

Computational Design and Directed Evolution of Novel Enzymes

Matthew D. Smith

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor Of Philosophy

University of Washington

2013

Reading Committee:

David Baker, Chair

Dustin J. Maly

Forrest Michael

Program Authorized to Offer Degree:

Molecular and Cellular Biology

©Copyright 2013

Matthew D. Smith

University of Washington

Abstract

Computational Design and Directed Evolution of Novel Enzymes

Matthew D. Smith

Chair of the Supervisory Committee:

Professor David Baker

Biochemistry

Enzymes are the most specific and active catalysts found in nature, and offer unique chemical properties suited to solving important industrial, medical, and environmental problems. In this work I present new methods for designing and evolving enzymes. In Chapter 1, I describe the theory and method of enzyme design, using the example of designing an enzyme for the Diels-Alder cycloaddition. In Chapter 2, I present the improvement of a computationally designed ester hydrolase enzyme through directed evolution and rational design. This work illustrates the insights that can be gained by performing directed evolution on computationally designed enzymes. In Chapter 3, I present the computational design and experimental characterization of novel alkyltransferase proteins. These proteins are particularly amenable to directed evolution, through which insight can be gained into possible improvements to future design efforts.

Table of Contents

List of Figures	vi
List of Tables.....	viii
Glossary	ix
Acknowledgements	x
Dedication	xi
Chapter 1: Theory and Method of Computational Enzyme Design	1
Introduction.....	1
Preparation of Enzyme Specification and Scaffold Set	4
Matching.....	7
Post-design Filtering.....	13
Conclusion	15
Notes.....	15
Chapter 2: Directed evolution and rational design of a computationally designed ester hydrolase.....	21
Introduction.....	21
A yeast display selection for improved acyl-enzyme intermediate formation	23
Evolved variants are improved catalysts	27
Rational design to improve catalytic turnover	31
Discussion.....	33
Chapter 3: Computational Design and Directed Evolution of Novel Alkyltransferases	59
Introduction.....	59

Computational design.....	60
Screening for alkyltransferase activity	62
Directed evolution through yeast display	63
Discussion.....	64
Bibliography.....	78

List of Figures

Figure 1.1: Substrates for Diels-Alder enzyme design project.....	17
Figure 1.2: Theozyme structure for the Diels-Alder reaction	18
Figure 1.3: Example of a theozyme interaction.....	19
Figure 1.4: Progress of design from theozyme, to match, to final design.....	20
Figure 2.1: Selection strategy to optimize an ester hydrolase.	36
Figure 2.2: Improvement in labeling on yeast.....	37
Figure 2.3: Michaelis-Menten kinetics show improvement in K_{cat} and K_m for evolved variants..	38
Figure 2.4: A rationally designed point mutant accelerates turnover.....	39
Figure 2.5: The five step predicted mechanism for ester hydrolysis.....	40
Figure 2.6: Library creation and selection scheme to improve ester hydrolase acylation.....	41
Figure 2.7: Location of regions randomized to create six doped libraries.	42
Figure 2.8: Improved acylation is dependent on active site cysteine.	45
Figure 2.9: Combination library (CL) variants show convergence on four sequences.	46
Figure 2.10: Areas targeted for error-prone PCR mutagenesis.	47
Figure 2.11: Initial, rough measurement of enzyme activity with free-ester	49
Figure 2.12: Mass spectrometry confirms hypothesized mechanism.....	50
Figure 2.14: Protein melt shows only small difference in stability	51

Figure 2.15: Three mutations contribute majority of improvement for variant B2A.....	53
Figure 2.16: One mutation patch contributes the majority of improvement for variant 5A2D.....	54
Figure 2.17: Mutations from B2A improve ECH19 KPPW activity only at high [substrate].....	55
Figure 2.18: Mutated position adjacent to catalytic cysteine is near optimal.....	56
Figure 2.19: Michaelis-Menten kinetics may show further improvement	58
Figure 3.1: Quenched and activated substrates for alkyltransferase design	66
Figure 3.2: Cysteine-histidine-acid catalytic triad and proposed mechanism	67
Figure 3.3: Initial <i>de novo</i> designed proteins purify solubly	68
Figure 3.4: <i>De novo</i> designed and repurposed triad proteins purify solubly.....	69
Figure 3.5: Design batch 1: <i>De novo</i> designed and repurposed triad proteins.	70
Figure 3.6: Design batch 2: <i>De novo</i> designed and repurposed triad proteins	71
Figure 3.7: Repeat of purification and test confirms activity of best designed variants.	74
Figure 3.8: Cysteine and histidine dyad knockouts confirm mechanism	75
Figure 3.9: Library construction strategy for alkyltransferase selection.....	76
Figure 3.10: Regions randomized in three designed alkyltransferases.	77

List of Tables

Table 2.1: Sequences and activities for clones from DL selection.....	43
Table 2.2: Template genes and randomized regions used to construct error-prone libraries	48
Table 2.3: Kinetic parameters for mutations on evolved esterase variants.	52
Table 2.4: Activity of variants with combinations of mutations from enzyme variants	57
Table 3.1: Identities of batch 1 designs.	72
Table 3.2: Identities of batch 2 designs.	73

Glossary

ROSETTA: Software for predicting and designing protein structure and function.

SDS: Sodium Dodecyl Sulfate

PAGE: Polyacrylamide Gel Electrophoresis

ORF: Open Reading Frame

Acknowledgements

First, I'd like to thank my advisor David Baker, whose support, guidance, expert knowledge, and patience made my PhD and this work possible. I would also like to thank my thesis committee:

Barry Stoddard, Michael Gelb, Dustin J. Maly, Forrest Michael, and Eric Klavins. Their guidance has played a large role in shaping and directing this research. Thanks also to the members of my reading committee: David Baker, Dustin J. Maly, and Forrest Michael.

Many thanks to all the current and past members of the Baker Lab, whose support, advice, and mentorship helped me enormously during my PhD research; especially Eric Althoff, Sagar Khare, Jeremy Mills, Justin Siegel, Florian Richter, Summer Thyme, Cassie Bryan, Eva-Maria Strauch, Liz Kellogg, and Olga Khersonsky. Thanks also to the Baker lab information technology administrators, Darwin O.V. Alonso and Keith Laidig, for their tireless work to keep the lab's computing fast and reliable. I'd also like to thank Michelle Scalley-Kim and Barabara Lovseth for administrative support; and Jasmine Gallaher for her incredible efforts keeping the Baker Lab efficient, effective, and well-supplied. Many thanks also to all the staff of the UW Biochemistry and Molecular and Cell Biology offices, especially MaryEllin Robinson and Diane Darling.

Thanks also to my coauthors. Chapter 1 coauthors are Alexandre Zanghellini and Daniela Grabs.

Chapter 2 coauthors are Cassie Bryan, Alex Kuzin, Zbigniew Lech, Gert Kiss, Scott

Lew, Florian Richter, Rochelle Nguyen, Greg Kornhaber, Liang Tong, Vijay Pande, Don Hilvert, and David Baker. Chapter 3 coauthors are Olga Khersonsky, Luc Reymond, Alberto Schena, Kai Johnsson, and David Baker.

Dedication

To my parents, Don and Michele Smith, for fostering in me a love of science and the belief that I can do extraordinary things; and also to Summer Thyme, without whose advice and support this work would not have been completed.

Chapter 1: Theory and method of computational enzyme design

In press in the Humana Press book series “Methods in Molecular Biology” as:

Computational design of novel enzymes without cofactors

Matthew D. Smith, Alexandre Zanghellini, Daniela Grabs

Chapter 2: Computational design of novel alkyltransferase enzymes

This chapter has not been published previously.

Chapter 3: Directed evolution and rational design of a computationally designed ester hydrolase

This chapter has not been published previously.

Chapter 1: Theory and Method of Computational Enzyme Design

Introduction

Computational design of novel protein catalysts is a truly interdisciplinary endeavor that brings together the fields of biochemistry, chemistry, biophysics and the power of computational methods. The design of enzymes for arbitrary chemical reactions has the potential to greatly impact many fields and industries by allowing the creation of valuable chemicals and breakdown of pollutants or toxins. Furthermore, missing links in biochemical pathways could be filled in, bringing us closer to the dream of the cell as a customizable, miniature chemical factory. I review here a recent method for developing novel protein catalysts. This method has been successfully applied to the design of novel catalysts for the retro-aldol cleavage reaction (**1**), the Kemp elimination reaction (**2**), ester hydrolysis (**3**), and the example I review here, the Diels-Alder 2+4 cycloaddition reaction (**4**).

In unpublished work presented in this thesis, this method has also been studied through directed evolution of a de novo computationally designed ester hydrolase (Chapter 2) and applied to design novel alkyltransferase proteins (Chapter 3).

Before starting the computations to design a novel enzyme, there are several points to consider. First, the ease of enzyme design can vary greatly for different chemical reactions, but it is often difficult to determine this in advance as the current methods (such as QM/MM) to assess

Reproduced in part with permission from Humana Press publication Methods in Molecular Biology, in press under the chapter title “Computational design of novel enzymes without cofactors”.

chemical reactivity *in silico* have shown moderate accuracy and are extremely computationally intensive. The enzyme design strategy presented here relies primarily on the preferential binding and stabilization of the transition state(s) of the reaction by providing hydrogen bond donors or acceptors, stabilizing charges, and/or pre-organizing the orientation of the substrate(s). The enzyme thereby reduces the energy barrier and catalyses the reaction. While we see a great range of uncatalyzed rates for the substrates of natural enzymes, we work under the assumption that the higher the activation energy of a reaction, the more difficult it will be to create an enzyme for it. Additionally, it has been found empirically that design is more successful when applied to substrates that contain both hydrophobic and hydrogen-bonding groups.

Second, the readout of the enzymatic reaction and the sensitivity of the assay must be considered. In the successful application of *de novo* enzyme design I present here, a variety of readouts—colorimetry, fluorimetry, as well as mass spectrometry have been used. These methods vary in their ease and speed, which affects the throughput and extent of downstream characterization. It is expected that initial computationally designed lead enzymes will exhibit low activities. In order to detect these low activities, assays that can measure low turnover numbers or product formation in the low μM range are preferred. Figure 1.1 depicts the substrates, transition state, and product for the Diels-Alder reaction, the example reaction I chose to illustrate computational enzyme design. In this case, there was no colorimetric or fluorimetric readout, so a liquid chromatography assay coupled to a mass spectrometry (for greater sensitivity) was used to measure both the formation and the stereochemistry of the reaction products.

Third, a critical step in enzyme design is to determine the transition states of the reaction and the core catalytic machinery that can catalyze the reaction. In the case of reactions with multiple

transition states, such as the retro-aldol reaction (*1*), one typically picks the transition state with the highest activation-energy barrier, although building “consensus” transition state models is a possibility (*1*). One way to elucidate and model transition states is through careful study of the literature on the reaction of interest and/or obtaining transition state structures from computational chemistry experiments. Once a reasonable transition state model has been made, the next challenge is to determine the arrangement of amino acid side chains around the transition state to best catalyze the reaction of interest (this arrangement of amino acid sidechains and transition state is called a theozyme) (*5*). We have had success working with quantum chemists, studying transition state models and enzyme mechanisms, and applying general quantum chemistry principles (stabilization/destabilization of orbitals of interest as they evolve along the reaction path and stabilization of charges as they develop in the transition state) as well as general principles from the vast literature on enzyme mechanisms in nature. In cases where small molecule catalysts have been found for the reaction of interest, their structures may be used to guide theozyme design. In more than one example from our work designing enzymes, we have seen it necessary to try multiple theozymes, as no activity was seen for the first models (*1, 3*).

For the Diels-Alder reaction in general and the particular example I review here (Figure 1.1), much was already known about the mechanism of catalysis (*6, 7*). There are many small molecule catalysts for the Diels-Alder reaction (*8*), and extensive mechanistic studies have elucidated relatively confident models of the orbital dynamics of the reaction. In addition, catalytic antibodies had been selected which carried out this Diels-Alder reaction (*9*). In this case, we looked to frontier molecular orbital theory to guide the active site design. Transition state stabilization can be achieved by raising the energy of the HOMO (highest occupied molecular orbital) and lowering the energy of the LUMO (lowest unoccupied molecular orbital),

thereby narrowing the energy gap between the two orbitals and reducing the free energy of activation. The energy of the HOMO is raised by positioning a hydrogen bond acceptor (such as a carbonyl group of glutamine or asparagine) to stabilize the developing positive charge on the carbamate NH of the diene, and the energy of the LUMO is lowered by positioning a hydrogen bond donor (such as serine, threonine or tyrosine) to stabilize the developing negative charge on the carbonyl of the dienophile. Quantum mechanical (QM) calculations were carried out to determine the geometry of the lowest free energy barrier transition state between substrates and product in the presence of these hydrogen bonding groups (4) which resulted in a three dimensional representation of the theozyme, depicted in Figure 1.2.

Once the initial development of a sensitive assay and the calculation of an optimal theozyme have been completed, one can then start the Rosetta enzyme design protocol which is divided into four stages:

1. Preparation of enzyme specification and scaffold set
2. Matching of the theozyme into the scaffold set
3. Design of the active site
4. Post-design filtering

Preparation of Enzyme Specification and Scaffold Set

To describe the theozyme we require a specification of the geometric relationships between the desired active site residues and the transition state. This specification is called a constraint file. From the theozyme structure of the Diels-Alder reaction (Figure 1.2), it is easy and convenient to extract the geometry of interactions between all catalytic sidechains (glutamine and tyrosine in this case) and the transition state. An important point to consider is the conformational

flexibility in the transition state structure. Rotatable bonds should be “sampled” in the transition state structure to allow more variability in compatible active sites (see Supplementary material in (4), and Supplementary material in (1) for examples of the ensemble used for matching). For Rosetta enzyme design, the convention is to name the ensemble of transition state structures the **downstream partner** and each catalytic sidechain in the theozyme as the **upstream partner**. Their interaction geometries can be uniquely specified by the six degrees of freedom between three atoms of either partner (Figure 1.3).

A constraint file contains one block (starting with CST::BEGIN and ending with CST::END) for each catalytic sidechain. For the Diels-Alder reaction we would need two such blocks (one for glutamine and one for tyrosine) to make up the full theozyme definition (note that the text in curly brackets throughout the chapter indicates comments and should not be included in the actual files):

CST::BEGIN

TEMPLATE:: ATOM_MAP: 1 atom_name: N1 C8 O2 {downstream partner}

TEMPLATE:: ATOM_MAP: 1 residue3: TRS

TEMPLATE:: ATOM_MAP: 2 atom_name: OE1 CD CG {upstream partner}

TEMPLATE:: ATOM_MAP: 2 residue3: GLN

CONSTRAINT:: distanceAB: 2.80 0.20 100. 0 2

CONSTRAINT:: angle_A: 113.5 10.0 10 360.2

CONSTRAINT:: angle_B: 120. 10.0 10 360.1

```
CONSTRAINT:: torsion_A: 0. 10.0 10 360.2
```

```
CONSTRAINT:: torsion_AB: 0. 30.0 0.0 90.1
```

```
CONSTRAINT:: torsion_B: 180. 20.0 10 360.3
```

```
CST::END
```

Each degree of freedom is defined by five numbers:

```
CONSTRAINT:: torsion_AB: 0. 30.0 0.0 90.1
```

The first number is the optimal value of the degree of freedom: either a distance, angle, or torsion angle. Here, torsion_AB (the torsion angle of D2-D1-U1-U2) is defined to be 0° . The next number, 30.0, defines the allowed deviation, for a final allowable range of $0^\circ \pm 30^\circ$. The third number, 0.0, defines the “weight” on this constraint—the magnitude of the energetic penalty for deviating from the allowed range, which is used during minimization and design but not during matching (*10*). In our example of the catalytic geometry between the glutamine and the transition state, the weight is zero as this degree of freedom is free to rotate and should not be penalized during design. The fourth number, 90, defines the periodicity of this value. A periodicity of 90 means that this degree of freedom is allowed every 90° starting from the initial value. In this example, the angles $0^\circ \pm 30^\circ$, $90^\circ \pm 30^\circ$, $180^\circ \pm 30^\circ$, and $270^\circ \pm 30^\circ$ are allowed catalytic geometries. The final number, 1, specifies the fineness of sampling to perform in the matching stage. The matching algorithm will sample this degree of freedom at $2*n + 1$ points between the allowed deviation from the specified value (with this much sampling at each point of periodicity). This gives a total of twelve samples, one at every 30° , for even sampling of this degree of freedom.

Next, we need to prepare a set of scaffolds of protein structures into which we will attempt to build our enzymes. In general, we choose protein structures with a reasonable resolution (3 Ångstroms or better), that contain small molecules (to ensure the presence of a cleft or pocket) and for which the protein is expressed in *E. coli* (for ease of expression and purification). There are a number of other criteria (like multimerization state, number of residues, protein origin, or protein fold) that can be applied as well. Proteins from thermophilic organisms are good candidates for the scaffold set and have been preferred in *de novo* enzyme design due to the fact that they are usually more tolerant to destabilizing mutations arising in the design of novel catalytic activity (**II**). Once a set of protein structures has been obtained for use as scaffolds for enzyme design, we must prepare them for use with Rosetta. For each protein structure we generate two files: a coordinate file (encoded in the Protein Data Bank – PDB – format (**I2**)) with just the amino acids of the protein present and a position file which contains a list of residue numbers (or amino acid positions) which will be considered as catalytic residue placements during matching. While theozyme placement in the core of the protein will most probably unfold the protein, theozyme placement on the surface of the protein will be difficult as only a limited number of residues are available for interactions with the transition state. Therefore, residues lining a pocket or a cleft have the highest chance to successfully build an active design and should be included in the positions file. A straightforward way to identify these residues is to select residues within 5 to 7 Ångstroms of the natural ligand in the crystal structure.

Matching

The goal of the matching stage is to efficiently find scaffolds where we can place the theozyme (including both amino-acid catalytic sidechains and transition state structure) in a geometric

orientation consistent with catalysis. In this section, we go through the matching protocol to give a broad overview of how it works and to provide an understanding of the options and parameters which can be set or changed (for a detailed technical discussion see (10) and (13)). To keep track of the location of the transition state, the matching algorithm first divides the active site pocket of the scaffold into geometrical boxes, or “bins”. The bin size can be set with the `–euclid_bin_size` option. It then places each discrete conformation of the catalytic residue (termed rotamer) independently at each allowed position in the scaffold (defined in position file). Then the transition state is placed according to the defined catalytic geometries (defined in constraint file) for this catalytic residue. If the atoms of the catalytic residue or the transition state structure do not overlap with the rest of the protein main chain, the bin in which the transition state structure resides, the orientation of the transition state structure, the rotamer, and the position of the catalytic residue are recorded in a “hash table”. A hash table is an efficient way to store and resolve possible theozyme placements. At the end of matching, the entries in the hash table are compared to identify successful theozyme placements. A successful placement is defined as one where all catalytic residues, originating from different positions, place the transition state into the same spatial bin in roughly the same orientation. How exact the orientation has to overlap is defined by the combination of `–euler_bin_size` and `–euclid_bin_size` options, and are key parameters for a matching run. For each such placement or “match”, a PDB file is written out which is then used as input for the design stage.

The match executable located in your Rosetta directory (Note 1) carries out the matching using the following command line options (refer to the documentation and manual for a complete set of command line options):

```
-database your_directory/rosetta/rosetta_database/ {pointing to Rosetta database}
```


-extra_res_fa TRS.params {params file for transition state}
 -in:file:s scaffold.pdb {scaffold file used for matching}
 -match:scaffold_active_site_residues scaffold.pos {position file for scaffold}
 -match:lig_name TRS {name of transition state as defined in params file}
 -match:orientation_restype TRS {use transition state to calculate orientation during matching}
 -match:orientation_atoms N1 C8 O2 {use these three atoms to calculate orientation}
 -match:geometric_constraint_file TRS.cst {describes catalytic geometry of theozyme}
 -match:euclid_bin_size 1.50 {bin size in Ångstrom}
 -match:euler_bin_size 40.0 {angle deviation of orientation}
 -match:bump_tolerance 0.6 {allowed overlap between transition state and catalytic side chain/backbone}
 -packing:ex1 {add extra rotamers at ± 1 standard deviation for chi1}
 -packing:ex2 {add extra rotamers at ± 1 standard deviation for chi2}
 -packing:ex3 {add extra rotamers at ± 1 standard deviation for chi3}
 -packing:ex4 {add extra rotamers at ± 1 standard deviation for chi4}
 -nstruct 1 { repeat this protocol once}

The more hits recorded in the hash table and compared at the end the more memory is used and the more files are written as successful matches (which takes time and disc space). There is thus a fine balance between finding enough matches (in the hundreds or thousands) and running out of memory or disc space. There are two categories of parameters one can adjust: Those that control the precision of the match (euler and euclid bin) and those that sample the different degrees of freedom (conformations of the transition state, allowed catalytic geometries, extra rotamers for catalytic side chains, etc) (Note 2). The pragmatic approach is to take two or three scaffolds with different folds, perform a test run and look at the results. If you get too few matches or none at all, you may want to loosen the precision of the match parameters,

increase the sampling, or both. If you get too many matches you do the opposite. When you are happy with the result you can then launch a production run (matching on all scaffolds) and carry the output forward to the design stage.

