable, since they're very similar to the wild-type conformation), it shows that structure prediction is absolutely necessary when using this protocol to design new backbone conformations. Second, while the rate of designing productive backbone conformations is low, it is possible, and thus this protocol can be used to design novel backbone conformations making desired interactions, provided there are enough computational resources.

4.6 Experimental characterization of alternative sequences designed for the ECH19 ABL

Of the 20 designs for the ECH19 ABL where the designed sequences was shown *in silico* to feature the desired backing-up interaction, 17 were selected for experimental testing. Expression and purification was carried out as described for ECH19 and other designs in Chapter 3. An overview of the designs is given in the table below:

Design	Backup res	length	solubility	Activity (x base design)
LD19_1	Q	+0	++	0.41
LD19_2	Q	+0	++	0.64
LD19_3	Q	+0	~	0.20
LD19_4	Q	+0	~	0.20
LD19_5	Q	+0	++	0.42
LD19_6	Q	+0	+++	0.42
LD19_7	Q	+0	+	0.12
LD19_8	E	+0	+++	0.43
LD19_9	E	+0	+++	0.37
LD19_10	E	+0	++	0.16
LD19_11	E	+2	+++	0.57
LD19_12	D	+4	++	0.30
LD19_13	E	+0	+++	0.51
LD19_14	D	+4	+	0.20
LD19_15	D	+3	++	0.12
LD19_16	E	+1	+++	0.55
LD19_17	E	+0	+++	0.6

Table 4.6 Experimental results of the ECH19 ABL redesigns

Unfortunately, none of the designs showed increased catalytic activity. The reason for this lack of catalytic activity is not clear. Either the ABL is in the designed conformation, and somehow the backed up histidine has less catalytic activity than the original design, or the

ABL is not in the designed conformation. The fact that most of the designs show robust expression and solubility would suggest that they are well folded, meaning that the redesigned ABL is probably structured and not just flexible in solution. However, it is entirely possible that the solution structure of the loop is different from what was designed and predicted by the loop modeling. There are two main shortcomings of relying on the loop modeling protocol for structure prediction: first, only the redesigned ABL is modeled, yet it is conceivable that regions of the protein outside of the ABL also change structure and thus influence the conformation of the loop. Second, the loop modeling protocol, like all Rosetta protocols, uses an implicit solvation model. While this is faster to evaluate computationally than explicit solvent approaches, it is also less accurate. This lack of accuracy might be particularly detrimental when it comes to modeling hydrogen bonds between protein residues on the protein surface or at the interface between the surface and the protein core, as these hydrogen bonding residues will be surrounded by water molecules and the hydrogen bonds are thus more prone to competition from protein-solvent interactions. As the hydrogen bond between the histidine and the backing-up residue falls into this class, it is possible that it might only be formed a fraction of the time.

Thus, the loop modeling stage of the protocol could be replaced by an explicit solvent Molecular Dynamics simulation step. However, MD simulations are far more computationally expensive and at the moment it would be unrealistic to run hundreds of simulations for a time long enough to judge whether the ABL stays in the designed conformation. However, for the 17 designs that were experimentally tested, 50ns MD simulations were run to examine whether the backing up conformation is realized. In 10 out of the 17 designs, the designed triad does not withstand solvent-bombardment, and the backing up residue reorients to form a hydrogen-bond with water. In the other seven cases, the triad stays together as designed. However, in these seven cases, the catalytic cysteine reorients. While the cysteine still interacts with the histidine in this alternative conformation, it is no longer in the same relative orientation to the backbone-NH residue that is providing the oxyanion stabilization. The two figures below (Figure 4.11) illustrate this: one shows the designed conformation of LD19_8, while the other shows the conformation of this design after 47ns of simulation, clearly depicting that the cysteine is rotated around its chi1, and thus can no longer interact with a substrate that is bound in the oxyanion hole.

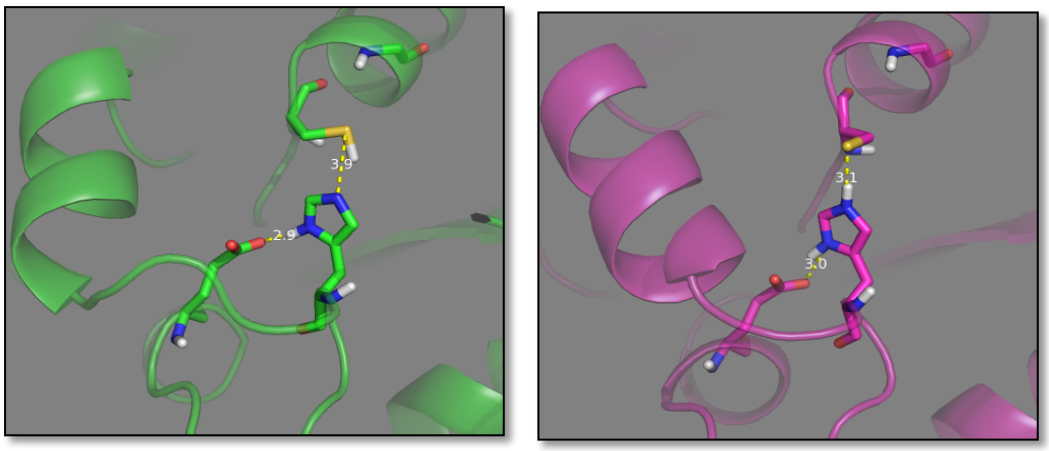


Figure 4.11 a) LD19_8 design model **b) LD19_8 after 47ns of MD simulation**

Bibliography / List of References

Chapter 1:

[1]Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **437**, 579–583 (2005).
[2] A. R. Leach, Molecular Modelling: Principles and Applications, 2001, ISBN 0-582-38210-6
[3] Boas, F Edward, and Pehr B Harbury. "Potential Energy Functions for Protein Design." *Current Opinion in Structural Biology* 17, no. 2 (April 2007): 199–204.
[4] C Pabo, "Molecular technology. Designing proteins and peptides," Nature 301, no. 5897 (January 20, 1983): 200
[5] Roland L Dunbrack, "Rotamer libraries in the 21st century," Current Opinion in Structural Biology 12, no. 4 (August 2002): 431-440
[6] C A Voigt, D B Gordon, and S L Mayo, "Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design," Journal of Molecular Biology 299, no. 3 (June 9, 2000): 789-803
[7] Holm, Lisa, and Chris Sander. "Fast and Simple Monte Carlo Algorithm for Side Chain Optimization in Proteins: Application to Model Building by Homology." *Proteins: Structure, Function, and Genetics* 14, no. 2 (1992): 213–223.
[8] Desmet, Johan, Jan Spriet, and Ignace Lasters. "Fast and Accurate Side-chain Topology and Energy Refinement (FASTER) as a New Method for Protein Structure Optimization." *Proteins* 48, no. 1 (July 1, 2002): 31–43.
[9]Janin, J. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst* **6**, 2351–2362 (2010).
[10] Roberts, K. E., Cushing, P. R., Boisguerin, P., Madden, D. R. & Donald, B. R. Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity. *PLoS Comput. Biol.* **8**, e1002477 (2012).
[11] Lippow, S. M., Wittrup, K. D. & Tidor, B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.* **25**, 1171–1176 (2007).
[12] Haidar, J. N. *et al.* Structure-based design of a T-cell receptor leads to nearly 100-fold improvement in binding affinity for pepMHC. *Proteins* **74**, 948–960 (2009).
[13]Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A generic program for multistate protein design. *PLoS ONE* **6**, e20937 (2011).
[14]Kapp, G. T. *et al.* Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5277–5282 (2012).
[15] Grigoryan, G., Reinke, A. W. & Keating, A. E. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859–864 (2009).
[16]Yosef, E., Politi, R., Choi, M. H. & Shifman, J. M. Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. *J. Mol. Biol.* **385**, 1470–1480 (2009).
[17] Karanicolas, John, and Brian Kuhlman. "Computational Design of Affinity and Specificity at Protein-protein Interfaces." *Current Opinion in Structural Biology* 19, no. 4 (August 2009): 458–463.
[18] Fleishman, Sarel J., Jacob E. Corn, Eva-Maria Strauch, Timothy A. Whitehead, John Karanicolas, and David Baker. "Hotspot-Centric De Novo Design of Protein Binders." *Journal of Molecular Biology* 413, no. 5 (November 11, 2011): 1047–1062.
[19] Fleishman, Sarel J, Timothy A Whitehead, Damian C Ekiert, Cyrille Dreyfus, Jacob E Corn, Eva-Maria Strauch, Ian A Wilson, and David Baker. "Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin." *Science* 332, no. 6031 (May 13, 2011): 816–821.
[20]Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9 (1998).

[21] Davis, Ian W, and David Baker. “RosettaLigand Docking with Full Ligand and Receptor Flexibility.” *Journal of Molecular Biology* 385, no. 2 (January 16, 2009): 381–392.
[22]Jha, R. K. *et al.* Computational Design of a PAK1 Binding Protein. *Journal of Molecular Biology* **400**, 257–270 (2010).
[23]Stranges, P. B., Machius, M., Miley, M. J., Tripathy, A. & Kuhlman, B. Computational design of a symmetric homodimer using β-strand assembly. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20562–20567 (2011).
[24]Sievers, S. A. *et al.* Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **475**, 96–100 (2011).
[25] Der, B. S. *et al.* Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J. Am. Chem. Soc.* **134**, 375–385 (2012).
[26] Salgado, E. N. *et al.* Metal templated design of protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1827–1832 (2010).
[27] Sia, S. K. & Kim, P. S. Protein Grafting of an HIV-1-Inhibiting Epitope. *PNAS* **100**, 9756–9761 (2003).
[28]Azoitei, M. L. *et al.* Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* **334**, 373–376 (2011).
[29] Garcia-Viloca, Mireia, Jiali Gao, Martin Karplus, and Donald G Truhlar. “How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations.” *Science (New York, N.Y.)* 303, no. 5655 (January 9, 2004): 186–195.
[30] Bhabha, G. *et al.* A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **332**, 234–238 (2011).
[31] Richter, Florian, Andrew Leaver-Fay, Sagar D. Khare, Sinisa Bjelic, and David Baker. “De Novo Enzyme Design Using Rosetta3.” *PLoS ONE* 6, no. 5 (May 16, 2011): e19230.
[32] Chen, Cheng-Yu, Ivelin Georgiev, Amy C Anderson, and Bruce R Donald. “Computational Structure-based Redesign of Enzyme Activity.” *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 10 (March 10, 2009): 3764–3769.
[33] Ashworth, Justin, James J Havranek, Carlos M Duarte, Django Sussman, Raymond J Monnat, Barry L Stoddard, and David Baker. “Computational Redesign of Endonuclease DNA Binding and Cleavage Specificity.” *Nature* 441, no. 7093 (June 1, 2006): 656–659.
[34] Gao, Huirong, Jeff Smith, Meizhu Yang, Spencer Jones, Vesna Djukanovic, Michael G Nicholson, Ande West, et al. “Heritable Targeted Mutagenesis in Maize Using a Designed Endonuclease.” *The Plant Journal: For Cell and Molecular Biology* 61, no. 1 (January 2010): 176–187.
[35] Windbichler, Nikolai, Miriam Menichelli, Philippos Aris Papathanos, Summer B Thyme, Hui Li, Umut Y Ulge, Blake T Hovde, et al. “A Synthetic Homing Endonuclease-based Gene Drive System in the Human Malaria Mosquito.” *Nature* 473, no. 7346 (May 12, 2011): 212–215.
[36] Murphy, Paul M, Jill M Bolduc, Jasmine L Gallaher, Barry L Stoddard, and David Baker. “Alteration of Enzyme Specificity by Computational Loop Remodeling and Design.” *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 23 (June 9, 2009): 9215–9220.
[37] Lippow, Shaun M., Tae Seok Moon, Subhayu Basu, Sang-Hwal Yoon, Xiazhen Li, Brad A. Chapman, Keith Robison, Daša Lipovšek, and Kristala L.J. Prather. “Engineering Enzyme Specificity Using Computational Design of a Defined-Sequence Library.” *Chemistry & Biology* 17, no. 12 (December 22, 2010): 1306–1315.
[38] Khare, Sagar D, Yakov Kipnis, Per Jr Greisen, Ryo Takeuchi, Yacov Ashani, Moshe Goldsmith, Yifan Song, et al. “Computational Redesign of a Mononuclear Zinc Metalloenzyme for Organophosphate Hydrolysis.” *Nature Chemical Biology* 8, no. 3 (2012): 294–300.
[39] Tantillo, D J, J Chen, and K N Houk. “Theozymes and Compuzymes: Theoretical Models for Biological Catalysis.” *Current Opinion in Chemical Biology* 2, no. 6 (December 1998): 743–750.

