



Next-generation enzyme design:
Incorporation of multiple conformational and chemical states

Niayesh Zarifi

A thesis submitted in partial fulfillment of the requirements for the
Doctorate in Philosophy degree in Chemistry

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

ABSTRACT

The design of tailor-made enzymes for new-to-nature reactions holds significant industrial potential. Despite advancements in *de novo* enzyme design, designed enzymes often exhibit catalytic efficiencies far lower than their natural counterparts, necessitating directed evolution to enhance their activities. This thesis aims to improve enzyme design methods to enable the creation of efficient enzymes without reliance on directed evolution by addressing key challenges in current approaches. Firstly, existing methods focus solely on the active site and do not include distal sites, whose contribution to catalysis is not well understood. In Chapter 2, we demonstrate that while active-site mutations alone can create ordered, efficient active sites, distal mutations can reorganize surface loops and widen tunnels to the active site, facilitating substrate binding and product release. The second key challenge is the availability of suitable scaffolds for enzyme design. In Chapter 3, we investigate minimal, *de novo*, and functionless TIM barrels as versatile templates customizable for any reaction. Using physics-based and machine-learning tools, we modified a simple TIM barrel and introduced structural segments to enclose the active site, creating a functional Kemp eliminase. The third challenge we address is the design of efficient enzymes for multistep reactions, which requires stabilizing multiple states along the reaction pathway. In Chapter 4, we present the first instance of multistate enzyme design, where we stabilized multiple chemical and conformation states to create *de novo* Michaelases. Our results suggest that to design efficient artificial enzymes, a holistic approach is required wherein the design procedure includes the whole reaction coordinate. This thesis advances computational enzyme design by unravelling the impact of distal mutations, evaluating the utility of machine-learning tools for enzyme design, and incorporating multistate approaches to design enzymes for multistep reactions. Our work lays the foundation for the design of highly efficient enzymes for diverse applications.

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Dr. Roberto Chica, for his support, guidance, and motivation throughout my graduate studies. Your mentorship has been invaluable, and working with you has truly been a pleasure.

I also thank our collaborators from UCSF, Dr. James Fraser and Dr. Pooja Asthana, for crystallizing Core and Shell variants. I sincerely thank Dr. Sílvia Osuna and Hiva Doostmohammadi for performing MD simulations on the HG series and Dr. Marc Garcia-Borràs for his mentorship on DFT calculations. I would also like to thank our collaborators from the University of Bayreuth, Dr. Birte Höcker, Julian Beck, and Emily Freund, for their contributions to designing Kemp eliminases on TIM barrels.

A heartfelt thank you goes to my incredible lab members, Dr. Rojo Rakotoharisoa, Serena Hunt, Dr. Hang Pham, Benjamin Smith, Jack Miller, and Dr. Cindy Klaus. Your support and encouragement made this journey truly enjoyable.

Lastly, I would like to extend my deepest appreciation to my family. To my parents, Ensieh Barmakhgou and Davoud Zarifi, your unconditional love and support have been my greatest source of strength. To my sisters, Setayesh and Aramesh, your love, humour, and constant belief in me have kept me motivated. I am incredibly fortunate to have you all by my side.

TABLE OF CONTENTS

Abstract.....	ii
Acknowledgements	iii
Table of Contents	iv
List of Figures.....	x
List of Tables.....	xv
List of Abbreviations.....	xviii
Chapter 1. Introduction.....	1
1.1. Enzymes	1
1.1.1. Enzyme engineering.....	1
1.2. Computational enzyme design	2
1.2.1. Computational protein design	3
1.2.1.1. Rotamer libraries.....	4
1.2.1.2. Energy function.....	5
1.2.1.3. Sequence optimization and search algorithm	8
1.2.2. Transition state theory	9
1.2.3. Theozyme calculation	11
1.2.4. Backbone selection and preparation	15

1.2.5. Theozyme placement	17
1.3. Successes in computational enzyme design.....	20
1.3.1. First example of <i>de novo</i> enzyme design.....	20
1.3.2. Kemp elimination.....	22
1.3.3. Retro-aldol reaction	27
1.3.4. Engineering of a Michaelase.....	30
1.3.5. Ester hydrolysis.....	31
1.3.6. Diels-Alder reaction.....	32
1.3.7. Morita–Baylis–Hillman reaction.....	33
1.3.8. Chemiluminescence reaction	34
1.4. Challenges in computational enzyme design.....	35
1.4.1. The challenge of predicting distal mutations	36
1.4.2. The challenge of identification of suitable scaffolds	40
1.4.3. The challenge of stabilizing multiple transition states.....	41
1.5. Machine learning methods in enzyme design.....	41
1.6. Thesis objectives	44
1.7. References.....	46
Chapter 2. Distal mutations enhance catalysis in designed enzymes by facilitating substrate binding and product release.....	60
2.1. Foreword	61
2.2. Statement of contribution.....	62

2.3. Abstract	63
2.4. Introduction.....	64
2.5. Results.....	66
2.5.1. Core and shell variants of <i>de novo</i> Kemp eliminases	66
2.5.2. Functional effects of active-site and distal mutations.....	69
2.5.3. Structural effects	71
2.5.4. Mechanistic effects	76
2.5.5. Dynamical effects	80
2.6. Discussion	84
2.7. Methods.....	86
2.7.1. Protein expression and purification	86
2.7.2. Circular dichroism and thermal denaturation assays	87
2.7.3. Steady-state kinetics.....	87
2.7.4. Kinetic solvent viscosity effects (KSVE)	88
2.7.5. Stopped-flow kinetics	89
2.7.6. Crystallization	89
2.7.7. X-ray data collection and processing.....	90
2.7.8. Structure determination.....	90
2.7.9. Ensemble refinement	91
2.7.10. Molecular dynamics simulations	91

2.7.11. Bottleneck radius analysis.....	93
2.8. Supplementary information	94
2.9. References.....	116
Chapter 3. Customizing the structure of a <i>de novo</i> TIM barrel to introduce enzymatic activity	123
3.1. Foreword.....	124
3.2. Statement of contribution.....	125
3.3. Abstract	126
3.4. Introduction.....	127
3.5. Design strategy and <i>in silico</i> results.....	131
3.5.1. Theozyme placement	132
3.5.2. Lid generation	134
3.5.3. Active site repacking and filtering	138
3.6. Results.....	140
3.7. Discussion.....	145
3.8. Methods.....	148
3.8.1. Computational enzyme design.....	148
3.8.2. Lid generation	149
3.8.3. Active-site repacking	149
3.8.4. Library design and filtering	150
3.8.5. Protein expression and purification	150

3.8.6. Steady-state kinetics.....	151
3.8.7. Circular dichroism and thermal denaturation assays	152
3.8.8. Molecular dynamics.....	152
3.9. Supplementary information	154
3.10. References.....	164
Chapter 4. <i>De novo</i> design of Michaelases using a multistate enzyme design approach....	170
4.1. Foreword.....	171
4.2. Statement of contribution.....	172
4.3. Abstract	173
4.4. Introduction.....	174
4.5. Characterization of the reaction mechanism.....	176
4.6. Computational design procedure	187
4.7. Results.....	193
4.8. Discussion	201
4.9. Methods.....	204
4.9.1. DFT calculations	204
4.9.2. Ensemble refinement	204
4.9.3. Ligand placement.....	205
4.9.4. Matching	206
4.9.5. Single-state design	206

4.9.6. Multistate design	207
4.9.7. Analysis and filtering	207
4.9.8. Protein expression and purification	207
4.9.9. Steady-state kinetics.....	208
4.9.10. Circular dichroism and thermal denaturation assays	209
4.9.11. Molecular dynamics	209
4.10. Supplementary information	211
4.11. References	222
Chapter 5. Discussion & Conclusions	227
5.1. Summary of Findings.....	227
5.2. Discussion	231
5.3. Future Directions	235
5.4. A paradigm shift in enzyme design.....	238
5.5. References.....	241

LIST OF FIGURES

Chapter 1

Figure 1.1. Two primary methods of enzyme engineering.	2
Figure 1.2. Computational Protein Design (CPD) algorithm.	4
Figure 1.3. Computational enzyme design.	11
Figure 1.4. Geometry optimization.	14
Figure 1.5. Ensemble refinement.	17
Figure 1.6. Theozyme placement methods.	19
Figure 1.7. Enzyme design timeline.	20
Figure 1.8. Ester hydrolysis reaction.	20
Figure 1.9. Kemp elimination reaction.	22
Figure 1.10. Active site structures of <i>de novo</i> designed and evolved Kemp eliminases.	26
Figure 1.11. Retro-aldol reaction.	27
Figure 1.12. Active site structures of designed and evolved retroaldolases.	29
Figure 1.13. Michael addition reaction.	30
Figure 1.14. Ester hydrolysis.	31
Figure 1.15. Diels-Alder.	32
Figure 1.16. Morita–Baylis–Hillman reaction.	33
Figure 1.17. Distribution of designed and evolved positions in <i>de novo</i> enzymes.	39
Figure 1.18. Challenges in enzyme design.	44

Chapter 2

Figure 2.1. Core and Shell variants of Kemp eliminases.....	68
Figure 2.2. Crystal structures.....	73
Figure 2.3. Ensemble refinement of Kemp eliminases.....	75
Figure 2.4. Kinetic solvent viscosity effects.....	79
Figure 2.5. Conformational landscapes of Kemp eliminases by molecular dynamics.....	83
Supplementary Figure 2.1. Far-UV circular dichroism.....	104
Supplementary Figure 2.2. Thermal denaturation of Kemp eliminases.....	105
Supplementary Figure 2.3. Electron density of Kemp eliminase active sites.....	106
Supplementary Figure 2.4. Active-site and distal mutations do not cause substantial changes to the backbone structure of Kemp eliminases.....	107
Supplementary Figure 2.5. Evaluation of Trp110 preorganization in 1A53-Core.....	108
Supplementary Figure 2.6.....	109
Supplementary Figure 2.7. Steady-state kinetics at various viscosities.....	110
Supplementary Figure 2.8. Kinetics of 6NBT binding to HG3 enzymes measured by stopped-flow Trp fluorescence quenching experiments.....	111
Supplementary Figure 2.9. Kinetics of 6NBT binding to 1A53 enzymes measured by stopped-flow Trp fluorescence quenching experiments.....	112
Supplementary Figure 2.10. Conformational flexibility of Kemp eliminases during molecular dynamics.....	113
Supplementary Figure 2.11. Dominant conformational changes in each enzyme family revealed by principal component analysis.....	114
Supplementary Figure 2.12. Distal mutations widen active-site tunnels.....	115

Chapter 3

Figure 3.1. De novo enzyme design methodology.....	127
Figure 3.2. Natural versus <i>de novo</i> TIM barrels.	130
Figure 3.3. Kemp elimination of 5-nitrobenzoxazole (5-NBZ).	132
Figure 3.4. Theozyme placement.....	134
Figure 3.5. Lid generation.....	137
Figure 3.6. Filtering of sequences.....	139
Figure 3.7. Structures of the experimentally characterized designs.	140
Figure 3.8. Circular dichroism (CD) and thermal melting results. A)	142
Figure 3.9. Kinetic characterization of KempTIM 4.	143
Figure 3.10. Conformational flexibility and active site accessibility of designed Kemp eliminases.	144
Supplementary Figure 3.1. Structural segments generated by RFdiffusion.	163

Chapter 4

Figure 4.1. Michael addition and retro-aldol reactions catalyzed by RA95.5-8.....	177
Figure 4.2. Reaction mechanism for Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (1a) and ethyl cyanoacetate (2a).....	178
Figure 4.3. Deprotonation of ethyl cyanoacetate.....	180
Figure 4.4. Step 1 of the reaction.....	181
Figure 4.5. Step 2 of the reaction.....	182
Figure 4.6. Step 3 of the reaction.....	183
Figure 4.7. Step 4 of the reaction.....	184
Figure 4.8. Step 5 of the reaction.....	185
Figure 4.9. Step 6 of the reaction.....	186
Figure 4.10. Energy profile for Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (1a) and ethyl cyanoacetate (2a).....	187
Figure 4.11. Schematic of the matching algorithm.....	190
Figure 4.12. Schematic representation of protein and enzyme design strategies.....	191
Figure 4.13. Michaelis-Menten graphs of designed Michaelases.....	194
Figure 4.14. Comparison of kinetic parameters for Michaelase variants.....	195
Figure 4.15. Structural models of transition states.....	197
Figure 4.16. Energy diagram and loop distances in Michaelase variants.....	199
Figure 4.17. Circular dichroism (CD) spectra and thermal melting graphs.....	200

Chapter 5

Figure 5.1. Summary of findings. 227

Figure 5.2. Future directions. 235

LIST OF TABLES

Chapter 2

Table 2.1. Properties of various Kemp eliminases.....	70
Supplementary Table 2.1. Amino acid sequences of various Kemp eliminases	94
Supplementary Table 2.2. Active-site and distal mutations	95
Supplementary Table 2.3. Purification yields	96
Supplementary Table 2.4. Crystallization conditions	97
Supplementary Table 2.5. Crystallographic data and refinement statistics	98
Supplementary Table 2.6. Ensemble refinement of Kemp eliminase structures with bound 6NBT	99
Supplementary Table 2.7. 6-nitrobenzotriazole (6NBT) binding kinetics.....	100
Supplementary Table 2.8. DNA Sequences	101

Chapter 3

Supplementary Table 3.1. Positions in the TIM barrel F15C tested for Asp placement.....	154
Supplementary Table 3.2. Geometric definitions for generation of transition-state poses off the side chain of catalytic base	155
Supplementary Table 3.3. Geometric definitions for generation of transition-state poses off the side chain of hydrogen bond donor	156
Supplementary Table 3.4. Geometric constraints used to define catalytic contacts during computational design	157
Supplementary Table 3.5. Amino acid sequences of Kemp eliminase designs	158
Supplementary Table 3.6. Criteria for design filtering: Designs were excluded if they failed to meet all the criteria listed below.	160
Supplementary Table 3.7. Design filtering – F15C template.....	161
Supplementary Table 3.8. Design filtering – CP1 and CP2 templates	162

Chapter 4

Table 4.1. Filtering parameters for the 12 experimentally tested designs	193
Table 4.2. Apparent steady-state parameters of Michaelase designs	194
Supplementary Table 4.1. Ensemble refinement of template scaffolds used in design of Michaelases.....	211
Supplementary Table 4.2. Lysine positions and tyrosine candidate positions in each template scaffold.....	212
Supplementary Table 4.3. Geometric definitions for generation of TS ₄ poses off the lysine side chain.....	213
Supplementary Table 4.4. Geometric definitions for generation of TS ₆ poses off the lysine side chain.....	214
Supplementary Table 4.5. Geometric constraints used to define TS ₄ catalytic contacts	215
Supplementary Table 4.6. Geometric constraints used to define TS ₆ catalytic contacts	216
Supplementary Table 4.7. Amino-acid sequences of Michaelase designs.....	217
Supplementary Table 4.8. DNA sequences of Michaelase designs	218
Supplementary Table 4.9. Amino-acid positions optimized during computational design of Michaelases.....	221

LIST OF ABBREVIATIONS

5NBZ	5-nitrobenzisoxazole
6NBT	6-nitrobenzotriazole
CD	Circular Dichroism
CPD	Computational Protein Design
DEE	Dead-End Elimination
DFT	Density Functional Theory
DMSO	Dimethyl Sulfoxide
DTZ	Diphenylterazine
FASTER	Fast and Accurate Side-Chain Topology and Energy Refinement
GA	Genetic Algorithms
GMEC	Global Minimum Energy Conformation
H-bond	Hydrogen Bond
HF	Hartree-Fock
IPTG	Isopropyl β -D-1-thiogalactopyranoside
IRC	Intrinsic Reaction Coordinate
K _D	Dissociation equilibrium constant
KSVE	Kinetic Solvent Viscosity Effects
LJ	Lennard-Jones
LB	Luria-Bertani
MAO	Monoamine Oxidase
MBH	Morita–Baylis–Hillman reaction
MC	Monte Carlo
MD	Molecular Dynamics
MES	2-(N-Morpholino)ethanesulfonic acid
MM-PEFs	Molecular Mechanics Potential Energy Functions
MPNN	Message Passing Neural Network
MSA	Multiple Sequence Alignment
MSD	Multistate Design
NMR	Nuclear Magnetic Resonance

OD	Optical Density
OPC	Optimized Point Charge (water model)
PCA	Principal Component Analysis
PCM	Polarizable Continuum Model
PDB	Protein Data Bank
PES	Potential Energy Surface
pLDDT	Predicted Local Distance Difference Test
PNP	p-Nitrophenol
PNPA	p-Nitrophenyl Acetate
QM	Quantum Mechanics
RIF	Rotamer Interaction Field
RMS	Root Mean Square
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent Accessible Surface Area
SCMF	Self-Consistent Mean Field
SPM	Shortest Path Map
SSD	Single-State Design
TIM	Triosephosphate Isomerase
TLS	Translation-Libration-screw
T _m	Melting Temperature
TS	Transition State
TSA	Transition-State Analogue

CHAPTER 1. INTRODUCTION

1.1. Enzymes

Enzymes are nature's biocatalysts, capable of accelerating reaction rates by up to 26 orders of magnitude.¹ Their high specificity, mild reaction conditions, and environmental friendliness have led to their use in various sectors, including food, agriculture, and pharmaceuticals. However, the industrial use of enzymes is limited by the finite repertoire of natural enzymes. Many desired new-to-nature reactions lack naturally occurring enzymes able to catalyze them, necessitating enzyme engineering to tailor enzymes for industrial applications.

1.1.1. Enzyme engineering

In enzyme engineering, a preexisting enzyme is modified to achieve new and improved characteristics. This is typically accomplished through one of two methods: directed evolution or rational design. In directed evolution, multiple rounds of generating mutant libraries are performed using random mutagenesis techniques, such as error-prone PCR or DNA shuffling, followed by high-throughput screening to identify improved variants (Figure 1.1A). Rational design leverages structural and functional information, aided by computational tools, to introduce mutations at specific sites using site-directed mutagenesis. This approach can be employed when the structure and function of the protein are well-known (Figure 1.1B).² These methods have produced many improved enzymatic activities;³⁻⁷ however, both have limitations. Directed evolution requires a starting enzyme with at least minimal activity and a high-throughput assay to screen for improved activity. Rational design requires extensive knowledge of the enzyme's structure and function, which is not always available. Therefore, developing a computational framework for designing

enzymes to catalyze any desired chemical reaction on demand will be very useful in industrial biocatalysis.

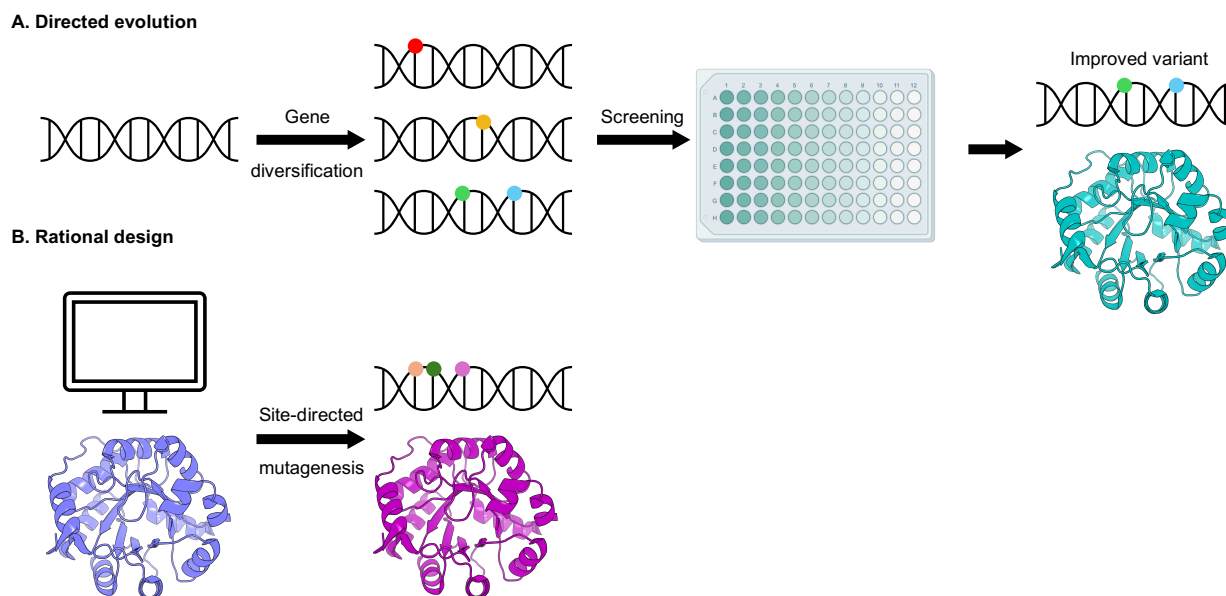


Figure 1.1. Two primary methods of enzyme engineering. (A) Directed evolution involves gene diversification through techniques such as random mutagenesis, error-prone PCR, or DNA shuffling, followed by high-through screening to identify improved variants. (B) Rational design uses site-directed mutagenesis to introduce specific mutations based on detailed knowledge of the enzyme's structure and function, allowing targeted modifications to enhance enzyme characteristics.

1.2. Computational enzyme design

In computational enzyme design, a *de novo* enzyme is created to catalyze a reaction of interest using structure-based computational protein design (CPD) methods.⁸ CPD algorithms search for amino acid sequences whose lowest free energy state adopts a specific target three-dimensional structure.⁹ The first instance of CPD was carried out in 1985 by DeGrado *et al.* In this pioneering study, they successfully designed, synthesized, and characterized the tightest-binding peptide inhibitors known at the time.¹⁰ Since then, computational advances, the rapid growth of biological databases, and significant increases in computing power have dramatically

enhanced our ability to design the three-dimensional structures of proteins. Notably, the 2024 Nobel Prize in Chemistry was awarded, in part, to Dr. David Baker for pioneering work in the computational design of proteins with novel shapes and functions. In the following section, we will expand on different aspects of computational protein design.

1.2.1. Computational protein design

Most CPD algorithms consist of four steps (Figure 1.2):

1. Template selection: a fixed protein backbone template is selected.
2. Sidechain placement: discrete sidechain rotamers are placed at user-specified positions on the backbone template.¹¹
3. Energy calculation: the interaction energies between each rotamer and the backbone template (one-body energies) and between pairs of rotamers (two-body energies) are calculated using an energy function.^{12–15}
4. Sequence optimization: a search algorithm explores both rotamer and sequence space to identify sequences that yield the lowest energy values.^{16–19}

At the conclusion of this process, a list of ranked sequences, along with their corresponding structures, is generated. Each sequence is ranked based on an energy value that represents its predicted stability on the backbone template. The top sequences are then selected for experimental characterization. The following sections will provide an in-depth look at rotamer libraries, energy functions, and sequence optimization algorithms.