Design of Active Site

Once the theozyme has been matched into scaffolds, the next step is to redesign the rest of the active site pocket away from its previous function and towards activity for the reaction of interest. Quite often the introduced transition state structure and catalytic sidechains do not fit well into the pocket or cleft provided by the wild-type sequence. There can be steric overlap (“clashes”) between the placed theozyme and the rest of the protein sidechains of the original scaffold, or cavities where the protein sidechains provide no interaction with the transition state structure. The design protocol relieves these clashes by moving or mutating the impinging residues and introduces interactions that further bind and stabilize the transition state, which acts both to create affinity for the substrate as well as accelerate the rate of the reaction. Furthermore, the design approach works toward native-like “backing-up” of the catalytic residues, trying to achieve good interactions to stabilize these residues in the desired, catalytically productive orientation. All of this is accomplished simultaneously with the Metropolis criterion Monte Carlo sampling approach used in RosettaDesign (*14, 15*) to identify the most energetically favorable enzyme active site.

While it is in theory feasible to redesign the entire scaffold on which the theozyme has been placed, one typically restricts redesign to the active site pocket. This is done to make the problem more computationally tractable and to minimize mutations that may destabilize the protein scaffold. Therefore, we set cutoffs for amino acid positions that are to be mutated (6-8 Ångstroms from the placed transition state structure, with catalytic residues excluded from

design) and for those residues that are to be repacked, meaning to have their rotamer but not identity changed (10-12 Ångstroms from the transition state). These cutoffs can be set with the options `-enzdes:cut1` (all residues closer to the transition state structure than cut1 can be mutated), `-enzdes:cut2` (all residues between cut1 and cut2 and pointing towards the transition state structure can be mutated, the others repacked), `-enzdes:cut3` (all residues between cut2 and cut3 are repacked), and `-enzdes:cut4` (all residues between cut3 and cut4 and pointing toward the transition state structure are repacked, all others stay in their crystal structure conformation). Since during the matching stage we allowed for some deviation from the ideal catalytic geometry, the enzyme design protocol first optimizes all constraints (`-cst_opt` option) by minimizing the transition state structure position (done automatically), the chi angles of the catalytic side chains (`-chi-min` option) and the backbone conformation (`-bb_min` option). It then cycles between designing the active site (`-enzdes:design` option) and minimizing it (`-enzdes:cst_min` option) for a given number of times (`-enzdes:design_min_cycles <value>` option). The core of the design process is Monte Carlo sampling of amino acid rotamers. For those positions to be mutated, all rotamers for all allowed amino acids are available during sampling. For those that are to be repacked, only rotamers for the current amino acid are accessible during sampling. If different conformations for the transition state structure are defined, then those are sampled at this stage as well (Note 3).

Monte Carlo sampling requires an evaluation function. For design, Rosetta uses a full-atom scoring function, modified for use with enzymes. For information about the energetic terms in this scoring function (which include electrostatics, the Lennard-Jones potential, hydrogen bonding, and other terms), refer to (**16**). In addition, energetic penalties are computed to assess when the theozyme's degrees of freedom are out of the range specified in the constraint file.

The balance between these two contributors to total score (which determines selection of amino acid identity as well as their rotameric state) can be changed by increasing or decreasing the weight for each degree of freedom within the constraint file.

The executable EnzdesFixBB performs the active site design. Typical command line options for enzyme design are:

```
-database /your_directory/rosetta/rosetta_database/  {pointing to Rosetta database}
-extra_res_fa TRS.params          {params file for transition state}
-enzdes:cst_opt                    {optimize all constraints}
-enzdes:cst_design                 {design active site}
-enzdes:cst_min                   {minimize active site}
-enzdes:chi_min                   {minimize chi angles of side chains}
-enzdes:bb_min                    {minimize backbone during cst_opt and cst_min}
-enzdes:bb_min_allowed_dev 0.1    {allowed deviation of Ca before penalty applies}
-enzdes:cstfile TRS.cst           {describes catalytic geometry}
-enzdes:detect_design_interface   {define active site residues around transition state}
-enzdes:cut1 6.0                  {all residues within 6A of the transition state}
-enzdes:cut2 8.0                  {all residues within 8A of the transition state}
-enzdes:cut3 10.0                 {all residues within 10A of the transition state}
-enzdes:cut4 12.0                 {all residues within 12A of the transition state}
-enzdes:lig_packer_weight 1.8     {increase ligand weights during scoring by 1.8}
-enzdes:design_min_cycles 3        {cycle between design and minimization 3 times before output a
                                   structure}
-packing:ex1                      {add extra rotamers at  $\pm 1$  standard deviation for chi1}
-packing:ex2                      {add extra rotamers at  $\pm 1$  standard deviation for chi2}
-packing:ex1aro                   {add extra rotamers for chi1 for aromatic residues}
```

-packing:extrachi_cutoff 1	{add extra rotamers for residues with a neighbor count of 1 or more}
-packing:use_input_sc	{include rotamer from original pdb structure}
-packing:soft_rep_design	{decrease overlap penalty during design (and not minimization) stage}
-nblist_autoupdate	{update neighborlist during minimization}
-linmem_ig 10	{use linear memory interaction graph}
-nstruct 100	{perform this entire protocol 100 times and output 100 structures}

EnzdesFixBB produces two types of output: design structures and score data. The design structures are PDB files containing the coordinates of the design, showing the protein, redesigned active site, and transition state. The score data contains per-residue (contained in the corresponding PDB structure file) and total (contained in a separate file, the score file) score data (with these scores determined by the score terms and weighting described above) for the output design.

Post-design Filtering

After collecting all the output from the design stage one is left with the nontrivial task of selecting active designs and filtering out the inactive ones. Despite significant progress in the field of enzyme design this is still essential, especially if the number of designs that can be experimentally tested is limited. A wide variety of filtering criteria can be used to rank or filter the designs. While there are some general rules (low total energy of the protein indicates likely expression and solubility, low transition state energy indicates good binding, good theozyme

geometry (as constraint score) is necessary for catalytic activity), each theozyme will require its unique set of filters. In order to filter designs, one first has to define a list of features or characteristics that are important for the reaction to happen (a specific distance, angle, hydrogen bond interaction, etc). After defining a metric for each of these features one can look at a handful of randomly chosen designs and determine a threshold for acceptance/rejection of a design.

Applying these filters to the entire data set should result in the ‘best’ designs. Depending on the assay throughput and cost to synthesize genes, one may want to be more or less stringent in their filtering.

As an illustrative example, I present here the concrete cases of filtering on total protein score and catalytic geometry. When filtering total protein score it is important to consider each scaffold independently as the Rosetta score for each scaffold can vary significantly. The total protein score can be found in the first column of the `enzdes.scores` file, labeled ‘total_score’. Sorting this column, one typically keeps the best 10% to 50% of the designs. A similar approach is taken for the catalytic geometry. The total score for the catalytic geometry can be found in the fourth column of the `enzdes.scores` file, labeled ‘all_cst’. In this case, the aim is to keep all designs that have a catalytic geometry as close as possible to the geometry specified by the theozyme. The appropriate threshold for this value can be defined by looking at a handful of designs spanning the entire score range. Designs with a cst score higher than the threshold are discarded.

Finally, the sequence of the selected genes is synthesized and cloned into an expression vector with appropriate tags for affinity purification (Note 4). Standard protocols for expression and purification of recombinant proteins can be used to obtain pure protein, and this protein can then be tested for the enzyme activity of interest.

Conclusion

The *de novo* computational design method presented in this review emphasizes the importance of quantum effects in enzymes (through the concept of the theozyme) and of transition state stabilization (as implemented through the design process). While this method of enzyme design has proved successful in a number of cases, I believe that other considerations may be necessary to continue improving the method. These include treatment of quantum mechanics within the design state, evaluating the effect of protein motions, or the effect of long-range electrostatic interactions. Enzyme design is a fast growing field and has the potential to create commercially valuable enzymes as well as useful tools to probe biology. Finally, this method leads us closer to the dream of the cell as a customizable chemical factory, allowing cheap and green chemical processing.

Notes

Note 1: Rosetta is available for licensing and download at

<http://www.rosettacommons.org>

It is free for academic and non-profit users and is available at a competitive licensing rate for commercial users. Included with the distributed source code are instructions and manuals for downloading, building, and installing the software. The RosettaCommons website is also a good resource, with manuals and support on how to run the Rosetta software.

Note 2: For catalytic side chain to transition state interactions that have multiple degrees of freedom free to move (and therefore require extensive sampling), a complementary matching approach called “secondary matching” is available in the RosettaMatch application. A detailed

description and discussion of the advantages and disadvantages of this approach can be found in (13).

Note 3: Designing a large number of matches can be quite computationally intensive. If computing power is limited, a two-stage approach to design may be used. In a first design round minimize high-intensive sampling (for example, by omitting extra rotamers, backbone minimization, or extra design cycles) and filter the output. In a second design round, use only the best designs as input instead of the original matches and follow the described protocol.

Note 4: It is good practice to check the original scaffold for missing density or compare the final design sequence with the sequence of the original scaffold from UniProt or GenBank before ordering or synthesizing designed sequence. Rosetta will omit residues that are missing in the crystal structure without a warning or comment in the output PDB. A protein with a missing surface loop may be output as a design in this way, and such a protein might not express solubly, precluding experimental testing of the computationally designed novel enzyme.

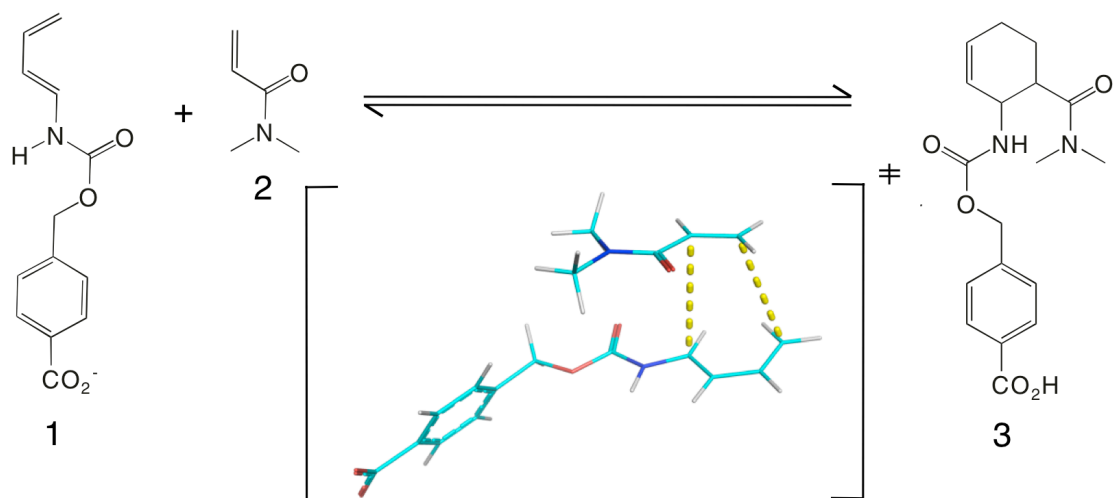


Figure 1.1: Substrates for Diels-Alder enzyme design project. Substrates (diene 1 and dienophile 2) and product (3) for Diels-Alder enzyme design project. The transition state structure of the reaction is depicted in cyan.

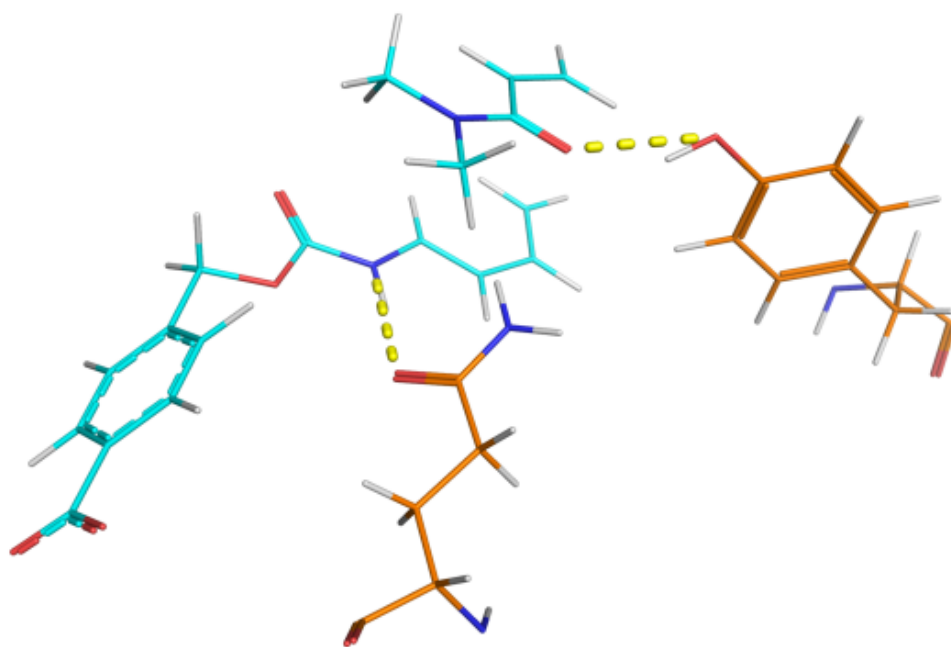
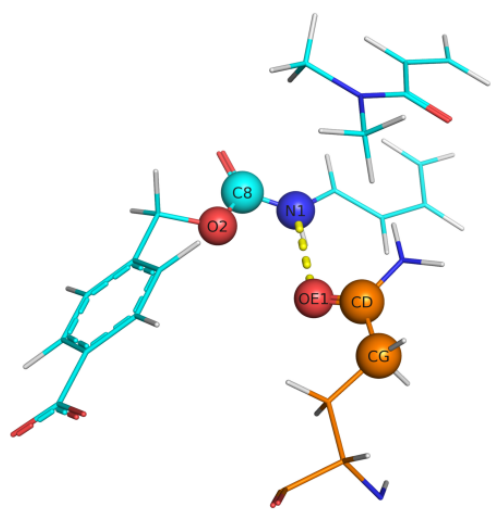


Figure 1.2: Theozyme structure for the Diels-Alder reaction used in computational enzyme design(4). A glutamine and a tyrosine serve as catalytic residues (in orange) to stabilize the transition state (in cyan).



Atom names: N1-C8-O2 -- CG-CD-OE1
 Generalized names: D1-D2-D3 -- U3-U2-U1
 Object names: ----A--- ----B---

Degree of freedom:	Definition:
Distance	Distance D1-U1
Angle A	Angle D2-D1-U1
Angle B	Angle D1-U1-U2
Torsion A	Torsion angle D3-D2-D1-U1
Torsion AB	Torsion angle D2-D1-U1-U2
Torsion B	Torsion angle D1-U1-U2-U3

Figure 1.3: Example of a theozyme interaction. Here we show the interaction from a glutamine (in orange) to the diene (in cyan) as well as the definition of the degrees of freedom used in the constraint file.

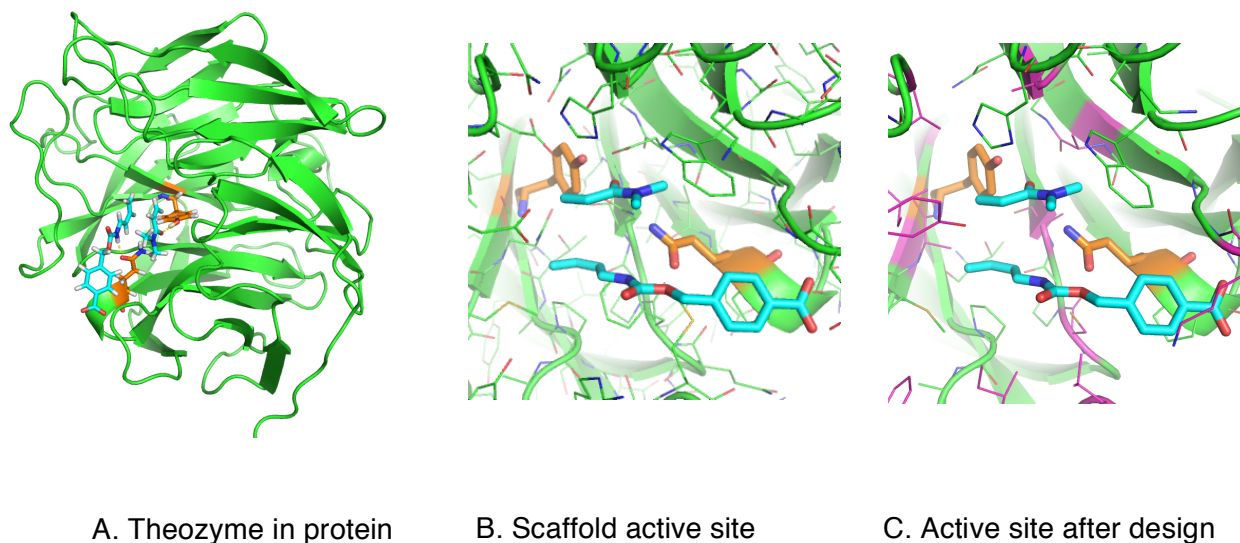


Figure 1.4: Progress of design from theozyme (A) to match (B), to final design (C). During the design process, wild-type residues (in green lines) are mutated (pink lines) if more favorable interactions can be made.

Chapter 2: Directed evolution and rational design of a computationally designed ester hydrolase

Introduction

De novo computational design of enzymes is a powerful technique for the development of novel catalysts. The repertoire of computationally designed enzymes continues to expand, with members falling into two classes. In the first class are true *de novo* enzymes—those designed on protein scaffolds with no activity for the reaction of interest or related reactions. In this class are a Retro-Aldol enzyme(**17**), a Kemp elimination(**2**) enzyme, the Diels-Alderase(**4**), an ester hydrolase(**3**), and an enzyme for the Morita-Baylis-Hillman(**18**) reaction. In the second class are enzymes designed on a scaffold with low or related activity (members of this subclass are typically referred to as specificity switches). Rosetta designed enzymes in this class include an alpha-gliadin peptidase(**19**), an organophosphate hydrolase(**20**), an anthrax coat protein peptidase(**21**), and a nucleoside kinase(**22**). The *de novo* computationally designed enzymes (the first class) have unexpectedly low activity, comparable to other novel enzymes such as catalytic antibodies(**23**) but much lower than natural enzymes. It is clear that for computational enzyme design to be medically or industrially useful, either the enzymes or the method itself must be improved.

One way to improve the enzymes as well as gain insight into possible method improvements is through directed evolution. However, directed evolution for enzymes can be difficult to accomplish. In the directed evolution experiments performed thus far to improve computationally designed enzymes, selection has been through segregation of clones as

plated cell colonies or (either as cell lysate or purified protein) in 96 or 384 well plates(24)(25)(26)(27). In the case of the ester hydrolase, these methods were especially difficult, as proteins in E coli can perform the same reaction as the enzyme, and so all enzyme variants must be purified away from cell lysate. Particulate screening technologies such as mRNA, ribosome, yeast surface, and phage display are often used to achieve high throughput screening of proteins. However, enzymes can be difficult to evolve using these methods, as a link must be maintained between the genotype (the particle containing the enzyme and genetic material) and the phenotype, which for enzymes is typically transformation of a diffusible small molecule. These techniques have however been used previously to improve enzyme activity, but only for specific mechanisms or through the use of complex selective strategies(28)(29)(30)(31). However, in the ester hydrolase mechanism (figure 2.5) there existed a possible solution: a long-lived acyl-enzyme intermediate (stable enough to be measured with mass spectrometry in previous work). This long-lived intermediate could act as a handle to isolate variants with improved efficiency for the first step of the ester hydrolysis mechanism.