[40] Zhang, Xiyun, Jason DeChancie, Hakan Gunaydin, Arnab B Chowdry, Fernando R Clemente, Adam J T Smith, T M Handel, and K N Houk. “Quantum Mechanical Design of Enzyme Active Sites.” *The Journal of Organic Chemistry* 73, no. 3 (February 1, 2008): 889–899.
[41] Hellinga, H W, and F M Richards. “Construction of New Ligand Binding Sites in Proteins of Known Structure. I. Computer-aided Modeling of Sites with Pre-defined Geometry.” *Journal of Molecular Biology* 222, no. 3 (December 5, 1991): 763–785.
[42] Zanghellini, Alexandre, Lin Jiang, Andrew M Wollacott, Gong Cheng, Jens Meiler, Eric A Althoff, Daniela Röthlisberger, and David Baker. “New Algorithms and an in Silico Benchmark for Computational Enzyme Design.” *Protein Science: A Publication of the Protein Society* 15, no. 12 (December 2006): 2785–2794.
[43] Malisi, Christoph, Oliver Kohlbacher, and Birte Höcker. “Automated Scaffold Selection for Enzyme Design.” *Proteins* 77, no. 1 (October 2009): 74–83.
[44] Röthlisberger, Daniela, Olga Khersonsky, Andrew M Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L Gallaher, et al. “Kemp Elimination Catalysts by Computational Enzyme Design.” *Nature* 453, no. 7192 (May 8, 2008): 190–195.
[45]Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3790–3795 (2012).
[46] Jiang, Lin, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Röthlisberger, Alexandre Zanghellini, Jasmine L Gallaher, et al. “De Novo Computational Design of Retro-aldol Enzymes.” *Science (New York, N.Y.)* 319, no. 5868 (March 7, 2008): 1387–1391.
[47] Siegel, Justin B, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St Clair, Jasmine L Gallaher, et al. “Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction.” *Science (New York, N.Y.)* 329, no. 5989 (July 16, 2010): 309–313.
[48]Dahiyat, B. I. In silico design for protein stabilization. *Curr. Opin. Biotechnol.* **10**, 387–390 (1999).
[49]Gao, J., Bosco, D. A., Powers, E. T. & Kelly, J. W. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nat Struct Mol Biol* **16**, 684–690 (2009).
[50] Patrick L. Wintrode and Frances H. Arnold, “Temperature adaptation of enzymes: Lessons from laboratory evolution,” in *Evolutionary Protein Design*, vol. 55 (Academic Press, 2001), 161-225.
[51]Russell, R. J., Gerike, U., Danson, M. J., Hough, D. W. & Taylor, G. L. Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure* **6**, 351–361 (1998).
[52] Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural & Molecular Biology* **5**, 470–475 (1998).
[53] Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460 (2003).
[54] Korkegian, A., Black, M. E., Baker, D. & Stoddard, B. L. Computational thermostabilization of an enzyme. *Science* **308**, 857–860 (2005).
[55]Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proceedings of the National Academy of Sciences* (2012).doi:10.1073/pnas.1115172109
[56] Dahiyat, B I, and S L Mayo. “De Novo Protein Design: Fully Automated Sequence Selection.” *Science (New York, N.Y.)* 278, no. 5335 (October 3, 1997): 82–87.
[57] Kuhlman, Brian, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. “Design of a Novel Globular Protein Fold with Atomic-level Accuracy.” *Science (New York, N.Y.)* 302, no. 5649 (November 21, 2003): 1364–1368.
[58]Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol* **268**, 209–225 (1997).

[59] Hill, R. B., Raleigh, D. P., Lombardi, A. & DeGrado, W. F. De novo design of helical bundles as models for understanding protein folding and function. *Acc. Chem. Res.* **33**, 745–754 (2000).
[60] Zaccai, N. R. *et al.* A de novo peptide hexamer with a mutable channel. *Nature Chemical Biology* **7**, 935–941 (2011).

Chapter 2:

[1] Justin B Siegel et al. (2010) “Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction,” Science (New York, N.Y.) 329, no. 5989: 309-313.
[2] Daniela Röthlisberger et al. (2008) “Kemp elimination catalysts by computational enzyme design,” Nature 453, no. 7192: 190-195.
[3] Lin Jiang et al. (2008) “De novo computational design of retro-aldol enzymes,” Science (New York, N.Y.) 319, no. 5868: 1387-1391.
[4] Mireia Garcia-Viloca et al. (2004) “How enzymes work: analysis by modern rate theory and computer simulations,” Science (New York, N.Y.) 303, no. 5655: 186-195.
[5] D J Tantillo, J Chen, and K N Houk (1998) “Theozymes and compuzymes: theoretical models for biological catalysis,” Current Opinion in Chemical Biology 2, no. 6: 743-750.
[6] Alexandre Zanghellini et al. (2006) “New algorithms and an in silico benchmark for computational enzyme design,” Protein Science: A Publication of the Protein Society 15, no. 12: 2785-2794
[7] Andrew Leaver-Fay et al. (2011) “ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules,” Methods in Enzymology 487: 545-574.
[8] Victor Guallar et al. (2004) “Computational Modeling of the Catalytic Reaction in Triosephosphate Isomerase,” Journal of Molecular Biology 337, no. 1: 227-239.
[9] Xiyun Zhang et al. (2008) “Quantum mechanical design of enzyme active sites,” The Journal of Organic Chemistry 73, no. 3: 889-899.
[10] Philip Bradley and David Baker (2006) “Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation,” Proteins 65, no. 4: 922-929.
[11] Gerwald Jogl et al. (2003) “Optimal alignment for enzymatic proton transfer: Structure of the Michaelis complex of triosephosphate isomerase at 1.2-Å resolution,” Proceedings of the National Academy of Sciences of the United States of America 100, no. 1: 50-55
[12] R K Wierenga (2001) “The TIM-barrel fold: a versatile framework for efficient enzymes,” FEBS Letters 492, no. 3: 193-198
[13] Will Sheffler and David Baker (2009) “RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation,” Protein Science: A Publication of the Protein Society 18, no. 1: 229-239
[14] Gert Kiss et al. (2010) “Evaluation and ranking of enzyme designs,” Protein Science: A Publication of the Protein Society 19, no. 9: 1760-1773
[15] James J. Havranek and David Baker (2009) “Motif-directed flexible backbone design of functional interactions,” Protein Science 18, no. 6: 1293-1305.
[16] Adam J T Smith et al. (2008) “Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle,” Journal of the American Chemical Society 130, no. 46: 15361-15373.
[17] Olga Khersonsky et al. (2009) “Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series,” Journal of Molecular Biology 396, no. 4: 1025-42.

Chapter 3:

1. Clark, J. D.; Schievella, A. R.; Nalefski, E. A.; Lin, L. L. *J. Lipid Mediat. Cell. Signal.* ***1995,*** *12*, 83-117.
2. Mignatti, P.; Rifkin, D. B. *Enzyme Protein* **1996**, *49*, 117-137.
3. DeClerck, Y. A.; Imren, S.; Montgomery, A. M.; Mueller, B. M.; Reisfeld, R. A.; Laug, W. E. *Adv. Exp. Med. Biol.* **1997**, *425*, 89-97.
4. MacDonald, T.; DeClerck, Y.; Laug, W. *Thromb. Haemostasis* **1997**, S1541-S1541.
5. Gorrell, M. D. *Clin. Sci. (Lond)* **2005**, *108*, 277-292.
6. Panke, S.; Held, M.; Wubbolts, M. *Curr. Opin. Biotechnol.* **2004**,*15*, 272-279.
7. Sio, C. F.; Quax, W. J. *Curr. Opin. Biotechnol.* **2004**, *15*, 349-355.
8. Riva, S.; Koskinen, A. M. P.; Klibanov, A.M. *Enzymatic Reactions in Organic Media* **1995**, p 140, Springer.
9. Patel, R. N. *Annu. Rev. Microbiol.* **1998**, *52*, 361-395.
10. Vanetten, R. L.; Clowes, G. A.; Sebastian, J. F.; and Bender, M. L. *J. Am. Chem. Soc.* **1967**, *89*, 3253-3262.
11. Vanetten, R. L.; Sebastian, J. F.; Clowes, G. A.; Bender, M. L. *J. Am. Chem. Soc.* **1967**, *89*, 3242-3253.
12. Breslow, R.; Dong, S. D. *Chem. Rev.* **1998**, *98*, 1997-2012.
13. Tanaka, F. *Chem. Rev.* **2002**, *102*, 4885-4906.
14. MacBeath, G.; Hilvert, D. *Chem. Biol.* **1996**, *3*, 433-445.
15. Janda, K. D.; Schloeder, D.; Benkovic, S. J.; Lerner, R. A. *Science* **1988**, *241*, 1188-1191.
16. Stewart, J. D.; Krebs, J. F.; Siuzdak, G.; Berdis, A. J.; Smithrud, D. B.; Benkovic, S. J. *Proc. Nat.l Acad. Sci. USA* **1994**, *91*, 7404-7409.
17. Bolon, D N; Mayo, S. L. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274–14279.
18. Smith, A.J.T; Muller, R.; Toscano, M.D.; Kast, P.; Hellinga, H.W.; Hilvert, D; Houk, K.N. *J. Am. Chem. Soc.* **2008**, *130*, 15361-15373.
19. Simón, L; Goodman, J. M. *J. Organic Chem.* **2010**, *75* (6), 1831-1840
20. Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. *Nature* **2008**, *453*, 190-195.
21. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Rothlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F. 3rd; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. *Science* **2008**, *319*, 1387-1391.
22. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. *Science* **2010**, *329*, 309-313.
23. 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.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popovic, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. *Methods Enzymol.* **2011**, *487*, 545-574.
24. Richter, F.; Leaver-Fay, A.; Khare, S.D.; Bjelic, S.; Baker, D. *PLoS ONE* **2011**, 6(5): e19230.
25. Hedstrom, L. *Chem. Rev.* **2002**,*102*, 4501-4523.
26. Otto, H. H.; Schirmeister, T. *Chem. Rev.* **1997**, *97*, 133-171.
27. Ma, S.; Devi-Kesavan, L. S.; Gao, J. *J. Am. Chem. Soc.* **2007**,*129*, 13633-13645.
28. Kast, P.; Asif-Ullah, M.; Jiang, N.; Hilvert, D. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 5043-5048.
29. Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Rothlisberger, D.; Baker, D. *Protein Sci.* **2006**, *15*, 2785-2794.
30. Vernet, T.; Tessier, D. C.; Chatellier, J.; Plouffe, C.; Lee, T. S.; Thomas, D. Y.; Storer, A. C.; Menard, R. *J. Biol. Chem.* **1995**, *270*, 16645-16652.
31. Tantillo, D. J.; Chen, J. G.; Houk, K. N. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743-750.

32. Zhang, X.; DeChancie, J.; Gunaydin, H.; Chowdry, A.B.; Clemente, F.R.; Smith, A.J.T.; Handel, T.M.; Houk, K.N.; *J. Org. Chem.* **2008**, 73, 889–899.
33 Abbott, D. W.; Boraston, A. B. (2007), *J. Mol. Biol.* **2007**, *369*, 759-770.
34. Kiss, G.; Rothlisberger, D.; Baker, D.; Houk, K. N. *Protein. Sci.* **2010**, *19*, 1760-1773.
35. Kellogg, E. H.; Leaver-Fay, A.; Baker, D. *Proteins* **2011**, *79*, 830-838.
36. Lowe, G.; Williams, A. *Biochem. J.* **1965**, *96*, 199.
37. Cheng, Y.; Prusoff, W. H. *Biochem. Pharmacol.* **1973**, *22*, 3099-3108.
38. Carter, P.; Abrahmsen, L.; Wells, J. A. *Biochemistry* **1991**, *30*, 6142-6148.
39. Chu, S. H.; Mautner, H. G. *J. Org. Chem.* **1966**, *31*, 308-&.
40. Retailleau, P.; Huang, X., Yin, Y.; Hu, M.; Weinreb, V.; Vachette, P.; Vonrhein, C.; Bricogne, G.; Roversi, P.; Ilyin, V.; Carter, C. W., Jr. *J. Mol. Biol.* **2003**, *325*, 39-63.
41. Weerapana, E.; Wang, C.; Simon, G. M.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F. *Nature* **2010**, *468*, 790
42. Carter, P.; Wells, J. A. *Nature* **1988**, *332*, 564-568.
43. Hinkle, P. M.; Kirsch, J. F. *Biochemistry* **1971**, *10*, 2717
44. Brocklehurst, K.; Malthouse, J.P. *Biochemical J.* **1981**,193, 819-823.
45. Patai, S. (ed) *The Chemistry of the Thiol Group*, **1974**, John Wiley & Sons, Ltd, Chapter 8
46. Craik, C. S.; Roczniak, S.; Largman, C.; Rutter, W.J. *Science* **1987**, 237, 909–913.
47. Bryan, P.; Pantoliano, M. W.; Quill, S. G.; Hsiao, H. Y.; Poulos, T. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3743-3745.
48. Bostrom, J.; Greenwood, J. R.; Gottfries, J. *J. Mol. Graph. Model.* **2003**, *21*, 449-462.
49. Sheffler, W.; Baker, D. *Protein Sci.* **2009**,*18*, 229-239.
50. Vigers, G. P. A.; Rizzi, J. P. *J. Med. Chem.* **2004**, *47*, 80-89.
51. Davis, I. W.; Baker, D. *J. Mol. Biol.* **2009**, *385*, 381-392.
52. Charbonnier, J. B.; Carpenter, E.; Gigant, B.; Golinelli-Pimpaneau, B.; Eshhar, Z.; Green, B.S.; Knossow, M. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 11721–11725.
53. Lesley, S. A.; Patten, P. A.; Schultz, P. G. *Proc. Natl. Acad. Sci. USA* **1993**, 90, 1160–1165.
54. Guo, J.; Huang, W.; Scanlan, T. S. *J. Am. Chem. Soc.* **1994**, *116*, 6062–6069.
55. Stewart, J. D.; Roberts, V. A.; Thomas, N. R.; Getzoff, E. D.; Benkovic, S. J. *Biochemistry* **1994**, *33*, 1994–2003.

Chapter 4:

[1] P T Ravi Rajagopalan and Stephen J Benkovic, “Preorganization and protein dynamics in enzyme catalysis,” Chemical Record (New York, N.Y.) 2, no. 1 (2002): 24-36.
[2]K T Simons et al., “Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions,” Journal of Molecular Biology 268, no. 1 (April 25, 1997): 209-225.
[3] Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. Progress in modeling of protein structures and interactions. *Science* **310**, 638–642 (2005).
[4]Colin A Smith and Tanja Kortemme, “Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction,” Journal of Molecular Biology 380, no. 4 (July 18, 2008): 742-756.
[5]Davis, I. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* **14**, 265–274 (2006).
[6] Evangelos A. Coutsias et al., “A kinematic view of loop closure,” Journal of Computational Chemistry 25, no. 4 (2004): 510-528.
[7]Carol A Rohl et al., “Protein structure prediction using Rosetta,” Methods in Enzymology 383 (2004): 66-93.
[8] Paul M Murphy et al., “Alteration of enzyme specificity by computational loop remodeling and design,” Proceedings of the National Academy of Sciences of the United States of America 106, no. 23 (June 9, 2009): 9215-9220.
[9] James J. Havranek and David Baker, “Motif-directed flexible backbone design of functional interactions,” Protein Science 18, no. 6 (2009): 1293-1305.
[10] Philip Bradley and David Baker, “Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation,” Proteins 65, no. 4 (December 1, 2006): 922-929.
[11] Huang, P.-S. *et al.* RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS ONE* **6**, e24109 (2011).
[12] Wang C, Bradley P, Baker D (2007) Protein-protein docking with backbone flexibility. Journal of Molecular Biology 373: 503–519
[13] Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with rosetta. Proteins-Structure Function and Bioinformatics 55: 656–677
[14] Canutescu AA, Dunbrack RL (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Science 12: 963–972.
[15] W Kabsch and C Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” Biopolymers 22, no. 12 (December 1983): 2577-2637.
[16] Fleishman, S. J. *et al.* Hotspot-Centric De Novo Design of Protein Binders. *Journal of Molecular Biology* **413**, 1047–1062 (2011).
[17] Wang, C., Bradley, P. & Baker, D. Protein-protein docking with backbone flexibility. *J. Mol. Biol.* **373**, 503–519 (2007).
[18] Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).