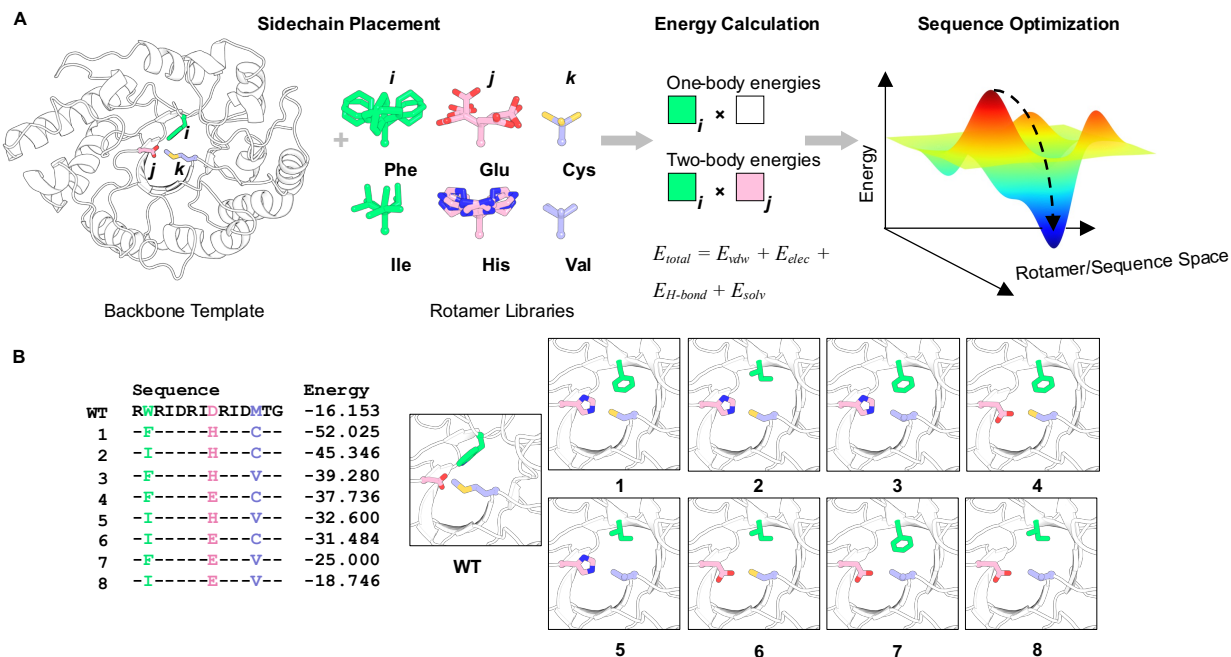


Figure 1.2. Computational Protein Design (CPD) algorithm. (A) The general CPD workflow consists of four steps: 1. Selection of a fixed protein backbone template; 2. Sidechain rotamer placement where discrete sidechain rotamers are placed at the designed positions on the backbone template (Phe and Ile at position i, Glu and His at position j, and His and Val at position k); 3. Interaction energies between rotamers and backbone (one-body energies) and between pairs of rotamers (two-body energies) are calculated according to an energy function including van der Waals (vdw), electrostatic (elec), hydrogen bonding (H-bond), and solvation (solv) terms; 4. A search algorithm explores the rotamer and sequence space to identify the sequences with the lowest energy values, indicative of their predicted stability; (B) a list of sequences and their corresponding structures is generated and ranked by their energy values. Figure adapted from St-Jacques *et al.*²⁰

1.2.1.1. Rotamer libraries

A rotamer, or ‘rotational isomer’, refers to a single amino acid sidechain conformation defined by a set of values for each dihedral angle degree of freedom. Given the small variances in protein bonds’ angles and lengths, they are usually excluded from the rotamer definition. A rotamer library is a compilation of rotamers for each amino acid, and it includes data on both the conformation and the frequency of each conformation. These libraries also include the variance around the mean dihedral angles. Rotamer libraries are derived from statistical analysis of

sidechain conformations in known protein structures. There are three types of rotamer libraries. Backbone-dependent rotamer libraries include rotamer conformations that are dependent on the local backbone conformation, represented by backbone dihedral angles ϕ and ψ . Backbone-independent libraries do not consider the backbone conformation and are derived from all the sidechains of a specific residue. Secondary-structure-dependent libraries provide different rotamer conformations and frequencies for specific secondary structures like α helices, β sheets, or coils.¹¹

1.2.1.2. Energy function

The stability of a protein sequence on a backbone template is quantified using an energy function. There are different types of energy functions, including molecular mechanics potential energy functions (MM-PEFs) and knowledge-based energy functions. MM-PEFs include both bonded and non-bonded terms. Bonded terms ensure that bond lengths, angles, and torsions remain near their equilibrium values. Non-bonded terms include the Lennard-Jones (LJ) potential, which describes van der Waals forces, and Coulomb's law, which describes electrostatic interactions. In contrast, knowledge-based energy functions are derived from statistical analysis of known protein structures. These functions calculate the probabilities of residues adopting specific or pairs of residues forming particular geometries. These probabilities are then converted into potential energy values using the Boltzmann equation: $\Delta G = -RT \ln(p_{obs}/p_{exp})$, where p_{obs} is the observed probability of a structural element and p_{exp} is the expected probability of that element occurring by chance.²¹

Potential energy functions used in computational protein design often include a combination of MM-PEF and knowledge-based terms. A prominent example is the Rosetta All-Atom energy function, which combines 19 energy terms, each with specific weight coefficients.

These terms include van der Waals interactions, electrostatic interactions, solvation effects, hydrogen and disulfide bonds, and backbone and side-chain torsional preferences.²² In this thesis, the sequences are evaluated using the PHOENIX energy function,^{23–26} which includes terms for van der Waals interactions (vdw), electrostatics (elec), hydrogen bonding (H-bond), solvation (solv), and secondary structure propensities (ss) terms.

Equation 1.1. PHOENIX energy function

$$E_{total} = E_{vdw} + E_{elec} + E_{H-bond} + E_{solv} + E_{ss}$$

The van der Waals energy term is calculated using the Lennard-Jones 12-6 potential, which includes a long-range attractive component proportional to $-(1/R)^6$ and a short-range repulsive component proportional to $(1/R)^{12}$. The Lennard-Jones 12-6 potential energy function is represented in Equation 1.2 where $D_{i,j}$ is the mean well depth of atoms i and j , α is the van der Waals radius scaling factor, $r_{i,j}$ is the interatomic distance between atoms i and j and $R_{l,2}$ is the sum of van der Waals radii of atoms i and j .²⁶

Equation 1.2. Lennard-Jones 12-6 van der Waals energy term

$$E_{vdw} = D_{i,j} \left[\left(\frac{\alpha R_{i,j}}{r_{i,j}} \right)^{12} - 2 \left(\frac{\alpha R_{i,j}}{r_{i,j}} \right)^6 \right]$$

The electrostatics term is computed using Coulomb's law equation expressed in Equation 1.3, where q_i and q_j represent the charges of the interacting atoms, ϵ is the dielectric constant that accounts for the medium between the charges, and $r_{i,j}$ is the distance between the two atoms.²⁷

Equation 1.3. Coulomb's law equation

$$E_{\text{elec}} = \frac{q_i q_j}{\epsilon r_{i,j}}$$

The hydrogen bond potential scoring function includes both a distance-dependent term and an angle-dependent term. In Equation 1.4, D_0 is the hydrogen bond well depth, typically around 8 kcal/mol, r_{ad} is the distance between hydrogen bond donor and acceptor heavy atoms, and R_{ad} is the equilibrium hydrogen bond donor-acceptor distance, generally around 2.8 Å. The angle-dependent term, $F(\theta, \phi, \psi)$, varies based on the hybridization states of the donor and acceptor atoms and ensures that only geometrically favourable hydrogen bonds are considered.²⁸

Equation 1.4. Hydrogen bond potential energy term

$$E_{\text{h-bond}} = D_0 \left[5 \left(\frac{R_{ad}}{r_{ad}} \right)^{12} - 6 \left(\frac{R_{ad}}{r_{ad}} \right)^{10} \right] F(\theta, \phi, \psi)$$

Modelling solvent explicitly is complicated by the high computational cost associated with exploring numerous water configurations.²⁹ Therefore, in PHOENIX, an implicit solvation potential is used instead. This can be achieved using either a surface area-based model^{12,30,31} or an occlusion-based model.^{25,28,32} The surface area-based model calculates solvation energy based on the solvent-accessible surface area of amino acid side chains, with favourable interactions for polar residues and unfavourable interactions for nonpolar residues. In contrast, the occlusion-based model evaluates atoms based on the percentage of their volume occluded by other atoms within the protein structure, rather than their solvent-accessible surface area. The secondary structure term in the PHOENIX energy function quantifies the propensity of amino acids to adopt specific backbone conformations, such as α -helices, β -sheets, or random coils, based on their ϕ (phi) and ψ (psi) backbone dihedral angles. This term is derived from statistical analysis of high-resolution

protein structures, capturing the experimentally observed ϕ/ψ preferences for each amino acid type. Propensities are computed as the conditional probability of observing a particular ϕ/ψ angle given an amino acid type, which are then converted to energy values using the Boltzmann hypothesis.

1.2.1.3. Sequence optimization and search algorithm

CPD aims to find protein sequences that stably fold to the target structure. To achieve this, a search algorithm must search the sequence-rotamer energy landscape to find the global minimum energy conformation (GMEC).³³ There are two types of search algorithms: stochastic and deterministic. Stochastic algorithms incorporate randomness as part of their logic. These algorithms use probabilistic decisions to explore the sequence space, and finding GMEC is not guaranteed. For instance, Monte Carlo (MC) simulated annealing,^{34,35} Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER),³⁶ and genetic algorithms (GA). In contrast, deterministic algorithms will always generate the same output given the same initial conditions; examples include dead-end elimination (DEE) and self-consistent mean-field (SCMF). MC algorithms begin by randomly selecting rotamers for a sequence. The algorithm then replaces an initial rotamer with a new one, either for the same residue or a different one. If the energy of this new configuration E_{new} is lower than the previous energy E_{old} , the new configuration is accepted. If E_{new} is higher, the substitution is accepted with a probability $p = \exp\left(\frac{-\Delta E}{kT}\right)$, where $\Delta E = E_{new} - E_{old}$, k is Boltzmann's constant, and T is the temperature.³⁷ Multiple cycles are performed where, in each cycle, the temperature is raised and lowered between designated high and low temperatures. The temperature serves to overcome local minima in the energy landscape by allowing large jumps. In contrast to MC, where a single or set of solutions are optimized iteratively,

DEE is a deterministic algorithm that systematically eliminates unfavourable rotamers and combinations of rotamers, therefore reducing the search space until a single solution remains, which is the GMEC.¹⁶

Section 1.2.1 explored various aspects of computational protein design, including rotamer libraries, energy functions, and sequence optimization algorithms. Computational enzyme design includes an additional step where an active site, comprising catalytic residues and a transition state, is placed on a pre-existing scaffold according to a set of predefined geometric constraints believed to be crucial for the reaction of interest. By leveraging transition state theory, computational enzyme design aims to stabilize the transition state, thereby lowering the activation energy barrier and increasing the reaction rate. In the following section, we will delve into the fundamentals of transition state theory.

1.2.2. Transition state theory

Transition state theory was developed by Henry Eyring in 1935 and is used to determine the rate of elementary chemical reactions.³⁸ According to this theory, a quasi-equilibrium is formed between reactants and an activated complex at the saddle point of the potential energy surface.³⁹ In order for a chemical reaction to occur, the reactant molecules must acquire enough energy to reach this high-energy transition state. The activation energy corresponds to the energy required to overcome this barrier, and it is directly related to the reaction rate as described by the equation:

Equation 1.5. Eyring equation

$$k(T) = \gamma \frac{k_B T}{h} e^{-\frac{\Delta G^\ddagger}{RT}}$$

Where k is the rate constant, γ is the transmission coefficient, k_B is the Boltzmann constant, T is the temperature, h is the Planck constant, R is the gas constant, and $\Delta G^\ddagger$ is the Gibbs energy of activation. Enzymes accelerate reaction rates and lower the activation energy by binding tightly to the substrate distorted towards the transition state.⁴⁰

Computational enzyme design leverages transition state theory by first constructing a theozyme, short for theoretical enzyme, which consists of a transition state structure and the minimally required catalytic residues arranged in an optimal orientation to stabilize the transition state (Figure 1.3A).⁴¹ Following theozyme calculation, a high-resolution crystal structure capable of accommodating the computed theozyme is selected and prepared by removing water molecules, non-protein atoms, while adding hydrogens and any missing residues. (Figure 1.3B). Next, during the theozyme placement step, the active site is constructed by identifying positions on the template backbone where the catalytic contacts can be made to the transition state (Figure 1.3C). Finally, during active site repacking, residues in the active site are optimized to stabilize the transition state further (Figure 1.3D). The following sections will explore each of these steps in greater detail.

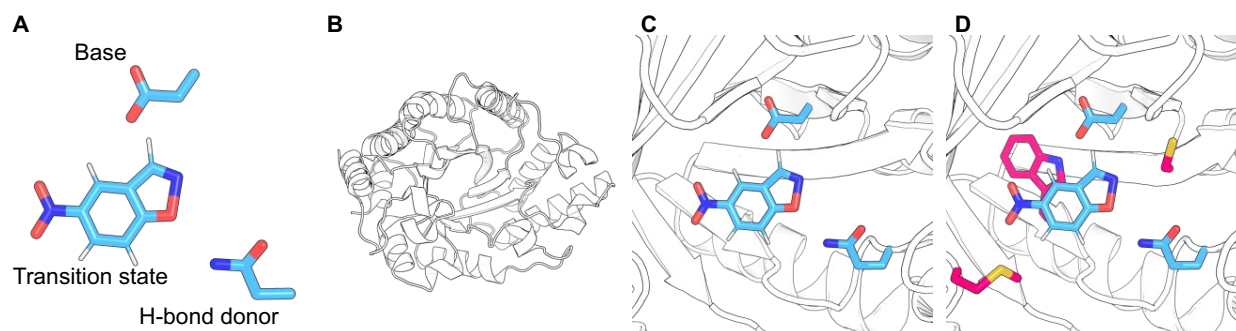


Figure 1.3. Computational enzyme design. (A) A theozyme, comprising the transition state (TS) and catalytic residues optimally arranged for catalysis, is calculated using high-level quantum mechanical theory. In this example, a base catalyzes the reaction while a hydrogen bond (H-bond) donor stabilizes the negative charge at the transition state and properly orients the substrate. (B) A crystal structure from the Protein Data Bank (PDB) is selected to accommodate the calculated theozyme within the active site (PDB ID: 5RGE). (C) In the theozyme placement step, the theozyme is placed onto the target backbone template. (D) During active site repacking, the surrounding residues are optimized to further stabilize the transition state, enhancing catalytic efficiency.

1.2.3. Theozyme calculation

The structure of a theozyme is typically calculated using quantum mechanical (QM) methods. A fundamental challenge in quantum chemistry is solving the Schrödinger equation for a many-body electronic system comprising nuclei and electrons. Traditional QM calculations involve solving the Schrödinger equation for a given molecular system. One of the most notable approaches to this problem is the Hartree-Fock (HF) method, which approximates the exact N-body wave function of the system using a single Slater determinant of N spin-orbitals.⁴³ A Slater determinant is a mathematical expression that ensures the anti-symmetry of the wave function in accordance with the Pauli exclusion principle, which states that no two electrons in an atom can have the same set of quantum numbers. The main limitation of the HF approach is that it assumes that electrons are moving independently from each other, thus ignoring electron correlation.

Finding approximate solutions to the Schrödinger equation for N interacting electrons is highly complex and computationally expensive. Consequently, most modern QM calculations now utilize density functional theory (DFT). Unlike traditional methods based on the wavefunction, DFT is based on two theorems proposed by Hohenberg and Kohn. The first theorem asserts that all ground-state properties of an electronic system are uniquely determined by the electron density, which is a function of the three Cartesian coordinates. The second theorem states that the energy of an electronic system can be expressed as a functional of electron density, and this functional reaches its minimum at the ground-state density.^{43,44} A functional is a mathematical construct that maps a function—in this case, the electron density—to a single value, the energy of the system. Hybrid functionals are a class of functionals used in DFT that combine a portion of the exact exchange energy from Hartree-Fock theory with the rest of the functional from other sources. One of the most widely used hybrid functionals is B3LYP, which is used in this thesis.

To accurately implement these methods, choosing an appropriate basis set is crucial. A basis set is a set of functions used to build the quantum mechanical wavefunction for a molecular system modelling molecular orbitals. By transforming the partial differential equations of the model into algebraic equations, basis sets make the problem suitable for computer calculations. Specifying a basis set effectively restricts each electron to a particular region of space. Larger basis sets are more accurate approximations of the wavefunction as they place fewer restrictions on electrons, although they require more computational resources. In this thesis, the 6-311+G(2d,p) basis set is used. The “6-311” defines the composition of the basis set, with “6” indicating that six Gaussian functions are used to describe core orbitals. The “311” portion signifies that the valence orbitals are split into three components: a single set with three Gaussian functions, a second set with one Gaussian function, and a third set with one Gaussian function. The “+” symbol denotes

the inclusion of diffuse functions, which are essential for accurately modelling electron density far from the nucleus. The “(2d,p)” part describes the polarization functions added to the basis set, where “2d” indicates two sets of d-type polarization on heavy atoms, and “p” indicates a single set of p-type polarization on hydrogen atoms.⁴⁵

With an accurate basis set defined, we can effectively explore the potential energy surface (PES) of a molecular system, which describes the relationship between the energy of a system and its geometry. During geometry optimization, the calculation aims to find stationary points which can either be a minimum or a transition state. At stationary points, the first derivative of energy or the gradient, which is negative of the forces, is zero. A minimum is located at the bottom of a valley where energy rises in all directions. A transition state is located at the saddle point, which is a maximum in one direction and a minimum in all others.⁴⁶ Geometry optimization calculations start from a specified input structure and proceed by stepping along the potential energy surface. At each point, both the energy and gradient are computed (Figure 1.4). An optimization calculation is considered complete when it meets the convergence criteria: (1) the forces must be nearly zero, (2) the root-mean-square (RMS) of the forces must be lower than a defined threshold of 0.0003, (3) the maximum displacement, representing the largest change in any coordinate of the structure must fall below a tolerance of 0.0018, and (4) the RMS of the displacement for the next step must be lower than the cutoff of 0.0012.⁴⁵

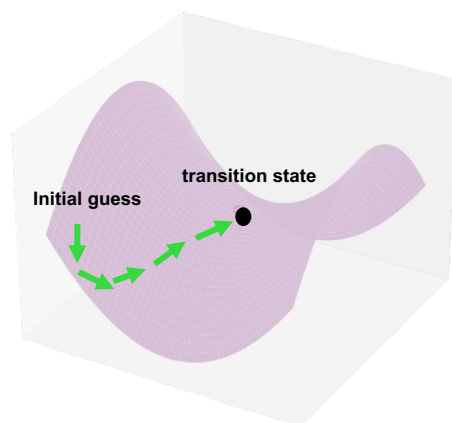


Figure 1.4. Geometry optimization. Illustration of a geometry optimization pathway on a potential energy surface, depicting the initial guess and the transition state.

In order to construct a theozyme, first, a reasonable guess is specified for the transition state and catalytic groups. Next, the geometry is optimized to locate the first-order saddle point, characterized by having one imaginary frequency, which means that in one direction on the potential energy surface, the energy is a maximum, while in all other directions, the energy is a minimum. Locating a stationary point with one imaginary frequency does not necessarily guarantee that the computed transition state connects the reactants and products of interest. To identify the reactants and products corresponding to a transition structure, an Intrinsic Reaction Coordinate (IRC) calculation must be performed. An IRC calculation begins at the saddle point and examines the reaction pathway descending from a transition state structure in both directions. Through this process, an IRC calculation connects two minima on the potential energy surface via a pathway that passes through the transition state.^{45,47} The resulting theozyme contains all the sidechains optimally oriented for catalysis but lacks a protein backbone.

1.2.4. Backbone selection and preparation

Prior to performing a theozyme placement calculation, a protein backbone template capable of accommodating the computed transition state must be selected and prepared. This initial structure can either be retrieved from the Protein Data Bank (PDB),^{48,49} which contains structures experimentally solved by X-ray crystallography or nuclear magnetic resonance (NMR) or be generated *de novo*.^{50,51} The selected backbone template must first be prepared by adding any missing atoms and hydrogens. However, using a single static structure is not suitable for enzyme design because it neglects protein dynamics.⁵² Proteins do not exist as a single rigid structure; they exhibit conformational flexibility. Designing an enzyme using a fixed backbone template or single state design (SSD) overlooks this dynamic nature. To address the limitations of SSD, backbone flexibility can be incorporated by using an ensemble of protein backbone templates.⁵³ This ensemble can be generated through various methods.

In the backrub method, inspired by the conformational heterogeneity observed in crystal structures, local backbone adjustments are incorporated through “backrub moves”, where atoms between the first and last alpha carbons rotate around a virtual axis. A backrub ensemble is generated by performing backrub Monte Carlo simulations. During these simulations, the starting template backbone is subjected to these small-angle rotations around axes defined by alpha carbons while the corresponding side chains sample rotamers from the Dunbrack rotamer library.¹¹ Moves are rejected or accepted according to the Metropolis criterion. Typically, multiple rounds of backrub moves and sidechain adjustments are conducted, ultimately generating an ensemble of low-energy structures.^{54–56}

Another method of ensemble generation is PertMin, which leverages the sensitivity of minimization to initial coordinates. In this approach, the coordinates of the starting backbone template are randomly perturbed. Subsequently, the energies of the set of perturbed structures are minimized, causing each to diverge to a different local minimum.^{53,57}

In this thesis, ensemble refinement has been used to generate ensembles of structures. Initially developed to improve the modelling of atomic disorder in crystal structures, this method exploits the fact that protein X-ray diffraction patterns are influenced by protein dynamics. Molecular dynamics simulations, restrained by diffraction data, are performed to generate an ensemble of Boltzmann-weighted structures representing the protein's conformational space (Figure 1.5). Ensembles generated through ensemble refinement have been shown to fit experimental data better, as evaluated by the R-free value. R-free is a measure used in X-ray crystallography to assess model accuracy. It is calculated by comparing how well the model predicts a subset of the data (about 10%) that was not used during refinement.⁵⁸

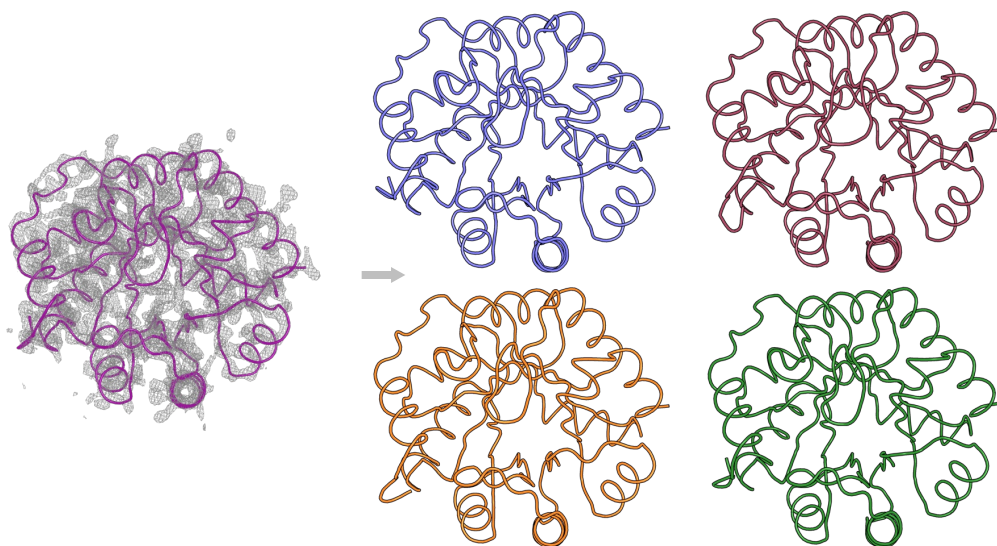


Figure 1.5. Ensemble refinement. Ensemble refinement performs molecular dynamics simulation restrained by diffraction data to generate an ensemble of structures

1.2.5. Theozyyme placement

Active site construction begins by identifying positions on the template backbones where the necessary catalytic contacts can be made to the transition state (TS). This can be achieved using either phoenixMatch,⁵⁹ RosettaMatch,⁶⁰ or rotamer interaction field (RIF) docking.⁶¹ Initially, a protein backbone template and a theozyyme must be provided as inputs (Figure 1.6A). Catalytic sidechain rotamers and TS binding poses are sampled during the calculation. Rotamers are generated using user-defined torsional degrees of freedom, while TS poses are created from the catalytic sidechains, utilizing torsional, rotational, and translational degrees of freedom to maintain the user-defined catalytic contacts.