Through computational design the Baker Lab was previously able to create a family of ester hydrolase enzymes(3). One of the best enzymes identified, ECH19, was scaffolded on a periplasmic binding protein (PBP) and had an active site highly amenable to evolution. ECH19 was designed on the closed conformation of this PBP scaffold, but a crystal structure showed that the enzyme instead adopted the open conformation. In that work, ECH19 was also subjected to low-throughput screening, through which two mutations to improve the enzyme's activity were identified (K354P and P364W). Here I use this double mutant ECH19 KPPW as a starting place for my directed evolution experiments. I performed three serial directed evolution experiments that improved the acylation activity of this ester hydrolase. By then isolating the evolved

variants through cloning, expression, and purification I was able to explore the enzyme kinetics and biophysical properties of these evolved variants. I found improved kinetics for these enzymes and performed detailed characterization on the two best enzyme variants. This improvement in kinetics was only for the first phase (acylation) of the ester hydrolysis mechanism, so I turned to structure-informed design to predict mutations that could speed the second slow phase of the enzyme mechanism (deacylation). I was able to identify a mutation that sped the deacylation phase, I hypothesize by stabilizing the designed cysteine-histidine catalytic dyad. This success points to the need for hybrid methods to improve computationally designed enzymes, combining insight from directed evolution with structure-informed redesign.

A yeast display selection for improved acyl-enzyme intermediate formation

The hypothesized mechanism of the ester hydrolase enzymes proceeds through an acyl-enzyme intermediate, which previous work has shown to be long-lived. I hypothesized that if I could select for faster and more complete formation of this acyl-enzyme intermediate, I could develop an enzyme with a faster burst (initial acylation) rate. This strategy would however not select for faster catalytic turnover. In figure 2.1A, I show the original substrate for which these enzymes were developed. To create our selective bait, we removed the methyl group and added a biotin separated by a linker to the acylating half of the ester molecule (figure 2.1B). With this substrate I could select for formation of the acyl-enzyme intermediate by first mixing yeast cells displaying a library of esterase variants with the biotinylated ester compound, and then labeling acylated proteins with fluorescently tagged streptavidin. Variants with very high activities (in this case, high activity being that of binding low-concentration substrate and then forming the acyl-enzyme intermediate) would have higher fluorescence and therefore be selectable using a

flow cytometer/cell sorter. In order to control for differing levels of protein display, I added a Myc tag to the C terminal tail of the esterase gene, which could be detected by use of a fluorescently labeled anti-Myc antibody in a second color.

To generate libraries from which to select improved variants, I used three separate library construction strategies in series, shown in figure 2.6. This complex library construction and selection strategy was necessitated by the large size of the ancestor protein ECH19 KPPW, my limited information as to the importance to protein activity of various regions on the protein, and finally the limited throughput of my selection strategy (with 10^7 - 10^8 the reasonable upper limit of library size addressable with yeast display). First, I selected six regions of contiguous sequence on the protein for partial randomization (figure 2.7). I selected regions that were either in the active site and able to make contacts to the substrate, near the catalytic dyad and able to refine the context around same, or in the "hinge" region between the two domains of the PBP and could conceivably induce closure or altered dynamics.

To generate diversity in these regions I used doped oligonucleotides to create contiguous randomized regions along the ECH19 KPPW gene. By using doped oligonucleotides I was able to precisely specify the desired mutation rate in order to create libraries with the greatest possible complexity. The doped oligonucleotides were used to construct the gene, both with and without Y250F and Y250W mutant oligonucleotides spiked in (see Results section "Rational design to improve catalytic turnover"). These mutations were designed to correct a problem seen in the crystal structure of ECH19, where the histidine in the catalytic dyad was seen to make an interaction with tyrosine 250 instead of the designed interaction with the cysteine. The libraries were then cloned into a yeast display vector and transformed into yeast cells. I called these

libraries DL, for Doped Library.

With these six libraries I performed a selection to improve acylation with the ester-biotin substrate, through use of the yeast display and fluorescence activated cell sorting (FACS). After labeling a library of displaying yeast cells with the ester-biotin substrate I added Anti Myc-FITC (to label for protein expression) and streptavidin-PE (phycoerythrin, to fluorescently label the biotin-acylated enzymes). I sorted for cells with improved signal both in display (Anti Myc-FITC) and acylation (streptavidin-PE) and found steady improvement in both over three iterative rounds of sorting, growth, and induction (Figure 2.2A, “Initial doped libraries”). However, after sequencing I found that not all libraries had converged, showing many different sequences with varying activity when tested individually. The sequences and numeric activity values from these individual clones are reported in table 1.

While I saw improved acylation for these evolved variants, I wanted to verify that the improved activity came from improvements in the designed active site, and not from new active sites or even from non-catalytic binding of the ester-biotin ligand. To test this, I made cysteine to alanine knockouts at the active-site cysteine for three of the most improved variants. By testing this knockout, I could verify that the activity was dependent on the active-site cysteine (ensuring that the activity I observed was in fact acylation on the designed cysteine). When I constructed, cloned, and tested these variants with yeast display, I in fact saw complete knockout of the improved activity (figure 2.8). This verified that activity was improved for the reasons I hypothesized, and that the selection was working as expected.

While the selection of each of the doped libraries could be continued to converge on the optimal sequence for each, we decided instead to shuffle their sequences together to obtain a globally

optimal sequence. We did this by PCR amplifying DNA purified from each of the six yeast libraries and then stitching them together using standard DNA assembly. All diversity in this library came from the assembled DNA and not from low-fidelity PCR or mutagenic oligonucleotides. With this combination library (abbreviated “CL”) we continued the selection, now in one pool. We performed three rounds of selection at increasing stringency by dropping the concentration of biotinylated substrate. Sequencing the selected pool at round 3 we began to see sequence convergence in selected clones, and in figure 2.9 we show a tree of the variants. We call the four most common clones found in sequencing the selected pool CL1-4.

While the selected variants from the combination library showed improved acylation, the library creation technique could only explore a small part of sequence space, and only coarsely (with combinations of already selected variants). I reasoned that further improvement could be found by more finely exploring sequence space around the selected combination library variants. To achieve this, I used error-prone PCR to sample mutations on top of the converged mutants CL1-4 (separately and as a pool). Again, because of the large size of the ECH19 esterase gene, I was not able to mutate the entire gene. I selected two different regions within the gene (a smaller region and a larger region as shown in figure 2.10) and performed error-prone PCR on these regions separately (details of each library in table 2.2). As before, I then constructed the rest of each gene around these mutant regions and transformed into yeast cells to create six error-prone libraries (abbreviated EPL). I performed 3-5 rounds of selection for display and acylation on these libraries at very stringent conditions (even lower [ester-biotin]) and then sequenced and tested individual clones as before. In figure 2.2A I show the progression of improvement from the ancestral esterase ECH19 KPPW over the course of the three library strategies, with acylation measured at two ester-biotin substrate concentrations. Figure 2.2B shows the ancestor

ECH19 KPPW, the variant CL1 from the combination library selection, and variant B2A from the error-prone library selection.

Evolved variants are improved catalysts

While I was able to improve acylation with the new biotinylated ester probe, it was not yet clear if I had achieved my goal of an enzyme with improved kinetics for the original free-ester substrate. To test this, I cloned evolved variants from the yeast display system into a bacterial expression system. I tested a number of variants for their activity as purified proteins over a range of free-ester substrate concentrations from 3.1 to 100 μM . Initial screens were performed in 96 well plates, due to the labor involved in activity screens using cuvettes (the method used later to obtain data for conclusive Michaelis-Menten kinetics). Plots of relative velocity (in arbitrary units) vs. [free-ester] for all purified enzyme variants (along with the ancestor ECH19 KPPW) can be found in figure 2.11. However, when using 96 well plates for characterization I cannot observe all of the fast initial burst phase (due to instrument dead time) and so cannot very accurately characterize kinetics.

From these initial screens I selected the two apparently most improved enzymes for further study: variants B2A and 5A2D. Variant B2A is named for being clone A from the third round of error-prone library (EPL) B2. Variant 5A2D was named for being clone D from the fifth round of EPL A2. My next goal was to precisely determine the change in catalytic efficiency as measured by the bimolecular rate constant V_{max}/K_m for these variants. I purified them as before and obtained kinetics for the burst phase (this time using a spectrophotometer with < 1 second dead time, allowing precise and full measurement of the burst phase) and then fitted them to

Michaelis-Menten kinetics as previously reported and as described in the Methods section. These kinetics, fits, and Michaelis-Menten parameters are shown in figure 2.3.

The free-ester substrate I used to obtain Michaelis-Menten parameters is only soluble up to 100 μM in aqueous solutions, even with supplemented acetonitrile to 5% v/v. This is not surprising, as it is a very hydrophobic molecule (figure 2.1A). Over a range of substrate concentrations from 3.13 to 100 μM , I saw no indication of saturation (which would allow determination of V_{max} and K_m) in the ancestor esterase ECH19 KPPW. However, the two most improved variants B2A and 5A2D did show saturation, allowing us to calculate their V_{max} and K_m (figure 2.3). The V_{max}/K_m for B2A was found to be $749.0 \text{ M}^{-1}\text{s}^{-1}$, and for 5A2D was found to be $772.3 \text{ M}^{-1}\text{s}^{-1}$. While I cannot compare these V_{max}/K_m values to that of ECH19 KPPW, I observe that over the range of substrate concentrations I tested, B2A was between 1.9 to 3.7 times faster than ECH19 KPPW with a mean rate enhancement of 2.9x. The enzyme variant 5A2D was between 1.8 to 3.3 times faster than ECH19 KPPW, with a mean of 2.5x.

I wanted to further characterize the variant B2A, as it had the highest V_{max} and greatest average rate enhancement over ECH19 KPPW of the variants I tested. While I showed that for DL variants labeling activity on yeast was dependent on the cysteine in the enzyme active site (Figure 2.8), I wanted to verify that acylation proceeded as expected for this variant from later in the selection. I purified a cysteine knockout of B2A (C161A) as well as ECH19 KPPW (C161A). I tested acylation with two chemicals: the free-ester and phenyl phenylacetate (PPA) and then verified the mass of these proteins with LC-MS analysis. I used PPA to see if any differences could be seen in acylation with a less activated ester compound but did not find any. The results (in figure 2.12) show that the cysteine-containing enzymes acylate as expected; at one location

on the enzyme, with the expected mass difference, and dependent on the active-site cysteine.

Our crystal structure of the ancestor ECH19 KPPW indicated that some mutations in variant B2A were likely in its core. It was not immediately apparent if these would be stabilizing or destabilizing to the protein, so I set out to determine the relative stability of ECH19 KPPW and B2A. I purified these proteins and performed a CD melting experiment, and determined that the evolved variant was only modestly destabilized relative to the ancestor, if at all (figure 2.13). This indicates that structural stability was maintained to some extent in the directed evolution experiment, and that these core mutations largely retain stability.

To better understand the origin of improved catalysis for the evolved variants, we performed molecular dynamics on models of the variants B2A and 5A2D. We also did this to direct efforts toward testing additional mutations that to further improve activity. These molecular dynamics studies indicated that key to the improvement seen in the evolved variants was the mutation T165A observed in almost every evolved variant from the CL and EPL. In MD simulations, this mutation was observed to change dynamics of the catalytic helix (containing the catalytic cysteine at its N terminus) such that the cysteine was in a catalytically competent conformation more of the time. Through analysis of these simulations, we predicted that the mutation to valine at this position may further improve activity through the same mechanism. I tested this mutation along with a series of other small amino acids and the A165T mutation (reversion to ECH19 KPPW identity at this position) for activity with the free-ester substrate (figure 2.17). I found that the mutation to valine did not significantly improve activity, and that the alanine at this position is near optimal.

In order to better understand the mechanism of improvement I set about dissecting the relative importance and nature of effect of each mutation on the evolved variants. To do this I constructed reversion mutants for each patch of mutations on variants B2A and 5A2D. I show the location of these patches on the ancestor crystal structure in figure 2.14A for B2A and figure 2.15A for 5A2D. I also show the exact location and identities of these mutations in table 2.3. By testing at two substrate concentrations (one above and one below) around the K_m of the evolved variants, I hoped to dissect to what degree each mutation was responsible for the improvements in V_{max} and K_m . I tested these variants with free-ester substrate as before and report the results in figure 2.14B/C for B2A and figure 2.15B/C for 5A2D. I was able to identify the most important mutations, as measured by the magnitude of their impact on activity when reverted. For B2A the most important mutations seem to be B2 and B5, and for 5A2D the majority of improvement seems to come from 5.4. Some reversions however appear to be close to neutral or even detrimental (eg B3, 5.1, 5.2). I also performed the reverse experiment, adding mutations from B2A to the ancestor ECH19 KPPW in three patches found to be the most important (figure 2.16). Surprisingly, this improved activity only at 75 μ M free-ester and not at 10 μ M. This detailed dissection of mutant activity can both serve as benchmarking data for computational design as well as guide further improvements to the enzymes themselves.

I used this information to guide the combination of B2A and 5A2D mutations with the goal of combining the higher V_{max} of B2A and lower K_m of 5A2D. B2A and 5A2D share three identical mutations from ECH19 KPPW but each has additional mutations not found in the other: five additional in B2A and three additional in 5A2D. I decided to test three mutants of B2A to try to incorporate mutations found to be important on 5A2D: C1, C2, and C3. I also made one mutant of 5A2D, called 5B.1, with the mutations found to be most important on B2A. The exact

combination of mutations in each of these variants is detailed in table 2.4. I then constructed, expressed, and purified these variants as before, and tested activity with free-ester. The activities of these variants are presented in table 2.4, with the most active combination variant being C3. In order to verify this improved activity, I performed full Michaelis-Menten characterization as before, in triplicate at eight substrate concentrations between 3.13 and 100 μM free-ester (figure 2.18). I hypothesized that C3 would have higher K_m than 5A2D (intermediate between B2A and 5A2D) and the same or slightly lower $V_{\max}$ as B2A. Surprisingly, C3 was found to have an even lower K_m ($50.9 \pm 7.7 \mu\text{M}$) but also near identical $V_{\max}$ ($0.051 \pm 0.003 \text{ s}^{-1}$) than low K_m / low $V_{\max}$ variant 5A2D ($0.053 \pm 0.004 \text{ s}^{-1}$ and $68.2 \pm 12.6 \mu\text{M}$, respectively). However, the ratio $V_{\max}/K_m$ for C3 may in fact be improved. However, the changes are small and not significantly greater than SD values. While this variant may represent further improvement over ancestor ECH19 KPPW, all second-order rate constants given here are for the initial burst phase. In order to further improve this enzyme, I looked to the second, steady state phase of the reaction.

Rational design to improve catalytic turnover

In our original paper on the computational design of ester hydrolase enzymes, we report that the enzymes show biphasic kinetics. There is an initial “burst” phase that I believe corresponds to initial acylation of all available activated cysteines, and then a second, much slower phase that corresponds to the deacylation step (which reforms the active cysteine). This is not a property of natural esterase enzymes, which use additional catalytic machinery to more rapidly deacylate the active site cysteine. The slow deacylation step in our designed esterase enzymes was a surprise, as we designed a cysteine-histidine dyad that we expected to give the rapid deacylation seen in natural esterases. In the previous work, we also report the crystal structure of ECH19 (PDB

accession code 3U1O), the ancestor of the evolved variants in the present work. The cysteine-histidine dyad of that enzyme is not in the designed conformation; rather the histidine is in an alternate rotamer, positioning it away from the catalytic cysteine and toward tyrosine 250 (with which it makes a hydrogen bond). I show this crystal structure along with the (computationally modeled) designed histidine conformation in figure 2.4A.

Using this crystal structure of ECH19, I designed two mutations which I hypothesized could help reform the designed cysteine-histidine dyad. The first mutation, Y250F, was based on the hypothesis that by removing the group interacting with the histidine (the hydroxyl on tyrosine 250), the histidine would no longer be able to make the observed hydrogen bond and would take its designed position in the dyad. The second mutation I designed was Y250W. Here I hypothesized that by increasing steric bulk at position 250 I could simply occlude the histidine from the core, again to form the designed dyad. I show my model of this tyrosine to tryptophan mutation in figure 2.4B. I constructed, expressed, and purified these proteins and tested them as before, at two concentrations of free-ester. In initial tests I found that the Y250W mutation had better kinetics than the Y250F mutation, so I continued characterization only with Y250W. I found that the Y250W mutation increases the rate for protein ECH19 KPPW both in the burst phase and in the steady state (figure 2.4C). In order to verify that this effect was dependent on the histidine, I constructed a knockout of the dyad histidine (H226A) on ECH19 KPPW Y250W. The activity of this knockout was significantly lower than that of ECH19 KPPW or its Y250W mutant, indicating that the improved activity is dependent on the histidine. Through the identification and characterization of this mutant I demonstrate that structure-based rational design can allow activity improvements that would be otherwise inaccessible. I hypothesize that this mutation was not found in my yeast-display selections because it enhances the enzyme

rate after the burst phase. This would presumably accelerate removal of the biotin handle through hydrolysis of the acyl-enzyme intermediate, an enzyme activity that would be strongly selected against.

Discussion

In this work I report a high-throughput particulate screening selection to improve the activity of a computationally designed enzyme. This sort of selection technique is especially useful, as very large libraries (10^7 - 10^8 variants) can be screened easily and iteratively(32)(33), as opposed to the large amount of labor required for well plate or colony selections. Selections to improve enzyme activity are difficult to engineer, as special difficulty lies in maintaining a link between genotype and phenotype (given that the products of most enzymes diffuse away). There have been reports of enzyme selections mediated by in vitro compartmentalization and cell-in-emulsion sorting(34)(35) as well as selections for enzyme reactions with products that become intracellularly trapped(36). In addition, ligase-type enzymes are more amenable to selections, with the possibility of linking genotype to phenotype through the ligation activity itself(37). Here, I exploited the existence of a transient acyl-enzyme intermediate in the ester hydrolysis mechanism and created a new ester-biotin probe to make a selection strategy for improved ester hydrolase enzymes.

By dissecting and combining the most active variants from the selection, I was able to gain valuable insights into improvements needed to obtain higher activity designed enzymes. While work remains to determine the mechanism of improvement for these mutations, I was able to show that they contribute to improvement of both K_m and V_{max} , likely to different extents (from the varying magnitude of effects at 10 and 75 μ M free-ester seen).

I was also able to improve kinetics in the steady state phase by structure-guided rational reengineering of the active site. I accomplished this by mutating away a residue that I found in a previously solved crystal structure to be interacting counterproductively with the histidine in the cysteine-histidine dyad. With this mutation I hypothesize that I essentially pushed the histidine toward the cysteine, both by removing its interaction partner (a hydroxyl on a tyrosine residue) and by occluding it from the core of the protein (with the steric bulk of the tryptophan). This sort of structure-guided redesign can be extremely valuable in efforts to improve designed proteins, but is often not possible due to lack of a crystal structure. In those cases, molecular dynamics is an attractive alternative, allowing modeling of possible undesired states. Furthermore, a method of assessing rotamer stabilization has been developed for the Rosetta computational design software(38). My work here demonstrates the importance of such methods, which may be able to predict mutations to stabilize and reshape designed proteins toward higher activity.

Ultimately I hope that directed evolution experiments can provide feedback to guide efforts to improve computational design techniques. In my analysis of the improved variants, I find it difficult to assign definitive roles to each mutation or to understand how I could have predicted them. Further study (especially through solution of crystal structures) of these variants is needed in order to give the feedback we desire. However, this work demonstrates that computationally designed enzymes can be good starting places for directed evolution and the creation of high-efficiency catalysts. By continuing to use hybrid methods and designing new enzymes amenable to high-throughput selection, we can extend and strengthen the insights from this work.