Appendix A: Supporting Information for Chapter 2

Table S1: Theozyme geometries

parameter	ideal value +- tolerance	values sampled by matcher	num matcher samples	$k_{constraint}$
Interaction 1 **Glu/Asp - DHAP** (total number of ligand placements per rotamer: 1*3*3*3*3*7= 567)				
distanceAB/Å (E/D:Ocarboxy-DHAP:C1)	3.06 +- 0.2	3.06	1	100
angleA /º (E/D:Ocarboxy - DHAP:C1 - DHAP:C2)	73.6 +- 10.0	63.6, 73.6, 83.6	3	80.0
angleB /º (E/D:Cy/g - E/D:Ocarboxy - DHAP:C1)	120.0 +- 15.0	105.0, 120.0, 135.0	3	80.0
torsionA /º (E/D:Ocarboxy - DHAP:C1 - DHAP:C2 - DHAP:O2)	-101.2 +- 15.0	-86.2, -101.2, -116.2	3	60.0
torsionB /º(E/D:Cγ/β - E/D:Cδ/γ - E/D:Ocarboxy - DHAP:C1)	180.0 +- 15.0	165.0, 180.0, -165.0	3	0.0
torsionAB /º (E/D:Cδ/γ - E/D:Ocarboxy - DHAP:C1- DHAP:C2)	180.0 +- 90.0	90.0, 120.0, 150.0, 180.0, -150.0, -120.0, -90.0	7	0.0
Interaction 2 **His - DHAP** (total number of ligand placements per rotamer: 1*3*3*3*3*10= 810)				
distanceAB /Å (DHAP:O2 - His:Nε2)	2.72 +- 0.20	2.72	1	100.0
angleA /º (DHAP:C2 - DHAP:O2 - His:Nε2)	111.2 +- 10.0	101.2, 111.2, 121.2	3	50.0
angleB /º (DHAP:O2 - His:Nε2 - His:Cε1)	120.3 +- 15.0	105.3, 120.3, 135.3	3	50.0
torsionA /º (DHAP:C1 - DHAP:C2 - DHAP:O2 - His:Nε2)	0.0 +- 10.0	-10.0, 0.0, 10.0	3	50.0
torsionB /º (DHAP:O2 - His:Nε2 - His:Cε1 - His:Nδ1)	180.0 +- 15.0	165.0, 180.0, -165.0	3	0.0
torsionAB /º (DHAP:C2 - DHAP:O2 - His:Nε2 - His:Cε1)	0.0 +- 30.0; 180.0 +- 30.0	-30.0, -15.0, 0.0, 15.0, 30.0 150.0, 165.0, 180.0, -165.0, -150.0	10	0.0
Interaction 3				

Lys - DHAP (secondary match algorithm used)				
distanceAB/Å (DHAP:O2 - Lys:Nζ)	2.90 +- 0.2	n/a	n/a	100.0
angleA /º (DHAP:C2 - (DHAP:O2 - Lys:Nζ)	109.3 +- 20.0	n/a	n/a	50.0
angleB /º (DHAP:O2 - Lys:Nζ - Lys:Cε)	109.8 +- 20.0	n/a	n/a	50.0
torsionA /º (DHAP:C1 - DHAP:C2 - DHAP:O2 - Lys:Nζ)	-100.0 +- 30.0	n/a	n/a	0.0
torsionB /º (DHAP:O2 - Lys:Nζ - Lys:Cε - Lys:Cδ)	any (0-360)	n/a	n/a	0.0
torsionAB /º (DHAP:C2 - DHAP:O2 - Lys:Nζ - Lys:Cε)	any (0-360)	n/a	n/a	0.0

Appendix B: Supporting information for Chapter 3

Methods

Protein Production. The cell pellets were resuspended in sonication buffer (25 mM Hepes (pH 7.5), 300 mM NaCl, 1 mM TCEP) containing 10 mM imidazole. Cell lysis was achieved by the addition of 1 mg/mL lysozyme and subsequent sonication. The soluble fraction was applied to Ni-NTA slurry (Qiagen), washed with 20 mM imidazole and then with 32.5 mM imidazole before elution with 250 mM imidazole in sonication buffer. The proteins were dialyzed into 20 mM Tris, 20 mM NaCl (pH 8.0) for 16 hours and then purified by anion exchange chromatography (MonoQ column, GE Healthcare) in the same buffer, eluting with a salt gradient (20 mM to 1,000 mM NaCl). The proteins were concentrated using Amicon Ultra-15 units (Millipore). Protein concentrations were determined by measuring the absorbance at 280 nm using the calculated extinction coefficients (ε (ECH13; ECH13 C45A; ECH13 C45A/H100A) = 30940 M^{-1} cm^{-1}, ε (ECH14; ECH14 C132A; ECH14 C132A/H104A) = 42860 M^{-1} cm^{-1}, ε (ECH19; ECH19 C161A; ECH19 C161A/H226A) = 95800 M^{-1} cm^{-1}, ε (ECH19 K354P/P364W) = 101300 M^{-1} cm^{-1}, ε (FR29; FR29 C10A/H126A; FR29 A44S/T112L/V151L) = 41830 M^{-1} cm^{-1}). Protein purity was confirmed by SDS-PAGE.

Initial activity screening. For all 55 designs, a mixture containing 20 μM purified protein and 100 μM of coumarin ester **2** was prepared and product formation was monitored in a fluorimeter at room temperature ((λ_{ex} = 340 nm, λ_{em} = 452 nm). Designs that showed at least a 20-fold increase over the background rate (100 μM substrate **2** in buffer taken from the dialysis bucket) were considered active, and the 4 designs that did were characterized further.

Kinetic measurements. For substrate **2**, product formation was monitored in a fluorimeter (Photon Technology International) at 29°C (λ_{ex} = 340 nm, λ_{em} = 452 nm). The signal was calibrated using a concentration series of 7-hydroxycoumarine (1 μM to 50 μM final concentration) in 25 mM Hepes buffer (pH 7.5), containing 100 mM NaCl and 5% acetonitrile. For substrate **3**, release of para-nitrophenol was monitored at 405 nm in a Lambda 35 UV/Vis spectrometer (PerkinElmer) at 29 °C.

***K*$_i$ determination.** The esterase variants (2 μM final concentrations) were pre-incubated with varying concentrations of the tyrosine ester **1** (0.001 – 300 μM final concentration) in 25 mM HEPES buffer (pH 7.5), 100 mM NaCl, 5% acetonitrile and the reactions were initiated by addition of the coumarin ester **2** (50 μM final concentration). Product formation was monitored as described previously. The IC_{50} value was determined by curve fitting (Hill-Slope model, v_i at infinite inhibitor concentration was set to zero) and subsequently converted into the corresponding K_i value using the Cheng-Prusoff equation $K_i = [IC_{50}]/(1+[S]/K_m)$ (*42*). The K_m value was determined independently under identical reaction conditions using the previously described protocol.

CD spectroscopy. The far-UV spectra of the protein samples (c (ECH13; ECH13 C45A; ECH13 C45A/H100A) = 10 μM, c (ECH19; ECH19 C161A; ECH19 C161A/H226A; ECH19 K354P/P364W) = 5 μM, c (ECH14; ECH14 C132A; ECH14 C132A/H100A) = 5 μM, c (FR29; FR29 C10A H126A; FR29 A44S/T112L/V151L) = 5 μM) were measured at 20 °C using an Aviv 202 spectropolarimeter (Aviv Associates, Lakewood, NJ). Thermal denaturation spectra of the proteins were monitored at 222 nm.

Generation of the optimized variants. The genes for the modified designs were generated by the Kunkel method[1] and the corresponding proteins were produced as described for the original designs. To confirm their identity, all variants were characterized by mass spectrometry (Tables 6S-9S). Their activities were evaluated by incubationg 5 or 10 μM of each enzyme with a tenfold excess of coumarin ester **2** and by subsequently monitoring fluorophore release as described above (Figures 1S and 2S).

Aminolysis experiments. The best third-generation ester hydrolase design, FR29 A44S/T112L/V151, was pre-incubated with a tenfold excess of the amines shown in Figure 6S (and one thiol) at pH 7.5 prior to addition of coumarin substrate **2**.

[1] Kunkel, T. A., Roberts, J. D., and Zakour, R. A. (1987) Rapid and efficient site-specific mutagenesis without phenotypic selection, *Methods Enzymol 154*, 367-382

Unfortunately, the deacylation rates remained identical to those observed in the absence of amines. Hence, we conclude that none of the added nucleophiles facilitates cleavage of acyl-enzyme intermediate**.**

Mass spectrometry. Protein samples for the characterization of the acyl-enzyme intermediates were prepared by incubating the designed hydrolases (10 μM final concentration in 25 mM Hepes (pH 7.5), 100 mM NaCl) with a 10-fold excess of tyrosine ester **1** or coumarin ester **2** (or no ester for the negative control) at room temperature for 6 hours or 60 min, respectively. The samples were then subjected to a shock-freeze in liquid nitrogen. Immediately before injection on the mass spectrometer the samples were thawed. For ESI-MS studies, the samples were desalted using a C_4 ZipTip and measured in 50% acetonitrile/0.2% formic acid (pH 2.0) on a Q-TOF Ultima mass spectrometer (Waters). For MALDI-MS/MS studies, 30 μl of the sample were digested with 10-20 μl trypsin (10 ng/ul in 10 mM Tris, 2 mM $CaCl_2$, pH 8.2) at 50° C for 2 hours and the resulting peptides were analyzed by MALDI. Comparison of the treated und untreated samples was used to identify the modified peptides, which were then further fragmented by collision-induced dissociation (CID). The resulting fragments were analyzed to identify the modified amino acid.

Molecular Dynamics Simulations. Molecular dynamics simulations were carried out on designs FR25 through FR32. Each was prepared for three independent MD runs: one apo-MD, one with substrate **1** bound to the active site, and one with substrate **2**. MD simulations were also carried out on ECH13, ECH14, and ECH19, which were prepared for apo-MD and with substrate **2**, only. FR26, FR29 and ECH14 were set up as dimeric systems Substrate parameters were generated within the antechamber module of AMBER 10[2] using the general AMBER force field, with partial charges set to

[2] Case DA, Darden TA, Cheatham III TE, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008), AMBER 10, University of California, San Francisco.

fit the electrostatic potential generated at HF/6-31G* by RESP.[3] The charges were calculated according to the Merz-Singh-Kollman scheme[4,5] using Gaussian 03.[6] Structures were immersed in a truncated octahedral box with a 10 Å buffer of TIP3P[7] water molecules. The systems were neutralized by addition of explicit counter ions. Water molecules were triangulated with the SHAKE algorithm such that the angle between the hydrogen atoms is kept fixed. After equilibration (SI), a 20 ns production MD simulation was performed for each of the systems using pmemd.[8] Geometries and velocities were saved every 100 steps (0.2 ps) which resulted in a total of 100,000 frames from each production run. Long-range electrostatic effects were modeled using the particle-mesh-Ewald method.[9] Post-MD data-extraction and analysis was performed using the ptraj module of AMBER 10 and the statistical analysis software OriginPro8.[10]

The active sites of FR25, FR26, and FR31 showed substantial instabilities that were deemed irreparable. FR27 and FR30 were designated as borderline cases that were expected to have weak activity at most due to high solvent accessibility (active site of

---

[3] Bayly CI, Cieplak P, Cornell WD, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. 97:10269-10280.

[4] Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J. Comput. Chem. 11:431-439.

[5] Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J. Comput. Chem. 5:129-145.

[6] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford CT.

[7] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79:926-935.

[8] Duke RE, Pedersen LG (2003) PMEMD. University of North Carolina, Chapel Hill.

[9] Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98:10089-10092.

[10] Origin, OriginLab, Northampton, MA.

FR30) or missing catalytic contacts (oxyanion hole of FR27). FR28 and FR32, the two most promising designs from MD analysis, turned out to express poorly. Efforts to improve their solubility were met with no success.

Structure determination

Sample preparation for crystallization experiments. The production of the four active designs was carried out as part of the high-throughput protein-production process of the Northeast Structural Genomics Consortium (NESG)[11]. Each of the designs was assigned and NESG identifier: OR49 for ECH19, OR51 for ECH13, OR52 for FR29, and OR54 for ECH14. *E. coli* BL21-GOLD (DE3) were transformed with those expression vectors. A single isolate was cultured in MJ9 minimal medium[12] supplemented with selenomethionine, lysine, phenylalanine, threonine, isoleucine, leucine and valine for the production of selenomethionine-labeled versions of the designs[13]. Initial growth was carried out at 37^{O}C until the OD_{600} of the culture reached 0.6–0.8. The incubation temperature was then decreased to 17^{O}C and protein expression was induced by the addition of IPTG (isopropyl β-D-1-thiogalactopyranoside) to a final concentration of 1 mM. Following overnight incubation, the cells were harvested by centrifugation. Cell pellets were resuspended in lysis buffer (50 mM Tris pH 7.5, 500 mM NaCl, 40 mM imidazole, 1 mM TCEP and 0.02% NaN_3) containing protease inhibitors (Complete, Mini, EDTA-free, Roche) and disrupted by sonication. The resulting lysate was clarified by centrifugation at 27,000 x g for 30 min at 4^{O}C, followed by filtering through a 0.2 mm filter. The supernatant was loaded onto an AKTAxpress system (GE Healthcare) with a two-step protocol consisting of IMAC (HisTrap HP) and gel-filtration (HiLoad 26/60

[11] Xiao, R.; Anderson, S.; Aramini, J.M.; Belote, R.; Buchwald, W.; Ciccosanti, C.; Conover, K.; Everett, J.K.; Hamilton, K.; Huang, Y.J.; Janjua, H.; Jiang, M.; Kornhaber, G.J.; Lee, D.Y.; Locke, J.Y.; Ma, L.-C.; Maglaqui, M.; Mao, L.; Mitra, S.; Patel, D.; Rossi, P.; Sahdev, S.;Sharma, S.; Shastry, R.; Swapna, G.V.T.; Tong, S.N.; Wang, D.; Wang, H.; Zhao, L.; Montelione, G.T.; Acton, T.B. J. Struct. Biol. 2010, 172: 21 - 33

[12] Jansson, M.; Li, Y.-C.; Jendeberg, L.; Anderson, S.; Montelione, G. T.; Nilsson, B. J. Biomol. NMR. 1996, 7: 131-141

[13] Doublie´, S.; Kapp, U.; Aberg, A.; Brown, K.; Strub, K.; Cusack, S. FEBS Lett. 1996, 384: 219-221

Superdex 75) chromatography. The purified designs were in a buffer containing 10 mM Tris-HCl , 100 mM NaCl, 5 mM DTT, pH 7.5, and concentrated to 6–10.5 mg/m1. Samples were flash-frozen in 50 µL aliquots using liquid nitrogen and stored at -80^{O}C until crystallization. The sample purity (>98%), molecular weight, oligomerization state were verified by SDS–PAGE, MALDI–TOF mass spectrometry, and analytic gel filtration followed by static light scattering, respectively.