In phoenixMatch, the ligand poses are generated from every rotamer of a single catalytic contact, followed by testing of rotamers of additional catalytic residues to identify poses that can

make all contacts. Next, pairwise energies for all interactions—sidechain/sidechain, sidechain/backbone, backbone/ligand and sidechain/ligand—are calculated using a force field that includes van der Waals, hydrogen-bonding, and electrostatic terms. An energy bias is then applied to favour sidechain/ligand contacts necessary for catalysis (Figure 1.6B).⁵⁹

RosettaMatch starts by identifying potential active site positions, which can include all positions in the protein or those specifically lining cavities and binding sites. Then, each catalytic sidechain rotamer and the TS model are individually placed on the protein backbone. If there are no clashes with the backbone, the position of the TS model for each rotamer is stored in a hash table. This hash table is subsequently analyzed to identify TS placements that are compatible with the placement of each catalytic residue (Figure 1.6C).⁶⁰

The RIF docking method first generates an ensemble of billions of discrete rotamers that interact with the ligand. It then docks protein scaffolds into these roamer ensembles using a hierarchical grid-base search. Top-scoring scaffold positions are refined through a Monte Carlo-based rotamer packing step to ensure stable interactions. Following docking, the sequences are optimized using Rosetta's design protocols, iterating between fixed-backbone and flexible-backbone design for ligand interaction and protein stability (Figure 1.6D).⁶¹

At the end of this process, residues in the active site are optimized to further stabilize the transition state (Figure 1.6E).

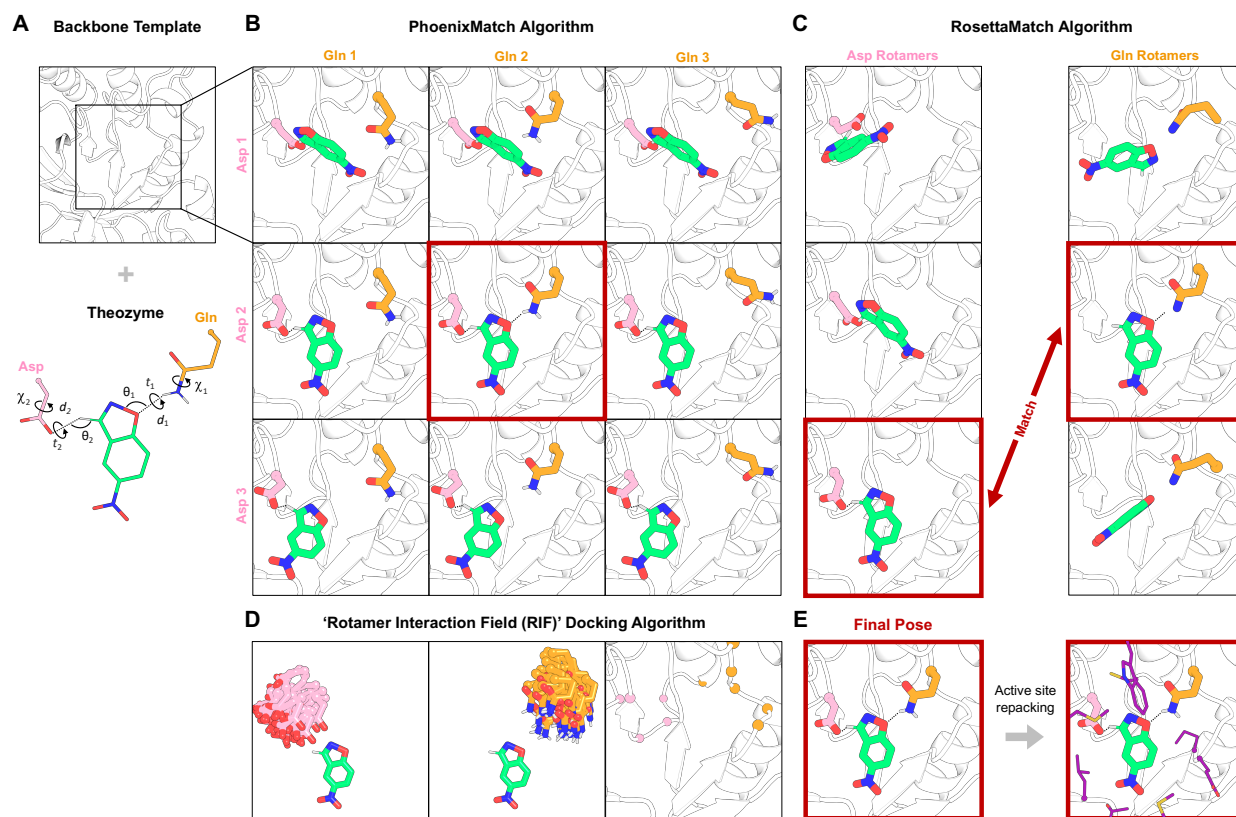


Figure 1.6. Theozyme placement methods. A protein backbone template and theozyme are required as inputs. Catalytic sidechain rotamers and transition state (TS) binding poses are sampled during the calculation. Rotamers are generated using user-defined torsional degrees of freedom (χ), while TS poses are generated from the catalytic sidechains using torsional (t), rotational (θ), and translational (d) degrees of freedom to maintain user-defined catalytic contacts (dashed lines). (B) In phoenixMatch, the TS pose is generated from each Asp rotamer according to the first catalytic contact (columns). Each Gln rotamer is then tested (rows) to identify combinations of Asp and Gln rotamers that satisfy both catalytic contacts (dashed lines). Active sites containing both catalytic contacts (red box) are chosen. (C) In RosettaMatch, the TS pose is generated from each Asp rotamer (left column) and Gln rotamer (right column) according to each associated catalytic contact. Overlapping TS poses that were generated by each of the contacts are identified as a match (red arrow and boxes). (D) RIF Docking generates a vast ensemble of rotamers interacting with the ligand, docks protein scaffolds into these ensembles, and refines top-scoring scaffold positions through Monte Carlo-based rotamer packing and Rosetta's design protocols. (E) Chosen active sites (red box) are repacked and ranked according to their predicted stability on the backbone template.

1.3. Successes in computational enzyme design

Over the past two decades, computational enzyme design has achieved significant milestones, successfully targeting a diverse range of chemical reactions. Notable examples include ester hydrolysis, Kemp elimination, and the Diels-Alder cycloaddition. Additionally, efforts have extended to more complex reactions, such as the retro-aldol reaction, the Morita-Baylis-Hillman reaction and chemiluminescence. In section 1.3, we will explore the advancements and achievements in computational enzyme design during this period (Figure 1.7).

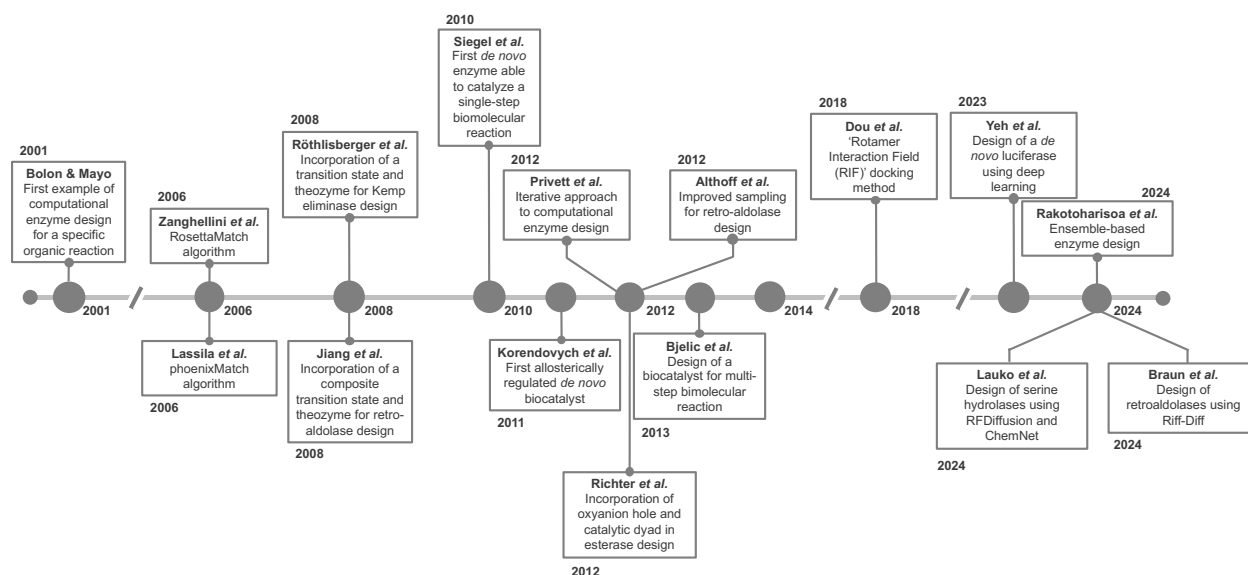


Figure 1.7. Enzyme design timeline. Timeline showcasing the computational enzyme design studies discussed in this section

1.3.1. First example of *de novo* enzyme design

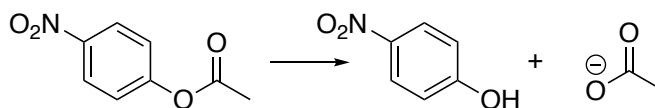


Figure 1.8. Ester hydrolysis reaction. Hydrolysis of *p*-nitrophenyl acetate catalyzed by PDZ2 at pH 7

In 2001, Bolon and Mayo⁶² reported the first computationally designed *de novo* enzyme. Using the ORBIT protein design software, they successfully introduced ester hydrolysis activity into the otherwise catalytically inert *E. coli* thioredoxin scaffold. The target reaction involved hydrolyzing *p*-nitrophenyl acetate (PNPA) into *p*-nitrophenol (PNP) and acetate, utilizing histidine as a nucleophile (Figure 1.8). In this reaction, a covalent acyl-enzyme intermediate is formed via the nucleophilic attack of the catalytic histidine. Bolon and Mayo modelled this tetrahedral intermediate as a proxy for the transition state. They performed an active site scan to identify favourable positions for the catalytic histidine, modelling each protein position with a set of high-energy state rotamers or a “superrotamer” library of the tetrahedral intermediate while allowing the other positions to sample either the wildtype residue or alanine. The optimal solution was then calculated using the Dead-End Elimination search algorithm, and the positions previously changed to alanine were mutated to stabilize the high-energy state rotamer further. Two sequences were characterized experimentally, and both displayed esterase activity. The more active variant, PDZ2, displayed a catalytic efficiency of $2.7 \text{ M}^{-1}\text{s}^{-1}$, well below those of natural enzymes, which exhibit, on average, $k_{\text{cat}}/K_{\text{M}}$ values in the range of $10^5 \text{ M}^{-1}\text{s}^{-1}$.⁶³ This could indicate that utilizing a high-energy intermediate as a proxy for the transition state may lead to suboptimal shape and charge complementarity. Nevertheless, this pioneering study demonstrated the feasibility of designing functional enzymes using computational enzyme design methods.

1.3.2. Kemp elimination

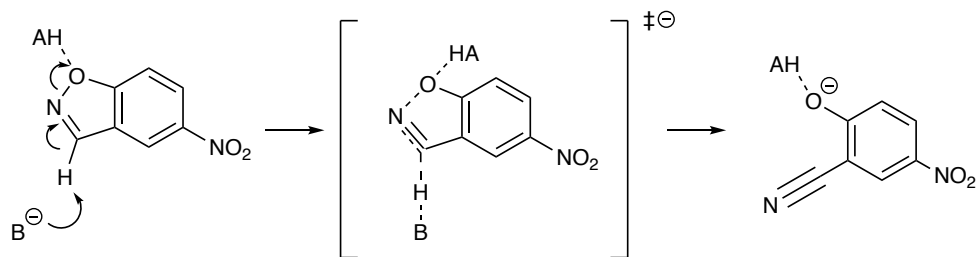


Figure 1.9. Kemp elimination reaction. Kemp elimination is a single-step bond-breaking reaction that proceeds through a concerted proton abstraction and ring-opening of 5-nitrobenzisoxazole

Kemp elimination has been established as a benchmark reaction for *de novo* enzyme design. During this reaction, a proton is abstracted from a carbon atom in the substrate, typically a benzisoxazole derivative, by a general base. Simultaneously, the N–O bond cleaves, and the resulting negative charge migrates to the oxygen atom, ultimately forming an ortho-hydroxybenzonitrile as the product (Figure 1.9). Kemp elimination proceeds through a single transition state, and the reaction's progress can be monitored spectroscopically, as the product has absorbance in the near-UV range.^{64,65}

In 2008, Röthlisberger *et al.*⁶⁶ reported a series of computationally designed Kemp eliminases. In this study, the authors first constructed the theozyme with a catalytic base, either Asp/Glu or a His-Asp dyad, a π -stacking aromatic residue to stabilize the transition state through charge delocalization, and a hydrogen bond donor to stabilize the negative charge on the oxygen of the substrate. Next, they utilized the RosettaMatch hashing algorithm to identify positions capable of accommodating the idealized active site. Finally, residues in the active site were mutated to further stabilize the transition state. Out of 59 designs tested, 8 displayed Kemp elimination activity. Notably, the KE07 enzyme, with Glu101 as the general base, Lys222 as the hydrogen-

bond donor, and Trp50 as the π -stacking residue (Figure 1.10A), showed a catalytic efficiency of $12 \text{ M}^{-1}\text{s}^{-1}$. This enzyme underwent seven rounds of directed evolution, which increased its activity by more than 200-fold through the incorporation of eight mutations. The identities of catalytic residues remained unchanged during evolution. Instead, mutations occurred in residues near the designed positions, likely fine-tuning the active site to enhance substrate binding. Specifically, the Ile7Asp mutation positions the Lys222, preventing its interaction with Glu10. This stabilization of the charge at the transition state increased the pK_a of Glu101 from less than 4.5 to 5.9, thereby enhancing the catalytic efficiency by optimizing the active site's electrostatic environment. Additionally, mutations on the enzyme's surface likely contributed to increased protein stability and solubility.

Another *de novo* Kemp eliminase, KE70, features a His17-Asp45 dyad as the catalytic base (Figure 1.10B). In this dyad, the imidazole of histidine is positioned by the carboxylate group of the aspartate to increase its basicity. The enzyme also includes Ser138 as the hydrogen bond donor and Tyr48 as the π -stacking residue. KE70 catalyzes Kemp elimination of 5NBZ with a catalytic efficiency of $78 \text{ M}^{-1}\text{s}^{-1}$. The enzyme was subjected to both computationally guided optimization and directed evolution, with beneficial mutations incorporated combinatorially. Nine rounds of optimization yielded an evolved variant with more than 400-fold improvement in catalytic efficiency, achieved through the introduction of eight mutations. Similar to KE07, in the evolved KE70 variant, the electrostatics in the active site were modified. In the designed enzyme, Arg69 interacted with the catalytic dyad, quenching its basicity. In the evolved variant, Gly101 is mutated to a serine which interacts with Arg69, redirecting it away from the catalytic dyad. Additionally, the mutations led to a deeper active site pocket with a narrower entrance, potentially improving substrate binding. Furthermore, molecular dynamics (MD) simulations suggested that in the

evolved KE70 variant, the active site residues were less flexible, and the catalytic dyad was more preorganized.⁶⁷

In 2012, Privett *et al.*²³ designed enzymes for Kemp elimination using an iterative approach. The initial design, HG1, was inactive. To investigate the cause of this inactivity, the authors employed X-ray crystallography and MD simulations. They discovered that HG1's active site was solvent-exposed and contained water molecules, and critical active site residues adopted unfavourable orientations and exhibited high flexibility. To address these issues, a second round of design was performed, shifting the active site deeper into the barrel to create a more hydrophobic environment, which can enhance substrate binding and increase the pK_a of the base. This second round of design yielded an active variant, HG2, which was further analyzed using MD simulations. It was proposed that a single S265T mutation could improve packing within the active site. Indeed, incorporating this mutation resulted in the HG3 variant, which showed a three-fold improvement in catalytic efficiency. HG3 utilizes Asp127 as the catalytic base, Lys50 as the hydrogen bond donor and Trp44 as the π -stacking residue (Figure 1.10C). HG3 was later evolved through 17 rounds, resulting in the evolved variant HG3.17 with a catalytic efficiency of $230,000 \text{ M}^{-1}\text{s}^{-1}$. HG3.17 was co-crystallized with the transition state analogue (TSA), 6-nitrobenzotriazole, revealing key insights into its enhanced activity.⁶⁸ Broom *et al.* used room temperature crystallography to study the evolutionary trajectory of HG3 variants, identifying three primary sources for the enhanced activity. Firstly, in HG3.17, the active site is better packed, leading to the rigidification of catalytic residues. Secondly, the active site exhibits improved pre-organization, favouring productive substrate binding. Thirdly, the entrance to the active site is wider in HG3.17, facilitating ligand entry and release.⁶⁹

Another Kemp eliminase designed by Privett and coworkers²³ is the 1A53-2 enzyme, originally designed to cleave 5-nitrobenzisoxazole. However, its crystal structure revealed that it could also react with 6-nitrobenzisoxazole, which was confirmed experimentally. This enzyme contains Glu178 as the base and two π -stacking residues, Trp110 and Trp210, to stabilize the transition state (Figure 1.10D). The optimization of 1A53-2 involved nine rounds of directed evolution, screening for 6-nitrobenzisoxazole cleavage. In the first five rounds, all first-shell residues were subjected to saturation mutagenesis, with beneficial mutations combined through DNA shuffling. In the final four rounds, the entire gene was subjected to error-prone PCR and DNA shuffling, resulting in the 1A53-2.9 variant, which showed a 3200-fold improvement in activity compared to the original design. Unfortunately, 1A53-2.9 failed to crystallize, so the structure of the intermediate variant 1A53-2.5 was solved in complex with the TSA 5-nitrobenzotriazole. The mutation A157Y was found to form a hydrogen bond with Glu178, positioning and stabilizing it, which caused a decrease in pK_a from 8.7 to 5.4. Additionally, mutations such as A157Y, A180C, and L184F improve ligand shape complementarity by filling the space surrounding the TSA.⁷⁰ In a recent study, Rakotoharisoa *et al.* reported a computational method to significantly enhance the catalytic efficiency of two *de novo* enzymes, HG3 and KE70. By using ensembles generated from dynamics-based refinement against X-ray diffraction data of the enzymes' crystal structures, they achieved a remarkable 100–250-fold increase in catalytic efficiency. This improvement is comparable to the results obtained through screening thousands of variants in multiple rounds of directed evolution.⁷¹

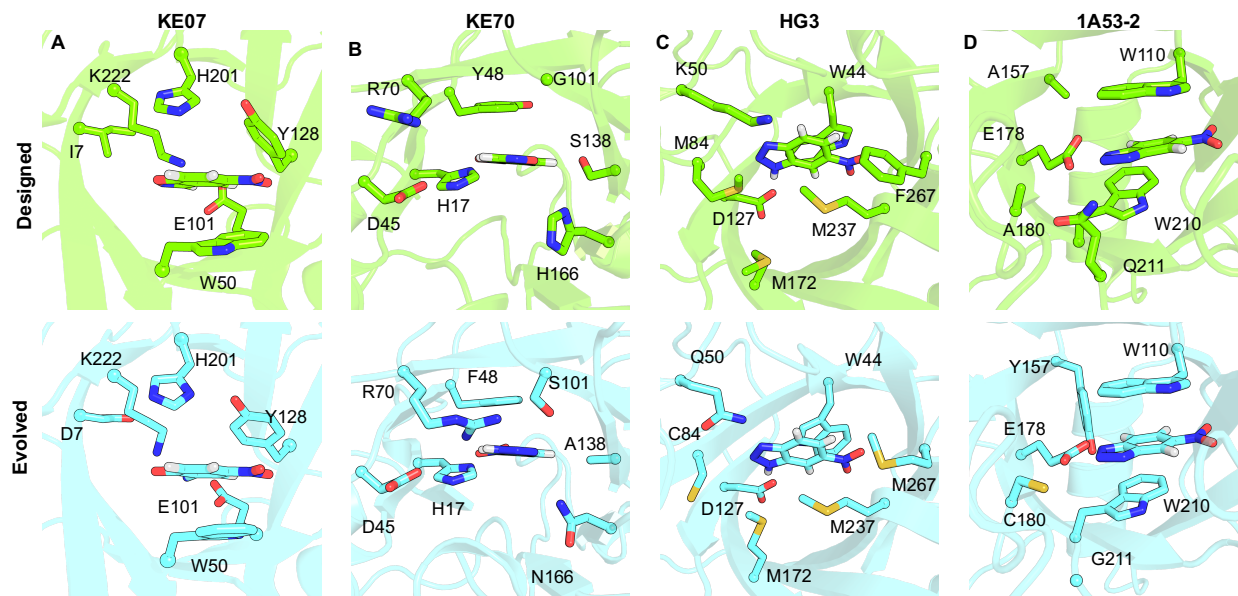


Figure 1.10. Active site structures of *de novo* designed and evolved Kemp eliminases. Active site structures of *de novo* designed (green) and laboratory-evolved (blue) Kemp eliminases. (A) KE07 features Glu101 as the catalytic base, Lys222 as the hydrogen bond donor, and Trp50 as the π -stacking residue. These residues are conserved in the evolved variant. KE07 designed (PDB ID: 2RKX⁶⁶) and evolved (PDB ID: 6CT3⁷²). (B) KE70 contains a His17-Asp45 dyad as the catalytic base, Tyr48 as the π -stacking residue (which gets mutated to Phe during evolution), and Ser138 as the hydrogen bond donor, (mutated to Ala in the evolved variant). KE70 designed (PDB ID: 3NPV⁶⁷) and evolved (PDB ID: 3Q2D⁶⁷). (C) HG3 includes Asp127 as the catalytic base and Lys50 as the hydrogen bond donor, which is mutated to Gln in the evolved variant. HG3 designed (PDB ID: 5RGA⁶⁹) and evolved (PDB ID: 5RGE⁶⁹). (D) 1A53-2 features Glu178 as the catalytic base with two π -stacking residues, Trp110 and Trp210. In the evolved variant, A157Y mutation forms a hydrogen bond with Glu178, lowering its pK_a. 1A53-2 designed (PDB ID: 3NZ1²³) and evolved (PDB ID: 6NW4⁷⁰).