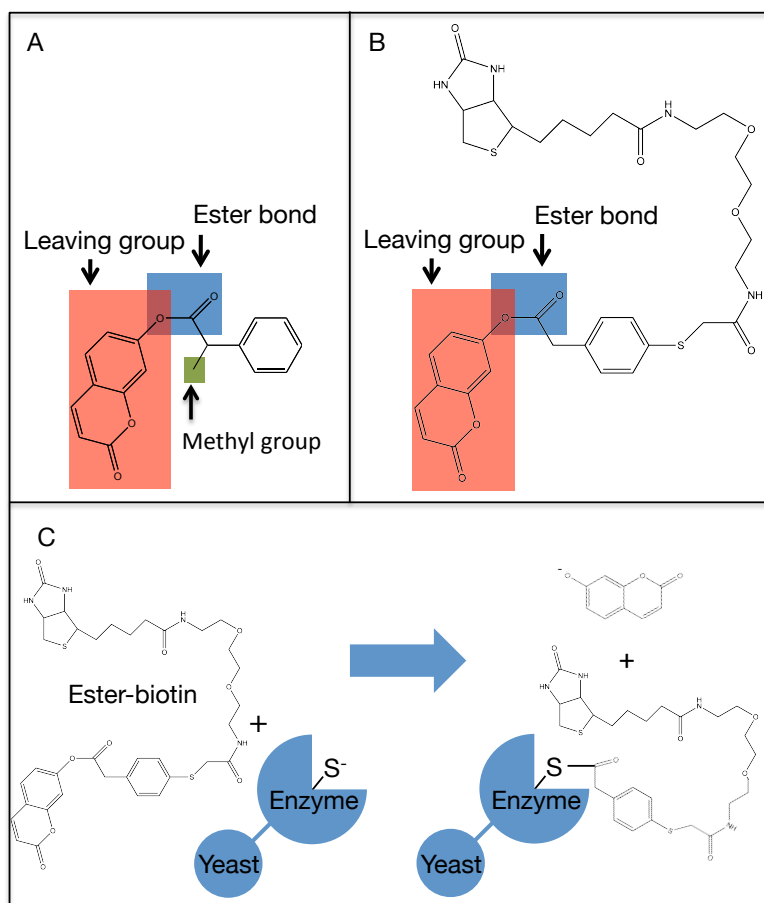


Figure 2.1: Selection strategy to optimize an ester hydrolase.

(A) Original ester substrate ("free-ester") used in previous work for design and testing of a novel ester hydrolase. (B) Ester-biotin substrate developed for yeast display evolution approach. (C) Selection scheme using ester-biotin substrate to label ester hydrolase enzymes.

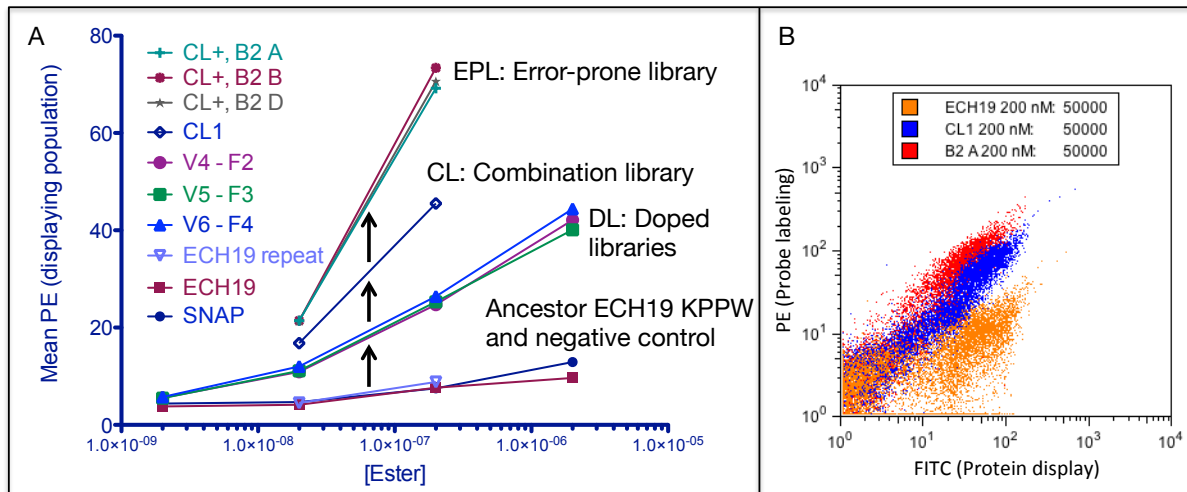


Figure 2.2: Improvement in labeling on yeast.

(A) Improvement of enzyme variant labeling over the course of over three serial directed evolution strategies. (B) Anti Myc-FITC (display) vs. streptavidin-PE (acylation) for ancestor protein ECH19 KPPW, combination library variant CL1, and error-prone library variant B2A with 200 nM probe.

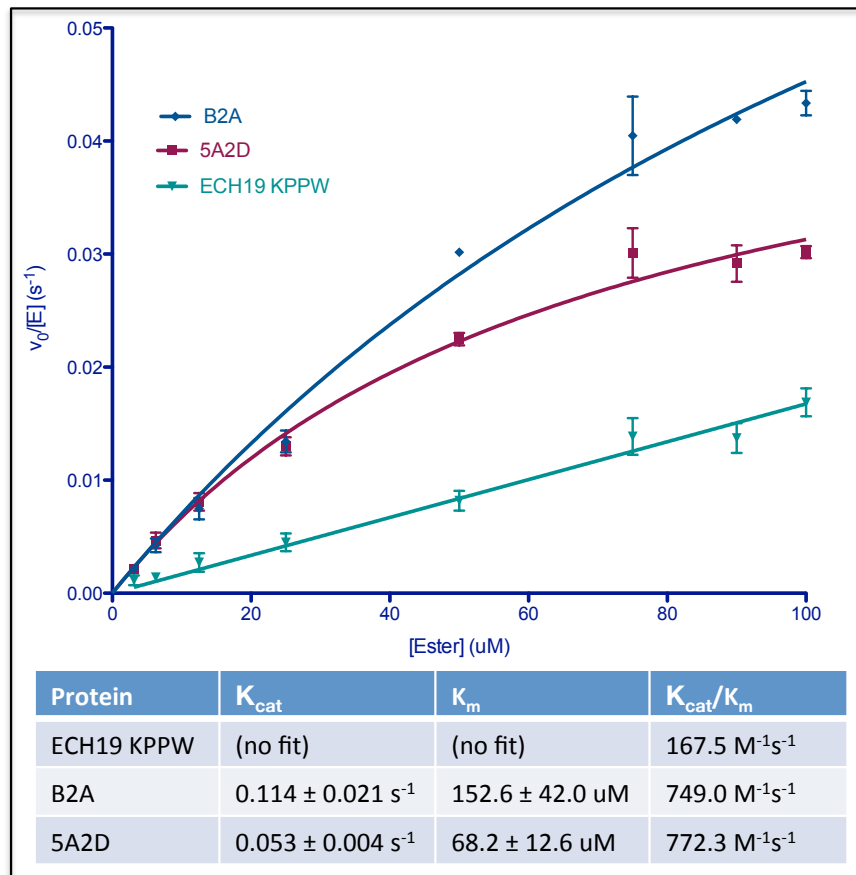


Figure 2.3: Michaelis-Menten kinetics show improvement in K_{cat} and K_m for evolved variants.

The data for ECH19 KPPW did not fit to a nonlinear regression model, and so linear regression was used to model first-order kinetics, giving a value for K_{cat}/K_m . All rates are for the initial burst phase of enzyme catalyzed reaction.

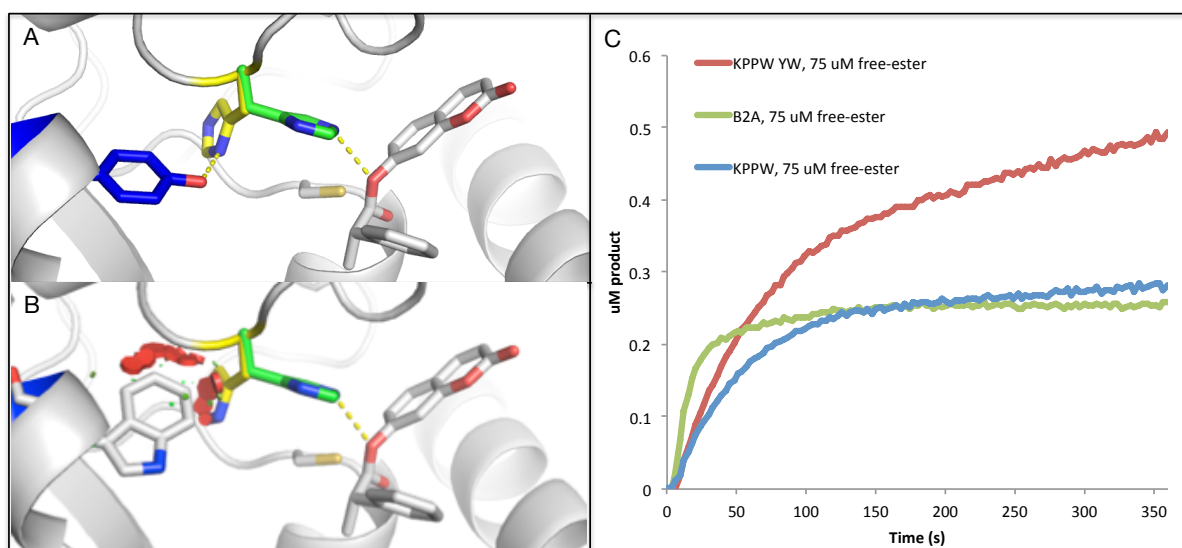


Figure 2.4: A rationally designed point mutant accelerates turnover.

(A) Crystal structure and computational model of ECH19 KPPW active site. Histidine in cys-his catalytic dyad was designed (green histidine rotamer) to activate cysteine and stabilize developing negative charge (as well as activate acyl-enzyme for turnover) was found in crystal structure to be flipped away from active site (yellow histidine rotamer), making a hydrogen bond to a tyrosine (in blue). This tyrosine was mutated to tryptophan (Y250W). (B) Hypothesized model of the substitution Y250W. I predict the histidine will adopt the designed conformation in green. (C) Product formation curves (buffer subtracted) for ECH19 KPPW, ECH19 KPPW Y250W, and evolved variant B2A, at 0.75 uM protein and 75 uM free-ester substrate.

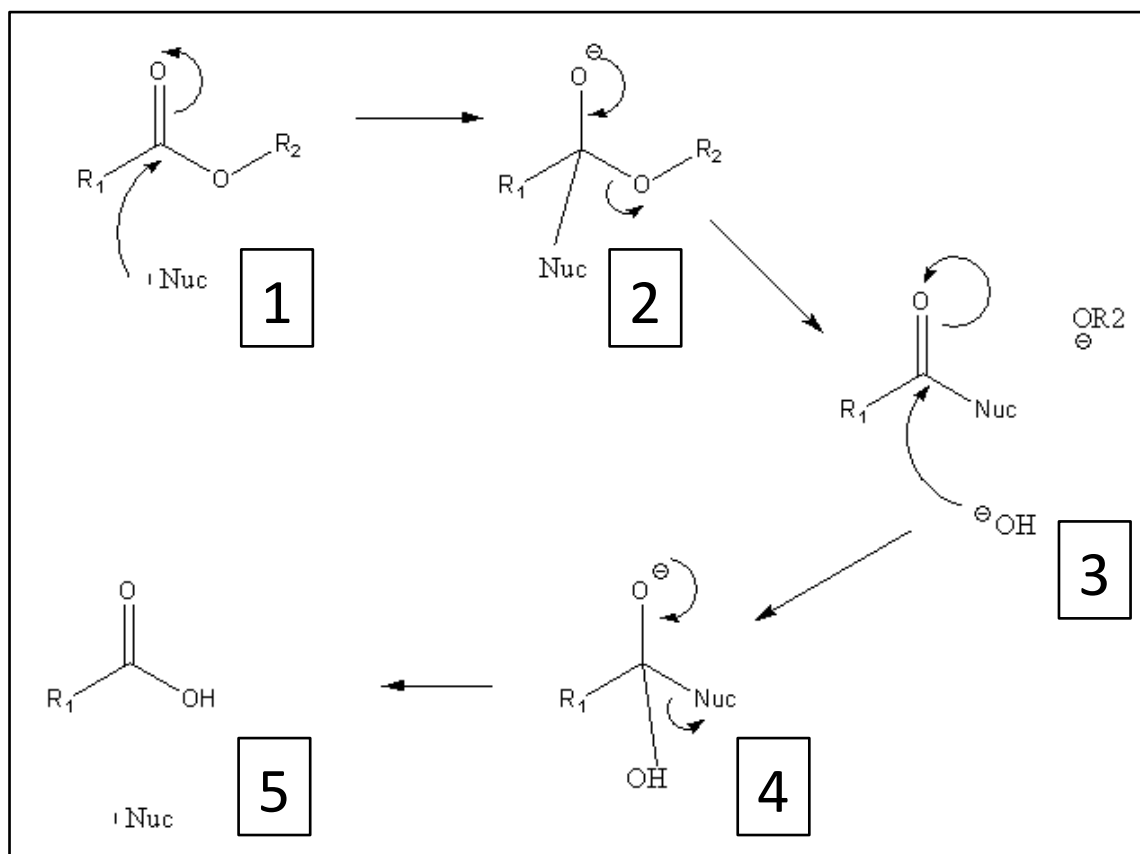


Figure 2.5: The five step predicted mechanism for ester hydrolysis by designed cysteine esterases.

Here “Nuc” is the nucleophilic cysteine sulfur atom in the cysteine-histidine catalytic dyad.

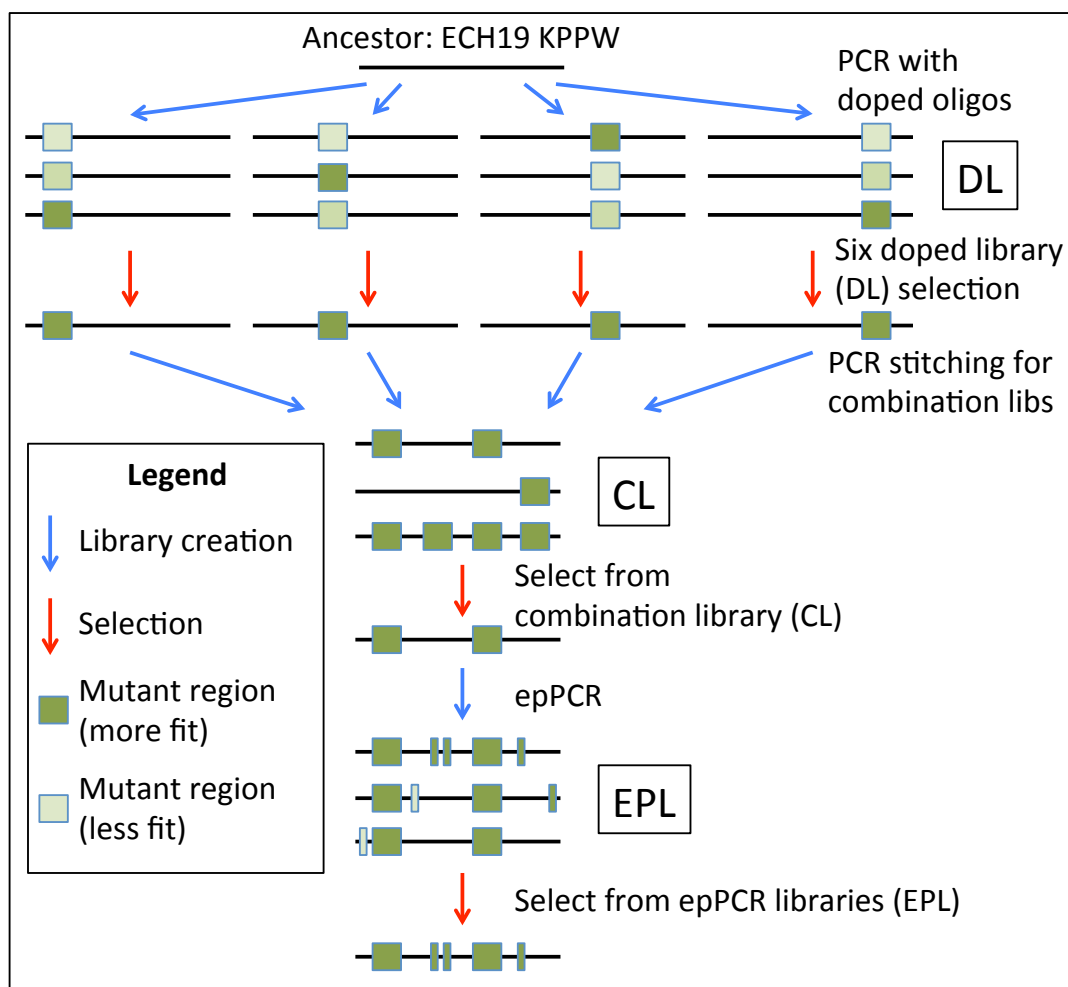


Figure 2.6: Library creation and selection scheme to improve ester hydrolase acylation.

This selection scheme is used due to the large size of the ECH19 KPPW gene as well as uncertainty as to which regions that are the most evolvable.

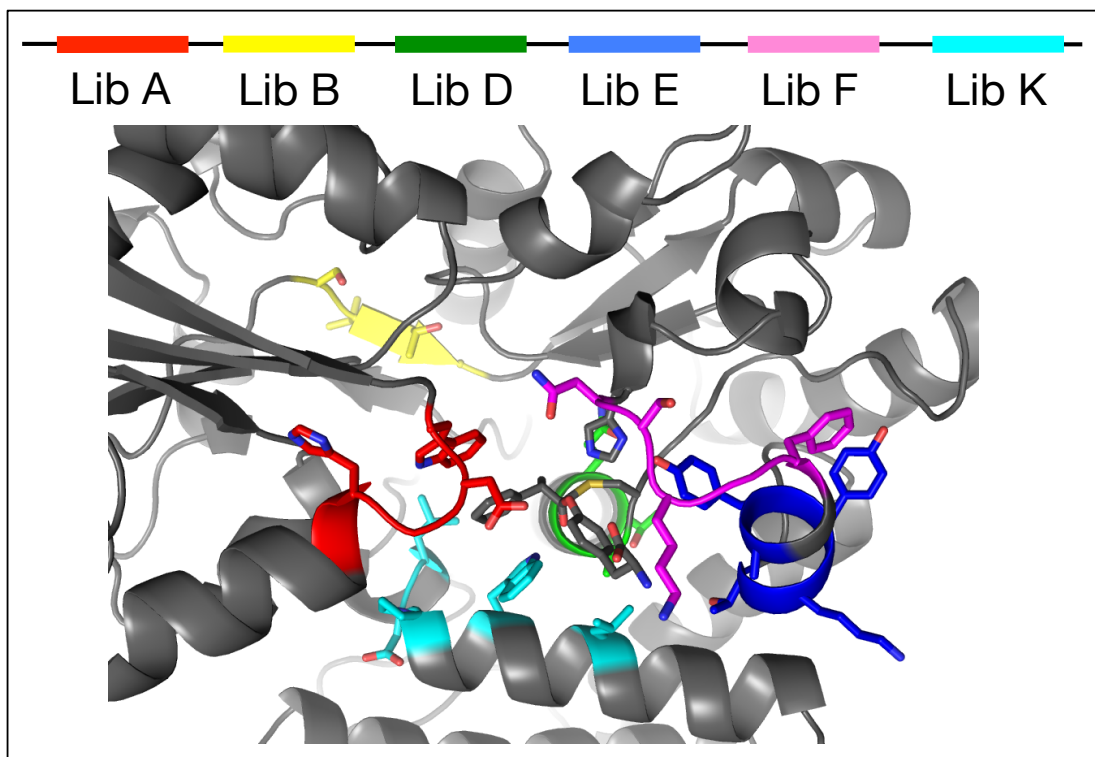


Figure 2.7: Location of regions randomized on ancestor protein ECH19 KPPW to create six doped libraries.

Each of these 4-5 amino acid regions were randomized separately and precisely using oligonucleotides with specified proportions of each base. The libraries made in this way are called “Doped Library” (DL).