Crystallization and data collection. Initial crystallization conditions for all four proteins were found by high-throughput robotic screening of 1536 different conditions at the Hauptmann-Woodward Institute, Buffalo, NY[14]. Crystallization and refinement statistics for all designs can be found in Table 10S.

ECH19 / OR49: Manual optimization of the initial crystallization conditions for ECH19 was performed by mixing of 10.1 mg/ml protein in buffer containing 100 mM NaCl, 5 mM DTT, 0.02% NaN_3, 10 mM Tris-HCl, pH 7.5 with 1 µl of the precipitant. The final precipitant solution contained 18% (w/v) PEG-3350, 0.15 M ammonium sulfate, 0.1 M Tris-HCl, pH 8.0. The crystals were cryoprotected with 15-20 % (w/v) ethylene glycol in the well solution before flash-freezing in liquid nitrogen. X-ray data were collected at the Brookhaven National Laboratory (BNL), beamline X4C. Crystal diffracted to 2.5Å and belongs to space group $P2_12_12$, with one molecule in the asymmetric unit [Table 10S].

ECH13/ OR51: Crystals of ECH13 protein were grown by microbatch method under mineral oil by mixing protein solution containing 100 mM NaCl, 5 mM DTT, 0.02% NaN_3, 10 mM Tris-HCl, pH 7.5 with reservoir solution consisted of 100 mM $NaHPO_4$, 12% PEG20K in 100 mM MES buffer pH 7.5. The complex with coumarin was prepared by a cocrystallization method. As fluorogenic coumarin ester is not water soluble, it was dissolved in DMSO at 200 mM, then mixed with protein solution at final concentration of 10 mM. The mixture was incubated on the ice overnight, and formed supernatant was used for crystallization by macrobatch method. Obtained crystals were flash-cooled at

[14] Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003). "A deliberate approach to screening for initial crystallization conditions of biological macromolecules.", Journal of Structural Biology, 142, 170-179

100K in a nitrogen stream with cryoprotectant, whereas the crystals were briefly soaked in solution containing 20% (v/v) ethylene glycol. X-ray data for native protein were collected on BL9-2 beamline, SSRL and were processed with the program DENZO.[15] Crystals diffracted to 1.6Å. Diffraction data for liganded form were collected on RAXIS-IV image plate detector, Rigaku rotating anode, at 1.5418Å wavelength. The crystals of unliganded form belong to space group $P4_32_12$ and diffracted to 1.6Å, but crystals of complex belong to space group C2, diffracted to 2.0Å. It's interesting that attempts to grow complex crystals by soaking of native crystals in a solution containing coumarin failed.

FR29 / OR52: The same approach as for ECH13 was used to get crystals for liganded/unliganded forms of FR29. The final concentration of crystallization cocktail for unliganded form was: 0.17M of NH_4acetate, 0.085M of Na_3citrate, 25% of PEG4K, 15% glycerol, pH 5.6. The crystal for complex with coumarin were grown from solution containing 0.5M of ammonium sulfate, 0.1M HEPES, 30% MPD, pH 7.5. Data for both crystal forms were collected at Brookhaven National Laboratory (BNL), beamlines X4A and X4C [Table 10S]. Both crystal forms belong to orthorhombic space group $P2_12_12_1$, 4 copies of the molecule in the asymmetric unit and diffracted to 2.8Å.

ECH14 / OR54: The crystals for ECH14 were grown by sitting drop under oil method at the 4^0C. Crystallization solution contained 0.1M of KH_2PO_4, 40% PEG 4000, 0.1M Tris, pH 8.0. X-ray data were collected from crystals maintained at 100 K using a wavelength of 0.979Å on beamline X4C at the National Synchtrotron Light Source at BNL. The diffraction images were integrated and merged using HKL2000 and SCALEPACK[14]. Orthorombic crystals (space group $P2_12_12_1$) diffracted to 3.2Å, with 2 molecules in the asymmetric unit [Table 10S].

Structure determination and refinement. ECH19 / OR49: Structures for unliganded cysteine esterases were determined by Molecular Replacement (MR) method (program BALBES[16]) using appropriate scaffold. The scaffold for ECH19 is periplasmic

[15] Otwinowski,Z. & Minor, W. (1997). Methods Enzymol. 276, 307-326
[16] F.Long, A.Vagin, P.Young & G.N.Murshudov (2008).

oligogalacturonide binding unsaturated hexuronate sugars from *Yersinia enterocolitica* (PDB id 2UVH) in its closed form, 430 amino acids residues length. Program automatically made corresponding mutations and refined preliminary model with REFMAC.[17] Further refinement was done with program package PHENIX[18] and refitting/remodeling was performed with program COOT.[18] Details of data collection and refinement are summarized in the Table 10S. The asymmetric unit contains two copies of the molecule (A and B labeled). Subunit A (residues 2-412) and subunit B (residues 3-409) are packed by 'face-to-face' mode. The RMS in CA atom position of two subunits is 0.20Å and maximum deviation is 1.89Å is observed in vicinity of residues Ser229. Overlay of the scaffold molecule A on the subunit A of target molecule showed bigger differences: 4.009Å for corresponding CA atoms with maximum 9.125Å for Asp47. When comparing to the open, unliganded form of the scaffold, PDB code 2UVG, the RMS deviation for all CA-atoms is 1.351Å, with a maximum of 7.045Å for Ser229 in subunit A, located close to mutated M231H. Subunit B is better matched: 1.33Å for all CA-atoms, and maximum difference 6.66Å for Phe226.

ECH13/ OR51: Three-dimensional structure of the apo enzyme ECH13 was solved by MR, using coordinates of the scaffold, human mitochondrial deoxyribonuletidase (PDB entry 1Q92) as search model for program BALBES. It was found 1 molecule (space group $P4_32_12$), and preliminary refined with REFMAC generated Rfree=0.287, R=0.255. After manual inspection and rebuilding with program COOT[19] and anisotropic refinement with solvent molecules with PHENIX, Rfree factor dropped to 0.195, and standard crystallographic R factor dropped to 0.178. The RMS deviation of the coordinates from ideal is 0.008Å. No amino acid residues are in the disallowed region,

---

"BALBES: a Molecular Replacement Pipeline", Acta. Cryst., D64, 125-132

[17] G.Murshudov, A.Vagin, & E.Dodson. (1997). " Refinement of Macromolecular Structures by the Maximum-Likelyhood Method", Acta Cryst, D53, 240-255

[18] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J.S. Richardson, T.C. Terwilliger & P.H.Zwart. (2010), Acta Cryst., D66, 213-221.

[19] Emsley, P.and Cowtan, K.D. (2004) "Coot: Model-building tools for molecular graphics". Acta Cryst., D 60, 2126-2132

91.8% are located in the most favored region on the Ramachandran map. The RMS deviation for CA atoms for target-scaffold molecules is 0.280Å. Biggest deviation 2.815Å is observed for the first amino acid in C-term, second one is for SeMet105, which is located in the vicinity of mutated amino acid I103H. It is interesting that the scaffold and unliganded target molecule have the same space group and almost the same cell parameters: a=b=73.758Å c=105.981Å (1Q92) vs a=b=73.436Å c=105.083Å (3U13, target molecule). The crystallization conditions are also similar: PEG 8K/20K, potassium/sodium phosphate, pH 5.3/6.0. The crystal of the search molecule diffracted to 1.4Å against the current resolution 1.6Å, but the completeness of the scaffold is low 64.8%, whereas the target completeness is 100%. The complex with ester coumarin was solved by MR using coordinates of the unliganded enzyme and was rebuilt/refined with COOT/PHENIX correspondingly.

FR29 / OR52: The three-dimensional structure of the apo-enzyme was determined by MR with the program BALBES using coordinates of the scaffold, tryptophanyl–tRNA synthase from *Bacillus Stearothermophilus* (PDB code 3FHJ). It was found 4 molecules and refined with program PHENIX [9]. Model was rebuilt/corrected with program COOT. Final model (Rfree=0.290, R=0.214 at the resolution 2.8Å, RMS deviation for bond lengths 0.009Å) was deposited to PDB (id 3U1V). Complex with ester coumarin was solved by MR using structure of the apo-enzyme. The RMS deviations from original molecule A are 1.536Å, 1.636Å, 1.618Å, and 1.737Å for target molecules A, B, C, D respectively.

ECH14 / OR54: The coordinates of the design scaffold, aspartate aminotransferase from *E.coli* (PDB id 1TOI, sequence identity 94.7%) was used as search model for MR to solve the structure of the ECH14. The initial R_{free}/R were 0.364/0.249 at the resolution 3.2 Å. After several cycles of rebuilding with program COOT and refinement with PHENIX, R_{free} reduced to 0.287, and R=0.197. 85% of the amino acid residues lie in the favored region on the Ramachandron plot, and 0.1% is in the disallowed region. RMS deviation of the protein atoms from ideal values is 0.009Å. The high average B-factor for protein atoms 74.2Å2 is correlated to poor diffraction of the crystals (3.2Å). Despite of low resolution data, all chains for both subunits (A and B) was traced,

whereas the high resolution 1.9Å scaffold molecule has gap 125-129. RMS deviation for subunit A (2616 atoms) between final model and scaffold, calculated with PHENIX (command Phenix.Superpose_pdbs) is 2.022Å, and 2.030Å for molecule B.

Substrate synthesis

Tyrosine ester 1 (SI, Figure 4S). *Benzyl 2-(S)-(N-fluorenyl-methoxycarbonyl)-amino-3-(p-hydroxyphenyl)-propanoate (Fmoc-Tyr(OH)-OBn)* (***1c***) – Fmoc-Tyr(O-*t*Bu)-OH (1.0 g, 2.18 mmol, 1.0 eq.) and HBTU (908 mg, 2.39 mmol, 1.1 eq.) were dissolved in anhydrous DMF (5 mL) and stirred under N_2 at 0^0C (ice/water bath). DIPEA (431 µL, 5.66 mmol, 2.6 eq.) was added drop wise. After stirring for 5 min, benzyl alcohol (338 µL, 3.26 mmol, 1.5 eq.) was added. After another 40 min, the reaction mixture was warmed up to room temperature and quenched with saturated aqueous NH_4Cl solution (100 mL) after incubation for another hour. Extraction with ethyl acetate (3x50 mL) and washing the combined organic fractions with saturated aqueous NH_4Cl (1x100 mL), brine (2x100 mL) followed by drying with anhydrous Na_2SO_4 and evaporation of solvents yielded the crude product. Purification was performed using flash chromatography on silicagel (most of benzyl alcohol elutes with hexane:ethyl acetate 20:1, the product was then eluted with hexane:ethyl acetate 3:1) resulting in 1.0 g (83% yield) of the benzyl ester.

TLC: R_f = 0.5 (cyclohexane:ethyl acetate 3:1); **LCMS (ESI):** RT = 13.93 min., *m/z* calculated for $C_{35}H_{36}NO_5$: 550.26 $[M+H]^+$, found: 550.24 $[M+H]^+$, *m/z* calculated for $C_{35}H_{35}NO_5Na$: 572.24 $[M+Na]^+$, found: 572.26 $[M+Na]^+$; **^{1}H NMR ($CDCl_3$) δ(ppm):** 7.78 (2H, d), 7.58 (2H, d), 7.29-7.44 (9H, m), 6.85-6.94 (4H, m), 5.32 (1H, d, *J* = 8.4 Hz), 5.17 (2H, m), 4.71 (1H, m), 4.34-4.47 (2H, m), 4.22 (1H, t, *J* = 7.2 Hz), 3.10 (2H, m), 1.34 (9H, s); **^{13}C NMR ($CDCl_3$) δ(ppm):** 171.0, 155.2, 154.2, 143.6, 143.5, 141.1, 134.9, 130.1, 129.6, 128.4, 128.3, 127.5, 126.9, 124.9, 123.9, 119.8, 78.4, 67.3, 67.0, 55.0, 47.3, 37.7, 29.0.

485 mg of the benzyl ester (Fmoc-Tyr(O-*t*Bu)-OBn, 0.882 mmol) was dissolved in trifluoroacetic acid (20 mL) and stirred at room temperature for 1h. The solvent was then

evaporated under reduced pressure and the resulting crude product was purified using flash chromatography (SiO_2, elution of impurities with hexane:ethyl acetate 5:1, the product's elution begins with hexane:ethyl acetate 3:1 (122 mg, 28% of the expected yield), however it is completely removed from the column with pure ethyl acetate followed by ethyl acetate:methanol 4:1). Fractions containing the product were combined and evaporated resulting in 422 mg (97% of the expected yield) of the pure Fmoc-Tyr(OH)-OBn.

TLC: R_f = 0.25 (cyclohexane:ethyl acetate 3:1); **LCMS (ESI):** RT = 12.35 min., *m/z* calculated for $C_{31}H_{28}NO_5$: 494.20 $[M+H]^+$, found: 494.24 $[M+H]^+$, *m/z* calculated for $C_{31}H_{27}NO_5Na$: 516.18 $[M+Na]^+$, found: 516.20 $[M+Na]^+$; **^{1}H NMR ($CDCl_3$) δ(ppm):** 7.67 (2H, d), 7.48 (2H, d), 7.19-7.33 (9H, m), 6.78 (2H, d, *J* = 8.4 Hz), 6.62 (2H, m), 5.05 (2H, m), 4.52 (1H, t, *J* = 5.7 Hz), 4.23-4.35 (2H, m), 4.10 (1H, t, *J* = 6.6 Hz), 2.72 – 3.00 (2H, m); **^{13}C NMR ($CDCl_3$) δ(ppm):** 171.4, 155.5, 143.4, 143.3, 140.9, 134.7, 130.0, 128.3, 128.3, 127.4, 126.8, 126.2, 124.8, 119.7, 115.2, 67.2, 66.9, 55.1, 47.1, 37.2.