1.3.3. Retro-aldol reaction

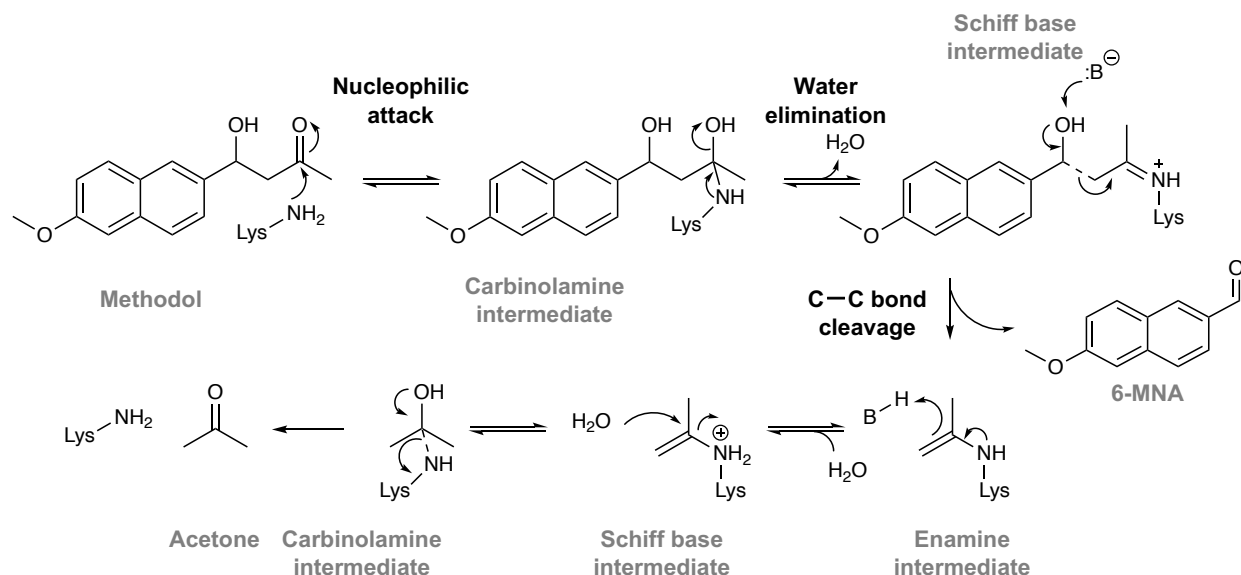


Figure 1.11. Retro-aldol reaction. Retro-aldol decomposition of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) catalyzed by RA95 using a nucleophilic lysine residue. The products 6-methoxy-2-naphthaldehyde (6-MNA) and acetone are produced.

Kemp elimination is an elementary reaction with a single transition state, unlike most useful organic reactions, which often involve multiple steps, transition states, and intermediates. Designing an enzyme to catalyze such complex reactions is more challenging. Jiang *et al.*⁷³ and Althoff *et al.*⁷⁴ addressed this by designing enzymes for the retro-aldol reaction. In this reaction, a catalytic lysine first attacks the ketone of the substrate, methodol, to form a carbinolamine. A water molecule is then eliminated, forming a Schiff base that acts as an electron sink. This facilitates the carbon-carbon bond cleavage triggered by the deprotonation of the β -alcohol. Finally, the enamine tautomerizes to form an imine, which is hydrolyzed to release the covalently bound product (Figure 1.11).

Jiang *et al.* used a composite transition state and four catalytic motifs to design retro-aldolase. In the composite transition state, models of multiple intermediates and transition states were superimposed, primarily represented by the carbinolamine and the transition state for the carbon-carbon bond-breaking step. The catalytic lysine, which requires a low pK_a to act efficiently as a nucleophile, was modelled either by introducing a second lysine in the active site or by placing the lysine in a hydrophobic environment. Among the four modelled active sites, the most successful designs featured a hydrophobic environment and a water molecule to stabilize the carbinolamine hydroxyl and β -hydroxyl groups. Out of 72 designs tested, 32 displayed retro-aldol activity.⁷³ Althoff *et al.* demonstrated the robustness of this method by designing new retro-aldolases on different scaffolds using the motif with a nonpolar environment and a water molecule. Initially, these designs exhibited low activity, which was subsequently enhanced through *in vitro* evolution.⁷⁴ Separately, Bjelic *et al.* replaced the water molecule with a glutamate, resulting in 5 out of 7 designs displaying activities comparable to previous successful designs.⁷⁵

Of note, the RA95 designed retro-aldolase featured an active site with Lys210 as the catalytic lysine for Schiff base formation and Glu53 to orient a water molecule (Figure 1.12A). RA95 displayed a rate enhancement of approximately 5,000 and a catalytic efficiency of $0.17 \text{ M}^{-1}\text{s}^{-1}$, far lower than that of natural enzymes. To enhance activity, directed evolution was employed. The first optimization phase generated RA95.5, introducing five active site mutations (V51Y, E53S, T83K, M180F, R182M) and one second-shell mutation (D183N). These changes increased k_{cat}/K_M by 94-fold.⁷⁴ The intermediate RA95.5-5, resulting from further rounds of random mutagenesis and screening, included six additional mutations (E53T, S110N, G178S, R23H, R43S, T95M). This variant showed increased loop mobility and active site plasticity, allowing for simultaneous labelling of Lys210 and the newly introduced Lys83. The subsequent

optimization phase yielded RA95.5-8, which displayed a catalytic efficiency of $1,600 \text{ M}^{-1}\text{s}^{-1}$.⁷⁶ Utilizing a droplet-based microfluidic screening platform, RA95.5-8 was subjected to six rounds of mutagenesis and screening, resulting in RA95.5-8F with a $>10^9$ rate enhancement. During evolution, the original catalytic lysine 210 switched to lysine 83 and a catalytic tetrad of Lys83-Tyr51-Asn110-Tyr180, which forms a hydrogen bond network, emerged (Figure 1.12B).⁷⁷

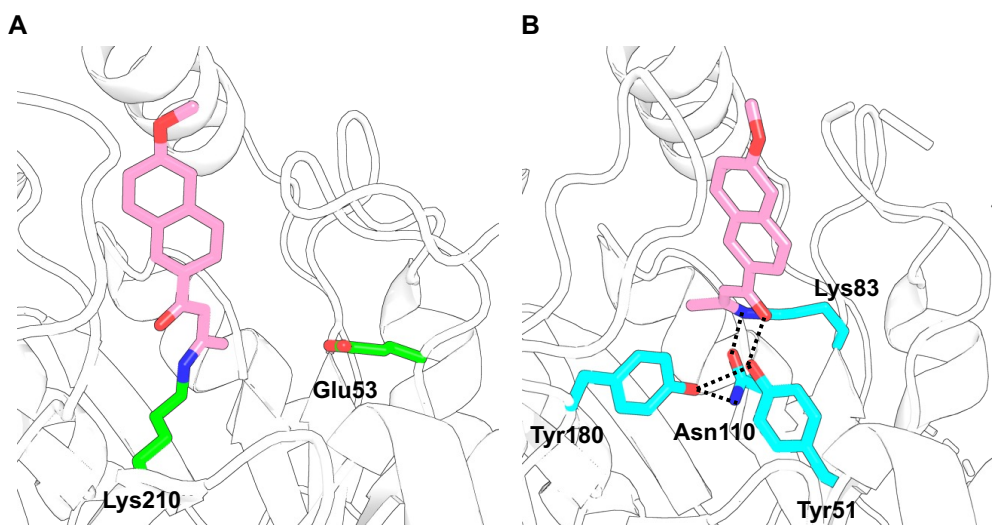


Figure 1.12. Active site structures of designed and evolved retroaldolases. Active site configurations of (A) RA95 (PDB ID: 4A29⁷⁶) and (B) RA95.5-8F (PDB ID: 5AN7⁷⁷). The diketone inhibitor used in crystallization is shown in pink. RA95 was designed to have a catalytic lysine (Lys210), and a glutamate (Glu53) positioned to orient a water molecule. During directed evolution, a catalytic tetrad including a new catalytic lysine (Lys83) and three additional residues forming a hydrogen bond network (Tyr51, Asn110, and Tyr180) were identified.

In a recent study, Braun *et al.*⁷⁸ used RFdiffusion,⁵¹ a machine-learning model for generation of protein structures, to design new catalysts for the retro-aldol reaction by using the highly optimized catalytic tetrad of evolved retro-aldolase RA95.5-8F obtained by multiple rounds of directed evolution. They developed the Rotamer Inverted Fragment Finder-Diffusion, Riff-Diff method, which generates artificial motif libraries and performs backbone diffusion and refinements. Using RFdiffusion, they generated alpha-helical scaffolds able to accommodate the

catalytic motif. Out of 35 experimentally characterized designs, 31 catalyzed the retro-aldol reaction, displaying higher catalytic efficiencies compared to previously designed retro-aldolases.

1.3.4. Engineering of a Michaelase

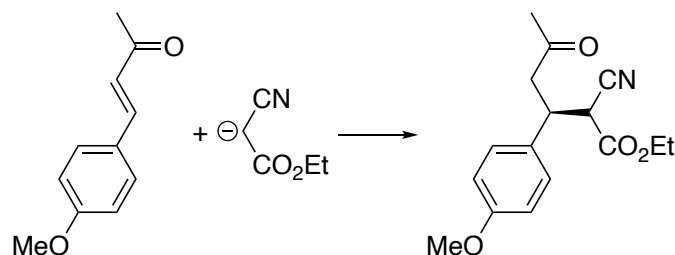


Figure 1.13. Michael addition reaction. C—C bond-forming Michael addition of (*E*)-4-(4-methoxyphenyl)but-3-en-2-one and ethyl cyanoacetate

It was hypothesized that the evolved retro-aldolase enzymes could also catalyze the C—C bond-forming Michael addition reaction. Consequently, RA95.5-8 was tested for activity towards a variety of unsaturated methyl ketones as potential Michael acceptors and a range of Michael donors. The reaction between (*E*)-4-(4-methoxyphenyl)but-3-en-2-one and ethyl 2-cyanoacetate proceeded rapidly and produced a high yield (Figure 1.13). Through random mutagenesis of nine active site residues, a variant with two mutations, T53L and K210H, was identified, exhibiting a threefold increase in catalytic efficiency. Crystallographic analysis of RA95.5-8 T53L/K210H complexed with a derivative of the Michael product revealed that the ligand was covalently bound to the catalytic Lys83 with its aromatic ring occupying the space where the naphthyl group of methodol binds. This study highlights the potential of computationally designed enzymes to exhibit catalytic promiscuity, which can be harnessed and optimized through directed evolution.⁷⁹

1.3.5. Ester hydrolysis

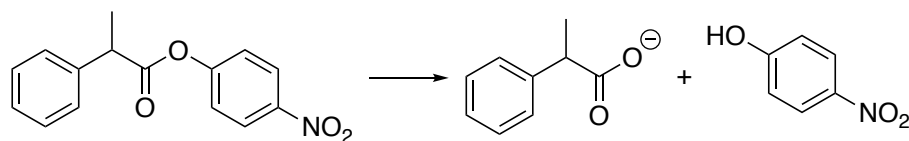


Figure 1.14. Ester hydrolysis. Hydrolysis of p-nitrophenyl-2-phenylpropanoate catalyzed by ECH13.

Richter *et al.*⁸⁰ designed enzymes for ester hydrolysis using QM-derived theozymes consisting of a Cys-His dyad and three different hydrogen bond donors to stabilize the oxyanion intermediate (Figure 1.14). These hydrogen bond donors included backbone amide groups, explicit water molecules, or sidechain functional groups. They employed the RosettaMatch theozyme placement algorithm to scan a set of 214 protein backbone templates to identify scaffolds capable of accommodating the theozymes. This process resulted in 207 unique matches, whose active sites were repacked and filtered down to 55 sequences. Four designs exhibited esterase activity, with the most active variant, ECH13, displaying a catalytic efficiency of 309 M⁻¹s⁻¹. ECH13 exhibited a rate enhancement one order of magnitude higher than PZD2, the hydrolase designed by Bolon and Mayo⁶², likely due to using a transition state model rather than a high-energy transition state proxy.

In a recent study, Lauko *et al.*⁸¹ employed RFdiffusion, a generative model for protein design, to generate protein structures to scaffold the active site of serine hydrolases, consisting of a Ser-His-Asp catalytic triad and oxyanion hole. They utilized ChemNet, a deep neural network trained on protein-small molecule complexes in the PDB, to evaluate the preorganization of active site residues. Iterative rounds of LigandMPNN for sequence design and Rosetta FastRelax to refine protein backbone and ligand pose were performed. Designs with catalytic residues positioned

within 1 Å of the AlphaFold2 predictions were selected for experimental testing. In the first round, 125 designs were tested, followed by 52 in the second round. Two designs from round one and 10 designs from round two displayed catalytic activity. However, their progress curves plateaued after producing one enzyme equivalent of product, suggesting they catalyzed the initial nucleophilic attack but failed to hydrolyze the acyl-enzyme intermediate. In the third round, a histidine-stabilizing catalytic acid and a second oxyanion hole hydrogen bond donor were incorporated, and ChemNet was used to evaluate preorganization of designs. Of the 132 designs in round three, two designs exhibited multiple turnover activity, with catalytic efficiencies higher than those of previously designed serine hydrolases.

1.3.6. Diels-Alder reaction

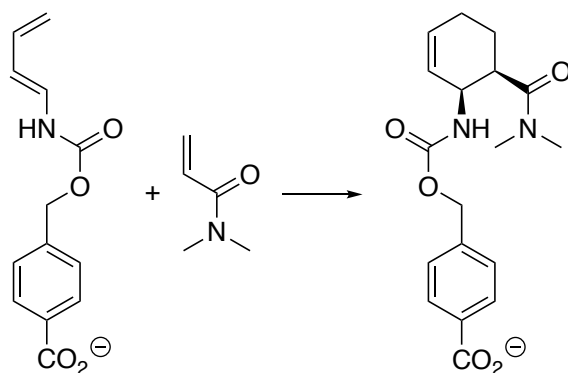


Figure 1.15. Diels-Alder. Diels–Alder [4 + 2] cycloaddition of 4-carboxy-benzyl-trans-1,3-butadiene-1-carbamate on N,N-dimethylacrylamide catalyzed by DA_20_10.

The reactions discussed thus far—Kemp elimination, retro-aldol, and ester hydrolysis—are all bond-breaking reactions. In contrast, the Diels-Alder reaction is a bimolecular bond-forming reaction where a conjugated diene and a dienophile undergo a cycloaddition to form an unsaturated six-membered ring (Figure 1.15). This reaction presents a greater design challenge because both substrates must be properly oriented to facilitate catalysis and ensure

stereoselectivity. Siegel and co-workers⁸² tackled this challenge by designing enzymes to catalyze the Diels-Alder reaction. They constructed a theozyme incorporating Asn/Gln as a hydrogen bond donor to the N-H group of the diene carbamate and Ser/Thr/Tyr as a hydrogen bond donor to the carbonyl oxygen of the dienophile amide. Out of 84 designs tested, two displayed Diels-Alderase activity: DA_42_00 and DA_20_00. These designs were further improved by individually mutating residues in contact with the transition state to amino acids predicted to enhance transition state binding, increasing their activities by 100-fold and 20-fold, respectively. The improved variant DA_20_10 was subjected to eight rounds of error-prone PCR and screening, resulting in DA_20_20, which exhibited a 5.5-fold higher catalytic efficiency.⁸³ Independently, Eiben *et al.*⁸⁴ designed a 24-residue helix-turn-helix to act as a lid for DA_20_10, constraining the substrates to a productive binding pose in a design named CE6. The beneficial mutations from DA_20_20 and the lid fragment in CE6 were combined to form CE11, which underwent further directed evolution. After eight rounds of mutagenesis and screening, CE20 was identified, demonstrating a 9,700-fold increase in chemical proficiency compared to the original DA_20_00.⁸³

1.3.7. Morita–Baylis–Hillman reaction

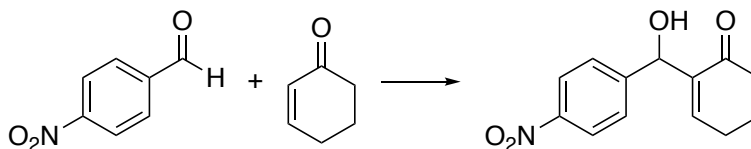


Figure 1.16. Morita–Baylis–Hillman reaction. Morita–Baylis–Hillman reaction between 4-nitrobenzaldehyde and 2-cyclohexanone catalyzed by BH32.

The Morita–Baylis–Hillman (MBH) reaction is another C–C bond-forming reaction for which no naturally occurring enzymes exist. In this reaction, a single C–C bond is formed between an α,β -unsaturated carbonyl and an electrophilic carbon, such as aldehyde (Figure 1.16). Bjelic *et*

*al.*⁸⁵ targeted the reaction between 2-cyclohexanone and 4-nitrobenzaldehyde, which contains three intermediates and four transition states. To catalyze this reaction, they designed a composite theozyme incorporating a nucleophile (His/Cys), hydrogen bond donors to stabilize intermediates, and aromatic residues for stacking interactions with benzaldehyde. Out of 48 designs tested, two exhibited MBH activity as determined by mass spectrometry. Efforts to enhance their activity included site saturation mutagenesis, leading to the most active design, BH32 Q128H, which showed modest activity with a rate enhancement of 54. Further optimization through 14 rounds of directed evolution resulted in BH32.14. This variant demonstrated a substantial increase in catalytic efficiency and enantioselectivity, achieving over a 710-fold improvement in activity compared to the initial design.⁸⁶

1.3.8. Chemiluminescence reaction

Yeh and coworkers aimed to design luciferase enzymes capable of catalyzing the oxidative chemiluminescence of two synthetic luciferin substrates, diphenylterazine (DTZ) and 2-deoxycoelenterazine.⁸⁷ They employed a deep-learning-based, family-wide hallucination approach to generate a large set of idealized protein structures with diverse pockets, providing shape complementarity to the substrates. Unlike traditional enzyme design methods, they did not construct an idealized theozyme. Instead, they focussed on designing a shape-complementary site for the anionic state of the substrates, given that computational and experimental data indicated that chemiluminescence proceeds through an anionic species. To stabilize this anionic state, the designed active sites placed an arginine guanidium moiety adjacent to it. They generated an ensemble of anionic DTZ conformers, created rotamer interaction fields (RIF) around each conformer, and then docked these into each scaffold. Out of 7,648 designs tested, three exhibited luciferase activity. These designs were further optimized by site saturation mutagenesis of the

substrate binding pocket residues. The most active enzyme displays a catalytic efficiency of $10^6 \text{ M}^{-1}\text{s}^{-1}$, matching natural luciferases but with a much greater substrate specificity.

1.4. Challenges in computational enzyme design

The above examples showcase the potential of computational enzyme design in successfully designing enzymes capable of catalyzing various model organic transformations. However, the overall success rate of computational enzyme design remains relatively low. Many designs either fail to express solubly^{73,80,82} or fail to bind the substrate.²³ The catalytic proficiency (K_{tx}^{-1}) of the active designed enzymes is often 11 orders of magnitude lower than naturally occurring enzymes. Interestingly, natural and *de novo* enzymes have comparable substrate binding affinities (K_M), but *de novo* enzymes possess lower rate enhancements (k_{cat}/k_{uncat}),⁸⁸ suggesting that the design algorithms can build active sites that bind ligands productively, but this alone does not guarantee high catalytic efficiency.

Equation 6. Catalytic proficiency

$$\text{Catalytic proficiency (M}^{-1}\text{)} = K_{tx}^{-1} = \frac{k_{cat}/K_M}{k_{uncat}}$$

The rate of enzymatic reactions is determined by the rate-limiting step within the entire catalytic process, which includes substrate binding, possible conformational changes, chemical transformation, and product release. Therefore, the rate-limiting step may not always correspond to the chemical reaction modelled by the theozyme. Studies have shown that for some enzymes, the rate-limiting step can be substrate binding,⁸⁹ product release,⁹⁰ or conformational changes between sub-states.⁹¹

Another challenge is related to the dynamic nature of enzymes. Enzymatic catalysis typically involves conformational changes, leading to deviations in backbone and sidechain conformations from those of the designed model. Since sub-Ångstrom level of precision is essential to achieve highly efficient catalysis, even slight deviations can significantly impact the enzyme's performance.^{92,93}

Furthermore, the inherent limitations of theozyme calculations in predicting the transition state structure can impede the design of efficient enzymes. Typically, these calculations are conducted in the gas phase with only minimal residues required for catalysis and without considering the effects of the protein environment. Catalytic residues often form hydrogen bonds to second-shell residues that are also excluded from the theozyme model.⁹⁴ The following sections will explore challenges to enzyme design in greater detail.

1.4.1. The challenge of predicting distal mutations

Given the low catalytic efficiencies of designed enzymes, directed evolution is often employed to enhance their activity. Enzymes catalyzing Kemp elimination,^{66–68,95,96} retro-aldol reaction,^{74,76,77} Dies-Alder reaction,⁸³ and Morita-Baylis-Hillman reaction⁸⁶ have been optimized by multiple rounds of directed evolution, increasing their efficiencies by up to 5 orders of magnitude. During evolution, activity-enhancing mutations are identified throughout the enzyme structure, including both active-site and distal positions⁹⁷ (Figure 1.17). The mechanisms by which mutations increase enzyme activity vary depending on their proximity to the active site. Notably, the effects of mutations on enzyme turnover (k_{cat}) do not directly correlate with their proximity to the active site, unlike specificity, which is strongly influenced by active site mutations.⁹⁸ Typically, mutations in inner shells tend to have larger effects on activity; however, several reported cases of

activity-enhancing distal mutations exist.⁹⁷ For example, a single mutation introduced 15 Å away from the active site of the LovD enzyme increased the enzyme's activity by twofold.⁹⁹ Similarly, distal mutations located more than 9 Å away from the active site in NOV1, a bacterial dioxygenase, resulted in a ~10-fold increase in activity.¹⁰⁰ In another study, a mutation 13 Å away from the active site increased hydrolysis of ceftazidime by a β -lactamase by more than 600-fold.^{97,101}

Active-site mutations can enhance activity by introducing new catalytic interactions,^{68,74,76} pre-organizing the active site,⁶⁹ or reshaping the binding pocket. The role of distal mutations in catalysis, however, is less understood. They have been shown to increase activity by conformational tinkering and fine-tuning of the active site. For instance, during the directed evolution of *N*-acyl-homoserine lactonase towards paraoxon, two distal mutations caused an active site residue, Phe68, to be displaced by 3 Å, enhancing interactions with the new substrate's leaving group.¹⁰² In another example, Jiménez-Osés *et al.* observed that laboratory evolution of the LovD enzyme for improved simvastatin synthesis led to distal mutations that restricted the conformational sampling of Tyr188, stabilizing a catalytically active conformation.⁹⁹ Similarly, during the evolution of KE07 kemp eliminase, distal mutations stabilized the most catalytically efficient active site configuration, optimizing the positioning of Trp50 relative to the substrate.⁷²

Distal mutations can also affect enzymes' conformational dynamics, causing population shifts or redistribution of enzymes' conformational substates. For instance, distal mutations in an evolved variant of monoamine oxidase (MAO), MAO-D5, containing mutations with a mean distance of 18-19 Å of the active site, altered the enzyme's conformational dynamics. These mutations made the open conformation kinetically and thermodynamically unfavourable, stabilizing closed and partially closed states.¹⁰³ In the directed evolution of TrpB, a non-

monomerically active subunit of the heterodimer tryptophan synthase (TrpS), the introduction of six distal mutations allowed TrpB to exhibit stand-alone activity. MD simulations revealed that these mutations enabled TrpB to sample open, partially closed and closed conformations similar to TrpS.¹⁰⁴

Despite the advancements highlighted by these studies, these findings are highly case-specific, with mutations and their effects often not generalizable across different enzymes or reaction conditions. Understanding the impact of distal mutations on catalytic efficiency is further convoluted by the multifaceted nature of natural enzyme evolution. Mutations are selected not only for increased activity but also for factors like specificity, stability, promiscuity, and other chemical constraints. Therefore, studying the effects of mutations on catalytic efficiency requires novel, creative approaches. Because the role of distal mutations in enzyme catalysis is poorly understood, predicting distal mutations in computational enzyme design is very challenging. Typically, enzyme design algorithms focus on building the active site and ignore distal residues, yet distal mutations have been shown to significantly impact the catalytic properties of many enzymes, making their prediction by computational enzyme design essential for creating highly active enzymes.