Ancestor:	Activity:																
ECH19 KPPW	8.4																
		Position:	47	48	49	50	51	126	131	180	187						
Library A:	Activity:	Native:	W	D	G	H	L	S	T	T	T						
A1	16.2		G	sil		Y	sil	F	M								
A2	19.6		G	sil		Y	sil	F	M								
A3	22.4				R												
A4	28.4		R														
A5	29.4	(No muts.)															
A6	~		S			P											
A7	~		R		D					A	M						
		Position:	47	48	49	50	51	126	131	119	120	121	122	89	52	30	165
Library B:	Activity:	Native:	W	D	G	H	L	S	T	S	V	T	A	S	S	E	V
B1	29										sil	L	T				
B2	34.5									sil	L	sil	S				
B3	~										sil						A
B4	~											I	sil				
B5	~										L	sil					
B6	~											I	sil				
B7	~		G	sil		Y	sil		M					F			
B8	~														P	G	
		Position:	168	169	170	171	184	150									
Library D:	Activity:	Native:	A	G	Q	T	T	G									
D1	32.1					M											
D2	21.5			V		S											
D3	18							S									
D4	25.3			V		S											
D5	19.2			V		S	S										
D6	18.7	(No muts.)															
D7	8.24					M											
D8	22.7				K	I											
		Position:	221	222	223	224	225	288	324	329	360						
Library E:	Activity:	Native:	T	K	Y	Y	A	Q	G	A	P						
E1	18.8		N				V		sil								
E2	35.7		sil							sil							
E3	23.7				R	F		R									
E4	27.5		S			S	T										
E5	17.1																
E6	24.1						S										
E7	12.8		sil				G				K						
E8	13.1																
		Position:	227	228	229	230	231					360					
Library F:	Activity:	Native:	F	G	K	S	N					P					
F1	12.5		I				K										
F2	22.6		I			Y	K										
F3	23.8		L		I		K										
F4	25.9		I			F	K										
F5	16.6		I				K										
F6	20		I														
F7	16.2				I	F	K						K				
F8	18.5		I			Y	K										
		Position:	368	369	370	373	377	365	207	332	221						
Library K:	Activity:	Native:	L	D	P	W	L	P	T	V	T						
K1	8.53			F			R										
K2	10.4		V	H													
K3	5.4						R						I				
K4	7.17			E								A					
K5	6.38		sil	N			Q										
K6	8.71			N	A												
K7	16.3		sil	N			R	sil	A								

Table 2.1: Sequences and activities for clones from DL selection after three rounds of sorting.

2-8 clones were isolated from each selected doped library and sequenced, and named for their library and clone number (eg. F2 is library F, clone 2). In addition, for most clones acylation was measured with flow cytometry, with 2 uM ester-biotin substrate and reported as mean PE

(acylation) signal for cells displaying protein (measurements in arbitrary units). The ancestor ECH19 KPPW is included as a control. The position and ancestral (“native”) identity of each mutation is given in the table. The green shaded regions indicate the areas targeted for mutation in each doped library, mutations outside these regions are likely due to low-fidelity *Taq* PCR or contamination between libraries (eg. variant B7). “Sil” indicated a silent mutation, and “No muts.” indicates a clone with no mutations seen in the sequenced region of the gene.

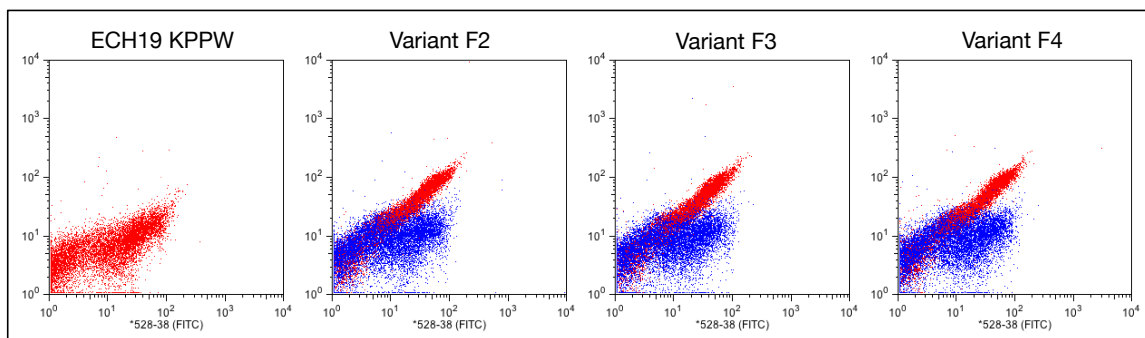


Figure 2.8: Improved acylation is dependent on active site cysteine.

Knockouts of variants F2, F3, and F4, with sequences as described in table XX) of proteins from first “doped library” selection show dependence of improved acylation on active-site cysteine.

Yeast displaying the variant (red dots) or its cysteine to alanine knockout (blue dots) were displayed on yeast surface, then acylated with 2 uM ester-biotin for 10 minutes, and labeled and analyzed as described.

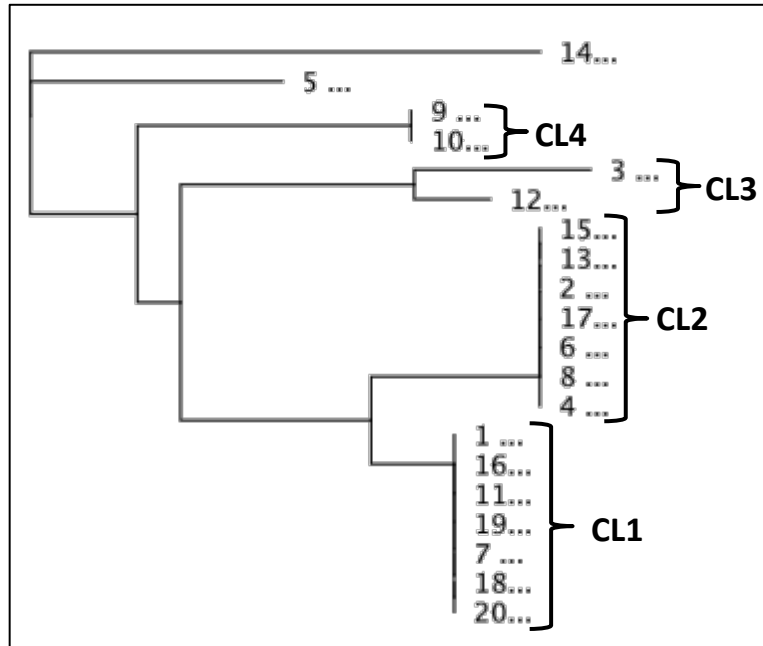


Figure 2.9: Combination library (CL) variants show convergence on four sequences.

Twenty clones from the combination library were sequenced after three rounds of selection.

They were aligned and pairwise distances computed to construct an unrooted tree. Convergence was found and four sequences predominated the selected pool: CL1 (7 identical clones), CL2 (7 identical clones), CL3 (2 very similar clones), and CL4 (2 identical clones).

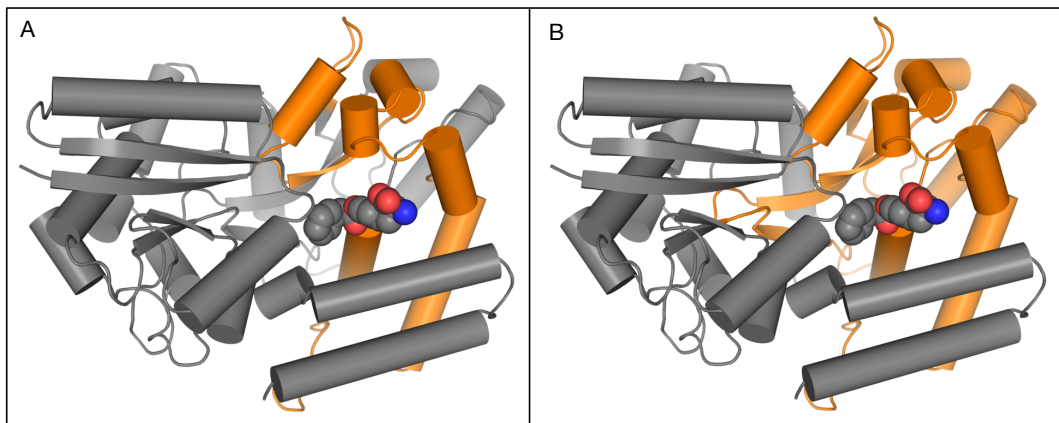


Figure 2.10: Areas targeted for error-prone PCR mutagenesis.

Modeled substrate is shown as spheres. Region targeted for mutagenesis shown in orange, rest of protein shown in gray. (A) The 99 amino acid “Ligand” region targeted in epPCR mutagenesis to give libraries T1, T2, A1, and A2. (B) The 173 amino acid “Hinge plus ligand” region targeted in epPCR mutagenesis to give libraries B1 and B2.

Library name	Template gene	epPCR region
B1	CL1	Hinge plus ligand
B2	CL1 + doped library A round 3 fragment	Hinge plus ligand
T1	CL1	Ligand
T2	CL4	Ligand
T3	CL round 3 pool	Ligand
A2	CL2	Ligand

Table 2.2: Template genes and randomized regions used to construct error-prone libraries (EPL).

Each gene was randomized using error-prone PCR in the region indicated (see figure 2.10 for details) and then the rest of the gene was constructed around this randomized region. The full library gene constructed in this way was then used to create the six EPL. In the case of library B2, additional diversity was incorporated by constructing the gene using diversity from the doped library A round 3 pool rather than the template gene CL1.

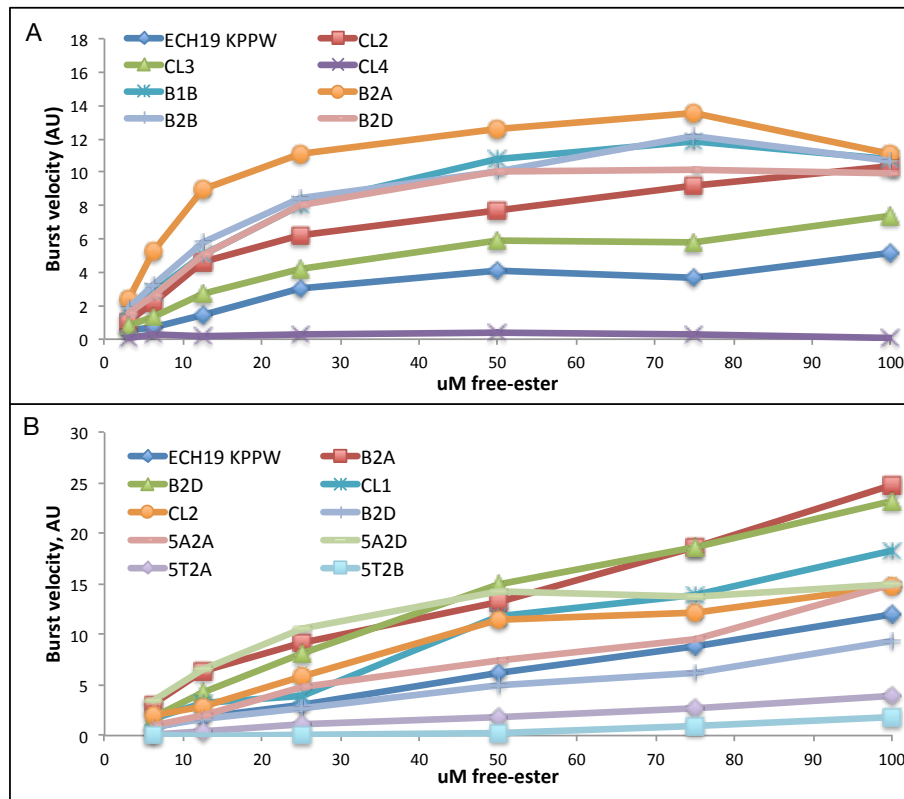


Figure 2.11: Initial, rough measurement of enzyme activity with free-ester identifies multiple improved variants.

This method of measuring enzyme activity (in 96 well plates) does not capture the entire burst phase due to instrument dead time and therefore these curves used to determine Michaelis parameters. In addition, these values are in arbitrary units and cannot be compared between the two experiments presented here, shown in (A) and (B).

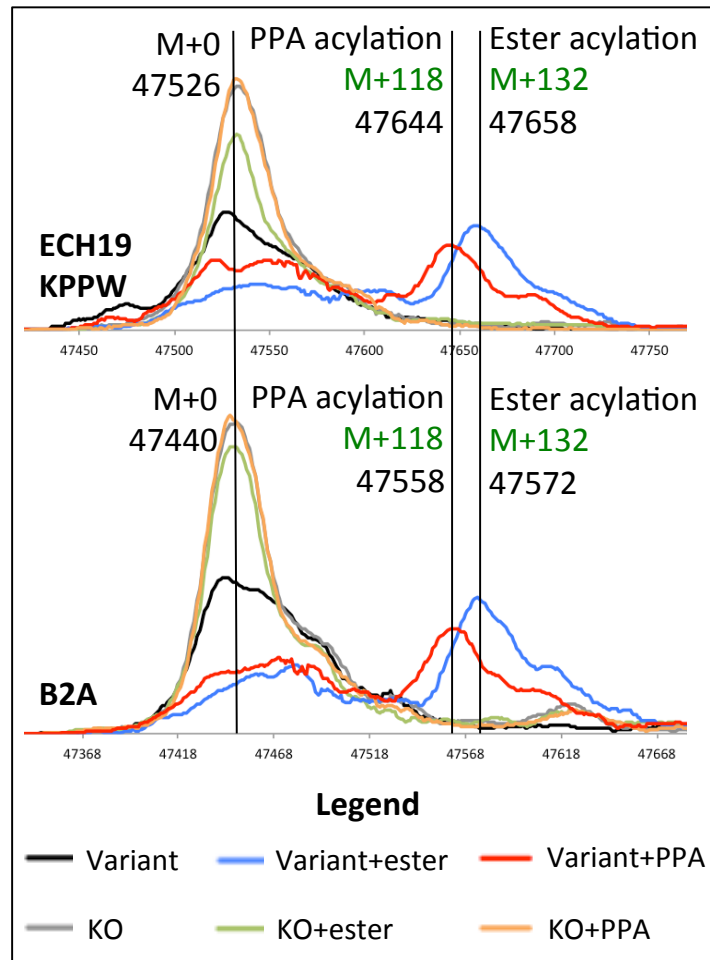


Figure 2.12: Mass spectrometry confirms hypothesized mechanism.

Each trace on graph represents a separately treated sample of esterase protein, masses are all in Da. In black is the variant with cysteine, not mixed with substrate. In blue is variant plus free-ester substrate, in red is variant plus PPA substrate. In gray is variant C->A knockout (KO), in green is the knockout plus free-ester substrate, in orange is the knockout plus PPA substrate. For oxidation (O_2 addition), I expect M+32 and see a shoulder at this mass (not labeled on chart.) For acylation by PPA substrate, I expect M+118. For acylation by free-ester substrate, I expect M+132.

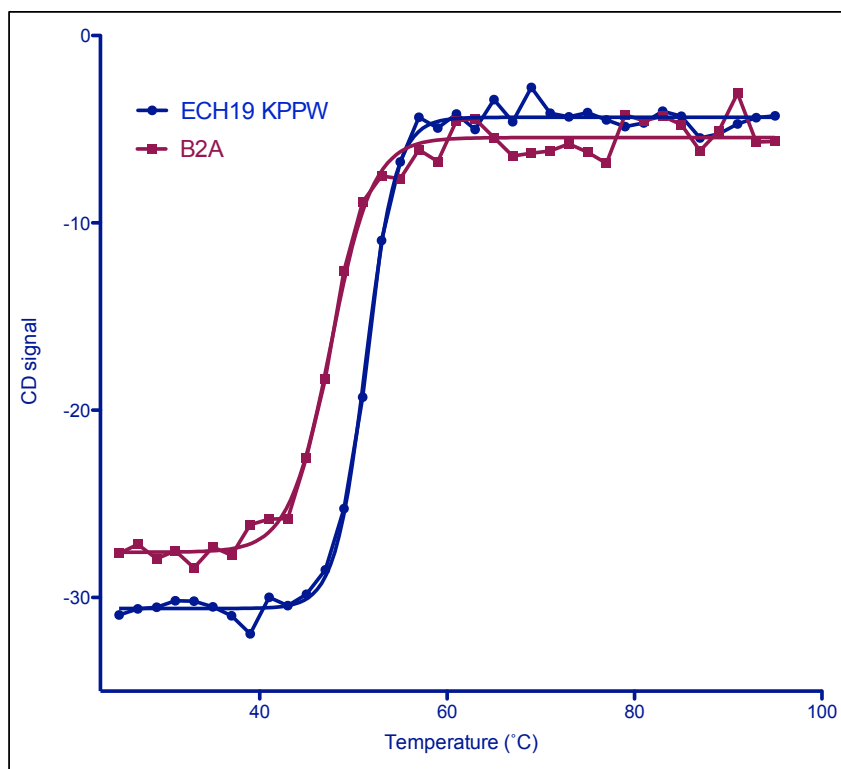


Figure 2.13: Protein melt shows only small difference in stability for evolved variant.

The protein melt was performed with 0.3 mg/mL pure protein. Calculated T_m for each protein was found to be: 51.3°C for ECH19 KPPW and 47.6°C for B2A.

	Enzyme variant	Mutation identity	$v_0/[E] \cdot 10^3 \text{ (s}^{-1}\text{)}$ at 10 uM sub.	$v_0/[E] \cdot 10^3 \text{ (s}^{-1}\text{)}$ at 75 uM sub.
	B2A	~	8.1	51.8
Reversion of mutations on B2A to KPPW identity	B1	R41W, N44H	5.0	37.5
	B2	N115T	2.4	24.1
	B3	A165T	5.9	55.8
	B4	I223K, K225N, I234N	5.7	53.7
	B5	S246T	2.3	21.6
MD informed interrogation of mutant position near catalytic cysteine	G2	A165V	8.3	56.6
	G3	A165S	4.0	28.3
	G4	A165G	2.4	19.9
	B3	A165T	5.9	55.8
	ECH19 KPPW	~	1.4	11.2
Forward mutations from KPPW to B2A	K1	W41R, H44N	1.9	26.4
	K2	T165A	1.9	19.3
	K3	K223I, N225K	2.0	20.2
	5A2D	~	3.7	24.7
Reversion mutations from 5A2D to KPPW	5.1	G116A	4.1	32.0
	5.2	A165T	3.3	26.1
	5.3	I223K, K225N	3.1	25.2
	5.4	K244N, A246T	1.6	14.7

Table 2.3: Kinetic parameters for mutations on evolved esterase variants.

The rate of the burst phase was measured in duplicate at 10 and 75 uM free-ester substrate and the average is reported here. All measurements with 0.75 uM protein.

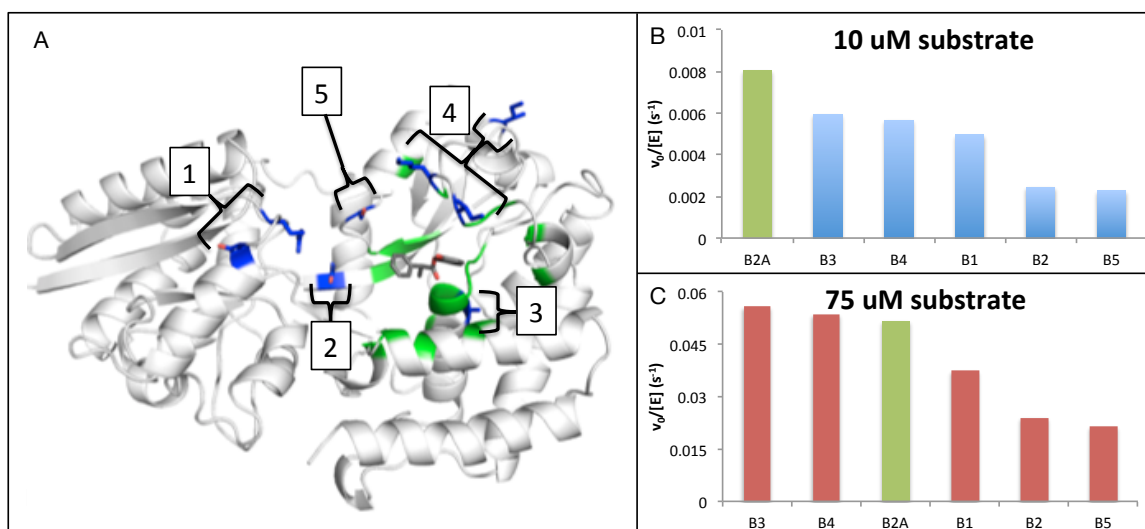


Figure 2.14: Three mutations contribute majority of improvement for variant B2A.

(A) Position of mutations from ECH19 KPPW to evolved variant B2A. Blue positions are mutated positions in B2A (from KPPW). Green are positions (not mutated) within 6 Å of modeled substrate. Black brackets and numbers indicate back-mutated regions. (B) Effect of back mutation (to ECH19 KPPW identity) at each of five positions at 10 uM free-ester substrate. (C) Activity of the same variants at 75 uM free-ester substrate.