*Benzyl 2-(S)-[(N-fluorenylmethoxycarbonyl)-amino]- 3-[p-(2'-phenyl-2'-methyl-acetyloxy)-phenyl] -propanoate(Fmoc-Tyr(O-2-Me-phenylacetyl)-OBn) (**1d**)* – 2-methyl-2-phenyl-acetic acid (107 mg, 0.71 mmol, 1.0 eq.) and HATU (283 mg, 0.74 mmol, 1.05 eq.) were dissolved in anhydrous DMF (3 mL) and stirred under N_2 at 0^0C (ice/water bath) until homogenization. DIPEA (129 µL, 0.78 mmol, 1.1 eq.) was added drop wise at 0^0C. After stirring for 5 min, the reaction mixture was warmed up to room temperature. Fmoc-Tyr(OH)-OBn (350 mg, 0.71 mmol, 1.0 eq.) was dissolved in anhydrous DMF (4 mL) and added to the activated acid. The mixture was then stirred overnight at room temperature in an N_2 atmosphere. The reaction mixture was quenched with saturated aqueous NH_4Cl solution (100 mL). Extraction with ethyl acetate (3x50 mL) and washing the combined organic fractions with saturated aqueous NH_4Cl (1x100 mL), brine (2x100 mL) followed by drying with anhydrous Na_2SO_4 and evaporation of solvents yielded the crude product. Purification was performed using flash chromatography (SiO_2, elution with hexane:ethyl acetate 10:1, then 5:1) resulting in 235 mg (53% yield) of the ester **1d**.

TLC: R_f = 0.65 (hexane:ethyl acetate 3:1); **^{1}H NMR ($CDCl_3$) δ(ppm):** 7.78 (2H, d), 7.58 (2H, d), 7.20-7.43 (14H, m), 6.97 (2H, d), 6.87 (2H, d), 5.33 (1H, d, *J* = 8.1 Hz), 5.16 (2H, m), 4.71 (1H, m), 4.35-4.49 (2H, m), 4.22 (1H, t, *J* = 6.9 Hz), 3.99 (1H, q^4, *J* = 6.9 Hz), 3.11 (2H, m), 1.65 (3H, d, *J* = 6.9 Hz); **^{13}C NMR ($CDCl_3$) δ(ppm):** 172.5, 170.8, 155.2, 149.6, 143.6, 141.1, 139.8, 132.9, 130.1, 128.4, 127.5, 126.9, 124.9, 121.3, 119.9, 67.4, 67.0, 54.8, 47.2, 45.9, 37.7, 18.7.

*Benzyl 2-(S)-amino- 3-[p-(2'-phenyl-2'-methyl-acetyloxy)-phenyl] -propanoate(H-Tyr(O-2-Me-phenylacetyl)-OBn) (**1e**)* – Compound **1d** (235 mg, 0.38 mmol, 1.0 eq.) was dissolved in DMF (4 mL). Piperidine (1 mL) was added and the mixture was stirred 20 min at room temperature. It was then evaporated to dryness under high vacuum. The residue was dissolved in ethyl acetate and then a 3-fold excess of hexane was added. Purification was performed using flash chromatography (SiO_2, elution with hexane:ethyl acetate 3:1 (the fluorenone-containing byproduct), then hexane ethyl acetate 1:3) resulting in 130 mg (86% yield) of the amine **1e**.

TLC: R_f = 0.15 (hexane:ethyl acetate 1:3); **LCMS (ESI):** RT = 8.30 min., *m/z* calculated for $C_{25}H_{26}NO_4$: 404.19 $[M+H]^+$, found: 404.16 $[M+H]^+$; **^{1}H NMR ($CDCl_3$) δ(ppm):** 7.28-7.44 (10H, m), 7.10 (2H, m), 6.90 (2H, m), 5.12 (2H, s), 3.96 (1H, q^4, *J* = 7.2 Hz), 3.74 (1H, t, *J* = 5.4 Hz), 2.85-3.15 (2H, m), 1.82 (2H, br s), 1.62 (3H, d, *J* = 6.9 Hz); **^{13}C NMR ($CDCl_3$) δ(ppm):** 174.7, 173.0, 149.8, 140.1, 135.5, 134.6, 130.4, 130.2, 128.9, 128.6, 128.5, 128.4, 127.6, 127.4, 121.4, 66.9, 55.8, 45.7, 40.4, 18.5;

*2-(S)-amino- 3-[p-(2'-phenyl-2'-methyl-acetyloxy)-phenyl]propanoic acid (H-Tyr(O-2-Me-phenylacetyl)-OH) (**1**)* **–** Compound **1e** (185 mg, 0.46 mmol, 1.0 eq.) was dissolved in methanol (10 mL). The solution was degassed. 5% Pd on charcoal (37 mg, 1 weight percent) was added. The flask was degassed and flushed with nitrogen (3 cycles). Then the reaction mixture was degassed again and filled with hydrogen gas under atmospheric pressure. After 90 min total substrate consumption was confirmed using TLC. The flask was degassed and flushed with nitrogen gas (3 cycles) before the mixture was filtered through a celite pad under reduced pressure. The pad was

subsequently washed with methanol (3 x 50 mL). The organic fractions were combined and evaporated yielding 141 mg (98% yield) of the expected compound **1**.

TLC: R_f = 0 (hexane:ethyl acetate 1:3); **LCMS (ESI):** RT = 7.12 min., *m/z* calculated for $C_{18}H_{20}NO_4$: 314.14 $[M+H]^+$, found: 314.08 $[M+H]^+$, *m/z* calculated for $C_{36}H_{39}N_2O_8$: 627.27 $[2M+H]^+$, found: 626.98 $[2M+H]^+$; **^{1}H NMR (d^6-DMSO+1 drop TFA) δ(ppm):** 8.27 (3H, br s), 7.39 (5H, m), 7.26-7.33 (3H, m), 6.99 (2H, m), 4.20 (1H, m), 4.08 (1H, q^4, *J* = 6.9 Hz), 3.08 (1H, d, *J* = 5.4 Hz), 1.50 (3H, d, *J* = 7.2 Hz); **^{13}C NMR (d^6-DMSO+1 drop TFA) δ(ppm):** 173.1, 152.5, 138.0, 135.2, 133.3, 131.5, 130.2, 130.0, 124.4, 55.8, 47.3, 37.9, 21.3;

Synthesis of coumarin ester 2. 2-Methyl-2-phenylacetic acid (250 mg, 1.66 mmol, 1.0 eq.) was dissolved in dry acetonitrile (2.5 mL). Dicyclohexylcarbodiimide (DCC, 377 mg, 1.83 mmol, 1.1 eq.) was added and the mixture cooled to 0^0C in ice/water bath. 7-hydroxycoumarin (324 mg, 2.00 mmol, 1.2 eq.) was added and stirred 15 min at 0^0C. The cooling bath was then removed and the mixture stirred at room temperature for 65 h. Next, the reaction mixture was quenched with saturated aqueous $NaHCO_3$ solution (50 mL). Extraction with ethyl acetate (3x25 mL) and washing the combined organic fractions with saturated aqueous $NaHCO_3$ (4x50 mL), brine (2x50 mL) followed by drying with anhydrous Na_2SO_4 and evaporation of solvents yielded the crude product. It was purified using flash chromatography on silicagel (25 g SiO_2, elution with mixture of cyclohexane:ethyl acetate 10:1) resulting in 303 mg (62% yield) of the expected product.

TLC: R_f = 0.3 (developed twice in cyclohexane:ethyl acetate 3:1); **LCMS (ESI):** RT = 13.04 min., *m/z* calculated for $C_{18}H_{15}O_4$: 295.10 $[M+H]^+$, found: 295.04 $[M+H]^+$; **^{1}H NMR ($CDCl_3$) δ(ppm):** 7.65 (1H, d, *J* = 9.3 Hz), 7.43 (1H, d, *J* = 8.4 Hz), 7.38 (4H, m), 7.33 (1H, m), 7.02 (1H, d, *J* = 2.4 Hz), 6.94 (1H, dd, *J* = 8.4 Hz, *J'* = 2.1 Hz), 6.37 (1H, d, *J* = 9.6 Hz), 4.00 (1H, q^4, *J* = 7.2 Hz), 1.63 (3H, d, *J* = 7.2 Hz); **^{13}C NMR ($CDCl_3$) δ(ppm):** 172.4, 160.3, 154.7, 153.4, 142.8, 139.6, 129.0, 128.5, 127.7, 127.5, 118.3, 116.7, 116.1, 110.3, 45.7, 18.4;

Synthesis of para-nitrophenol ester 3. 2-Methyl-2-phenylacetic acid (250 mg, 1.66 mmol, 1.0 eq.) was dissolved in dry acetonitrile (2.5 mL). Dicyclohexylcarbodiimide (DCC, 377 mg, 1.83 mmol, 1.1 eq.) was added and the mixture cooled to 0^0C in ice/water bath. *p*-Nitrophenol (278 mg, 2.00 mmol, 1.2 eq.) was added and stirred 15 min at 0^0C. The cooling bath was then removed and the mixture stirred at room temperature for 65 h. Next, the reaction mixture was quenched with saturated aqueous $NaHCO_3$ solution (50 mL). Extraction with ethyl acetate (3x25 mL) and washing the combined organic fractions with saturated aqueous $NaHCO_3$ (4x50 mL), brine (2x50 mL) followed by drying with anhydrous Na_2SO_4 and evaporation of solvents yielded the crude product. It was purified using flash chromatography on silicagel (25 g SiO_2, elution with mixture of cyclohexane:ethyl acetate 14:1) resulting in 407 mg (90% yield) of the expected product.

TLC: R_f = 0.35 (cyclohexane:ethyl acetate 10:1); **^{1}H NMR ($CDCl_3$) δ(ppm):** 8.22 (2H, m), 7.40 (4H, m), 7.34 (1H, m), 7.18 (2H, m), 4.00 (1H, q^4, *J* = 7.2 Hz), 1.64 (3H, d, *J* = 7.2 Hz); **^{13}C NMR ($CDCl_3$) δ(ppm):** 172.2, 155.6, 145.3, 139.4, 129.0, 127.7, 127.5, 125.2, 122.3, 45.7, 18.4

Supporting figures and tables

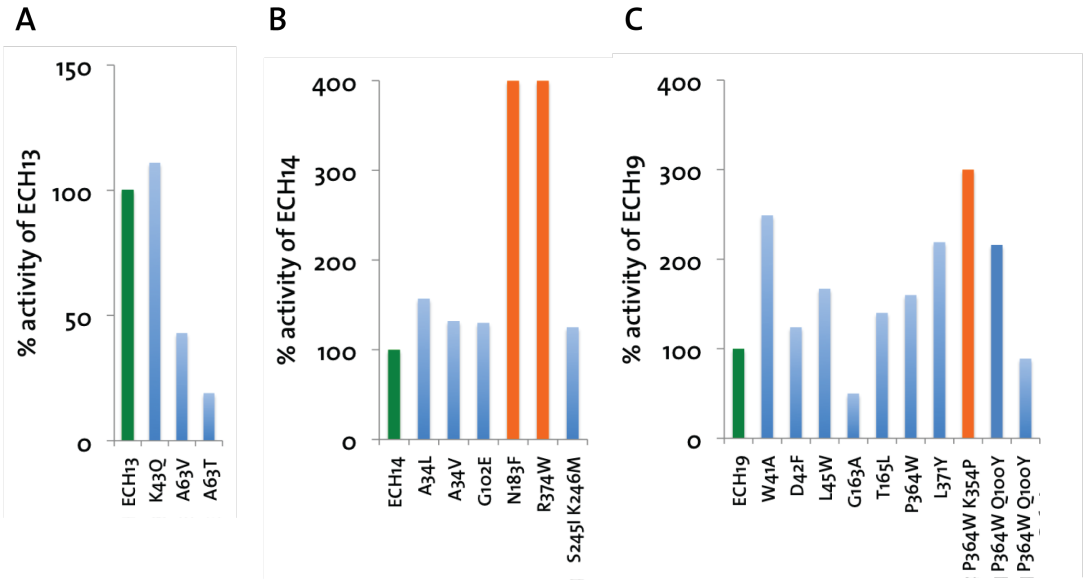


Figure 1S: Esterase activity of soluble **A)** ECH13, **B)** ECH14 and **C)** ECH19 variants. Coumarin release was measured in triplicate after addition of 100 µM substrate **2** to 5 µM of purified protein. Individual measurements deviated by approximately 5-15%. The plots show the relative activity of the variants compared to the parental design (green bar: parental enzyme, blue bars: 0 – 250 %, orange bars: 250 – 500 % of the parental design activity).