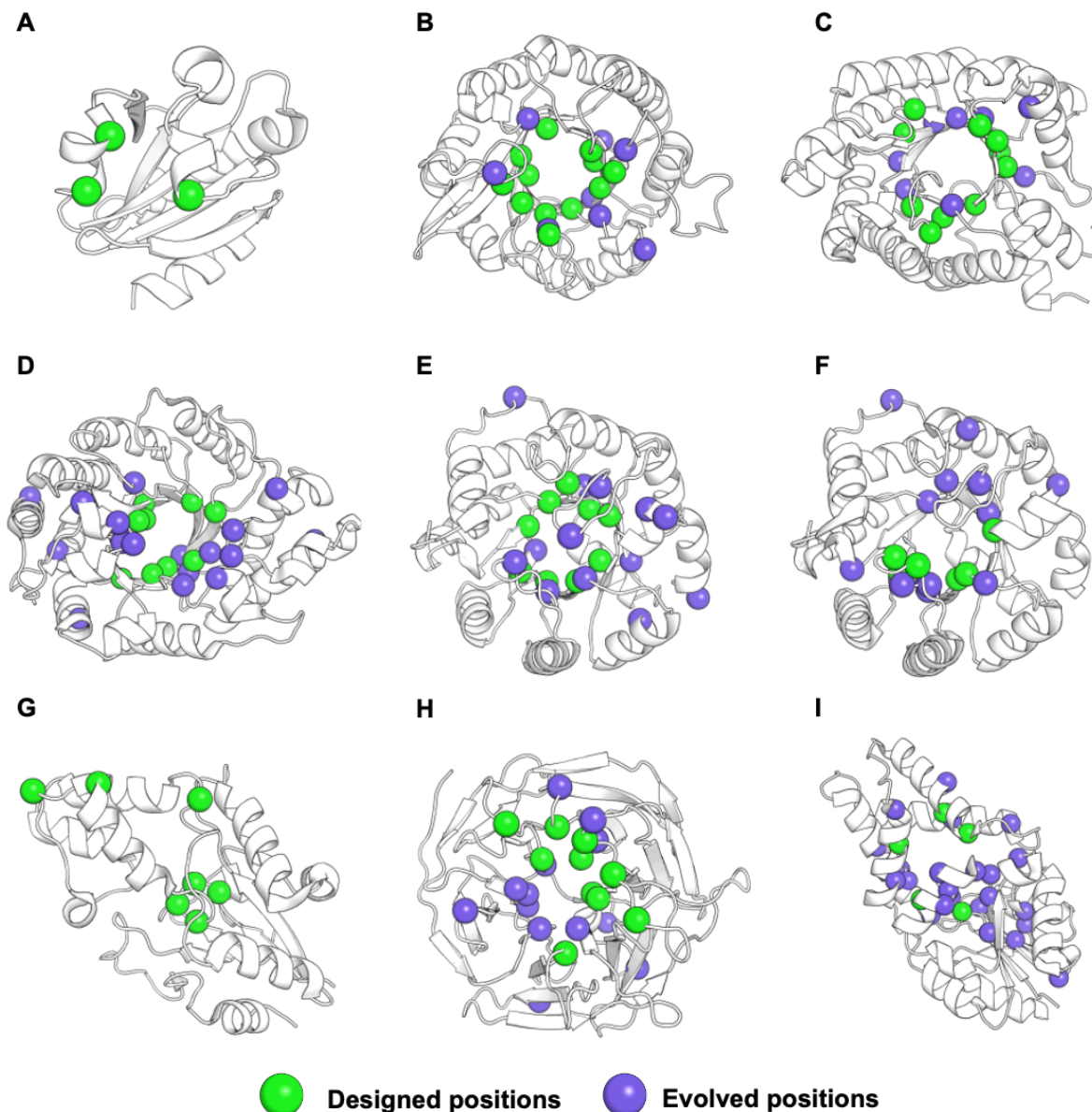


Figure 1.17. Distribution of designed and evolved positions in *de novo* enzymes. Positions that were computationally designed and mutated during directed evolution are shown as green and purple spheres, respectively. (A) *E.coli* thioredoxin scaffold (PDB ID: 2TRX¹⁰⁵) used in the design of PDZ2. (B) Imidazoleglycerolphosphate synthase from *T. Maritima* (PDB ID: 1THF¹⁰⁶) used in the design of KE07. (C) D-2-deoxyribose-5-phosphate aldolase (PDB ID: 1JCL¹⁰⁷) used in the design of KE70. (D) *T. auranticus* xylanase scaffold (PDB ID: 1GOR¹⁰⁸) used in the design of HG-2. (E) *S. solfataricus* indole-3-glycerolphosphate synthase scaffold (PDB ID: 1A53¹⁰⁹) used in the design of KE59. (F) *S. solfataricus* indole-3-glycerolphosphate synthase scaffold (PDB ID: 1LBL¹⁰⁹) used in the design of RA95. (G) *H. sapiens* mitochondrial deoxyribonucleotidase scaffold (PDB ID: 1Q91¹¹⁰) used in the design of ECH13. (H) *L. vulgaris* diisopropylfluorophosphatase scaffold (PDB ID: 1E1A¹¹¹) used in the design of DA_20_00. (I) *P. horikoshii* haloacid dehalogenase scaffold (PDB ID: 1X42¹¹²) used in the design of BH32.

1.4.2. The challenge of identification of suitable scaffolds

Another challenge in designing highly efficient enzymes is selecting an appropriate scaffold for the reaction of interest. This process typically involves repurposing native protein scaffolds identified through a search of the Protein Data Bank, which contains over 144,000 protein structures.¹¹³ Given the vast number of protein structures available, finding a suitable scaffold can be computationally expensive and time-consuming. Furthermore, natural proteins are highly evolved to perform their native functions, so there may not be a natural scaffold with a binding pocket well-suited to the novel reaction of interest. Natural scaffolds often contain structural segments such as flexible loops that are difficult to model *in silico*, and they are often marginally stable; therefore, introducing mutations can further destabilize them.^{114–116} An alternative to repurposing natural protein scaffolds is the *de novo* generation of scaffolds, as demonstrated in the design of luciferases⁸⁷, serine hydrolases⁸¹, and retro-aldolases⁷⁸. This approach theoretically allows for the creation of protein backbone geometries that are tailor-made to the reaction of interest. However, designing an enzyme requires a high level of precision that *in-silico* models alone cannot achieve without experimental validation.. For instance, in the case of luciferases, only three enzymes exhibited activity out of 7,648 designs tested.⁸⁷ Similarly, for serine hydrolases, hundreds of designs were tested, yet the successful variants displayed catalytic efficiencies orders of magnitude lower than natural hydrolases.⁸¹ For retro-aldolases, the turnover rate of the most active designed enzyme was 300-fold lower than that of previously evolved retro-aldolases.⁷⁸ These challenges underscore the need for alternative approaches to scaffold selection or generation.

1.4.3. The challenge of stabilizing multiple transition states

Creating an efficient *de novo* enzyme requires more than just stabilizing a single transition state. An efficient enzyme ensures low energy barriers across all reaction steps. Targeting only the rate-limiting step for barrier reduction can bring anti-catalytic effects on other steps. A notable example of this can be seen during the evolution of retro-aldolase enzymes. In the RA95.5-5 variant, the rate-limiting step corresponds to the third step, which consists of C–C bond cleavage of methodol. In the highly evolved RA95.5-8F variant, however, the third step is stabilized, and the rate-limiting step is no longer the third step, but it's the fourth, which corresponds to the formation of the Schiff base.¹¹⁷

During the design of enzymes for retro-aldol⁷² and MBH reactions,⁸² composite transition states were used. A composite transition state is generated by the superposition of multiple intermediates and transition states. This approach, however, does not address the problems arising from the use of a single transition state.

1.5. Machine learning methods in enzyme design

In 1959, Arthur Samuel, a pioneer in the field of artificial intelligence, defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed”.¹¹⁸ With recent advances in machine learning algorithms, the surge in computational power, and the increase in protein sequence and structure data, several new tools have emerged with significant implications for enzyme design.

One such tool is RFdiffusion, a guided diffusion model for protein structure generation. RFdiffusion uses denoising diffusion probabilistic models (DDPMs) that work by introducing

Gaussian noise to input data in several repeated steps and then learning to denoise the corrupted, noisy data to generate new samples. These models have been used to create high-quality images that never existed before.¹¹⁹ RFdiffusion applies the same principle to protein structures, adding 3D Gaussian noise to C α coordinates and Brownian motion to N-C α -C rotations and then learning to denoise to create diverse protein structures.⁵¹ RFdiffusion can be used to generate new protein structures of a desired length, secondary structure, fold, oligomeric structures, functional-motif scaffolding and protein-binding proteins. Therefore, in theory, RFdiffusion can generate new protein backbone structures that can serve as initial structures for enzyme design.

However, protein backbone structures generated by RFdiffusion lack sequences. This is where protein message passing neural network (ProteinMPNN), a deep learning tool, comes into play. ProteinMPNN generates amino acid sequences that can fold to a desired 3D structure.¹²⁰ Traditional methods like Rosetta rely on rotamer optimization and energy calculation and are computationally intensive. ProteinMPNN, however, achieved superior performance, recovering 52.4% of sequences on native protein backbones compared to Rosetta's 32.9%. It is broadly applicable to monomers, cyclic oligomers, nanoparticles and target-binding proteins. The method leverages deep learning to rapidly generate candidate sequences, bypassing the need for extensive computational resources. Experimental validation showed that ProteinMPNN can rescue previously failed designs, producing sequences that fold accurately and confidently into target structures. Recently, ProteinMPNN was applied to improve thermostability and yield of TEV protease and human myoglobin.⁵⁷

Another powerful machine learning tool that can aid enzyme design is AlphaFold2, developed by Alphabet's DeepMind. AlphaFold2 is a deep learning model that can predict protein 3D structures from their amino acid sequences, effectively addressing the long-standing protein

folding problem. In recognition of this breakthrough, the 2024 Nobel Prize in Chemistry was awarded, in half, to the creators of AlphaFold2, for their achievement in solving the challenge of predicting proteins' complex structures. The AlphaFold model has been trained on sequences and structures from the PDB, multiple sequence alignments (MSA), and known structures of homologous proteins or templates. It extracts features such as amino acid sequences, physical properties, and evolutionary relationships from these sources. The MSA and the template are then passed through a transformer, which captures evolutionary information and spatial relationships between residues. This data is subsequently passed on to the structure module that predicts the 3D coordinates of the proteins by iterative refinement, improving accuracy in each step.¹²¹ Recently, a new version, AlphaFold3, was released, extending its capabilities beyond proteins to include other biomolecules such as DNA, RNA and small molecule ligands.¹²² The above-mentioned tools, and other rapidly advancing ones, hold great potential in protein design.

1.6. Thesis objectives

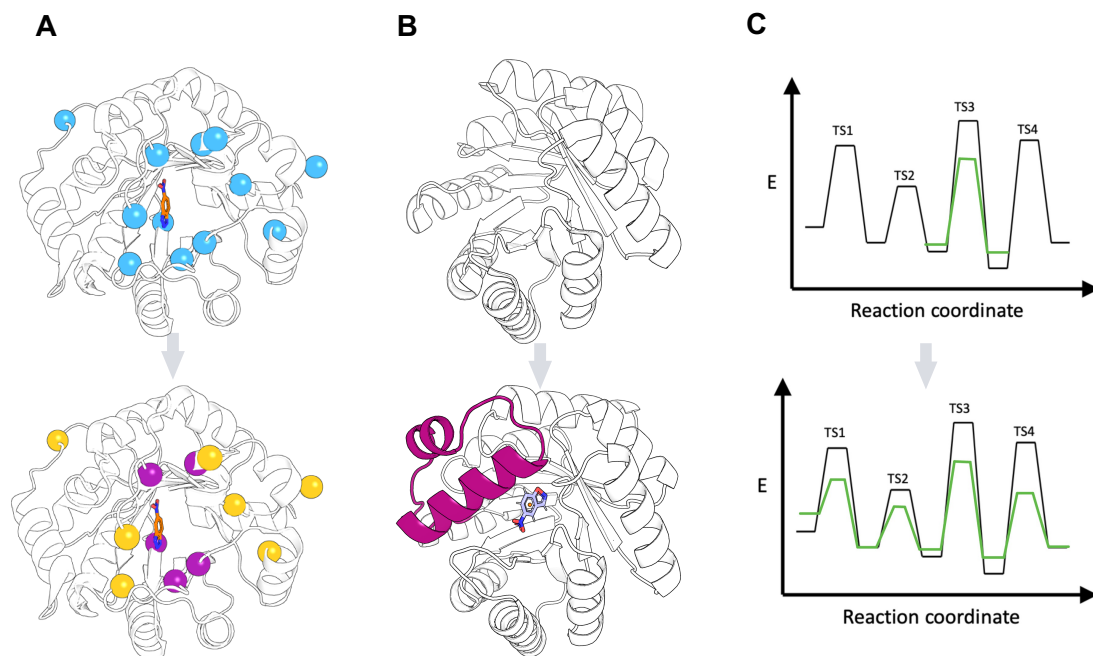


Figure 1.18. Challenges in enzyme design. The first objective of this thesis is to (A) investigate how distal mutations impact catalysis, (B) develop a workflow to customize a minimal scaffold for the reaction of interest, and (C) employ multistate enzyme design to design an enzyme for a multistep reaction.

In this thesis, our objective is to enhance enzyme design methodologies to reduce the current reliance on directed evolution by addressing several key challenges facing enzyme design. As outlined, during evolution, activity-enhancing mutations appear throughout the enzyme structure. How residues far from the active site contribute to catalysis is not well understood. Typically, in enzyme design, only the active site is designed, as identifying distal residues that contribute to catalysis is challenging. *De novo*-designed and laboratory-evolved enzymes offer a unique opportunity to unravel the effects of active-site and distal mutations since these mutations were selected solely for their ability to increase catalytic activity. In Chapter 2, our objective is to

study the role of distal mutations in enzyme catalysis, aiming to determine whether designing distal residues is necessary for creating efficient enzymes (Figure 1.18A).

A second challenge in enzyme design is the availability of protein scaffolds that are well-suited to the reaction of interest. Previous enzyme design efforts have focused on either repurposing natural scaffolds^{66,73,85} or generating new *de novo* scaffolds.^{81,87} Both approaches have limitations. While natural scaffolds offer a diverse range of shapes and topologies, they carry millions of years of evolutionary history, are marginally stable,¹¹⁴ and are often further destabilized by mutations that engineer new functions.¹²³ On the other hand, *de novo* scaffolds have well-characterized sequence-function relationships but lack the sub-ångstrom precision required for efficient catalysis. In Chapter 3, we present a new approach for scaffold selection where we develop a workflow to customize a minimal, *de novo* and functionless TIM barrel to create a functional enzyme for the Kemp elimination reaction (Figure 1.18B). This chapter also evaluates the suitability of machine learning methods for enzyme design.

The third challenge is the design of efficient enzymes for multistep reactions. Stabilizing a single transition state is not sufficient to create highly efficient enzymes for useful, multistep reactions. An efficient enzyme must maintain low energy barriers across the whole reaction pathway. Previous efforts focused on stabilizing a single^{23,66} or composite^{73,85} transition state. To address this, in Chapter 4, we employ a multistate enzyme design approach to design an enzyme to catalyze the Michael addition reaction, a C—C bond-forming reaction (Figure 1.18C).

1.7. References

1. Wolfenden, R. & Snider, M. J. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).
2. Sharma, A, Gupta, G., Ahmed, T. & Mansoor, S. Enzyme Engineering: Current Trends and Future Perspectives. *Food Rev. Int.* **37**, 122–124 (2021).
3. Gupta, R. D., Goldsmith, M., Ashani, Y., Simo, Y., Mullokandov, G., Bar, H., Ben-David, M., Leader, H., Margalit, R., Silman, I., Sussman, J. L. & Tawfik, D. S. Directed evolution of hydrolases for prevention of G-type nerve agent intoxication. *Nat. Chem. Biol.* **7**, 120–125 (2011).
4. Bornscheuer, U. T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol.* **5**, 137–143 (2001).
5. Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
6. Onuffer, J. J. & Kirsch, J. F. Redesign of the substrate specificity of Escherichia coli aspartate aminotransferase to that of Escherichia coli tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Sci. Publ. Protein Soc.* **4**, 1750–1757 (1995).
7. Chen, K. Q. & Arnold, F. H. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnol. Nat. Publ. Co.* **9**, 1073–1077 (1991).
8. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).
9. Pan, X. & Kortemme, T. Recent advances in *de novo* protein design: Principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).
10. DeGrado, W. F., Prendergast, F. G., Wolfe Jr., H. R. & Cox, J. A. The design, synthesis, and characterization of tight-binding inhibitors of calmodulin. *J. Cell. Biochem.* **29**, 83–93 (1985).
11. Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
12. Dahiyat, B. I. & Mayo, S. L. Protein design automation. *Protein Sci.* **5**, 895–903 (1996).

13. Street, A. G. & Mayo, S. L. Computational protein design. *Structure* **7**, 105–109 (1999).
14. Schmidt Am Busch, M., Lopes, A., Mignon, D. & Simonson, T. Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *J. Comput. Chem.* **29**, 1092–1102 (2008).
15. Rohl, C. A. & Baker, D. *De Novo* Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc.* **124**, 2723–2729 (2002).
16. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542 (1992).
17. Desmet, J., Spriet, J. & Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins Struct. Funct. Bioinforma.* **48**, 31–43 (2002).
18. Koehl, P. & Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Mol. Biol.* **2**, 163–170 (1995).
19. Patrice, Koehl & Marc, Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275 (1994).
20. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Chapter 4. Computational Enzyme Design: Successes, Challenges, and Future Directions. in *Catalysis Series* (eds. Williams, G. & Hall, M.) 88–116 (Royal Society of Chemistry, Cambridge, 2018). doi:10.1039/9781788010450-00088.
21. Boas, F. E. & Harbury, P. B. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* **17**, 199–204 (2007).
22. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O’Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta All-

- Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
23. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).
24. Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc. Natl. Acad. Sci.* **107**, 19838–19843 (2010).
25. Chica, R. A., Moore, M. M., Allen, B. D. & Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc. Natl. Acad. Sci.* **107**, 20257–20262 (2010).
26. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
27. Sitkoff, D., Sharp, K. A. & Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **98**, 1978–1988 (1994).
28. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10172–10177 (1997).
29. Lazaridis, T. & Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477–487 (1999).
30. Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **319**, 199–203 (1986).
31. Street, A. G. & Mayo, S. L. Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **3**, 253–258 (1998).
32. Gadda, G. & Sobrado, P. Kinetic Solvent Viscosity Effects as Probes for Studying the Mechanisms of Enzyme Action. *Biochemistry* **57**, 3445–3453 (2018).

33. Gainza, P., Nisonoff, H. M. & Donald, B. R. Algorithms for protein design. *Curr. Opin. Struct. Biol.* **39**, 16–26 (2016).
34. Voigt, C. A., Gordon, D. B. & Mayo, S. L. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design 1 Edited by J. Thornton. *J. Mol. Biol.* **299**, 789–803 (2000).
35. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* **97**, 10383–10388 (2000).
36. Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.* **31**, 904–916 (2010).
37. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
38. Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **3**, 107–115 (1935).
39. Laidler, K. J. & King, M. C. Development of transition-state theory. *J. Phys. Chem.* **87**, 2657–2664 (1983).
40. Molecular Architecture and Biological Reactions. *Chem. Eng. News Arch.* **24**, 1375–1377 (1946).
41. Tantillo, D. J., Jiangang, C. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).
42. Ryde, U. Chapter Six - QM/MM Calculations on Proteins. in *Methods in Enzymology* (ed. Voth, G. A.) vol. 577 119–158 (Academic Press, 2016).
43. Orio, M., Pantazis, D. A. & Neese, F. Density functional theory. *Photosynth. Res.* **102**, 443–453 (2009).
44. Sholl, D. S. & Steckel, J. A. *Density Functional Theory: A Practical Introduction*. (Wiley, Hoboken, N.J, 2009).
45. Foresman, J. B. & Frisch, A. *Exploring Chemistry with Electronic Structure Methods*. (Gaussian, Inc, Wallingford, CT USA, 2015).

46. Lewars, E. G. The Concept of the Potential Energy Surface. in *Computational Chemistry* 9–43 (Springer Netherlands, Dordrecht, 2011). doi:10.1007/978-90-481-3862-3_2.
47. Hratchian, H. P. & Schlegel, H. B. Finding minima, transition states, and following reaction pathways on ab initio potential energy surfaces. in *Theory and Applications of Computational Chemistry* 195–249 (Elsevier, 2005). doi:10.1016/B978-044451719-7/50053-6.
48. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980–980 (2003).
49. Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **51**, D488–D508 (2023).
50. MacDonald, J. T., Maksimiak, K., Sadowski, M. I. & Taylor, W. R. *De novo* backbone scaffolds for protein design. *Proteins Struct. Funct. Bioinforma.* **78**, 1311–1325 (2010).
51. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M. & Baker, D. *De novo* design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
52. Lassila, J. K. Conformational diversity and computational enzyme design. *Curr. Opin. Chem. Biol.* **14**, 676–682 (2010).
53. Davey, J. A. & Chica, R. A. Multistate Computational Protein Design with Backbone Ensembles. in *Computational Protein Design* (ed. Samish, I.) 161–179 (Springer, New York, NY, 2017). doi:10.1007/978-1-4939-6637-0_7.
54. Davis, I. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **14**, 265–274 (2006).

55. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A Simple Model of Backbone Flexibility Improves Modeling of Side-chain Conformational Variability. *J. Mol. Biol.* **380**, 757–774 (2008).
56. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* **380**, 742–756 (2008).
57. Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I. M., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., De La Cruz, J., Sankaran, B., Bera, A. K., Jiménez-Osés, G. & Baker, D. Improving protein expression, stability, and function with ProteinMPNN. *J Am Chem Soc* **146**, 2054–2061 (2024).
58. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).
59. Lassila, J. K., Baker, D. & Herschlag, D. Origins of catalysis by computationally designed retroaldolase enzymes. *Proc. Natl. Acad. Sci.* **107**, 4937–4942 (2010).
60. Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D. & Baker, D. New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
61. Dou, J., *et al.* De novo design of a fluorescence-activating β -barrel | *Nature* **561**, 485–491 (2018).
<https://www.nature.com/articles/s41586-018-0509-0>.
62. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* **98**, 14274–14279 (2001).
63. Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S. & Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).

64. Casey, M. L., Kemp, D. S., Paul, K. G. & Cox, D. D. Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **38**, 2294–2301 (1973).
65. Kemp, D. S. & Casey, M. L. Physical organic chemistry of benzisoxazoles. II. Linearity of the Broensted free energy relation for the base-catalyzed decomposition of benzisoxazoles. *J. Am. Chem. Soc.* **95**, 6670–6680 (1973).
66. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
67. Khersonsky, O., Röthlisberger, D., Wollacott, A. M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K. N., Baker, D. & Tawfik, D. S. Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *J. Mol. Biol.* **407**, 391–412 (2011).
68. Blomberg, R., Kries, H., Pinkas, D. M., Mittl, P. R. E., Grütter, M. G., Privett, H. K., Mayo, S. L. & Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
69. Broom, A., Rakotoharisoa, R. V., Thompson, M. C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J. S. & Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
70. Bunzel, H. A., Kries, H., Marchetti, L., Zeymer, C., Mittl, P. R. E., Mulholland, A. J. & Hilvert, D. Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **141**, 11745–11748 (2019).
71. Rakotoharisoa, R. V., Seifinoferest, B., Zarifi, N., Miller, J. D. M., Rodriguez, J. M., Thompson, M. C. & Chica, R. A. Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J Am Chem Soc* **146**, 10001–10013 (2024).