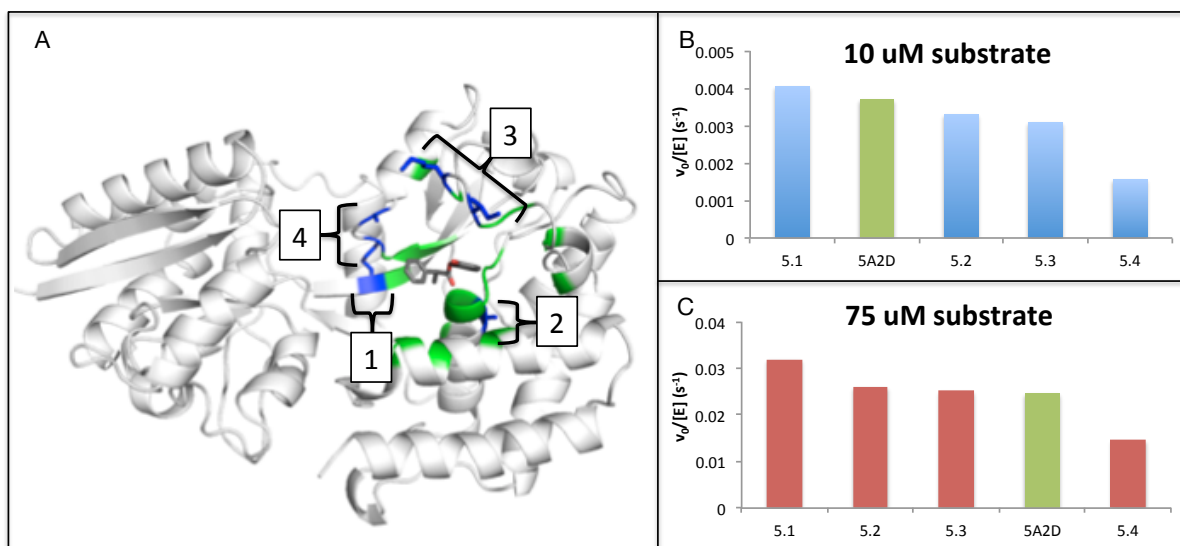


Figure 2.15: One mutation patch contributes the majority of improvement for variant 5A2D.

Mutation patch 5.4 is seen to have largest effect on activity at both low and high [free-ester]. (A) Mutations from ECH19 KPPW to evolved variant 5A2D. Blue positions are mutated positions in 5A2D (from KPPW). Green are positions (not mutated) within 6 Å of modeled substrate. Black brackets and numbers indicate back-mutated regions. (B) Effect of back mutation (to ECH19 KPPW identity) at each of four positions at 10 uM free-ester substrate. (C) Activity of same variants at 75 uM free-ester substrate.

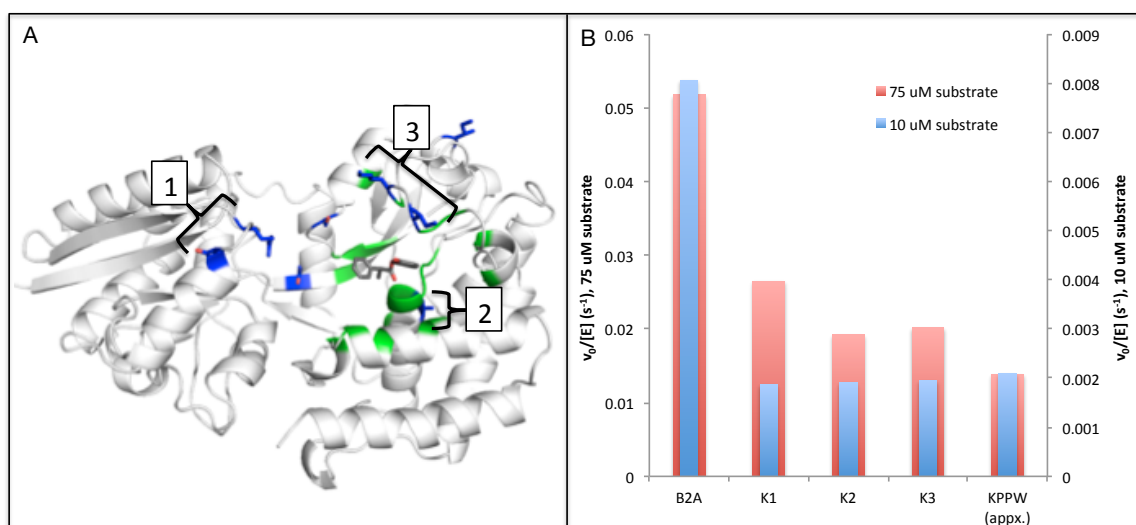


Figure 2.16: Mutations from B2A improve ECH19 KPPW activity only at high [substrate].

(A) Locations of mutations from ECH19 KPPW to evolved variant B2A. Blue positions are mutated positions in B2A (from KPPW). Green are positions (not mutated) within 6 Å of modeled substrate. Black brackets and numbers indicate forward-mutated regions. (B) Effect of forward mutation on ECH19 KPPW (to B2A identity) at each of three positions when protein is mixed with 75 uM (left Y axis, thick red bars) or 10 uM (right Y axis, thin blue bars) free-ester substrate.

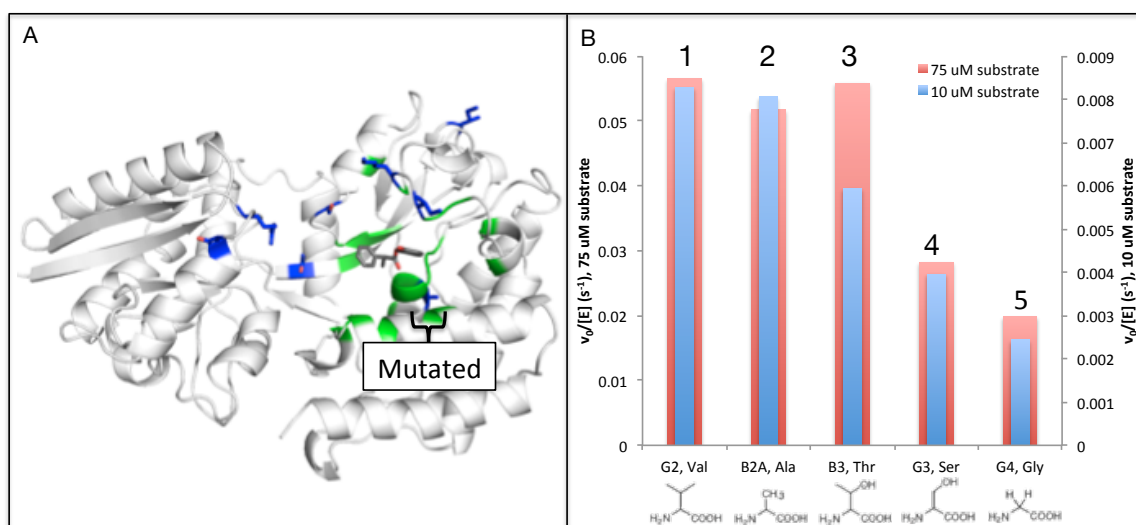


Figure 2.17: Mutated position adjacent to catalytic cysteine is near optimal.

(A) Locations of mutations from ECH19 KPPW to evolved variant B2A. Blue positions are mutated positions in B2A (from KPPW). Green are positions (not mutated) within 6 Å of modeled substrate. Black bracket indicates mutated position (165) directly below catalytic cysteine. (B) Effect of substitutions at mutated position on B2A activity when protein is mixed with 75 μM (left Y axis, thick red bars) or 10 μM (right Y axis, thin blue bars) free-ester substrate. Mutant 1 was predicted to improve activity by MD analysis. Mutant 2 is evolved variant B2A with no substitutions. Mutant 3 is B2A with substitution at mutant position to ECH19 KPPW identity. Mutants 4 and 5 are included to profile activity through series of small hydrophilic and hydrophobic amino acids.

	Enzyme variant	Mutation identity	$v_0/[E] \cdot 10^3 \text{ (s}^{-1}\text{)}$ at 10 uM sub.	$v_0/[E] \cdot 10^3 \text{ (s}^{-1}\text{)}$ at 75 uM sub.
A. Mutations from B2A to 5A2D	C1	N244K, S246A	6.19	34.9
	C2	N244K, S246A, N115T, A116G	6.14	34.7
	C3	N244K, S246A, A116G	7.37	38.8
B. Mutation from 5A2D to B2A	5B.1	W41R H44N	5.70	34.0

Table 2.4: Activity of variants with combinations of mutations from enzyme variants B2A and 5A2D.

(A) Activity of variants with mutations from 5A2D on the B2A template. (B) Activity of a variant with mutations from B2A on the 5A2D template.

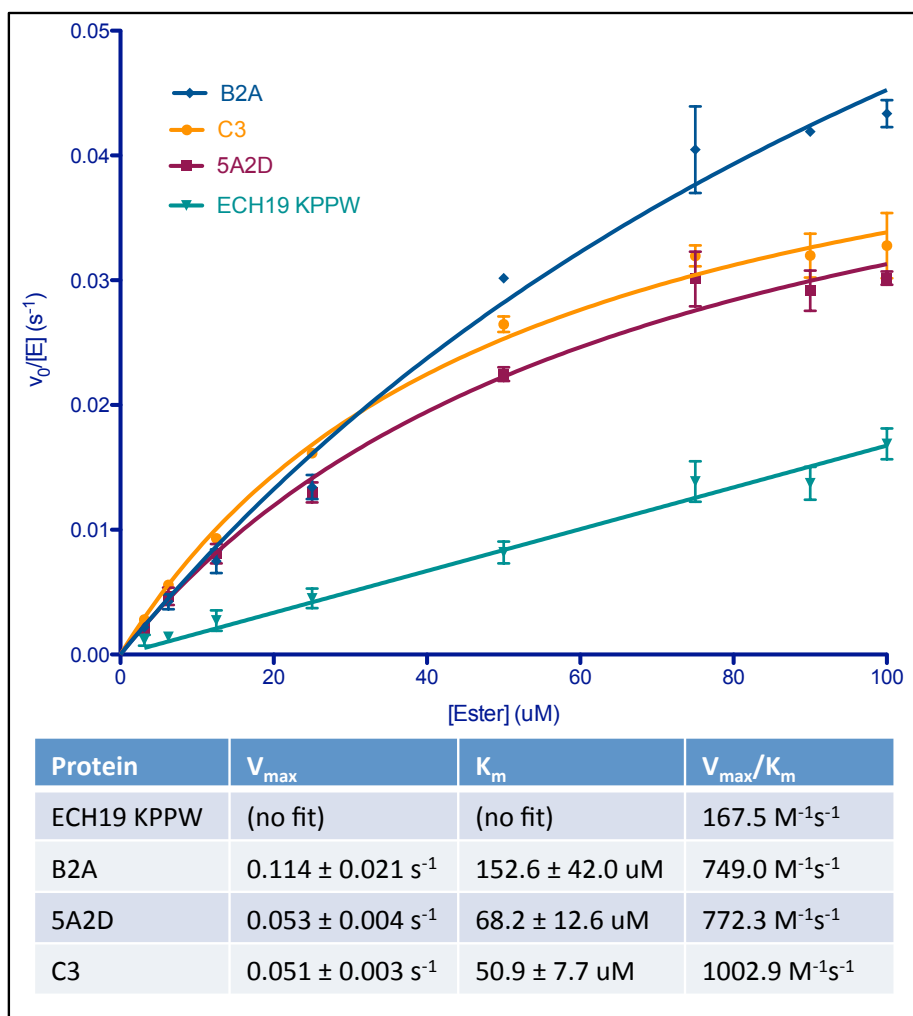


Figure 2.18: Michaelis-Menten kinetics may show further improvement in $V_{\max}/K_m$ for the combination of two evolved variants.

This combination mutant (including mutations from B2A and 5A2D) C3 is shown. $V_{\max}/K_m$ for the ancestor protein ECH19 KPPW was determined by linear regression. All rates are for the initial phase, kinetics measured as described in methods section.

Chapter 3: Computational Design and Directed Evolution of Novel Alkyltransferases

Introduction

One of the largest difficulties in moving enzyme design forward is the lack of simple and fast high-throughput selection technologies for enzymes. In sharp contrast is the case of protein-protein interface design—for this design problem, yeast surface display of very large libraries of design variants (containing 10^7 - 10^8 members) can be performed(32)(33). These libraries can then be selected, simply and rapidly, for binding to protein targets of interest(39). This has led to rapid feedback and improvement of the protein-protein interface design approach(40). I set out to identify an enzyme design target amenable to high-throughput selection to enable us to perform directed evolution experiments. By doing this I hope to be able to yield information which will allow us to advance the methods of computational enzyme design.

To identify a design target, I looked to previous directed evolution experiments on enzymes and enzyme-like proteins. One enzyme-like protein which has been the subject of extensive directed evolution is O6 alkylguanine DNA alkyltransferase(41). Its function in biology is to act as a sacrificial recipient of harmful DNA modifications. DNA becomes alkylated by various means and O6 alkylguanine DNA alkyltransferase (AGT) transfers this alkyl group off the DNA base and onto a nucleophilic cysteine in the AGT active site, repairing the DNA and covalently modifying the AGT protein (figure 3.4).

The self-modification ability of AGT has been used to create a number of very useful chemical biology tools. The Johnsson group has evolved AGT to self-label with arbitrary chemical groups,

most commonly fluorescent groups(42). This technology is known as the SNAP and CLIP tags. Instead of an alkylated base in genomic DNA, this protein's substrate is a modified DNA base. The AGT protein was evolved (using a yeast display selection and fluorescence-activated cell sorting as well as phage display) to self-label with a fluorescently labeled alkyl chain attached to a guanine base. In this work I present the design, characterization, and evolution of novel alkyltransferase proteins. I believe these proteins and this method for high-throughput selection will allow progress in improving enzyme design approaches.

Computational design

I set out to design a novel alkyltransferase protein with activity toward a new alkylating substrate. There are currently very few washless labeling techniques (where fluorescence is only generated in a sample when label contacts its target). Our collaborators in the Kai Johnsson lab developed an alkyltransferase substrate with both a fluorophore (Alexa 488) and a dye quencher (figure 3.1A). This substrate allows for washless labeling in cells and biological samples, due to a 50-fold increase in fluorescence upon protein alkylation (and leaving group loss). This more activated substrate allows for identification and improvement of low activity proteins that can then be evolved for activity with the quenched substrate. However, in previous enzyme design cases, the most success has been seen with very activated leaving groups on substrates, allowing identification of low-activity designs which can then be improved through redesign and directed evolution/screening. Working with the Johnsson group, we identified a synthetic intermediate (Figure 3.1B) would be an ideal activated substrate. With a paranitro-pyridol leaving group, this chemical would be both reactive and also have hydrogen bonding groups and aromatic groups to give additional binding energy for my designed proteins.

I studied the mechanism of the natural alkyltransferase protein DNA alkylguanine alkyltransferase (AGT), SNAP and CLIP, and also the general reaction these proteins catalyze, the SN2 displacement reaction. I reasoned that the reaction could be carried out by a cysteine-histidine-acid triad, rather than the water-bridged triad seen in the AGT structure. Studying the mechanism, I believe that the main factor in the high activity of AGT is the very nucleophilic cysteine residue and the extensive stabilization of negative charge on the DNA base leaving group. In my design approaches, I included features to address both of these factors.

My design approaches can be broken into a number of generations, each utilizing a cysteine as the nucleophile with a histidine placed to activate it. However, in each generation of design I took slightly different approaches to binding as well as transition state modeling, as I received feedback from my tests of designed proteins (design continued during and after the testing of initial designs). The combination of the transition state and catalytic machinery is called a theozyme. I used the RosettaMatch algorithm to match these theozymes into scaffolds culled from the PDB, selected to be monomers with reasonable resolution. I also filtered for proteins that were initially expressed in E coli, assuming this would improve my rate of soluble designs (by removing protein scaffolds that can only express in mammalian or fungal expression systems). I then matched the theozyme into the protein scaffold set and then performed Rosetta computational enzyme design as described in Chapter 1 of this dissertation and in previous work(*13*).

In a second case I set out to use already existing designed cysteine-histidine-acid triads from other enzyme design projects. This is a very common mechanism in our work (as well as in nature), and I hypothesized that triad-based designed enzymes with activity for other substrates

would be reasonable starting places for alkyltransferase design. I collected design structures of active triad-based designed enzymes and used RosettaDock for small molecules to test compatibility with the NBP ligand. For those with complementary active sites, I then either redesigned the active site using Rosetta design, or performed RosettaMatch again to place the triad and the substrate in the active site, and then performed design. All designs were filtered for interface and theozyme constraint scores, and the best were tested as purified proteins.

Screening for alkyltransferase activity

With the designed proteins I now had to test their expression, solubility, and activity. Genes containing the sequence for each of 87 genes were cloned into either pET29b+ (Novagen), pET-FLAG, pET21_NESG, or pET15_HE. The proteins were then expressed and purified using Ni-NTA affinity for the 6xHis tag. Because the activity of the protein is to covalently self-label with a fluorophore, I decided to test by reacting each purified design with the NBP substrate, then running an SDS-PAGE gel and imaging with fluorescence, as done in previous work to characterize the SNAP and CLIP tags. A normalized amount of each purified protein was mixed with NBP substrate, incubated, and run on an SDS-PAGE gel. To test activity, the gel was imaged for fluorescence from the fluorescein group on the NBP substrate and then stained to verify protein concentration in the reaction. The results of these tests can be seen in figures 3.3-6. Knockouts of early active variants can be seen in figure 3.4. I selected the proteins with highest apparent activity (as measured by relative labeling versus the quantity of protein) for testing alongside later designs, in order to ascertain activity across batches. In figure 3.7 I show the result of this testing the most active designs together.

These highest activity variants were retested along with their triad cysteine to alanine and histidine to alanine knockouts, to verify that the observed activity came from the designed active site. Figure 3.8 shows the results of this experiment, where 4 out of the 5 proteins tested showed full loss of activity with cysteine to alanine mutation. For the histidine to alanine mutation, much less loss of activity was seen, with only one of the 4 proteins with histidine knockouts purified showing any loss of activity. This led us to hope that significant improvements in activity could be yielded from directed evolution, especially focused around the active site. If mutations could be found to strengthen the interactions in the cysteine-histidine dyad, I hypothesized that significant improvements in activity would result.

Directed evolution through yeast display

The enzymes that were identified as the most active in fluorescent gel screens were then prepared for directed evolution. The *de novo* design AT44 as well as repurposed triads Ami6 and Ami18 were selected. I constructed focused libraries on regions around the triad and substrate. I used a library creation technique similar to that used to construct the doped libraries of ester hydrolase ECH19 KPPW in Chapter 2 of this dissertation. In short, I selected between 16 and 22 amino acid positions in these regions (as shown in figure 3.10), and then reconstructed the gene using doped oligos as well as the gene itself, as shown in figure 3.9. Library construction using doped oligos in this way allows a precise and sequence-specific level of control over the final library. In this way I can explore a greater area of sequence space in a region (with randomization focused around the substrate and catalytic triad) most likely to give improvements in activity. The libraries are then cloned into a yeast display vector, transformed into yeast, and selected for improved labeling with NBP using FACS (similar to the selection in Chapter 2 of

this thesis). The amount of protein displayed on the yeast surface is also monitored and selected for using a fluorescently labeled Anti-Myc antibody (that binds to a Myc-tag on each displayed protein). In this way I can select for improved alkyltransferase proteins.

While selection and testing of these libraries is ongoing, some improvement has been seen over the four rounds of selection for these three libraries. Notably, the library on Ami6 has shown improvement in labeling with the NBP substrate. Initial sequencing of the selected pool (data not shown here) indicates that this improvement is likely from mutations to the Ami6 gene at the targeted positions, and is not the result of contamination with the positive control, contaminating microorganisms, or other genes.

Discussion

In this work I present the computational design, experimental characterization, and directed evolution of novel alkyltransferase proteins. While characterization of the directed evolution experiment is ongoing, the proteins have already proved to be a useful source of insight into computational design. Specifically, the usefulness of the amidase triads demonstrates that reuse of catalytic machinery (even from designed enzymes) to design new enzymes is a good strategy. Indeed, we see this strategy in nature, where promiscuous enzymes are often repurposed toward new activities⁽⁴³⁾⁽⁴⁴⁾, and then evolved toward specificity (and higher activity) toward their new substrate.

The alkyltransferase proteins from this work may be evolved further toward activity with the quenched substrate. This would allow both insights into how specificity can be switched for computationally designed enzymes. A protein that allows washless labeling would also be useful as a new tool in chemical and molecular biology. By continuing to develop novel enzymes

that are amenable to high-throughput selection and directed evolution, we can generate the data and insights needed to improve computational enzyme design methods even further.