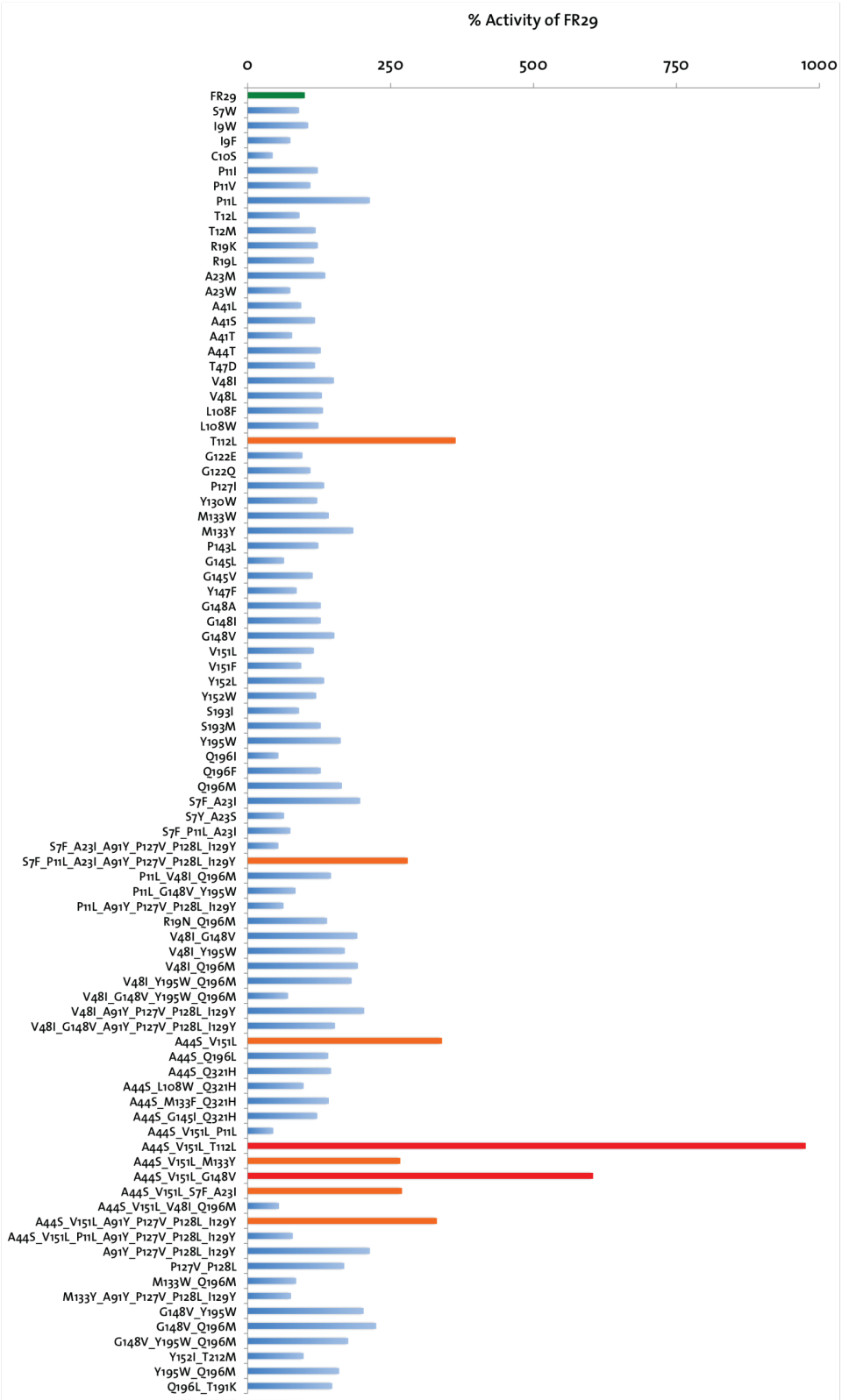

% Activity of FR29
0
250
500
750
1000
FR29
S7W
I9W
I9F
C10S
P11I
P11V
P11L
T12L
T12M
R19K
R19L
A23M
A23W
A41L
A41S
A41T
A44T
T47D
V48I
V48L
L108F
L108W
T112L
G122E
G122Q
P127I
Y130W
M133W
M133Y
P143L
G145L
G145V
Y147F
G148A
G148I
G148V
V151L
V151F
Y152L
Y152W
S193I
S193M
Y195W
Q196I
Q196F
Q196M
S7F_A23I
S7Y_A23S
S7F_P11L_A23I
S7F_A23I_A91Y_P127V_P128L_I129Y
S7F_P11L_A23I_A91Y_P127V_P128L_I129Y
P11L_V48I_Q196M
P11L_G148V_Y195W
P11L_A91Y_P127V_P128L_I129Y
R19N_Q196M
V48I_G148V
V48I_Y195W
V48I_Q196M
V48I_Y195W_Q196M
V48I_G148V_Y195W_Q196M
V48I_A91Y_P127V_P128L_I129Y
V48I_G148V_A91Y_P127V_P128L_I129Y
A44S_V151L
A44S_Q196L
A44S_Q321H
A44S_L108W_Q321H
A44S_M133F_Q321H
A44S_G145I_Q321H
A44S_V151L_P11L
A44S_V151L_T112L
A44S_V151L_M133Y
A44S_V151L_G148V
A44S_V151L_S7F_A23I
A44S_V151L_V48I_Q196M
A44S_V151L_A91Y_P127V_P128L_I129Y
A44S_V151L_P11L_A91Y_P127V_P128L_I129Y
A91Y_P127V_P128L_I129Y
P127V_P128L
M133W_Q196M
M133Y_A91Y_P127V_P128L_I129Y
G148V_Y195W
G148V_Q196M
G148V_Y195W_Q196M
Y152I_T212M
Y195W_Q196M
Q196L_T191K

Figure 2S: Esterase activity of soluble FR29 variants. Coumarin release was measured in triplicate after addition of 100 μM substrate **2** to 5 μM of purified protein. Individual measurements deviated by approximately 5-15%. The plot shows the relative activity of the individual variants compared to the parental design FR29 (green bar: parental design, blue bars: 0 – 250 %, orange bars: 250 – 500 %, red bars: > 500% of FR29 activity).

Table 1S: Overview of expressed designs. The theozyme numberings correspond to the theozymes as presented in Figure S10

Design	scaffold	# mutations to scaffold	soluble	Initial activity
Theozyme I				
ECH01	1eyn	6	yes	-
ECH02	1jcm	20	no	-
ECH03	1fp2	10	no	-
ECH04	1cil	12	yes	-
ECH05	1fp2	9	no	-
ECH06	1b4p	4	yes	-
ECH07	1ftx	17	yes	-
ECH08	1cjw	11	no	-
ECH09	1h1d	11	yes	-
ECH10	1dzu	11	yes	-
ECH11	1n9l	10	no	-
ECH12	1q11	14	yes	-
ECH13	1q91	9	yes	++
ECH14	1toi	13	yes	+
ECH15	1vhn	11	yes	-
ECH16	1vhn	11	yes	-
ECH17	1xjd	14	no	-
ECH18	2uvh	12	yes	-
ECH19	2uvh	11	yes	++
ECH20	4fua	14	no	-
FR25	1nah	14	yes	-
FR26	1qpr	13	yes	-
FR27	1pa9	11	no	-
FR28	1r5l	18	no	-
FR29	1mau	20	yes	+
FR30	1q91	21	yes	-
FR31	1q91	22	no	-
FR32	1cil	14	no	-
EA22	1eus	14	yes	-
EA29	1uyp	18	yes	-

EA30	2h13	16	no	-
Theozyme II				
EA23	1pt2	15	no	-
EA24	1v04	18	yes	-
EA27	1dzu	8	yes	-
EA34	2dri	19	no	-
EA35	1thf	15	yes	-
EA36	1is3	13	no	-
EA37	1dl3	13	no	-
EA28	1pii	13	yes	-
EA38	1c9u	31	no	-
EA39	1f5j	14	no	-
EA40	1st8	13	no	-
EA41	1yna	15	yes	-
Theozyme III				
1ajk_2	1ajk	15	yes	-
1ajk_3	1ajk	17	yes	-
1ukr_1	1ukr	20	no	-
1mac_3	1mac	14	yes	-
2ayh_1	2ayh	18	yes	-
1h0b_1	1h0b	18	yes	-
1dyp_1	1dyp	17	no	-
1mve_1	1mve	13	yes	-
1gbg_1	1gbg	15	yes	-
1f5j_1	1f5j	22	no	-
2jen_1	2jen	11	no	-
1mac_2	1mac	20	yes	-

Table 2S: Identification of the artificial ester hydrolases by ESI-MS.

Variant	$Mass_{calc}$ [Da]	$Mass_{exp}$ [Da]
FR29	38284.9	38284.7
C10A /H126A	38186.7	38186.0
A44S/T112L/V151L	38326.9	38326.5
ECH13	23836.1	23834.9
ECH13 C45A	23804.1	23803.3
ECH13 C45A/H100A	23738.0	23736.9
ECH14	44849.7	44851.6/44892.1
ECH14 C132A	44817.6	44817.1
ECH14 C132A/H104A	44751.5	44752.3
ECH19	47479.1	47479.7
ECH19 C161A	47447.1	47445.5
ECH19 C161A/H226A	47381.0	47379.7
ECH19 K354P/P364W	47537.2	47535.6

Table 3S: Melting temperatures of ester hydrolase designs were determined by curve fitting of the temperature dependent CD signals at 222nm with Sigma plot. Denaturation was irreversible in all cases.

ECH13		**ECH14**		**ECH19**		**FR29**	
Variant	[°C]	Variant	[°C]	Variant	[°C]	Variant	[°C]
wt-design	46	wt-design	42/50	wt-design	52	wt-design	71
C45A	45/49	C132A	47	C161A	52		
C45A/H100A	49	C132A/H104A	46	C161A/H226A	52	C10A/H126A	70
				K354P/P364W	47	A44S/T112L/V151L	69

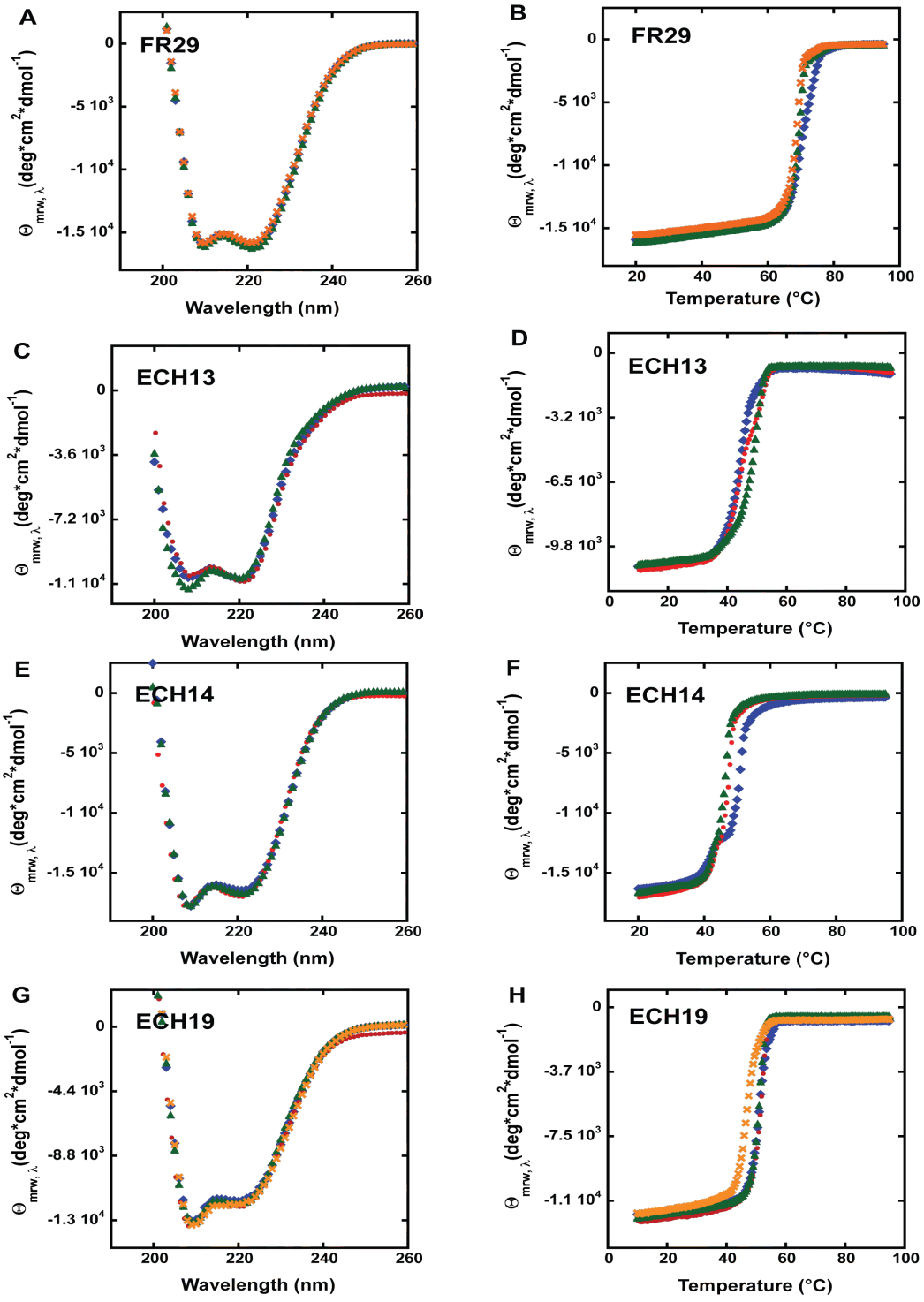


Figure 3S: CD spectra and melting curves of the *de novo* designed ester hydrolases, of their respective knockout mutants, and of the improved FR29 and ECH19 variants. The curves of the parental designs are depicted in blue, the traces of the single knockout variants (cysteine) are shown red, the traces of the double knockout variants (cysteine and histidine) are illustrated in green and the curves of the improved variants are highlighted in orange: **A-B)** FR29; **C-D)** ECH13; **E-F)** ECH14; **G-H)** ECH19.

Table 4S: Results of the mass spectrometric analysis of the *in silico* designs incubated with coumarin ester **2**.

	Before Incubation		After Incubation		
Design	Mass (Da) (ESI)		No. of modifications (ESI)	Location of modification (MALDI –MS/MS)	
	exp	calc		Peptide	Residue
FR29	38286.8	38284.9	1		
FR29 C10A/H126A	38187.8	38186.7	0		
FR29 A44S/V151L/M133Y	38346.3	38346.9	2	TIFSAICPTGVITIGR	Cys10 (AS)
ECH13	23837.1	23836.1	2	ACEQYGR RPCGSLEHHHHH	Cys45 (AS) Cys195
ECH13 C45A/H100A	23737.2	23738.0	n.d.		
ECH14	44851.6/ 44892.1	44849.7	7	NFGLYNESVGACT RGSAVAYVGFEERL TAQPGGHGALR (AS-His)* VWVYNPSSNCSK (AS-Cys)*	
ECH14 C132A/H104A	44752.3	44751.5	7		
ECH19	47480.0/ 47512.4	47479.1	1-2	MSWWGGNGR* AEYETGWDGHLSR* TVQETAEYFNKQGD*	
ECH19 C161A/H100A	47381.9	47381.0	0		

* Due to the low intensity of these peptides, the interpretations are speculative (AS stands for active site).

Table 5S: Results of the mass-spectroscopic analysis of the *in silico* designs after incubation with tyrosine substrate **3**.