72. Hong, N.-S., Petrović, D., Lee, R., Gryn'ova, G., Purg, M., Saunders, J., Bauer, P., Carr, P. D., Lin, C.-Y., Mabbitt, P. D., Zhang, W., Altamore, T., Easton, C., Coote, M. L., Kamerlin, S. C. L. & Jackson, C. J. The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.* **9**, 3900 (2018).
73. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. *De Novo* Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
74. Althoff, E. A., Wang, L., Jiang, L., Giger, L., Lassila, J. K., Wang, Z., Smith, M., Hari, S., Kast, P., Herschlag, D., Hilvert, D. & Baker, D. Robust design and optimization of retroaldol enzymes: Robust Design and Optimization of Retroaldol Enzymes. *Protein Sci.* **21**, 717–726 (2012).
75. Bjelic, S., Kipnis, Y., Wang, L., Pianowski, Z., Vorobiev, S., Su, M., Seetharaman, J., Xiao, R., Kornhaber, G., Hunt, J. F., Tong, L., Hilvert, D. & Baker, D. Exploration of Alternate Catalytic Mechanisms and Optimization Strategies for Retroaldolase Design. *J. Mol. Biol.* **426**, 256–271 (2014).
76. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N. & Hilvert, D. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
77. Obexer, R., Godina, A., Garrabou, X., Mittl, P. R. E., Baker, D., Griffiths, A. D. & Hilvert, D. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
78. Braun, M., Tripp, A., Chakatok, M., Kaltenbrunner, S., Totaro, M. G., Stoll, D., Bijelic, A., Elaily, W., Hoch, S. Y. Y., Aleotti, M., Hall, M. & Oberdorfer, G. Computational design of highly active *de novo* enzymes. Preprint at <https://doi.org/10.1101/2024.08.02.606416> (2024).
79. Garrabou, X., Beck, T. & Hilvert, D. A Promiscuous *De Novo* Retro-Aldolase Catalyzes Asymmetric Michael Additions via Schiff Base Intermediates. *Angew. Chem. Int. Ed.* **54**, 5609–5612 (2015).

80. Richter, F., Blomberg, R., Khare, S. D., Kiss, G., Kuzin, A. P., Smith, A. J. T., Gallaher, J., Pianowski, Z., Helgeson, R. C., Grjasnow, A., Xiao, R., Seetharaman, J., Su, M., Vorobiev, S., Lew, S., Forouhar, F., Kornhaber, G. J., Hunt, J. F., Montelione, G. T., Tong, L., Houk, K. N., Hilvert, D. & Baker, D. Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).
81. Lauko, A., Pellock, S. J., Anischanka, I., Sumida, K. H., Juergens, D., Ahern, W., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I. R., Jamieson, C., Kang, A., Brackenbrough, E., Bera, A. K., Sankaran, B., Houk, K. N. & Baker, D. Computational design of serine hydrolases. Preprint at <https://doi.org/10.1101/2024.08.29.610411> (2024).
82. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E. & Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
83. Preiswerk, N., Beck, T., Schulz, J. D., Milovnik, P., Mayer, C., Siegel, J. B., Baker, D. & Hilvert, D. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci.* **111**, 8013–8018 (2014).
84. Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Players, F., Stoddard, B. L., Popovic, Z. & Baker, D. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* **30**, 190–192 (2012).
85. Bjelic, S., Nivón, L. G., Çelebi-Ölçüm, N., Kiss, G., Rosewall, C. F., Lovick, H. M., Ingalls, E. L., Gallaher, J. L., Seetharaman, J., Lew, S., Montelione, G. T., Hunt, J. F., Michael, F. E., Houk, K. N. & Baker, D. Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita–Baylis–Hillman Reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).

86. Crawshaw, R., Crossley, A. E., Johannissen, L., Burke, A. J., Hay, S., Levy, C., Baker, D., Lovelock, S. L. & Green, A. P. Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
87. Yeh, A. H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S. J., Evans, D., Ma, P., Lee, G. R., Zhang, J. Z., Anishchenko, I., Coventry, B., Cao, L., Dauparas, J., Halabiya, S., DeWitt, M., Carter, L., Houk, K. N. & Baker, D. *De novo* design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
88. Mak, W. S. & Siegel, J. B. Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
89. Wang, L. H., Tsai, A. L. & Hsu, P. Y. Substrate binding is the rate-limiting step in thromboxane synthase catalysis. *J. Biol. Chem.* **276**, 14737–14743 (2001).
90. Chen, N., Zhao, Y., Lu, J., Wu, R. & Cao, Z. Mechanistic Insights into the Rate-Limiting Step in Purine-Specific Nucleoside Hydrolase. *J. Chem. Theory Comput.* **11**, 3180–3188 (2015).
91. Kerns, S. J., Agafonov, R. V., Cho, Y.-J., Pontiggia, F., Otten, R., Pachov, D. V., Kutter, S., Phung, L. A., Murphy, P. N., Thai, V., Alber, T., Hagan, M. F. & Kern, D. The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124–131 (2015).
92. Linder, M. COMPUTATIONAL ENZYME DESIGN: ADVANCES, HURDLES AND POSSIBLE WAYS FORWARD. *Comput. Struct. Biotechnol. J.* **2**, e201209009 (2012).
93. Li, X., Zhang, Z. & Song, J. COMPUTATIONAL ENZYME DESIGN APPROACHES WITH SIGNIFICANT BIOLOGICAL OUTCOMES: PROGRESS AND CHALLENGES. *Comput. Struct. Biotechnol. J.* **2**, e201209007 (2012).
94. DeChancie, J., Clemente, F. R., Smith, A. J. T., Gunaydin, H., Zhao, Y.-L., Zhang, X. & Houk, K. N. How similar are enzyme active site geometries derived from quantum mechanical theozymes to crystal structures of enzyme-inhibitor complexes? Implications for enzyme design. *Protein Sci. Publ. Protein Soc.* **16**, 1851–1866 (2007).

95. Khersonsky, O., Kiss, G., Röthlisberger, D., Dym, O., Albeck, S., Houk, K. N., Baker, D. & Tawfik, D. S. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci.* **109**, 10358–10363 (2012).
96. Khersonsky, O., Röthlisberger, D., Dym, O., Albeck, S., Jackson, C. J., Baker, D. & Tawfik, D. S. Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J. Mol. Biol.* **396**, 1025–1042.
97. Yang, G., Miton, C. M. & Tokuriki, N. A mechanistic view of enzyme evolution. *Protein Sci.* **29**, 1724–1747 (2020).
98. Currin, A., Neil, Swainston, Philip, D Jay, & Douglas B Kell. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev* **44**, 1172–1239 (2015).
99. Jiménez-Osés, G., Osuna, S., Gao, X., Sawaya, M. R., Gilson, L., Collier, S. J., Huisman, G. W., Yeates, T. O., Tang, Y. & Houk, K. N. The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **10**, 431–436 (2014).
100. De Simone, M., Alonso-Cotchico, L., Lucas, M. F., Brissos, V. & Martins, L. O. Distal mutations enhance efficiency of free and immobilized NOV1 dioxygenase for vanillin synthesis. *J. Biotechnol.* **391**, 92–98 (2024).
101. Vakulenko, S. B., Taibi-Tronche, P., Tóth, M., Massova, I., Lerner, S. A. & Mobashery, S. Effects on Substrate Profile by Mutational Substitutions at Positions 164 and 179 of the Class A TEMpUC19 β -Lactamase from *Escherichia coli*. *J. Biol. Chem.* **274**, 23052–23060 (1999).
102. Yang, G., Hong, N., Baier, F., Jackson, C. J. & Tokuriki, N. Conformational Tinkering Drives Evolution of a Promiscuous Activity through Indirect Mutational Effects. *Biochemistry* **55**, 4583–4593 (2016).

103. Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J. & Osuna, S. Hidden Conformations in *Aspergillus niger* Monoamine Oxidase are Key for Catalytic Efficiency. *Angew. Chem. Int. Ed.* **58**, 3097–3101 (2019).
104. Maria-Solano, M. A., Iglesias-Fernández, J. & Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **141**, 13049–13056 (2019).
105. Katti, S. K., LeMaster, D. M. & Eklund, H. Crystal Structure of Thioredoxin from *Escherichia coli* at 1.68 Å Resolution. *J Mol Biol* **212**, 167–184 (1990).
106. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. *Science* **289**, 1546–1550 (2000).
107. Heine, A., DeSantis, G., Luz, J. G., Mitchell, M., Wong, C.-H. & Wilson, I. A. Observation of Covalent Intermediates in an Enzyme Mechanism at Atomic Resolution. *Science* **294**, 369–374 (2001).
108. Lo Leggio, L., Kalogiannis, S., Eckert, K., Teixeira, S. C. M., Bhat, M. K., Andrei, C., Pickersgill, R. W. & Larsen, S. Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. *FEBS Lett.* **509**, 303–308 (2001).
109. Hennig, M., Darimont, B. D., Jansonius, J. N. & Kirschner, K. The Catalytic Mechanism of Indole-3-glycerol Phosphate Synthase: Crystal Structures of Complexes of the Enzyme from *Sulfolobus solfataricus* with Substrate Analogue, Substrate, and Product. *J. Mol. Biol.* **319**, 757–766 (2002).
110. Rinaldo-Matthis, A., Rampazzo, C., Balzarini, J., Reichard, P., Bianchi, V. & Nordlund, P. Crystal Structures of the Mitochondrial Deoxyribonucleotidase in Complex with Two Specific Inhibitors. *Mol. Pharmacol.* **65**, 860–867 (2004).
111. Scharff, E. I., Koepke, J., Fritzsche, G., Lücke, C. & Rüterjans, H. Crystal structure of diisopropylfluorophosphatase from *Loligo vulgaris*. *Structure* **9**, 493–502 (2001).

112. Arai, R., Kukimoto-Niino, M., Kuroishi, C., Bessho, Y., Shirouzu, M. & Yokoyama, S. Crystal structure of the probable haloacid dehalogenase PH0459 from *Pyrococcus horikoshii* OT3. *Protein Sci.* **15**, 373–377 (2006).
113. wwPDB consortium *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
114. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Bioinforma.* **46**, 105–109 (2002).
115. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
116. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).
117. De Raffe, D., Martí, S. & Moliner, V. Understanding the Directed Evolution of *De Novo* Retro-Aldolases from QM/MM Studies. *ACS Catal.* **10**, 7871–7883 (2020).
118. Awad, M. & Khanna, R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. (Apress, Berkeley, CA, 2015). doi:10.1007/978-1-4302-5990-9.
119. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J. & Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Proc. 36th Int. Conf. Neural Inf. Process. Syst.* 36479–36494 (2024).
120. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., De Haas, R. J., Bethes, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P. & Baker, David. Robust deep learning based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
121. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

122. Abramson, J. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
123. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4**, e1000002 (2008).

CHAPTER 2. DISTAL MUTATIONS ENHANCE CATALYSIS IN DESIGNED ENZYMES BY FACILITATING SUBSTRATE BINDING AND PRODUCT RELEASE

Niayesh Zarifi,^{1,2} Pooja Asthana,³ Hiva Doostmohammadi,^{4,5} Cindy Klaus,^{1,2} Janet Sanchez,^{4,5} Serena E. Hunt,^{1,2} Rojo V. Rakotoharisoa,^{1,2} Sílvia Osuna,^{4,5} James S. Fraser³ & Roberto A. Chica^{1,2,*}

¹ Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

² Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

³ Department of Bioengineering and Therapeutic Science, University of California, San Francisco, San Francisco, California 94158, United States

⁴ CompBioLab Group, Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Girona, Spain

⁵ ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain

2.1. Foreword

As discussed in Chapter 1, current enzyme design methods fail to achieve the rate enhancements seen in natural enzymes, and designing enzymes for industrially relevant reactions remains a significant challenge. We hypothesized that one contributing factor to the low catalytic efficiency of *de novo* designed enzymes is that residues distal to the active site are not considered in the design procedure. The role of residues far from the active site in catalysis remains poorly understood, as previous studies have often explored the impact of distal mutations alongside active-site mutations, obscuring the direct contribution of distal residues. In this chapter, we aimed to address this knowledge gap by studying functional, structural, and dynamic effects of distal mutations. Our goal was to understand the role that distal mutations play in catalysis and deconvolute their effects on enzymatic activity from those of active-site mutations. We investigated whether efficient enzymes can be created by focusing solely on the active site, or if distal mutations are essential for achieving high catalytic efficiency and current enzyme design methodologies need to be improved to also consider distal mutations.

2.2. Statement of contribution

Niayesh Zarifi, Cindy Klaus, Rojo V. Rakotoharisoa, and Serena E. Hunt performed enzyme kinetics. Niayesh Zarifi performed circular dichroism and thermal melting experiments. Niayesh Zarifi, Hiva Doostmohammadi, and Janet Sanchez performed molecular dynamics simulations. Niayesh Zarifi performed principal component analysis. Pooja Asthana crystallized proteins and performed X-ray diffraction experiments. Pooja Asthana and Roberto A. Chica performed refinements. Pooja Asthana and James S. Fraser designed X-ray crystallography experiments. Cindy Klaus performed kinetic solvent viscosity and binding assay experiments. Niayesh Zarifi performed stopped-flow experiments. Niayesh Zarifi and Roberto A. Chica conceived this project. Niayesh Zarifi and Roberto A. Chica wrote the chapter.

2.3. Abstract

The role of amino-acid residues distant from an enzyme's active site in facilitating the complete catalytic cycle—including substrate binding, chemical transformation, and product release—remains poorly understood. Here, we investigate how distal mutations promote the catalytic cycle by engineering mutants of three *de novo* Kemp eliminases containing either active-site or distal mutations identified through directed evolution. Kinetic analyses, X-ray crystallography, and molecular dynamics simulations reveal that while active-site mutations create preorganized catalytic sites for efficient chemical transformation, distal mutations enhance catalysis by facilitating substrate binding and product release through tuning structural dynamics to widen the active-site entrance and reorganize surface loops. These distinct contributions work synergistically to improve overall activity, demonstrating that a well-organized active site, though necessary, is not sufficient for optimal catalysis. Our findings reveal critical roles that distal residues play in shaping the catalytic cycle to enhance efficiency beyond what is achievable through active-site optimization alone.

2.4. Introduction

Enzymes are essential for life, catalyzing nearly all biochemical reactions with remarkable precision and efficiency. Understanding the molecular principles underlying enzyme catalysis has been a cornerstone of biochemistry, with decades of research providing a detailed understanding of how active-site residues orchestrate catalytic mechanisms.^{1,2} Structural, mutational, and computational studies have also demonstrated that residues located far from the active site contribute to catalysis by modulating conformational dynamics to enrich productive active-site configurations.^{3–12} Despite these insights, the specific contributions of distal residues in facilitating other aspects of the catalytic cycle, such as substrate binding and product release, remain poorly understood. Elucidating the role of distal residues in these processes in natural enzymes is challenging due to the intricate allosteric networks in their structures¹² and the presence of epistatic interactions¹³ shaped by millions of years of evolution. As a result, reliably predicting the functional impact of distal mutations remains a significant challenge,¹⁴ hindering our ability to fully understand and exploit enzyme function.

Over the past two decades, *de novo* enzyme design has emerged as an approach for engineering enzymes tailored to specific reactions.^{15–18} These methods focus on constructing active sites by incorporating catalytic residues and ligand-binding pockets into protein scaffolds,^{19,20} neglecting the role of distal residues. While this approach has successfully produced several *de novo* enzymes,^{15–18,21,22} these biocatalysts have displayed low activity, necessitating subsequent directed evolution to enhance catalytic efficiency.^{3,23–27} Throughout these evolutionary processes, numerous mutations that enhance activity were discovered, both within and far from the active site. Several studies have shed light on how evolution enhances catalysis in artificial enzymes.^{4–6,26,27} However, they have assessed the functional impacts of distal mutations alongside

active-site mutations, obscuring the direct contribution to catalysis of residues far from the active site. Moreover, these studies have focused on the role of mutations found by directed evolution in accelerating the chemical transformation,^{3–6,23,24,27–29} limiting insights into their role in substrate binding and product release—key catalytic cycle steps that can impact efficiency. Given the well-documented evolutionary paths of these *de novo* enzymes and the fact that mutations were specifically selected by directed evolution for their ability to boost catalytic efficiency, these artificial enzymes offer a unique opportunity to investigate the effects of distal mutations on the catalytic cycle without complications arising from the evolutionary history of the wild-type protein scaffold from which they are derived.

Here, we investigate the role of distal mutations in facilitating the enzyme catalytic cycle by studying three *de novo* Kemp eliminases that have undergone directed evolution. Using mutants containing either active-site or distal mutations, referred to as Core and Shell variants, respectively, we analyze their effects on enzyme activity, structure, and dynamics. Our results reveal that while active-site mutations establish highly organized active sites for efficient chemical transformation, distal mutations enhance catalysis by modulating energy barriers of substrate binding and product release to facilitate the overall catalytic cycle. This is achieved through altered structural dynamics that widen the active-site entrance and reorganize surface loops. Together, these mutation sets work synergistically to optimize enzyme activity. Our findings suggest that designing more efficient artificial enzymes will require novel strategies to balance the structural rigidity essential for precise active-site alignment with the flexibility needed for efficient progression through the catalytic cycle.

2.5. Results

2.5.1. Core and shell variants of *de novo* Kemp eliminases

To investigate how distal mutations contribute to the catalytic cycle, we generated variants of three computationally designed enzymes: HG3,¹⁸ 1A53-2,¹⁸ and KE70.¹⁵ These enzymes, generated from three distinct TIM barrel scaffolds with no sequence homology, were designed to catalyze the Kemp elimination (Figure 2.1A), a model organic transformation frequently used to benchmark enzyme design protocols. This reaction is ideal for assessing the influence of distal mutations, as it occurs in a single step, simplifying mechanistic evaluations of the chemical transformation. The original *de novo* Kemp eliminases display modest catalytic efficiencies ($k_{\text{cat}}/K_M \leq 10^2 \text{ M}^{-1} \text{ s}^{-1}$), requiring directed evolution to enhance activity.^{3, 24, 26, 27} Directed evolution improved catalytic efficiency by several orders of magnitude, mostly due to increases in k_{cat} ,^{3, 26, 27} resulting in Evolved variants containing mutations distributed throughout each enzyme's structure.

The original *de novo* enzymes, referred to as Designed variants here, lack the mutations introduced by directed evolution except for two cases: HG3-Designed, which is a single point mutant of HG3 (K50Q) that contains a catalytic hydrogen-bond donor not present in the original *de novo* enzyme but that appeared during its evolution,²⁶ and KE70-Designed, which includes two insertions in active-site loops (S20a and A240a) that were manually introduced during KE70's evolution.³ We employ these Designed variants instead of the original *de novo* enzymes for consistent comparisons within each enzyme family, ensuring identical catalytic residues and sequence lengths across all related variants (Supplementary Table 2.1).

In addition to the Designed and Evolved enzymes, we created Core and Shell variants by introducing non-overlapping sets of mutations derived from their respective Evolved forms (Figure

2.1B, Supplementary Table 2.2). Core variants contain mutations that occurred within the active site, defined as residues directly interacting with the transition-state analogue in their crystal structure (first shell), as well as those in direct contact with ligand-binding residues (second shell). These positions are those that are typically targeted during *de novo* enzyme design.^{15,18,30} By contrast, Shell variants contain mutations that occurred outside the active site. This process yielded a total of 12 enzymes containing two to 16 mutations across three distinct lineages (Table 2.1).

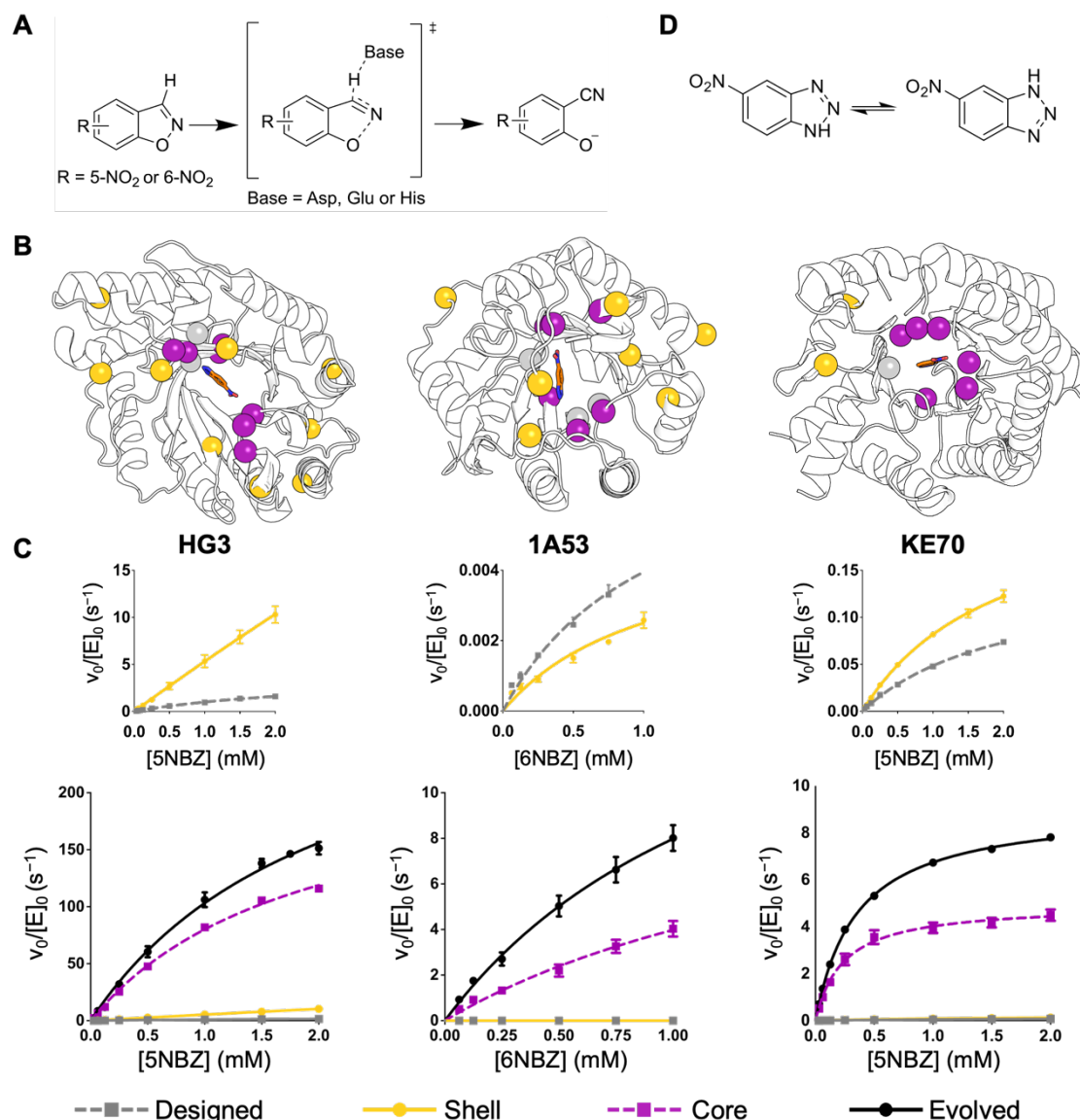


Figure 2.1. Core and Shell variants of Kemp eliminases. Kemp eliminases catalyze the concerted deprotonation and ring opening of benzisoxazoles using an Asp, Glu or His catalytic base. (B) Core and Shell variants from the HG3, 1A53 and KE70 families contain either active-site (purple) or distal (yellow) mutations found by directed evolution, respectively. Mutations are listed on Supplementary Table 2.2. Catalytic residues are shown as grey spheres. (C) Michaelis–Menten plots of normalized initial rates as a function of substrate concentration are shown. Data represent the average of six or nine individual replicate measurements from two or three independent protein batches, with error bars indicating the SEM (mean $\pm$ SEM). (D) The 6-nitrobenzotriazole transition-state analogue tautomerizes in water, which allows it to bind to enzymes that react with either 5- or 6-nitrobenzisoxazole.