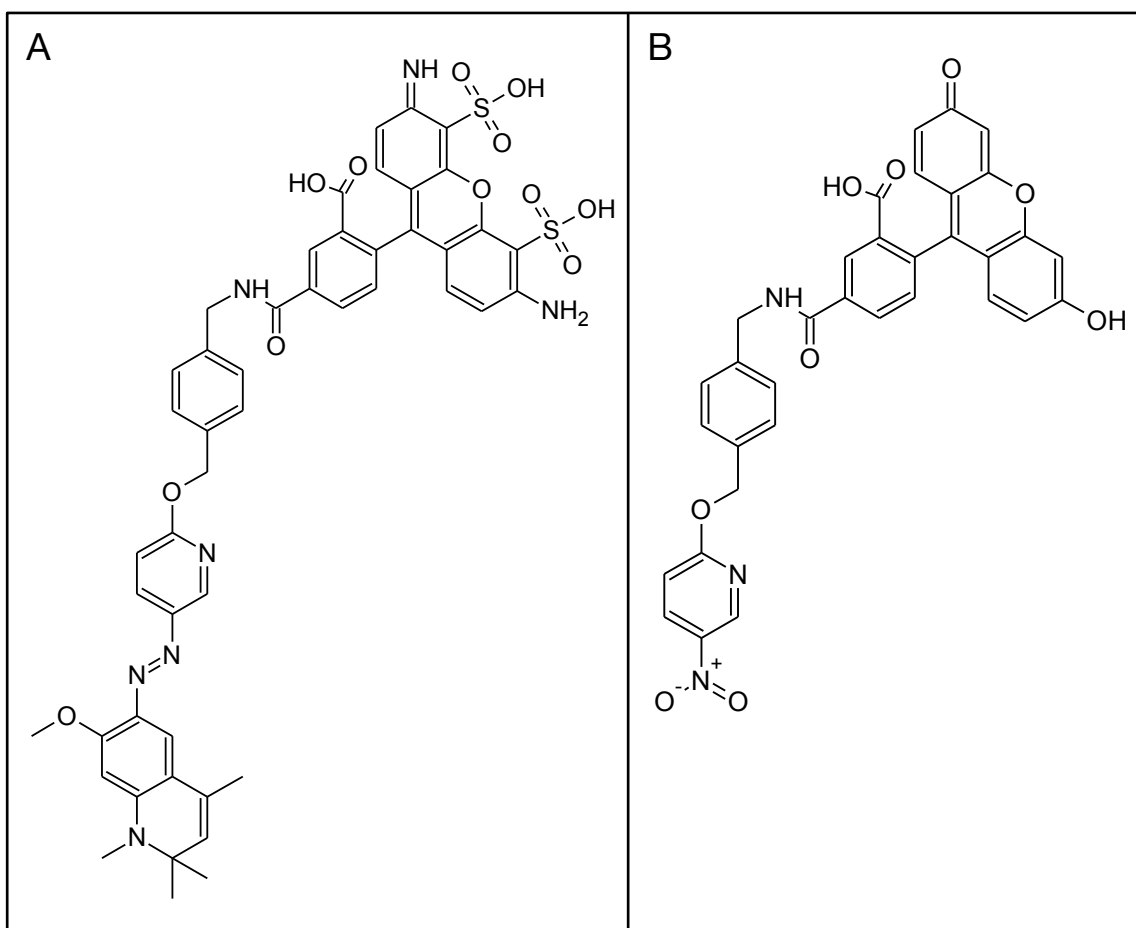


Figure 3.1: Quenched and activated substrates for alkyltransferase design and directed evolution.

(A) Quenched substrate (QBP) for alkyltransferase reaction. This substrate allows for washless labeling in cells and biological samples, due to a 50-fold increase in fluorescence upon protein alkylation (and leaving group loss). (B) Activated substrate (NBP) used in this work for design, testing, and evolution. This substrate has a fluorescein moiety for easy detection of alkyltransferase activity. This more activated substrate allows for identification and improvement of low activity proteins that can then be evolved for activity with the quenched substrate.

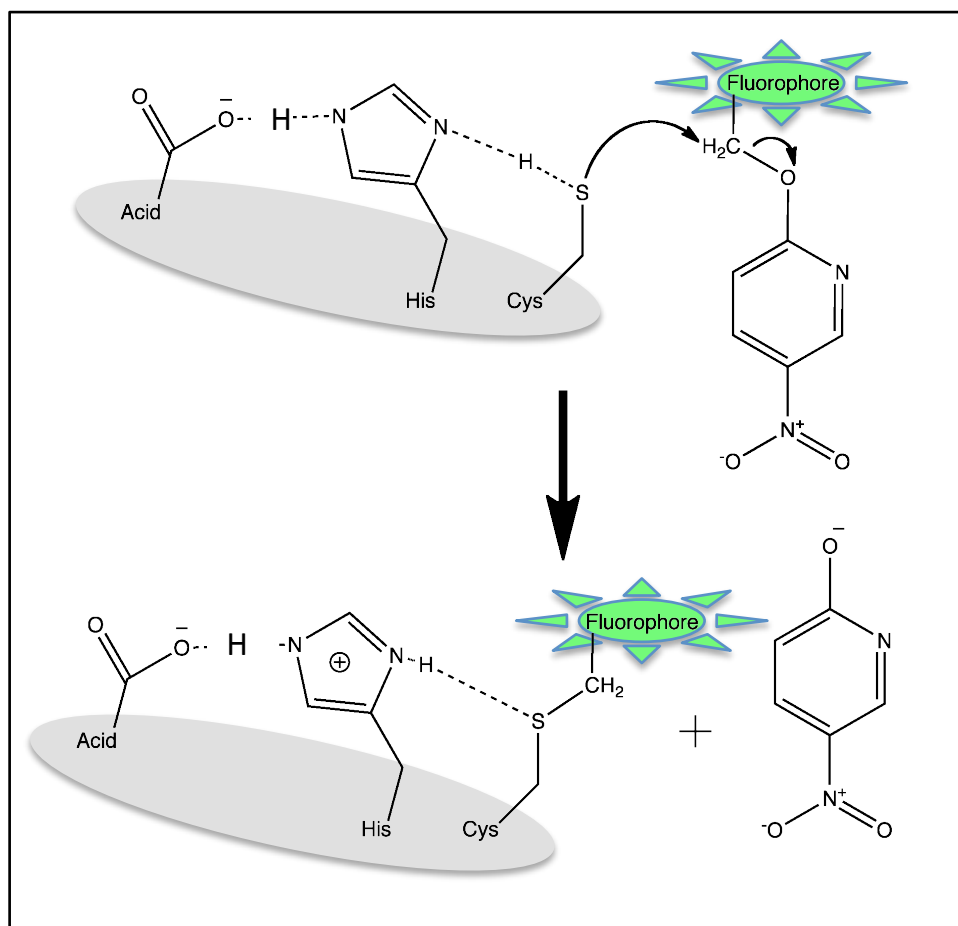


Figure 3.2: Cysteine-histidine-acid catalytic triad and proposed mechanism used in *de novo* design.

The nucleophilic sulfur in the cysteine is shown performing an S_N2 displacement reaction, the mechanism for alkyltransferase reactions. I used this mechanism as the basis of my theozyme and matching efforts.

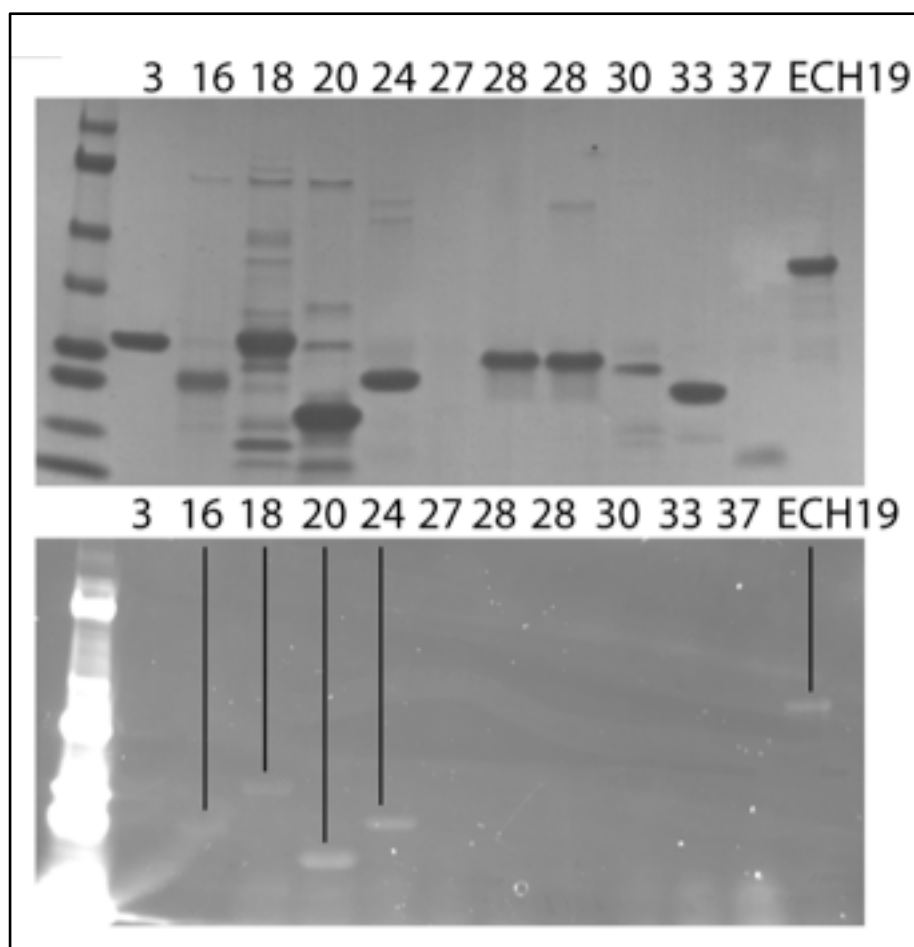


Figure 3.3: Initial *de novo* designed proteins purify solubly and some have low levels of activity, near that of a cysteine-containing negative control.

Numbers on gel correspond to *de novo* design numbering (eg. AT3, AT16) as used elsewhere. These gels show all proteins solubly expressed in the range of designs AT1-AT40. The top gel shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each design was tested with NBP substrate at 5 μ M, with labeling for 16 h. The cysteine esterase ECH19 KPPW was included as a negative control, a protein with an activated cysteine but which was not designed as an alkyltransferase.

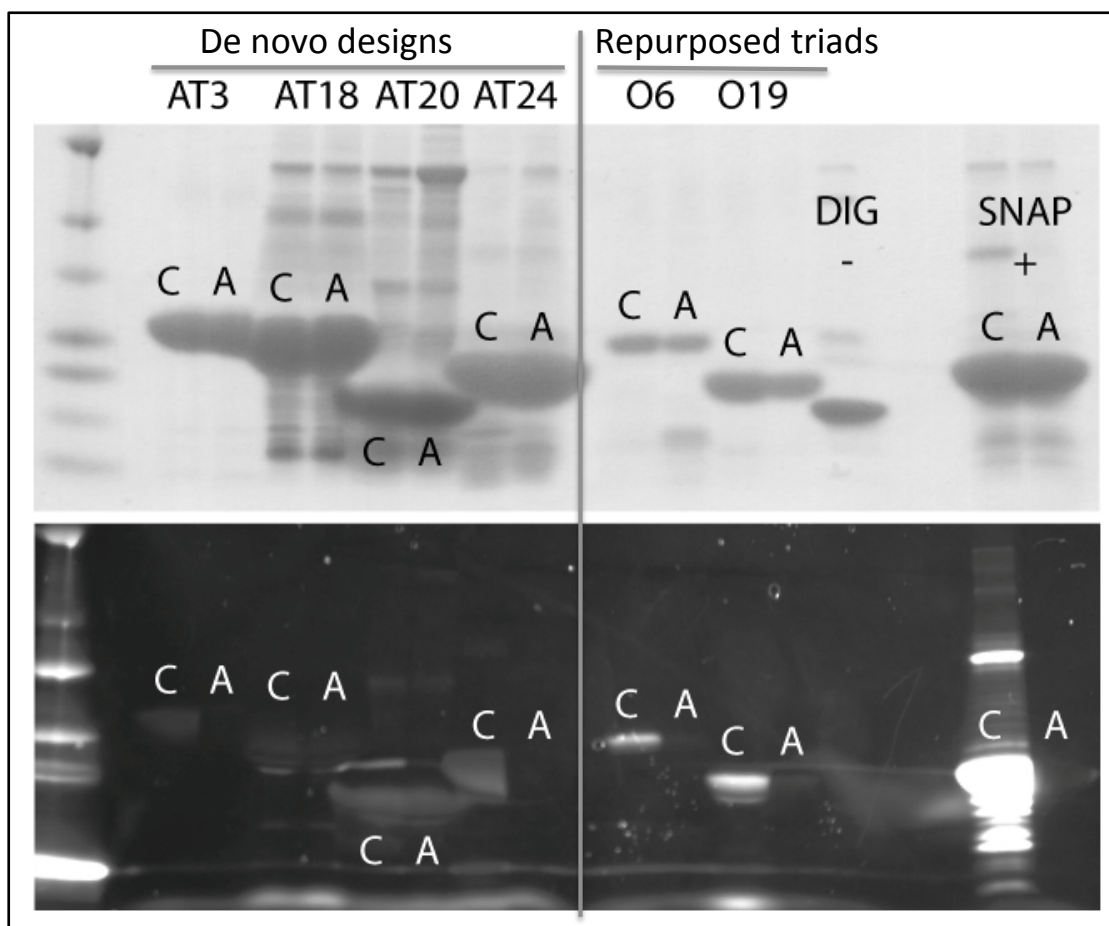


Figure 3.4: *De novo* designed and repurposed triad proteins purify solubly, most knock out with cysteine to alanine mutation.

Original design or its cysteine to alanine mutant is indicated by the letter “C” or “A” on the gel. The top gel of each pair shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each design was tested with NBP substrate at 10 uM, with labeling for 16 h. SNAP was included as a positive control.

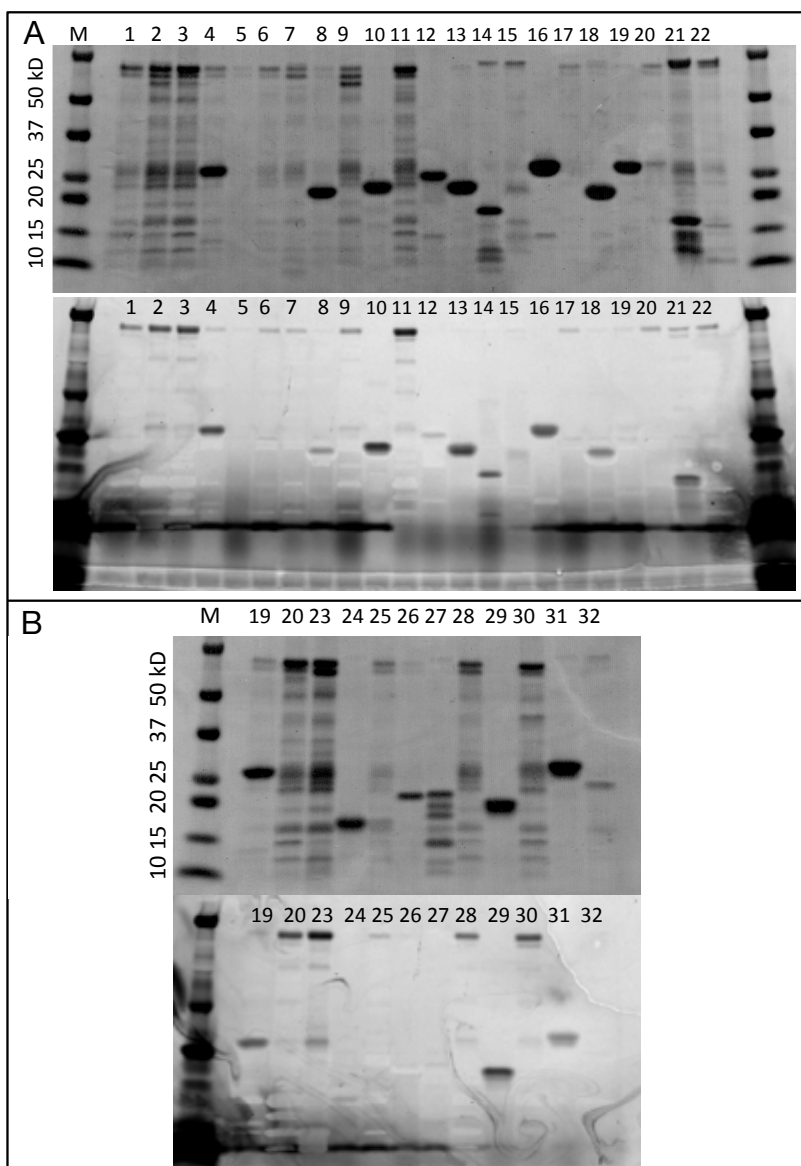


Figure 3.5: Design batch 1: *De novo* designed and repurposed triad proteins purify solubly, and some have enzyme activity.

Numbers on gel correspond to designs identified in table 3.1. The top gel of each pair shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each design was tested with NBP substrate at 10 μ M, with labeling for 16 h. Samples 19 and 20 are repeated from (A) to (B) as internal controls.

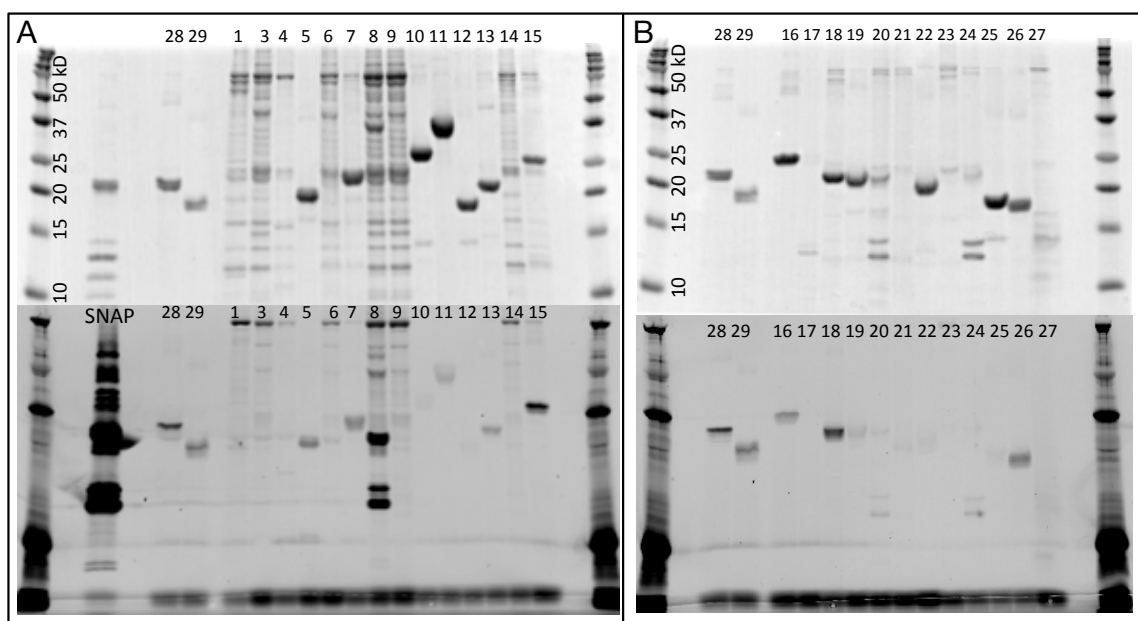


Figure 3.6: Design batch 2: *De novo* designed and repurposed triad proteins purify solubly, and some have enzyme activity.

Numbers on gel correspond to designs identified in table 3.2. The top gel shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each design was tested with NBP substrate. Sample 28 and 29 are repeated from (A) to (B) as internal controls. The alkyltransferase SNAP is included as a positive control. Variant 8, while having apparent activity as visualized with fluorescence, is not expected to be a true alkyltransferase, as it was later determined that the scaffold protein likely has related activity and is promiscuous towards NBP alkylation.

Gel number	Design name
1	AT55
2	AT63
3	AT24
4	AT28
5	AT41o
6	AT44
7	AT48
8	AT51
9	AT69_1via
10	AT74_1s17
11	AT78_1e4y
12	AT41
13	AT44
14	AT48
15	AT51
16	AT52
17	AT69_1via
18	AT76_2iyy
19	AT78_1e4y
20	AT84_1yre
21	AT42h
22	AT43
23	AT43h2
24	AT45
25	AT46h
26	AT47h
27	AT49
28	AT50h
29	Ami6
30	AT58
31	AT18
32	AT30

Table 3.1: Identities of batch 1 designs.