	Before Incubation		After Incubation
Design	Mass (Da) (ESI)		No. of modifications (ESI)
	exp	calc	
FR29 A44S/T112L/V151L	38326.5	38326.9	0
ECH19 K354P/P364W	47537.2	47535.6	0

Table 6S: Analysis of FR29 variants.

Mutation(s)	$Mass_{calc}$ [g/mol]	$Mass_{exp}$ [g/mol]	Δ	ε [M^{-1} cm^{-1}]	Activity (% FR29)
S7W	38384.9			47330	90
I9W	38357.9			47330	106
I9F	38318.9			41830	75
C10S	38268.8	38268.2	0.6	41830	44
P11I	38300.9	38300.4	0.5	41830	123
P11V	38286.9			41830	110
P11L	38300.9			41830	214
T12L	38296.9			41830	91
T12M	38314.9	38314.6	0.3	41830	119
R19K	38256.8			41830	123
R19L	38241.8	38241.5	0.3	41830	116
A23M	38345.0	38344.9	0.1	41830	136
A23W	38400.0			47330	75
A41L	38326.9			41830	94
A41S	38300.9	38301.3	0.4	41830	118
A41T	38314.9	38315.4	0.5	41830	78
A44T	38314.9	38314.5	0.4	41830	128
T47D	38298.8	38298.4	0.4	41830	118
V48I	38298.9	38299.4	0.5	41830	151
V48L	38298.9	38299.7	0.8	41830	130
L108F	38318.9	38319.7	0.8	41830	132
L108W	38357.9	38357.4	0.5	47330	124
T112L	38296.9	38296.3	0.3	41830	364

G122E	38356.9	38357.7	0.8	41830	96
G122Q	38355.9	38355.0	0.9	41830	110

Mutation(s)	$Mass_{calc}$ [g/mol]	$Mass_{exp}$ [g/mol]	Δ	ε [M^{-1} cm^{-1}]	Activity (% FR29)
P127I	38300.9	38301.5	0.6	41830	134
Y130W	38307.9	38308.8	0.9	45840	122
M133W	38339.9	38340.6	0.7	47330	142
M133Y	38316.8	38315.2	1.6	43320	185
P143L	38300.9	38299.9	1.0	41830	124
G145L	38341.0	38341.0	0	41830	64
G145V	38326.9	38327.1	0.2	41830	114
Y147F	38268.9	38267.6	1.3	40340	86
G148A	38298.9	38299.8	0.9	41830	128
G148I	38341.0	38342.1	1.1	41830	128
G148V	38326.9	38326.3	0.6	41830	152
V151L	38298.9	38298.2	0.7	41830	116
V151F	38332.9	38333.6	0.7	41830	94
Y152L	38234.8	38235.2	0.4	40340	134
Y152W	38307.9	38308.6	0.7	45840	120
S193I	38310.9	38309.4	1.5	41830	90
S193M	38329.0	38327.9	1.1	41830	128
Y195W	38307.9	38308.5	0.6	45840	163
Q196I	38269.9	38269.5	0.4	41830	54
Q196F	38303.9	38303.7	0.2	41830	128
Q196M	38287.9	38287.0	0.9	41830	165
S7F_A23I	38387.0	38385.9	1.1	41830	197

S7Y_A23S	38377.0	38376.4	0.6	43320	64
S7F_P11L_A23I	38403.1	38401.8	1.3	41830	75
S7F_A23I_A91Y_P127V_P128L_I129Y	38547.2	38546.8	0.4	44810	54
S7F_P11L_A23I_A91Y_P127V_P128L_I129Y	38563.2	38366.7	3.5	44810	280
P11L_V48I_Q196M	38318.0	38317.2	0.8	41830	146

Mutation(s)	**$Mass_{calc}$ [g/mol]**	**$Mass_{exp}$ [g/mol]**	**Δ**	**ε [M^{-1} cm^{-1}]**	**Activity (% FR29)**
P11L_G148V_Y195W	38366.0	38365.2	0.8	45840	84
P11L_A91Y_P127V_P128L_I129Y	38461.1	38459.5	1.6	44810	63
R19N_Q196M	38245.8	38244.7	1.1	41830	139
V48I_G148V	38341.0	38340.5	0.5	41830	192
V48I_Y195W	38321.9	38320.9	1.0	45840	170
V48I_Q196M	38301.9	38301.5	0.4	41830	193
V48I_Y195W_Q196M	38325.0	38323.6	1.4	45840	182
V48I_G148V_Y195W_Q196M	38367.1	38365.4	1.7	45840	71
V48I_A91Y_P127V_P128L_I129Y	38459.1	38457.7	1.4	44810	204
V48I_G148V_A91Y_P127V_P128L_I129Y	38501.1	38500.6	0.5	44810	153
A44S_V151L	38314.9	38314.4	0.5	41830	340
A44S_Q196L	38285.9	38285.4	0.5	41830	141
A44S_Q321H	38309.9	38310.6	0.7	41830	146
A44S_L108W _Q321H	38383.8	38383.4	0.4	47330	98
A44S_M133F_Q321H	38325.8	38324.9	0.9	41830	142
A44S_G145I_Q321H	38366.0	38364.7	1.3	41830	122
A44S_V151L_P11L	38330.9	38330.6	0.3	41830	45
A44S_V151L_T112L	38326.9	38326.5	0.4	41830	976
A44S_V151L_M133Y	38346.9	38346.6	0.3	43320	267

A44S_V151L_G148V	38357.0	38356.5	0.5	41830	604
A44S_V151L_S7F_A23I	37417.1	38416.9	0.2	41830	270
A44S_V151L_V48I_Q196M	38332.0	38332.0	0	41830	55
A44S_V151L_A91Y_P127V_P128L_I129Y	38475.1	38474.0	1.1	44810	331
A44S_V151L_P11L_A91Y_P127V_P128L_I129Y	38491.1	38490.9	0.2	44810	79
A91Y_P127V_P128L_I129Y	38445.0	38444.8	0.2	44810	214
P127V_P128L	38302.9	38301.6	1.3	41830	169
M133W_Q196M	38342.9	38342.4	0.5	47330	85

Mutation(s)	**$Mass_{calc}$ [g/mol]**	**$Mass_{exp}$ [g/mol]**	**Δ**	**ε [M^{-1} cm^{-1}]**	**Activity (% FR29)**
M133Y_A91Y_P127V_P128L_I129Y	38477.0	38477.0	0	46300	76
G148V_Y195W	38350.0	38349.5	0.5	45840	203
G148V_Q196M	38330.0	38329.6	0.4	41830	225
G148V_Y195W_Q196M	38353.0	38352.6	0.4	45840	176
Y152I_T212M	38264.9	38265.3	0.4	40340	98
Y195W_Q196M	38311.0	38309.9	1.1	45840	160
Q196L_T191K	38297.0	38297.9	0.9	41830	148

Table 7S: Analysis of ECH13 variants.

Mutation(s)	Soluble (y/n)	$Mass_{calc}$ [g/mol]	$Mass_{exp}$ [g/mol]	Δ	ε [M^{-1} cm^{-1}]	Activity (% ECH13)
D10W	n	23907.2			36440	
V16I	n	23850.1			30940	
E17M	n	23838.2			30940	
E17W	n	23893.2			36440	
K43Q	y	23836.1	23834.8	1.3	30940	111
A63V	y	23864.2	23864.0	0.2	30940	43
A63T	y	23866.1	23865.9	0.2	30940	19
F68W	n	23875.1			36440	

Table 8S: Analysis of ECH14 variants.

Mutation(s)	Soluble (y/n)	$Mass_{calc}$ [g/mol]	$Mass_{exp}$ [g/mol]	Δ	ε [M^{-1} cm^{-1}]	Activity (% ECH14)
A34L	y	44891.7	44892.6/ 44932.6	0.9/ 40.9	42860	157
A34V	y	44877.7	44877.2/ 44919.4	0.5/ 41.7	42860	132
G102E	y	44921.7	44921.8/ 44961.1	0.1/ 39.4	42860	130
N183F	y	44882.7	44881.0/ 44924.9	1.7 42.2	42860	400
R374W	y	44879.7	44879.5/ 44921.2	0.2/ 41.5	48360	400
S245I K246M	y	44878.8	44879.3	0.5	42860	125

The mass spectra of the ECH14 variants always exhibited a second peak of approximately +42 Da corresponding to acetylation of the proteins.

Table 9S: Analysis of ECH19 variants.

Mutation(s)	Soluble (y/n)	$Mass_{calc}$ [g/mol]	$Mass_{exp}$ [g/mol]	Δ	ε [M^{-1} cm^{-1}]	Activity (% ECH19)
W41A	y	47364.0	47364.0	0	90300	249
D42F	y	47511.2	47511.2	0	95800	124
L45W	y	47552.2	47552.2	0	101300	167
G163A	y	47493.2	47479.7	13.5	95800	50
T165L	y	47491.2	47490.1	1.1	95800	140
T215W	n	47564.3			101300	
P364W	y	47568.2	47567.4	0.8	101300	160
L371Y	y	47529.2	47528.5	0.7	97290	219
P364W K354P	y	47537.2	47535.6	1.6	101300	300
P364W Q100Y	y	47603.3	47601.8	1.5	101300	216
P364W Q100Y G163A	y	47617.3	47616.2	1.1	102790	89

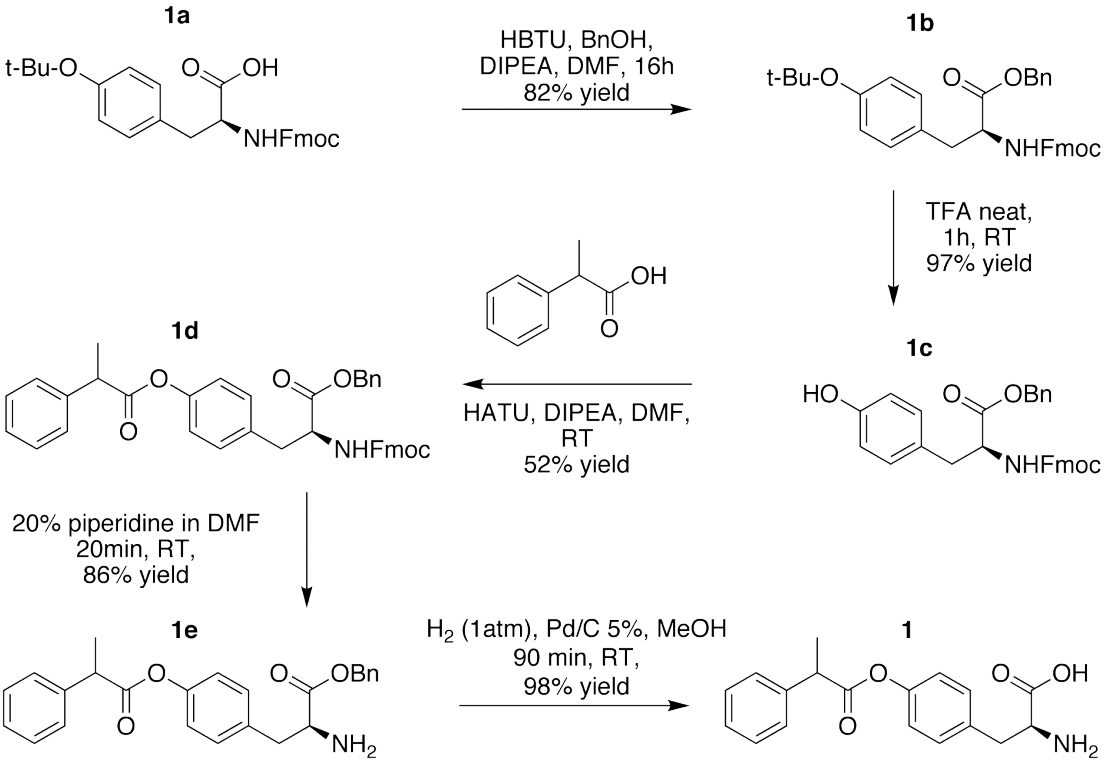


Figure 4S: Synthesis scheme of tyrosine ester **1.**

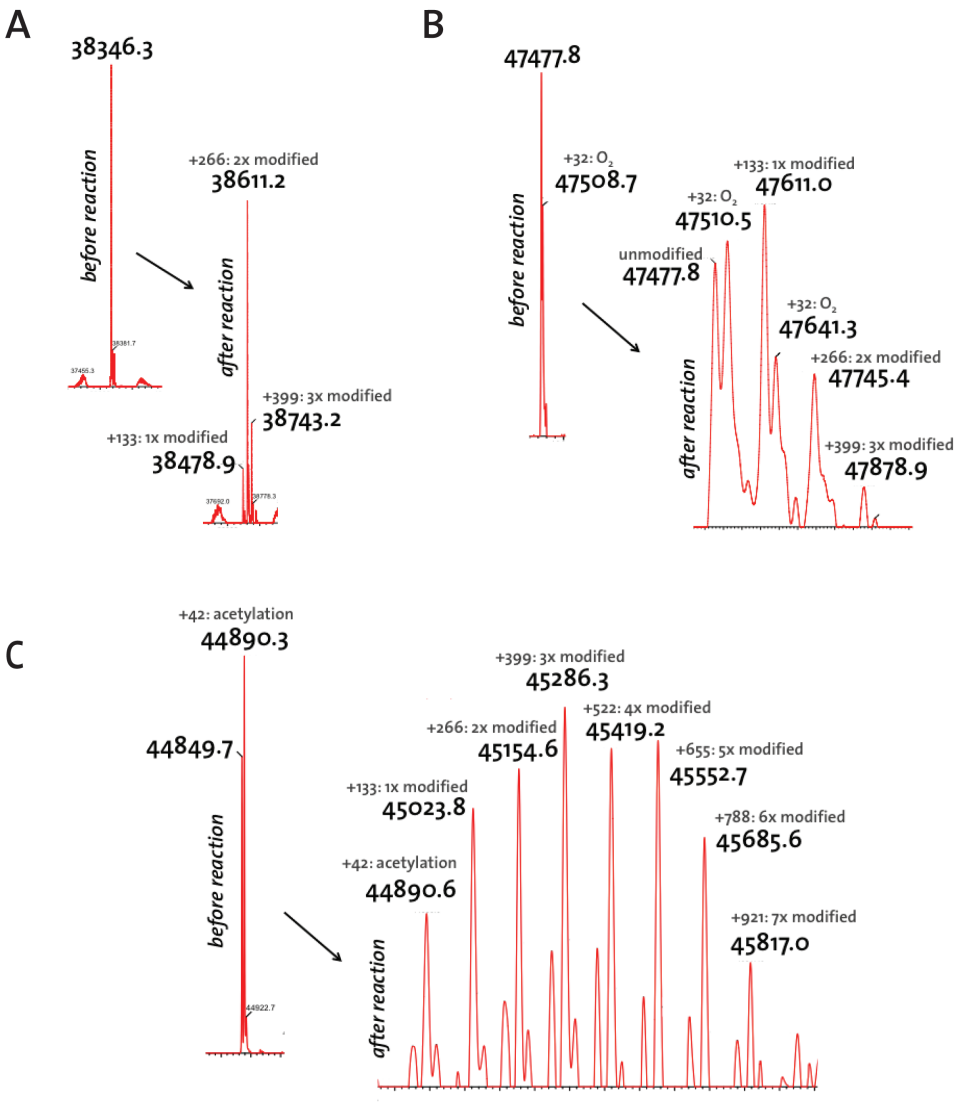


Figure 5S: Exemplary mass spectra before and after the addition of the coumarin ester 1 to **A)** FR29 A44S M133Y V151L, **B)** ECH19 and **C)** ECH 14.