2.5.2. Functional effects of active-site and distal mutations

All enzymes could be expressed and purified in good yields (Supplementary Table 2.3) except for 1A53-Shell, which afforded substantially lower yields and tended to precipitate when stored overnight at 4 °C or when concentrated. Due to this issue, we conducted enzyme kinetics immediately after purification. Core variants were 90 to 1500-fold more catalytically efficient than their corresponding Designed enzymes, and only slightly less active (1.2–2-fold) than Evolved variants (Figure 2.1C, Table 2.1). By contrast, Shell variants did not exhibit significant improvements in k_{cat}/K_M over their Designed counterparts except for HG3-Shell, which is 4-fold more catalytically efficient than HG3-Designed. These results demonstrate that active-site mutations are the primary drivers of enhanced activity, while distal mutations work synergistically with active-site mutations to further increase catalytic efficiency.

Additionally, the incorporation of active-site or distal mutations into Designed enzymes did not produce consistent effects on stability. Core and Shell variants displayed either increased or decreased stability relative to the Designed enzymes, with no clear trend (Table 2.1, Supplementary Figure 2.1–2.2). Similarly, combining both sets of mutations resulted in variable effects on stability, ranging from stabilization (HG3-Evolved) to destabilization (1A53-Evolved), or no substantial change (KE70-Evolved) relative to the Designed enzymes. These results, combined with the observation that Evolved variants are more active than Core variants, suggest that distal mutations are primarily selected during directed evolution to enhance catalytic efficiency rather than to mitigate trade-offs between activity and stability introduced by active-site mutations. This conclusion challenges prior hypotheses that distal mutations are primarily compensatory,^{31,32} instead highlighting their critical functional contributions to enzyme catalysis.

Table 2.1. Properties of various Kemp eliminases

Enzyme	Original name ^a	# mutations	K_M (mM) ^b	k_{cat} (s ⁻¹) ^b	k_{cat} / K_M (M ⁻¹ s ⁻¹) ^c	k_{cat} / K_M fold increase	T_m (°C)
HG3 series							
HG3-Designed	HG3-K50Q ^d	–	N.D.	N.D.	1300 ± 90	–	51
HG3-Shell	–	9	N.D.	N.D.	4900 ± 500	4	50
HG3-Core	HG4	7	1.9 ± 0.2	230 ± 20	120000 ± 20000	90	52
HG3-Evolved	HG3.17	16	2.1 ± 0.3	320 ± 30	150000 ± 40000	120	56
1A53 series							
1A53-Designed	1A53-2	–	N.D.	N.D.	4.6 ± 0.4	–	74
1A53-Shell	–	8	N.D.	N.D.	5.0 ± 0.7	1	65
1A53-Core	–	6	2.0 ± 0.5	13 ± 2	7000 ± 3000	1500	85
1A53-Evolved	1A53-2.9	14	1.4 ± 0.2	19 ± 2	14000 ± 3000	2800	61
KE70 series							
KE70-Designed	KE70-S20a/A240a ^e	–	N.D.	N.D.	150 ± 7	–	57
KE70-Shell	–	2	1.9 ± 0.3	0.24 ± 0.02	130 ± 30	1	60
KE70-Core	–	6	0.23 ± 0.03	5.0 ± 0.2	22000 ± 4000	150	55
KE70-Evolved	R6 6/10A	8	0.35 ± 0.02	9.1 ± 0.1	26000 ± 2000	170	58

^a Original names are those previously used in literature for these variants.

^b Kinetic parameters for all enzymes except the 1A53 series are reported for the 5-nitrobenzisoxazole substrate. For the 1A53 series, kinetic parameters are reported for 6-nitrobenzisoxazole. All reactions were performed at 27 °C in 50 mM sodium phosphate buffer (pH 7.0), supplemented with 100 mM NaCl and 10 % methanol. N.D. indicates that individual parameters K_M and k_{cat} could not be determined accurately because saturation was not possible at the maximum substrate concentration tested, which is the substrate's solubility limit (2 mM or 1 mM for 5-nitrobenzisoxazole or 6-nitrobenzisoxazole, respectively). Data represent the average of six individual replicate measurements from two independent protein batches. Errors of regression fitting, which represent the absolute measure of the typical distance that each data point falls from the regression line, are provided.

^c Catalytic efficiencies (k_{cat}/K_M) were calculated from individual K_M and k_{cat} parameters, or from the slope of the linear portion ($[S] \ll K_M$) of the Michaelis-Menten model ($v_0 = (k_{cat}/K_M)[E_0][S]$) when saturation was not possible.

^d HG3-K50Q is a point mutant of *de novo* enzyme HG3 that contains the Q50 hydrogen-bond donor identified by directed evolution. We added this mutation into HG3-Designed to allow for direct comparison with Shell, Core and Evolved HG3 variants, which all contain this mutation. HG3-K50Q is 9-fold more catalytically efficient than HG3 ($k_{cat}/K_M = 146 \text{ M}^{-1} \text{ s}^{-1}$)⁴.

^e KE70-S20a/A240a is a double mutant of *de novo* enzyme KE70 that contains two insertions in active site loops (S20a and A240a), which were manually introduced into KE70 during its directed evolution³. We added them into KE70-Designed to allow for direct comparison with Shell, Core and Evolved KE70 variants, which all contain these insertions. These insertions increase catalytic efficiency by approximately 2-fold when introduced on their own into KE70³⁰.

2.5.3. Structural effects

To investigate how active-site and distal mutations cause the observed effects on activity, we solved crystal structures of several Core and Shell variants from each lineage and compared them with available structures of their Designed,^{5,18} Core,⁴ and Evolved,^{3,4,27} counterparts. Crystals of 1A53-Core, KE70-Core, and HG3-Shell, both in the presence and absence of transition-state analogue 6-nitrobenzotriazole (6NBT, Figure 2.1D), were obtained under different conditions (Supplementary Table 2.4). The crystals diffracted at resolutions ranging from 1.44 to 2.36 Å (Supplementary Table 2.5). Unit cells of 1A53-Core (P31 2 1) and HG3-Shell (P21 21 21) contained a single protein chain and had similar dimensions regardless of the presence of 6NBT. By contrast, crystals of KE70-Core displayed variations in space group and unit cell dimensions with and without 6NBT (P1 21 1 and P21 21 21, respectively), accommodating six or two protein molecules in the asymmetric unit, respectively. All protein structures crystallized with 6NBT displayed clear density for this ligand within the active-site pocket (Supplementary Figure 2.3). In the case of 1A53-Core, the structure obtained in the absence of 6NBT shows clear density for a 2-(*N*-morpholino)ethanesulfonic acid (MES) molecule from the crystallization buffer bound in the active-site pocket, overlapping with the binding site of the transition-state analogue. In all cases, there were no substantial changes to the backbone conformation caused by the introduction of active-site or distal mutations (Supplementary Figure 2.4).

Comparison of bound and unbound structures reveals that the active-site configurations of Core and Shell variants, except for 1A53-Core, remain mostly unchanged upon 6NBT binding, with catalytic residues adopting nearly identical side-chain conformations (Figure 2.2A). These results imply that these active sites are preorganized for catalysis in the absence of the transition-state analogue. In 1A53-Core however, W110 adopts a non-productive conformation that would

prevent binding of 6NBT due to an approximately 120-degree rotation around χ_2 compared to its conformation in the bound structure. We hypothesized that this conformational change results from the presence of MES in the active site. To test our hypothesis, we performed molecular dynamics simulations of this enzyme structure in the absence of the buffer molecule, which showed that W110 readily adopts a conformation suitable for binding 6NBT via pi-stacking throughout the simulation (Supplementary Figure 2.5). These results suggest that the 1A53-Core active site is also preorganized for catalysis in the absence of 6NBT.

When comparing Core variants to their Evolved counterparts, both exhibit nearly identical active sites in the presence of the transition-state analogue (Figure 2.2B). These findings imply that active-site mutations alone are sufficient to form a binding pocket optimized for efficient chemical transformation in Core variants. On the other hand, HG3-Shell's active site closely resembles the suboptimal active site of HG3-Designed when 6NBT is present. These results demonstrate that introducing distal mutations in either Designed or Core variants does not substantially alter the active-site configuration. Yet, distal mutations enhance catalysis more effectively when introduced in Core variants than in Designed variants, suggesting that they do so through subtle effects that cannot be explained by analyzing average crystal structures alone.

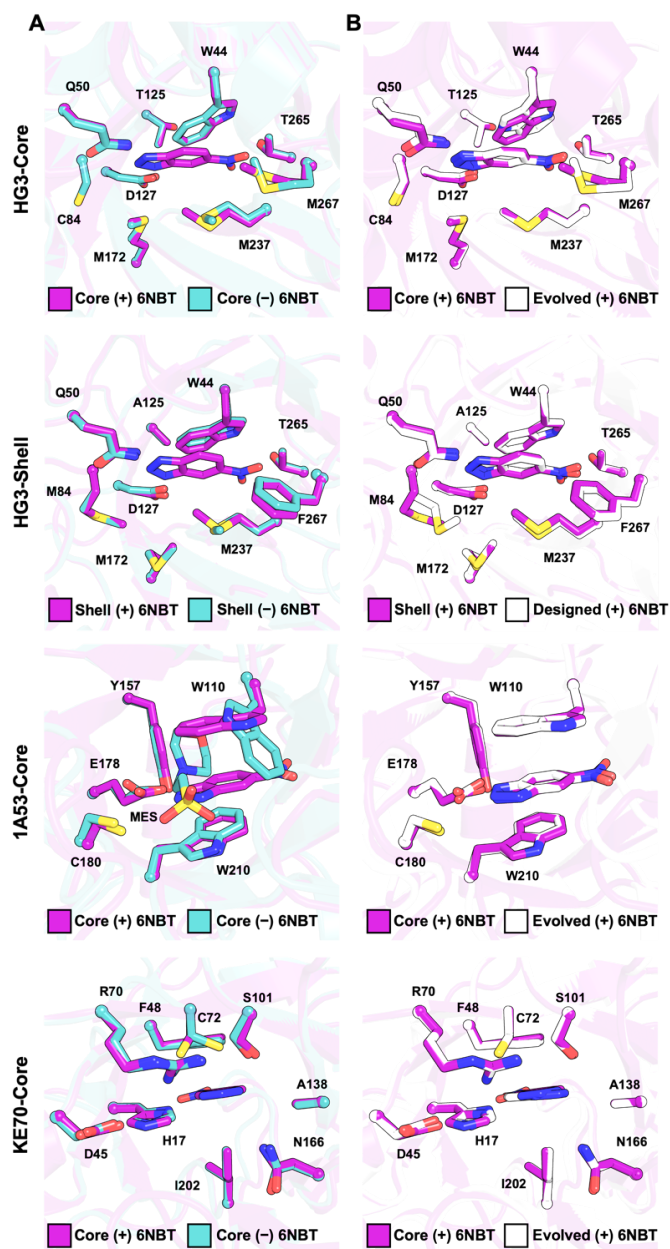


Figure 2.2. Crystal structures. (A) The active sites of HG3-Core, HG3-Shell and KE70-Core do not change substantially upon binding of the 6-nitrobenzotriazole (6NBT) transition-state analogue. In 1A53-Core, a 2-(*N*-morpholino)ethanesulfonic acid (MES) molecule from the crystallization buffer occupies the 6NBT binding site, inducing a conformational change in W110. (B) Core variants display active-site configurations nearly identical to those of their Evolved counterparts, while HG3-Shell displays a configuration similar to HG3-Designed. The 1A53-2.5 crystal structure (PDB ID: 6NW4)²⁷ was used for comparison with 1A53-Core, as 1A53-Evolved was unavailable. We re-refined the KE70-Evolved crystal structure with bound 6NBT (PDB ID: 3Q2D)³ to model missing residues (22–25 in chain A and 3–25 in chain B) into the available density and to flip 6NBT into a productive binding pose. In all cases, only the major conformer of each first-shell active-site residue is shown, with the 6NBT transition-state analogue at the center.

To explore this possibility, we assessed the impact of distal mutations on the enzyme conformational ensemble, as they have been proposed to enhance activity by enriching catalytically productive substates.¹⁴ To do so, we performed ensemble refinement of the 6NBT-bound structures (Supplementary Table 2.6) and examined the percentage of ensemble members capable of forming catalytic contacts (*e.g.*, hydrogen bonds and pi-stacking interactions) with the transition-state analogue. We found that in all cases, the incorporation of active-site mutations into Designed enzymes to yield Core variants greatly increases the proportion of substates within the conformational ensemble capable of establishing hydrogen bonds or pi-stacking interactions between catalytic residues and 6NBT, while simultaneously making catalytic residues more ordered (Figure 2.3, Supplementary Figure 2.6). Introduction of distal mutations into Designed or Core variants to yield Shell or Evolved variants, respectively, also enhances the proportion of productive substates and rigidifies catalytic residues but to a much lesser extent. Overall, crystallography revealed that active-site mutations enhance catalysis by structuring the active site for efficient chemical transformation, while distal mutations complement this effect by modestly increasing the proportion of productive substates in the conformational ensemble. The minimal impact of distal mutations on catalytic efficiency in Shell variants, compared to their greater effect in Evolved variants, suggests that they may play a more significant role in accelerating other catalytic cycle steps, such as substrate binding and product release, which can be efficiency bottlenecks.

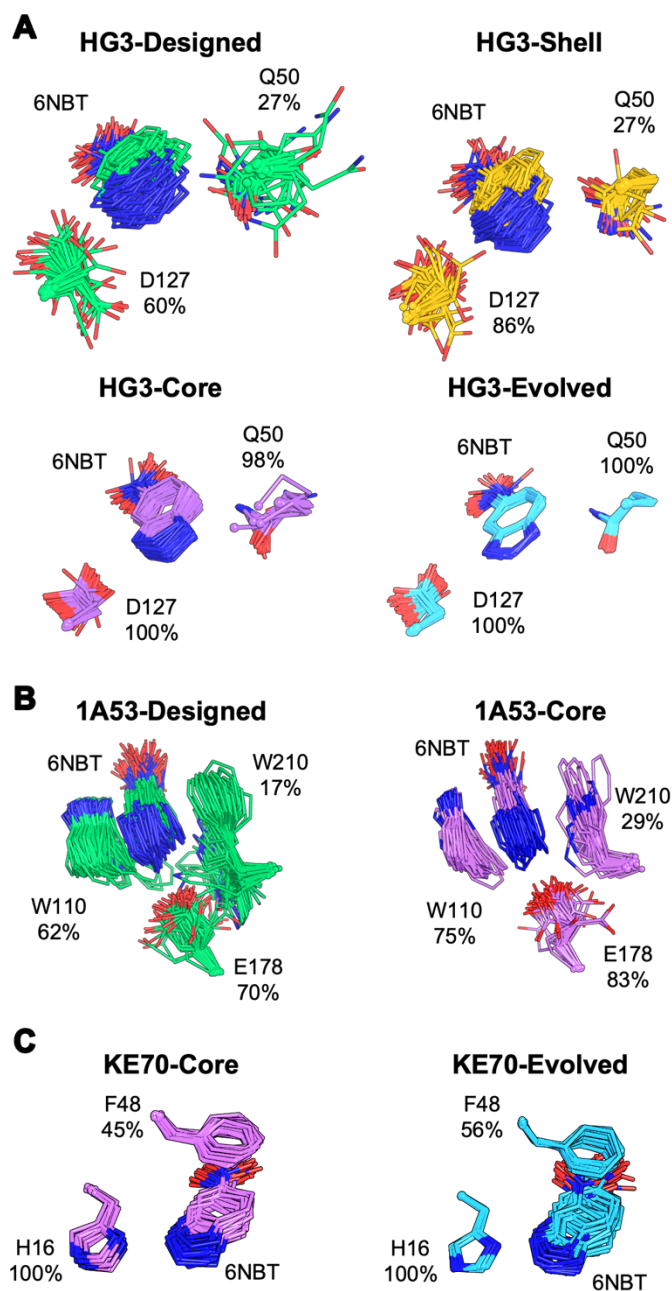


Figure 2.3. Ensemble refinement of Kemp eliminases. Ensemble refinements of 6NBT-bound crystal structures for (A) HG3, (B) 1A53, and (C) KE70 variants reveal that distal and active-site mutations increase the proportion of productive substrates in the conformational ensemble. Percentages next to each residue indicate the fraction of substrates that form hydrogen bonds or π -stacking interactions with 6NBT.

2.5.4. Mechanistic effects

Next, we examined the influence of distal mutations on substrate binding and product release (Figure 2.4A). To do so, we measured kinetic solvent viscosity effects (KSVE) on k_{cat}/K_M and k_{cat} , which probe whether productive substrate binding is limited by the rate of substrate diffusion and whether product release is the rate-limiting step in the catalytic cycle,³³ respectively. In these experiments, the rates of substrate binding, product release, and enzyme conformational changes are expected to be sensitive to solvent viscosity. By contrast, the chemical transformation itself is assumed to be independent of solvent viscosity, as it occurs within the active site, which is shielded from bulk solvent effects. For these analyses, we focused on Core and Evolved variants, as the activity of Designed and Shell variants was too low to yield reliable data when viscosity was increased.

KSVE on k_{cat} could not be measured for most enzymes due to lack of saturation within the substrate's solubility limit in the presence of viscogen (Supplementary Figure 2.7). However, KSVE on k_{cat}/K_M were measurable and revealed a value of approximately 1.0 for HG3-Evolved, indicating that productive substrate binding is much slower than the chemical transformation (Figure 2.4B). By contrast, HG3-Core exhibited a KSVE of 0.59, implying that substrate diffusion to form productive enzyme-substrate complexes and the chemical transformation occur at comparable rates, potentially due to faster diffusion than in HG3-Evolved or slower chemical transformation, or both. To explore these possibilities, we performed stopped-flow experiments to monitor the binding kinetics of the transition-state analogue 6NBT and reaction product 2-hydroxy-5-nitrobenzonitrile via quenching of intrinsic tryptophan fluorescence⁵. Due to strong inner filter effects, binding experiments with 2-hydroxy-5-nitrobenzonitrile could not be performed reliably. However, stopped-flow measurements with 6NBT (Supplementary Figure 2.8)

demonstrated that the transition-state analogue bound more rapidly to HG3-Evolved than to HG3-Core, while its unbinding rate remained unchanged (Supplementary Table 2.7). These results demonstrate that distal mutations accelerate substrate binding when incorporated into HG3-Core to form HG3-Evolved.

To investigate the influence of distal mutations on product release, we compared relative rate constants from various kinetic experiments. Given that K_M values for HG3-Core and HG3-Evolved are in the millimolar range (Table 2.1), while dissociation constants for 6NBT are in the micromolar range (Supplementary Table 2.7), substrate unbinding (k_2) is expected to be much faster than unbinding of the transition-state analogue. Since the rate of the chemical transformation (k_3) in HG3-Evolved is expected to be at least 100-fold higher than that of substrate unbinding (Figure 2.4), k_3 likely exceeds k_{cat} for this variant (Table 2.1). As a result, product release (k_4) becomes rate-limiting, since $k_{cat} = k_3 \times k_4 / (k_3 + k_4)$. This conclusion is consistent with previous kinetic isotope effect experiments on HG3-Evolved, which showed that turnover in this variant is limited by product dissociation³⁴. In HG3-Core, where the rate of the chemical transformation is only 1.4-fold higher than that of substrate unbinding ($m = k_3 / (k_2 + k_3) = 0.59$), product release may either be rate-limiting or faster than the chemical transformation. If product release is rate-limiting, it would likely be slower than in HG3-Evolved due to the lower k_{cat} value of the Core variant. Alternatively, if the chemical transformation remains rate-limiting in HG3-Core, this indicates that distal mutations shift the rate-limiting step from the chemical transformation in HG3-Core to product release in HG3-Evolved. In either case, this analysis reveals that distal mutations in HG3-Evolved not only enhance the rate of the chemical transformation, likely through the enrichment of productive substates observed in the crystallographic ensemble (Figure 2.3A), but

also modulate energy barriers related to substrate binding and product release, thereby optimizing the catalytic cycle for more efficient catalysis.

For both 1A53-Core and 1A53-Evolved, KSVE experiments on $k_{\text{cat}}/K_{\text{M}}$ revealed a slope of approximately 1.0, indicating that formation of productive enzyme-substrate complexes is much slower than the chemical transformation in both variants (Figure 2.4C). Despite stopped-flow experiments showing slower 6NBT binding to 1A53-Evolved (Supplementary Table 2.7, Supplementary Figure 2.9), its higher k_{cat} suggests that distal mutations facilitate product release. The higher K_{M} values of these enzymes (Table 2.1), compared to the transition-state analogue dissociation constants (Supplementary Table 2.7), suggest that substrate unbinding is much faster than 6NBT unbinding. Since the chemical transformation is at least 100-fold faster than substrate unbinding for these enzymes, k_3 is expected to be much larger than k_{cat} , indicating that product release is rate-limiting. These analyses show that distal mutations enhance catalysis in 1A53-Evolved primarily by accelerating product release.

For KE70-Core, KSVE on $k_{\text{cat}}/K_{\text{M}}$ revealed a near-zero slope, indicating that the diffusional encounter of substrate and this enzyme is much faster than the chemical transformation. In KE70-Evolved, a KSVE of 0.47 indicates near equivalent rates for substrate unbinding and the chemical step (Figure 2.4D). The nearly doubled k_{cat} in KE70-Evolved compared to KE70-Core (Table 2.1) reflects a faster chemical transformation in the Evolved variant rather than slower substrate capture, likely caused by the enrichment of productive conformational substates observed in its crystallographic ensemble (Figure 2.3C). While we could not perform binding assays with these enzymes because they lack a tryptophan reporter in the active site, we could measure KSVE on k_{cat} for KE70-Evolved, which revealed an inverse hyperbolic pattern (Figure 2.4E). This pattern suggests a solvent-sensitive internal isomerization of the enzyme-product complex into a more

active conformation that is favored in a more viscous medium.³³ Since k_{cat} comprises both the chemical transformation (k_3) and product release (k_4) rate constants, and no inverse hyperbolic pattern was observed for KSVE on k_{cat}/K_M , we propose that product release by KE70-Evolved is accelerated under viscous conditions by a diffusion-dependent conformational change of the enzyme-product complex. These findings highlight the critical role of conformational dynamics in accelerating product release during KE70-Evolved catalysis.