Protein numbers on gels in figure 3.5 correspond to designs identified in this table. Designs with a pdb code in their name are repurposed triad designs, all other AT prefixed designs are *de novo* designs.

Gel number	Variant name
1	AT41h
~	SNAP
3	AT53
4	AT54
5	AT56
6	AT57
7	AT60
8	AT61
9	AT62
10	AT64
11	AT65
12	AT33
13	AT67
14	AT71
15	Ami6
16	Ami35
17	Ami36
18	Ami42
19	Ami43
20	Ami44
21	Ami48
22	AT70_1s17
23	AT72_1via
24	AT73_1s17
25	AT33
26	AT27
27	AT16

Table 3.2: Identities of batch 2 designs.

Protein numbers on gels in figure 3.6 correspond to designs identified in this table. Variants with an underscore and pdb code in their name as well as “Ami” prefixed variants are repurposed triad designs, all other AT prefixed designs are *de novo* designs.

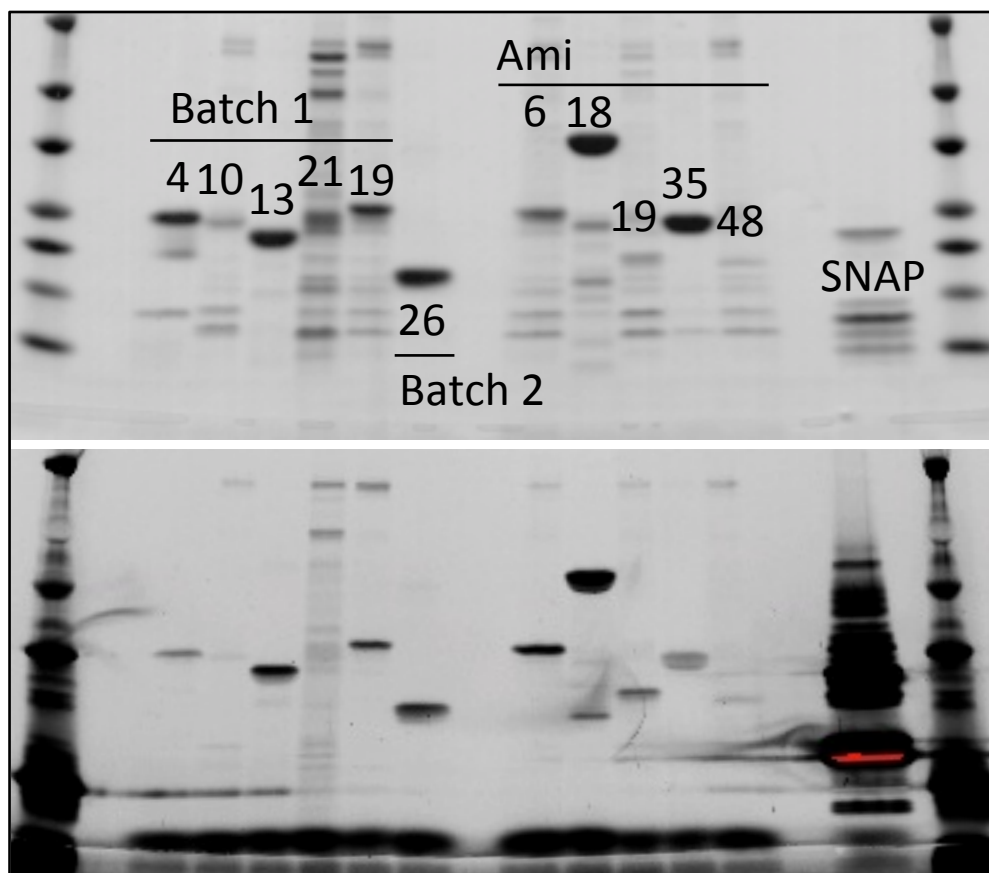


Figure 3.7: Repeat of purification and test confirms activity of best designed variants.

The top gel shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each design and the positive control SNAP protein was tested with NBP substrate. Batch 1 corresponds to designs with tests shown in figure 3.5 and identified in table 3.1; batch 2 corresponds to one design with test shown in figure 3.6 and identified in table 3.2. Some but not all of the “Ami” variants are also tested in figure 3.5 and 3.6.

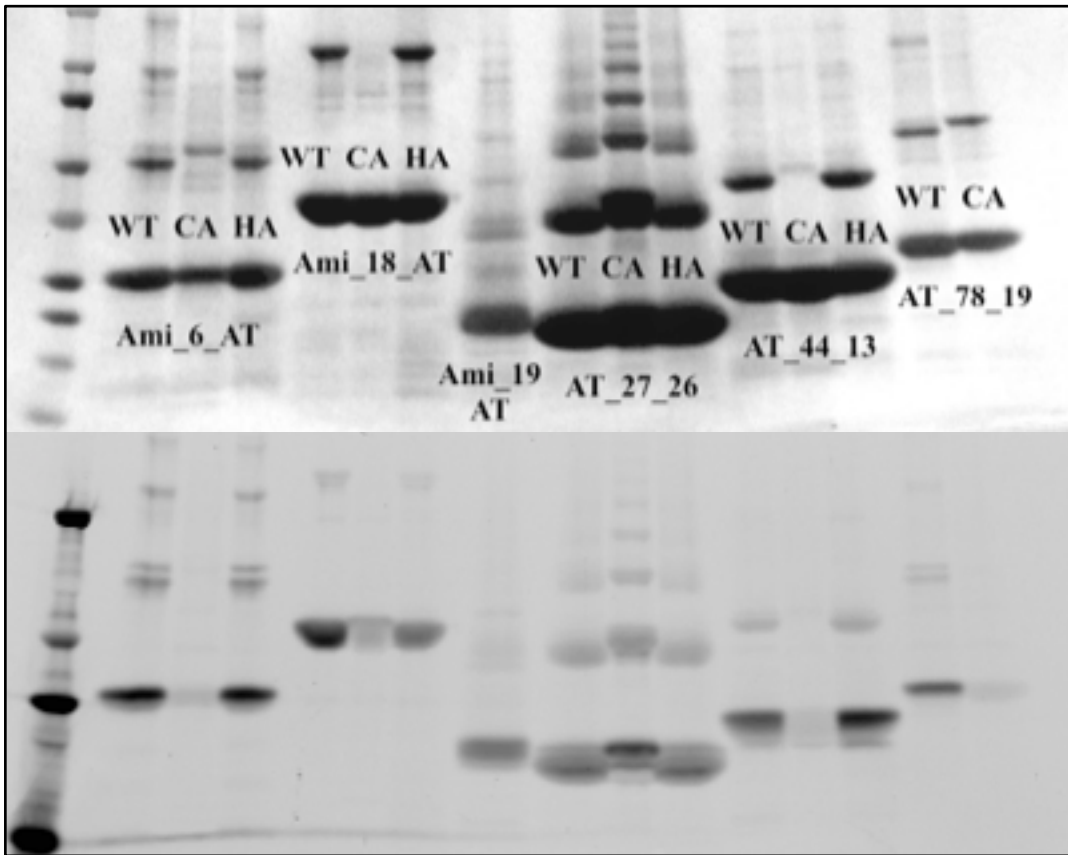


Figure 3.8: Cysteine and histidine dyad knockouts confirm mechanism works as expected for some variants.

The top gel shows coomassie (protein) staining and the bottom shows fluorescence (alkyltransferase activity). Each of the most active designs were tested along with alanine knockouts of the designed triad cysteine and histidine. Some cysteine (Ami 19) and histidine (Ami 19, AT78) knockouts did not express solubly. Ami 18 shows the expected pattern of loss of activity with cysteine knockout and diminished activity with histidine knockout.

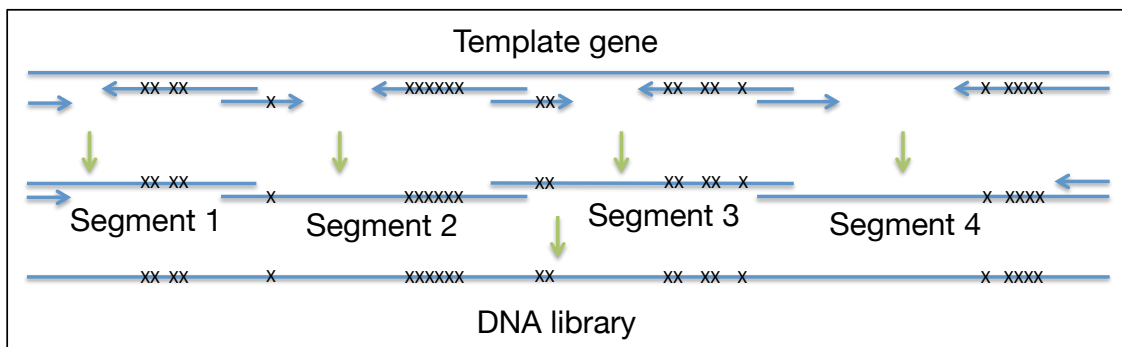


Figure 3.9: Library construction strategy for alkyltransferase selection.

Oligonucleotides with doped regions are used to construct the gene in overlapping segments through PCR. These segments are then assembled into the full gene with precisely specified randomization (made possible with the use of doped oligonucleotides). More than 4 segments are used to assemble the three libraries constructed here, but is presented in this manner here for the sake of simplicity.

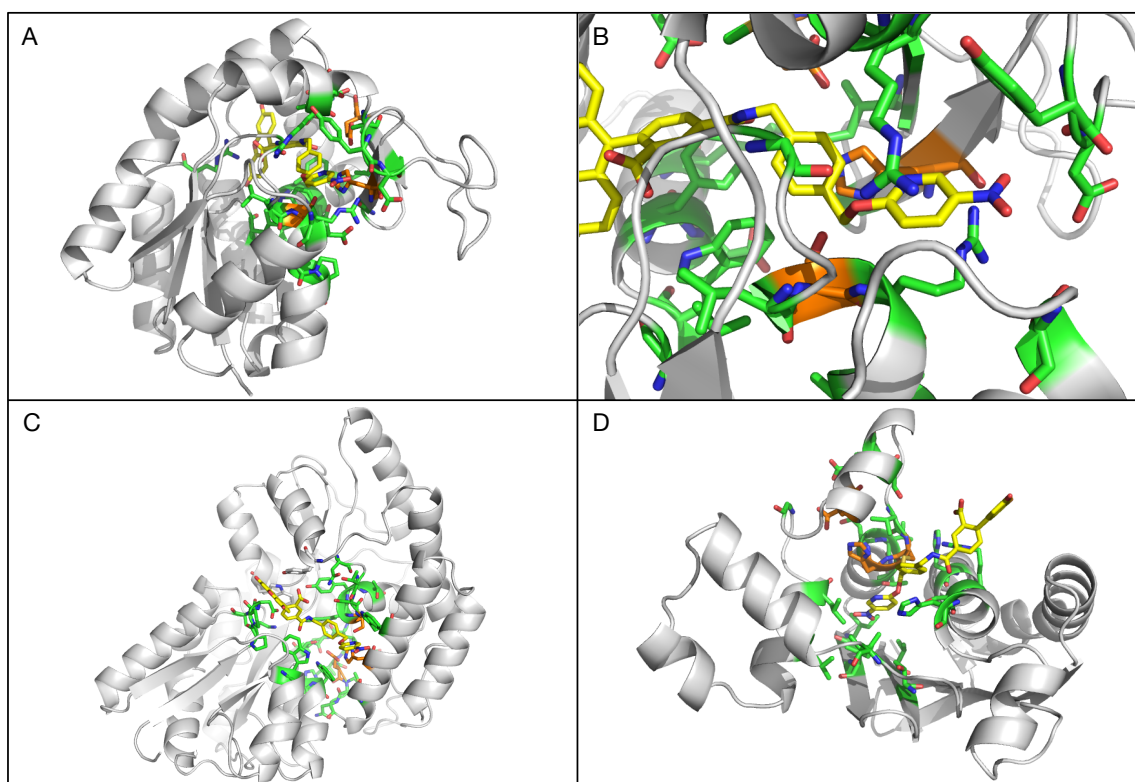


Figure 3.10: Regions randomized in three designed alkyltransferases.

16 to 22 amino acid positions around the substrate and catalytic triad in each protein were chosen for randomization through the use of doped oligo assembly. The catalytic triad is shown in orange, randomized positions in green, and the NBP substrate in yellow. (A) shows the randomized regions in Ami 6, (B) shows a detail of the Ami 6 active site, (C) shows the randomized regions in Ami18, and (D) shows the randomized regions in AT44.

Bibliography

1. E.A. Althoff, L. Wang, L. Jiang, et al. (2012) Robust design and optimization of retroaldol enzymes, *Robust design and optimization of retroaldol enzymes*, Protein Science, Protein Science. 21, 21, 717, 717–726, 726.
2. D. Röthlisberger, O. Khersonsky, A.M. Wollacott, et al. (2008) Kemp elimination catalysts by computational enzyme design, *Nature*. 453, 190–5.
3. F. Richter, R. Blomberg, S.D. Khare, et al. (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis, *Journal of the American Chemical Society*. 134, 16197–16206.
4. J.B. Siegel, A. Zanghellini, H.M. Lovick, et al. (2010) Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction, *Science*. 329, 309–313.
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, 743–750.
6. S.P. Kim, A.G. Leach, and K.N. Houk (2002) The origins of noncovalent catalysis of intermolecular Diels-Alder reactions by cyclodextrins, self-assembling capsules, antibodies, and RNAses, *The Journal of organic chemistry*. 67, 4250–4260.
7. J.F. Blake, D. Lim, and W.L. Jorgensen (1994) Enhanced Hydrogen Bonding of Water to Diels-Alder Transition States. Ab Initio Evidence, *The Journal of Organic Chemistry*. 59, 803–805.
8. R. Breslow and S.D. Dong (1998) Biomimetic Reactions Catalyzed by Cyclodextrins and Their Derivatives, *Chemical reviews*. 98, 1997–2012.
9. J.T. Yli-Kauhaluoma, J.A. Ashley, C.-H. Lo, et al. (1995) Anti-Metallocene Antibodies: A New Approach to Enantioselective Catalysis of the Diels-Alder Reaction, *Journal of the American Chemical Society*. 117, 7041–7047.
10. A. Zanghellini, L. Jiang, A.M. Wollacott, et al. (2006) New algorithms and an in silico benchmark for computational enzyme design, *Protein science : a publication of the Protein Society*. 15, 2785–94.
11. W. Besenmatter, P. Kast, and D. Hilvert (2007) Relative tolerance of mesostable and thermostable protein homologs to extensive mutation, *Proteins*. 66, 500–506.
12. H.M. Berman, T. Battistuz, T.N. Bhat, et al. (2002) The Protein Data Bank, *Acta crystallographica. Section D, Biological crystallography*. 58, 899–907.
13. F. Richter, A. Leaver-Fay, S.D. Khare, et al. (2011) De Novo Enzyme Design Using Rosetta3, *PLoS ONE*. 6, e19230.
14. B. Kuhlman, G. Dantas, G.C. Ireton, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy, *Science (New York, N.Y.)*. 302, 1364–8.
15. B. Kuhlman and D. Baker (2000) Native protein sequences are close to optimal for their structures, *Proceedings of the National Academy of Sciences*. 97, 10383–10388.

16. A. Leaver-Fay, M. Tyka, S.M. Lewis, et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, *Methods in Enzymology*. 487, 545–574.
17. L. Jiang, E.A. Althoff, F.R. Clemente, et al. (2008) De Novo Computational Design of Retro-Aldol Enzymes, *Science*. 319, 1387–1391.
18. S. Bjelic, L.G. Nivón, N. Çelebi-Ölçüm, et al. (2013) Computational design of enone-binding proteins with catalytic activity for the Morita-Baylis-Hillman reaction, *ACS chemical biology*. 8, 749–757.
19. S.R. Gordon, E.J. Stanley, S. Wolf, et al. (2012) Computational design of an α -gliadin peptidase, *Journal of the American Chemical Society*. 134, 20513–20520.
20. S.D. Khare, Y. Kipnis, P.J. Greisen, et al. (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis, *Nature Chemical Biology*. 8, 294–300.
21. S.J. Wu, C.B. Eiben, J.H. Carra, et al. (2011) Improvement of a potential anthrax therapeutic by computational protein design, *The Journal of Biological Chemistry*. 286, 32586–32592.
22. L. Liu, P. Murphy, D. Baker, et al. (2009) Computational design of orthogonal nucleoside kinases, *Chemical Communications*. 46, 8803–8805.
23. G. Kiss, N. Çelebi-Ölçüm, R. Moretti, et al. (2013) Computational Enzyme Design, *Angewandte Chemie International Edition*. 52, 5700–5725.
24. O. Khersonsky, G. Kiss, D. Röthlisberger, et al. (2012) Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59, *Proceedings of the National Academy of Sciences of the United States of America*. 109, 10358–10363.
25. E.A. Althoff, L. Wang, L. Jiang, et al. (2012) Robust design and optimization of retroaldol enzymes, *Protein Science*. 21, 717–726.
26. L. Giger, S. Caner, R. Obexer, et al. (2013) Evolution of a designed retro-aldolase leads to complete active site remodeling, *Nature chemical biology*.
27. O. Khersonsky, D. Röthlisberger, A.M. Wollacott, et al. (2011) Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution, *Journal of Molecular Biology*. 407, 391–412.
28. J.-P. Goddard and J.-L. Reymond (2004) Enzyme assays for high-throughput screening, *Current Opinion in Biotechnology*. 15, 314–322.
29. M. Sunbul, N.J. Marshall, Y. Zou, et al. (2009) Catalytic Turnover-Based Phage Selection for Engineering the Substrate Specificity of Sfp Phosphopantetheinyl Transferase, *Journal of Molecular Biology*. 387, 883–898.
30. G. Yang, J.R. Rich, M. Gilbert, et al. (2010) Fluorescence Activated Cell Sorting as a General Ultra-High-Throughput Screening Method for Directed Evolution of Glycosyltransferases, *Journal of the American Chemical Society*. 132, 10570–10577.
31. H. Leemhuis, R.M. Kelly, and L. Dijkhuizen (2009) Directed evolution of enzymes: Library screening strategies, *IUBMB Life*. 61, 222–228.

32. G. Chao, W.L. Lau, B.J. Hackel, et al. (2006) Isolating and engineering human antibodies using yeast surface display, *Nat. Protocols*. 1, 755–768.
33. L. Benatuil, J.M. Perez, J. Belk, et al. (2010) An improved yeast transformation method for the generation of very large human antibody libraries, *Protein Engineering Design and Selection*.
34. A. Rothe, R. Surjadi, and B. Power (2006) Novel proteins in emulsions using in vitro compartmentalization, *Trends in Biotechnology*. 24, 587–592.
35. A. Aharoni, G. Amitai, K. Bernath, et al. (2005) High-Throughput Screening of Enzyme Libraries: Thiolaconases Evolved by Fluorescence-Activated Sorting of Single Cells in Emulsion Compartments, *Chemistry & Biology*. 12, 1281–1289.
36. H. Shim, A. Karlström, S.M. Touami, et al. (2004) Flow cytometric screening of aldolase catalytic antibodies, *Bioorganic & Medicinal Chemistry Letters*. 14, 4065–4068.
37. B. Seelig and J.W. Szostak (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold, *Nature*. 448, 828–831.
38. S.J. Fleishman, S.D. Khare, N. Koga, et al. (2011) Restricted sidechain plasticity in the structures of native proteins and complexes, *Protein science: a publication of the Protein Society*. 20, 753–757.
39. S.J. Fleishman, T.A. Whitehead, D.C. Ekiert, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin, *Science (New York, N.Y.)*. 332, 816–821.
40. T.A. Whitehead, D. Baker, and S.J. Fleishman (2013) Computational design of novel protein binders and experimental affinity maturation, *Methods in enzymology*. 523, 1–19.
41. A.E. Pegg (2000) Repair of O6-alkylguanine by alkyltransferases, *Mutation Research/Reviews in Mutation Research*. 462, 83–100.
42. T. Gronemeyer, C. Chidley, A. Juillerat, et al. Directed evolution of O6-alkylguanine-DNA alkyltransferase for applications in protein labeling, *Protein Engineering Design and Selection*. 19, 309–316.
43. M. Ben-David, G. Wieczorek, M. Elias, et al. (2013) Catalytic metal ion rearrangements underline promiscuity and evolvability of a metalloenzyme, *Journal of molecular biology*. 425, 1028–1038.
44. O. Khersonsky and D.S. Tawfik (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective, *Annual review of biochemistry*. 79, 471–505.