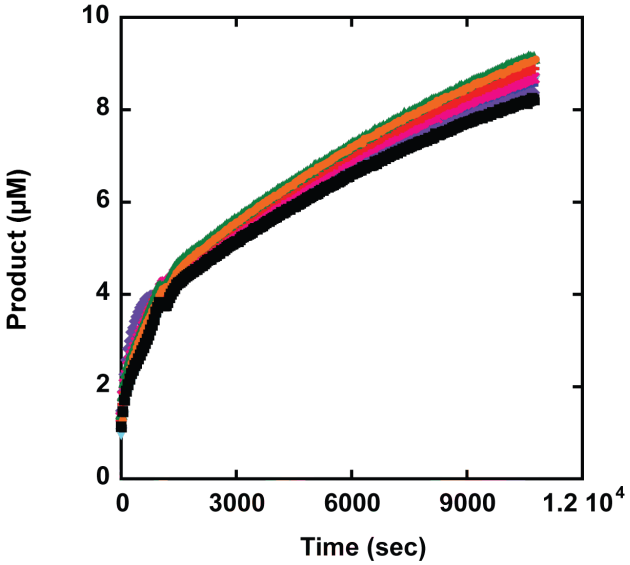


Figure 6S: Progress curve of the conversion of coumarin ester **2** by a FR29 variant in the presence of small nucleophiles. FR29 A44S/T112L/V151L (5 µM) was incubated with 50 µM of methoxyamine (cyan, pK_a = 4.6), aniline (blue; pK_a = 4.6), 2-aminobenzoic acid (purple; pK_a = 2.0), 4-aminobenzoic acid (pink; pK_a = 2.3), (*R*)-1-phenylethanamine (red), (*S*)-1-phenylethanamine (green), benzylthiol (orange) and no nucleophile (black) before 50 µM of coumarin ester **2** was added to the reaction mixture. The coumarin release was monitored in a plate reader at 29°C and pH 7.5. For all progress curves the corresponding background reactions without enzyme were subtracted. In the case of hydroxylamine the background reaction became dominating (data not shown).

Table 10S: X-ray crystallography and structure refinement statistics

General information				
PDB id	3U13	3U1O	3U1V	3UAK
NESG id	OR51	OR49	OR52	OR54
Gene	ECH13	ECH19	FR29	ECH14
Type	apo	apo	apo	apo
Crystal parameters				
Space group	$P4_32_12$	$P2_12_12$	$P2_12_12_1$	$P2_12_12_1$
a (Å)	73.44	109.14	97.75	67.65
b (Å)	73.44	129.20	100.76	81.81
c (Å)	105.08	72.18	188.23	159.71
α (°)	90	90	90	90
β (°)	90	90	90	90
γ (°)	90	90	90	90
Z (number mols./au)	1	2	4	2
Data quality				
Beam line/X-ray source	BL9-2	X4C	X4C	X4C
Resolution range (Å)	30-1.6	30-2.6	30-2.8	30-3.2
Total reflections	2254145	394063	600893	375667
Observed reflections	72527	36229	42465	26157
R_{merge}	0.05/0.054	0.146/0.433	0.124/0.417	0.122/0.645
Mean redundancy	18.0/17.1	4.9/4.9	2.5/2.4	3.1/2.6
Completenes (%)	100.0/100.0	99.5/100.0	90/9/63.8	84.8/80.7
<I>/<σI>	48.7/4.1	11.4/3.5	8.0/1.8	8.9/1.7
Refinement				
Resolution range (Å)	30-1.6	30-2.5	30-2.8	30-3.2
Number of reflections	36618	36117	42365	26157
R_{work}	0.178	0.203	0.214	0.210
R_{free}	0.195	0.258	0.290	0.287
Number of prot. atoms	1642	6583	9937	6146
Number of waters	248	441		
Number of ligand atoms				
Overall mean B factors:				
-protein	24.2	19.3	48.4	74.2
-water	36.5	23.4	28.1	
-ligand				
RMSD bond length (Å)	0.008	0.008	0.009	0.009
RMSD bond angles (o)	1.21	1.18	1.17	1.20
Ramachandran plot				
-most favoured (%)	91.8	93.8	88.3	85.0
-additional allowed (%)	7.6	5.8	11.1	13.7
-generously allowed (%)	0.6	0.3	0.6	1.3
-disallowed (%)	0	0.1	0	0

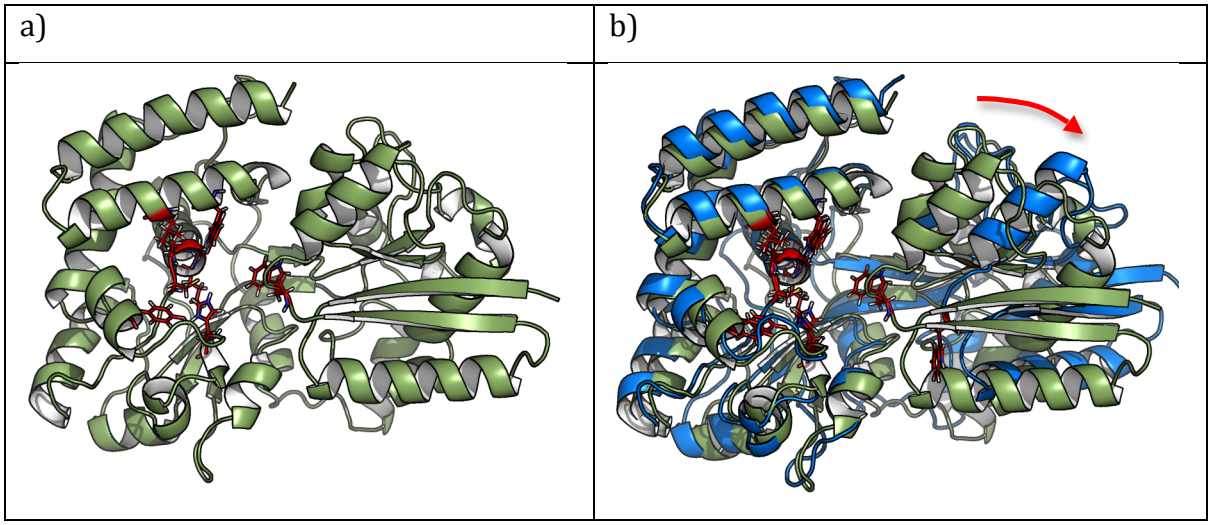


Figure S7. Molecular Dynamics of ECH19 in cartoon representations with the active site residues highlighted (red sticks). (a) The computational design. (b) The equilibrated MD structure superimposed onto the computational design in (a). The backbone RMSD is 3.5 Å.

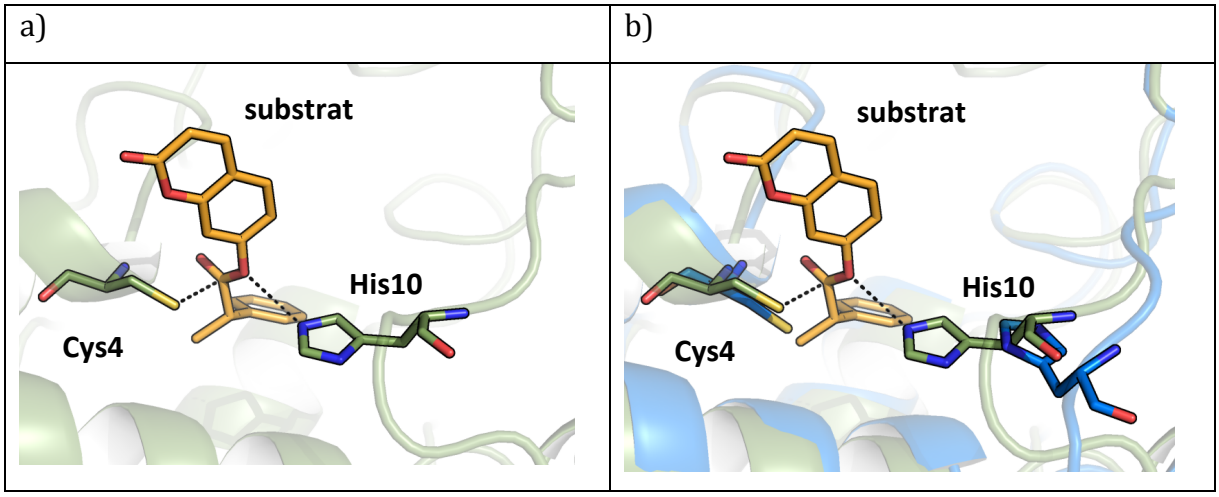


Figure S8. Molecular Dynamics of ECH13 (active site shown). (a) Computational design with the docked substrate in orange. (b) MD (blue) over computational design: His100 is part of a flexible loop and does not remain pre-oriented at the designed position.

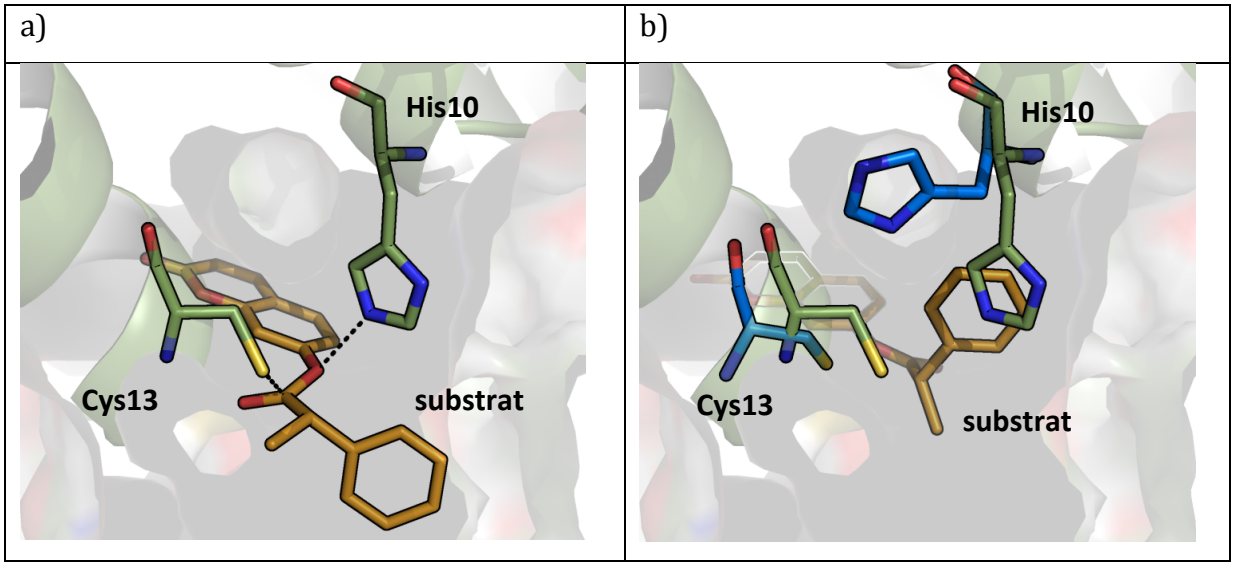
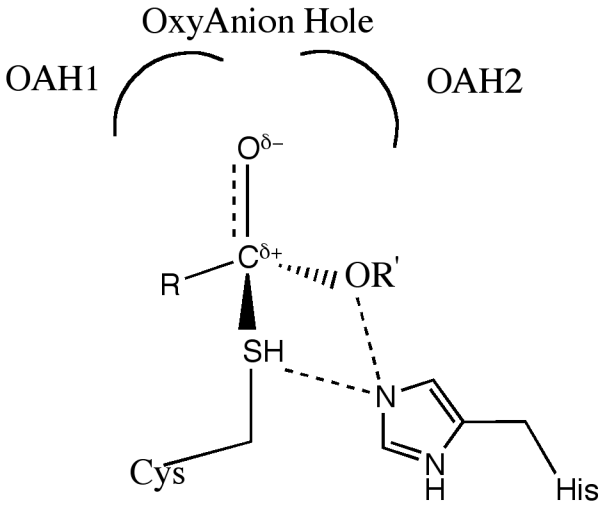


Figure S9. Molecular Dynamics of ECH14 (active site shown). (a) Computational design with the docked substrate in orange. (b) MD (blue) over computational design: His104 occupies a catalytically incompetent conformation.



Theozyme I: OAH1 = backbone-NH OAH2 = none

TheozymeII: OAH1 = sidechain Hbond donor OAH2 = sidechain Hbond donor

Theozyme III: OAH1 = explicit H_2O OAH2 = explicit H_2O/sidechain Hbond donor

Figure S10: Schematic representation of the three theozymes used during the matching stage of the computational design process. A sidechain Hbond donor could be any one of the following atoms: Ser/Thr-Og; Tyr-OH; Asn-Nd; Gln-Ne; His-Nd; His-Ne;

Vita

Florian Richter was born in Dortmund, Germany in November 1981. After attending High-School in Bonn, Germany and Colorado Springs, Co and graduating in 2000, he enrolled at the Swiss Federal Institute of Technology in Zurich for undergraduate studies, and obtained a Master's degree in biochemistry in 2006. After that, he moved to Seattle, Wa to pursue a doctoral degreee in biochemistry.