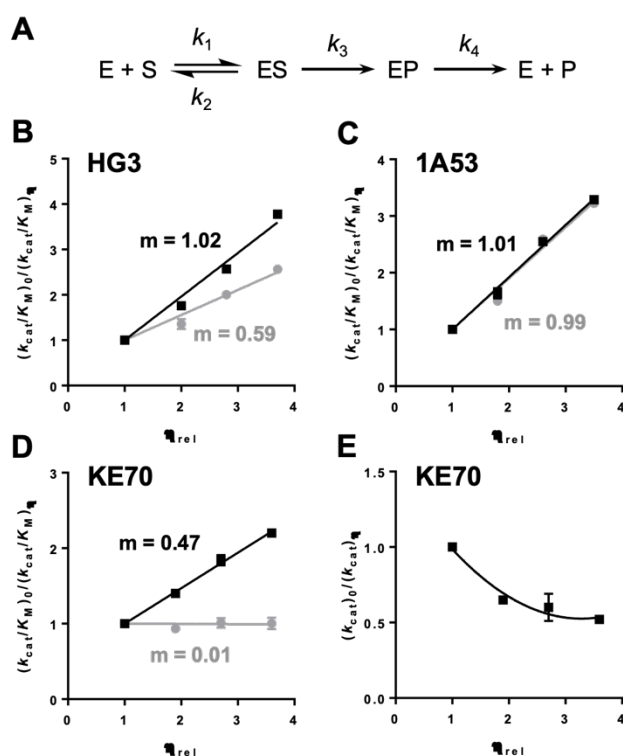


Figure 2.4. Kinetic solvent viscosity effects. (A) Kemp elimination reaction scheme. (B-E) Grey and black lines correspond to Core and Evolved variants, respectively. Kinetic solvent viscosity effects on k_{cat}/K_M for (B) HG3, (C) 1A53 and (D) KE70 variants show a linear dependence on relative viscosity (η_{rel}). The slope (m) corresponds to $k_3 / (k_2 + k_3)$, where k_2 is the rate constant of substrate dissociation and k_3 is the rate constant for the chemical transformation. A slope of 1 indicates that k_3 is much larger than k_2 ($m = k_3 / k_3$). A slope of 0 indicates that k_3 is much smaller than k_2 ($m = k_3 / k_2 = 0$). (E) Kinetic solvent viscosity effects on k_{cat} for KE70-Evolved show an inverse hyperbolic pattern, which is consistent with the presence of a solvent-sensitive internal isomerization of the enzyme-product complex. Data represent the average of 6 individual replicates from 2 independent protein batches, with error bars reporting the SEM.

2.5.5. Dynamical effects

KSVE and stopped-flow experiments showed that distal mutations facilitate substrate binding and/or product release, which could be achieved by promoting active-site opening. To test this hypothesis, we conducted microsecond-timescale molecular dynamics simulations of the enzymes in their unbound form (Methods). Root-mean-square fluctuations revealed that the most flexible regions in all enzymes were surface loops (Supplementary Figure 2.10). In 1A53 and KE70 enzymes, the most flexible loop is located at the entrance of the active site, whereas in the HG3 family, it is further away. Principal component analysis (PCA) showed that both distal and active-site mutations reshape the conformational landscape, allowing access to alternative conformations or depopulating others (Figure 2.5A). For 1A53 and KE70 enzymes, one or both principal components reflected movements of loops covering the active site, while in HG3, they involved distal loops (Supplementary Figure 2.11).

To assess active-site opening, we analyzed the distances between surface loops covering the active site (Figure 2.5B,C). In 1A53-Designed and 1A53-Shell, loop distance analysis revealed two primary conformational states: a closed state with a median loop distance of ~ 10 Å, corresponding to the crystallographic conformation, and an open state with a median distance of ≥ 17 Å. Distal mutations introduced into 1A53-Designed (to form 1A53-Shell) increased the openness of the open state, shifting the median loop distance by 2.5 Å. In 1A53-Core and 1A53-Evolved, a third, partially open state centered at ~ 15 Å was observed. Incorporating distal mutations into 1A53-Core further expanded the median open distance by more than 10 Å. These findings suggest that distal mutations enhance active-site accessibility in 1A53 by promoting more open loop conformations, which is consistent with faster product release in 1A53-Evolved.

For KE70-Designed and KE70-Shell, two states were observed: closed and open, with median loop distances at around 7 and 12.5 Å, respectively, the latter corresponding to the crystallographic conformation. Introducing distal mutations into KE70-Designed to generate KE70-Shell reduced the closed conformation population from 30% to 15%, but median loop distances remained unchanged. For KE70-Core and KE70-Evolved, a third conformational state is observed, which is partially open (median loop distance around 10.5 Å). While distal mutations introduced into KE70-Core to produce KE70-Evolved increased the closed conformation population, in contrast to their effect on KE70-Designed, the open state became more open by nearly 1 Å. These results suggest that distal mutations in KE70 facilitate opening of the active site by either increasing the proportion of open substates or shifting open conformations toward a more open state.

In contrast to 1A53 and KE70 enzymes, HG3 variants exhibited minimal changes in active-site loop conformations, consistent with the most flexible loop being distant from the active site (Supplementary Figure 2.10). Since the dynamics of loops near the active site were largely unaffected, we examined whether distal mutations influenced sidechain dynamics, potentially altering the bottleneck radius of the active site entrance. Previous studies linked a larger bottleneck radius to enhanced catalysis in HG3 variants.⁴ Caver3³⁵ analysis of bottleneck radii in molecular dynamics snapshots (Supplementary Figure 2.12) revealed that distal mutations increased the percentage of structures in the conformational ensemble with a bottleneck radius ≥ 0.9 Å, a threshold chosen because it matches the radius of small ions commonly bound in protein structures, and raised the median bottleneck radius. These findings suggest that distal mutations enhance active-site accessibility in HG3 enzymes by widening the entrance without significantly altering

loop conformations. This is consistent with the faster substrate binding observed in HG3-Evolved and suggests that product release is also accelerated.

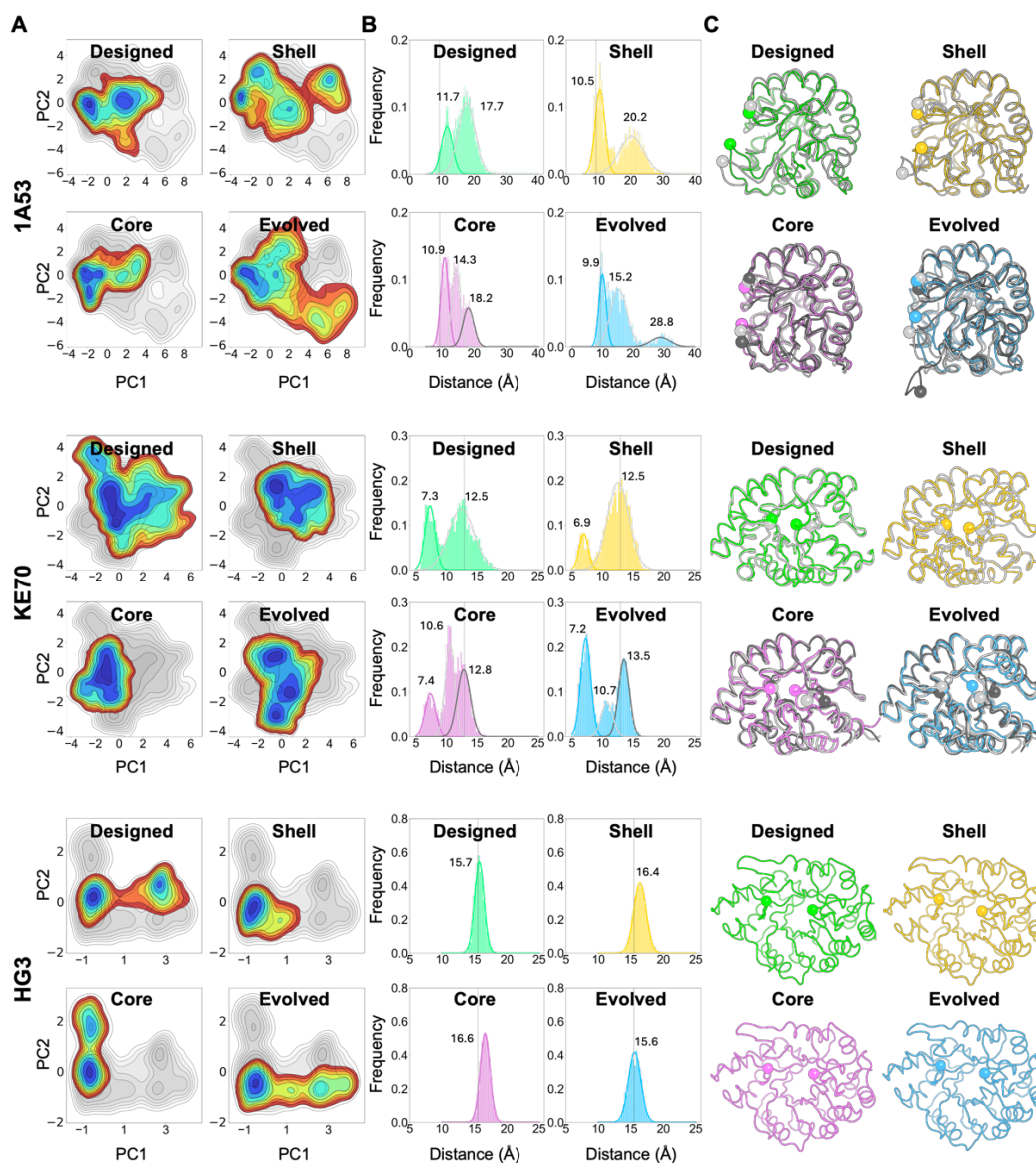


Figure 2.5. Conformational landscapes of Kemp eliminases by molecular dynamics. (A) Representation of the molecular dynamics trajectories projected into the two most important principal components (PC1, PC2) based on Ca contacts for each enzyme family. Color coding represents the relative population density of the states, with blue indicating the most populated wells and red representing the least populated regions. Details on conformational changes contributing to PC1 and PC2 are shown on Supplementary Figure 2.11. (B) Histograms of loop distances across the entire energy landscape for KE70 and 1A53 enzymes, or from the most populated well for HG3 enzymes. Loop distances were calculated between Ca carbons of residues 21 and 78 for KE70 variants, 58 and 188 for 1A53 variants, and 90 and 275 for HG3 variants (shown as spheres in panel c). The dashed line represents the average loop distance in crystal structures of each enzyme family. The multimodal distributions were deconvoluted using a Gaussian mixture model, with the median loop distances for each conformational state indicated in Å. (C) Representative snapshots for each conformational state sampled by the enzymes. These snapshots display loop distances equivalent to the median value for each population in the distribution. The Ca carbons of residues used to calculate loop distances are shown as spheres.

2.6. Discussion

In this study, we investigated the influence on the catalytic cycle of distal mutations introduced by directed evolution across three lineages of *de novo* enzymes designed to catalyze the same reaction. Unlike previous investigations of structural and dynamic changes along enzyme evolutionary trajectories,^{5,9,36} we sought to specifically isolate the effects of distal mutations from those occurring at the active site by generating Core and Shell variants. Our findings demonstrate that active-site mutations are the primary drivers of enhanced catalytic efficiency, organizing the active site for optimal enzyme-transition state interactions to accelerate the chemical transformation. These findings are consistent with the long-standing understanding that active-site residues play a dominant role in shaping enzyme activity.³⁷ However, distal mutations complement these effects by tuning energy barriers to substrate binding and product release, optimizing the catalytic cycle for efficient catalysis. This role is particularly critical when these steps are rate-limiting, as observed in many natural enzymes.^{38–40} In such cases, large-scale structural and dynamic rearrangements across the protein scaffold become necessary to achieve further efficiency gains. As a result, the benefits of distal mutations emerge only when the active site is already optimized, as improved substrate binding and product release cannot compensate for slow chemical transformation. Although we cannot rule out that distal mutations enhance catalysis through additional mechanisms, our strategy of generating Core and Shell variants provides a valuable approach for studying underexplored catalytic effects possibly mediated by distal mutations, such as alterations to heat capacity,²⁷ local electric fields,⁴¹ and conformational entropy.⁴²

Our results highlight important challenges in *de novo* enzyme design. Traditional enzyme design algorithms focus on carving active sites for specific reactions on preexisting protein

scaffolds without considering distal positions.^{15,16,18} Even recent deep-learning-based methods for enzyme design optimize distal residues for folding, stability and solubility rather than for their influence on catalysis.^{43,44} While both approaches have produced functional enzymes, the resulting catalytic activities remain low ($k_{\text{cat}} \leq 0.1 \text{ s}^{-1}$). Meanwhile, directed evolution consistently remodels designed active sites to enhance catalytic efficiency, sometimes even altering catalytic residues,^{25,26,29} while also fine-tuning conformational dynamics through mutations at distal positions.^{10,45} Our study shows that even if perfectly optimized active sites could be designed from scratch, these enzymes would still fall short of their full catalytic potential without the appropriate distal mutations. This finding underscores the importance of removing bottlenecks in the catalytic cycle, such as substrate binding and product release, by optimizing distal positions to adjust conformational dynamics and enhance activity.

Moving forward, we propose a framework for *de novo* enzyme design that optimizes the active site first, potentially through ensemble-based design³⁰ or deep-learning techniques^{4–6,26,27,43}, before introducing distal mutations to adjust conformational dynamics for facilitated ligand binding and release. Identifying beneficial distal mutations to precisely tune structural dynamics remains challenging, but computational methods capable of identifying dynamically-linked residues could help.⁴⁶ Balancing the structural rigidity needed for precise active-site alignment with the flexibility required for efficient substrate binding and product release will likely be essential for unlocking the full catalytic potential of *de novo* enzymes. The Core and Shell variants developed here offer valuable benchmarks for testing new predictive methods aimed at refining distal regions, potentially guiding future developments in enzyme design.

2.7. Methods

2.7.1. Protein expression and purification

Codon-optimized (*E. coli*) and his-tagged (C-terminus) genes for Kemp eliminases (Supplementary Table 2.8) cloned into the pET-29b(+) vector via *Nde*I and *Xho*I were obtained from Twist Bioscience. Enzymes were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth (LB) supplemented with 100 $\mu\text{g mL}^{-1}$ kanamycin. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of 0.3–0.7, after which protein expression was initiated with 1 mM isopropyl β -D-1-thiogalactopyranoside. Following incubation for 16 h at 16 °C with shaking (250 rpm), cells were harvested by centrifugation, resuspended in 8 mL lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer, pH 8.0), and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography using Ni-NTA agarose (Qiagen) pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Columns were washed twice, first with 10 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0), and then with the same buffer containing 20 mM imidazole. Bound proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and exchanged into 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride using Econo-Pac 10DG desalting pre-packed gravity-flow columns (Bio-Rad). Proteins used for crystallization were further purified by gel filtration in 50 mM sodium citrate buffer (pH 5.5) and 150 mM sodium chloride using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad). Purified samples were concentrated using Amicon Ultracel-10K centrifugal filter units (EMD Millipore) and quantified by measuring the absorbance at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPASy ProtParam tool (<https://web.expasy.org/protparam/>).

2.7.2. Circular dichroism and thermal denaturation assays

Circular dichroism measurements were performed with a Jasco J-815 spectrometer using 650- μ L aliquots of each Kemp eliminase in a 1-mm path-length quartz cuvette (Jasco) at a concentration of 30–40 μ M in 10 mM sodium phosphate buffer (pH 7) supplemented with 100 mM sodium chloride. For structural characterization of protein folds, circular dichroism spectra were acquired from 200 to 250 nm, sampled every 0.2 nm at a rate of 10 nm min⁻¹. Five scans were acquired and averaged for each sample. For thermal denaturation assays, samples were heated at a rate of 0.5 °C per minute, and ellipticity at 222 nm was measured every 0.2 °C. T_m values were determined by fitting a 2-term sigmoid function with baseline correction using nonlinear least-squares regression⁴⁷.

2.7.3. Steady-state kinetics

All assays were carried out at 27 °C in 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride. Triplicate 200- μ L reactions with varying concentrations of freshly prepared 5-nitrobenzisoaxazole (AstaTech) or 6-nitrobenzisoaxazole (Combi-blocks) dissolved in methanol (10 % final concentration, pH of reaction mixture adjusted to 7.0 after addition of methanol-solubilized substrate) were initiated by the addition of 10 μ M KE70-Designed/Shell, 100 nM KE70-Core/Evolved, 25 μ M 1A53-Designed/Shell, 1 μ M 1A53-Core/Evolved, 2 μ M HG3-Designed, 5 nM HG3-Core/Evolved and 50 nM HG3-Shell. Product formation was monitored spectrophotometrically in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax Plus plate reader (Molecular Devices) with measurements at 380 or 400 nm for reactions with 5-nitrobenzisoaxazole ($\epsilon = 15,800 \text{ M}^{-1} \text{ cm}^{-1}$)¹⁸ or 6-nitrobenzisoaxazole ($\epsilon = 2,870 \text{ M}^{-1} \text{ cm}^{-1}$)²⁷, respectively. Path lengths for each well were calculated ratiometrically

using the difference in absorbance of 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride and 10 % methanol at 900 and 975 nm (27 °C). Linear phases of the kinetic traces were used to measure initial reaction rates. For cases where saturation was not possible at the maximum substrate concentration tested (1 or 2 mM for 6-nitrobenzisoaxazole or 5-nitrobenzisoaxazole, respectively), data were fitted to the linear portion of the Michaelis-Menten model ($v_0 = (k_{\text{cat}}/K_M)[E_0][S]$) and k_{cat}/K_M was deduced from the slope. In all other cases, data were fitted to the Michaelis-Menten equation to calculate individual k_{cat} and K_M parameters.

2.7.4. Kinetic solvent viscosity effects (KSVE)

The effect of solvent viscosity on k_{cat}/K_M was determined at 27 °C in 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM NaCl, using sucrose as the viscogen at different concentrations (0, 20, 28, 33 %w/v). Solution viscosities were approximated from published viscosity data of sucrose solutions⁴⁸. Kinetic measurements were performed as described above using 5 nM of HG3-Core and HG3-Evolved, 1 μ M of 1A53-Core and 1A53-Evolved, and 150 nM of KE70-Core and KE70-Evolved. Initial rates were determined and fitted to the linear portion of the Michaelis-Menten model ($v_0 = (k_{\text{cat}}/K_M) [E_0] [S]$) as saturation was not possible for most reactions in the presence of sucrose. Reference values (k_{cat}/K_M or k_{cat}) in the absence of sucrose were divided by the values obtained at different sucrose concentrations and plotted against the relative buffer viscosity (η_{rel}) to give the corresponding slopes reported in Figure 2.4. For KSVE on k_{cat}/K_M , the slope (m) corresponds to $k_3 / (k_2 + k_3)$, where k_2 is the rate constant of substrate dissociation and k_3 is the rate constant for the chemical transformation (Figure 2.4A). A slope of 1 indicates that k_3 is at least 100-fold larger than k_2 ($m = k_3 / k_3$), while a slope of 0 indicates that k_3 is much smaller than k_2 ($m = k_3 / k_2 = 0$).

2.7.5. Stopped-flow kinetics

Stopped-flow experiments measuring intrinsic tryptophan fluorescence were used to monitor the kinetics of transition-state analogue (6NBT) binding to the Kemp eliminases. The measurements were performed using OLIS RSM 1000 rapid-scanning monochromator, equipped with a water bath to control the temperature. Changes in intrinsic Trp fluorescence upon binding and dissociation of 6NBT were monitored using an excitation wavelength of 295 nm and a long-pass 300 nm cut-off filter to detect emission. All experiments were performed at 20 °C. A solution of 10 μ M HG3-Core, HG3-Evolved, 1A53-Core and 1A53-Evolved in 50 mM Sodium phosphate buffer, 100 mM NaCl pH 7, was loaded and quickly mixed varying TSA concentrations in the same buffer containing 1% (v/v) DMSO (mixing ration 1:1 resulting in final enzyme concentration of 5 μ M. A significant decrease in fluorescence intensity was observed upon binding of 6NBT. Three replicate measurements were made for each 6NBT concentration and the data were fit to exponential equation $F = F_0 + Amp. \exp(-k_{obs} \cdot t)$ using GraphPad Prism where F is fluorescence intensity, F_0 is the initial fluorescence intensity, Amp is the amplitude of the change in fluorescence and k_{obs} is the observed rate constant.

2.7.6. Crystallization

Enzyme variants were prepared in 50 mM sodium citrate buffer (pH 5.5) at the concentrations listed in Supplementary Table 2.4. For samples that were co-crystallized with the transition-state analogue, a 100 mM stock solution of 6NBT (AstaTech) was prepared in dimethyl sulfoxide (DMSO) and diluted in the enzyme solutions for a final concentration of 5 mM (5 % DMSO). For all variants except 1A53-Core, crystallization was performed using SWISS CI 3-well sitting drop plates. For 1A53-Core, the hanging drop method was used. Crystallization drops were

prepared by mixing 1 μ L of protein solution with 1 μ L of the mother liquor and sealing the drop inside a reservoir containing an additional 500 μ L of the mother liquor solution. Mother liquor solutions contained various precipitants, and the specific growth conditions that yielded the crystals used for X-ray data collection are listed on Supplementary Table 2.5. For the structures bound to transition-state analogue, 6NBT was added directly on top of crystal drops using Acoustic dispensing with an Echo 650 liquid handler (Labcyte). The transition-state analogue was allowed to diffuse in the crystal drop for approximately 10 min.

2.7.7. X-ray data collection and processing

All crystals were flash frozen in liquid nitrogen using appropriate cryoprotectant and single-crystal X-ray diffraction data were collected on beamline 8.3.1 at the Advanced Light Source. The beamline was equipped with a Pilatus3 S 6M detector and was operated at a photon energy of 11111 eV. Data was processed and scaled using XDS⁴⁹ and AIMLESS⁵⁰, respectively.

2.7.8. Structure determination

We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser,⁵¹ as implemented in v1.13.2998 of the PHENIX suite.⁵² Available structures of Evolved variants were used as molecular replacement search models. Next, we performed iterative steps of manual model rebuilding followed by refinement of atomic positions, atomic displacement parameters, and occupancies using a riding hydrogen model and automatic weight optimization. For the 1A53-Core structure with and without transition-state analogue, and KE70-Core with transition-state analogue, refinement also includes a translation-libration-screw (TLS) model. All model building was performed using Coot 0.8.9.236⁵³ and refinement steps were performed with phenix.refine within the PHENIX suite (v1.13-2998).

Restraints for 6NBT were generated using phenix.elbow, starting from coordinates available in the Protein Data Bank (PDB ligand ID: 6NBT). Further information regarding model building and refinement, as well as PDB accession codes for the final models, are presented in Supplementary Table 2.5.

2.7.9. Ensemble refinement

Time-averaged ensembles were generated with phenix.ensemble_refinement implemented in PHENIX v.1.15.2-3472. In ensemble refinement,⁵⁴ local atomic fluctuations are sampled using molecular dynamics simulations accelerated and restrained by electron density to produce ensemble models fitted to diffraction data. Briefly, input crystal structures were edited to remove low-occupancy conformers and assign an occupancy of 1.0 to the remaining conformer. Following addition of riding hydrogens, parallel ensemble refinement simulations were performed using various combinations of the parameters p_{TLS} (0.6, 0.8, 0.9, 1.0), τ_x (0.5, 1.5, 2.0) and $w_{x\text{-ray}}$ (2.5, 5.0, 10.0), where p_{TLS} describes the percentage of atoms included in a translation-libration-screw (TLS) model use to remove the effects of global disorder, τ_x is the simulation time-step and $w_{x\text{-ray}}$ is the coupled b_{th} offset, which controls the extent to which the electron density contributes to the simulation force field such that the simulation runs at a target temperature of 300 K. The ensemble generated from each crystal structure that displayed the lowest R_{free} value was selected for analysis (Supplementary Table 2.6).

2.7.10. Molecular dynamics simulations

The Amber 2020 software (<http://ambermd.org/>) with the AMBER19SB forcefield⁵⁵ was used for all simulations. Crystal structures of the enzymes were used as initial structures, except for 1A53-Shell, 1A53-Evolved and KE70-Shell, which were generated in silico from the 1A53-

2.5 crystal structure (PDB ID: 6NW4) and KE70-Evolved structure (PDB ID: 3Q2D) using the protein design software Triad⁵⁶ (Protabit LLC, Pasadena, CA, USA). Amino-acid protonation states were predicted using the H++ server (<http://biophysics.cs.vt.edu/H++>). Each enzyme structure was immersed in a pre-equilibrated cubic box (10 Å edge length) filled with buffer, using the OPC water model,⁵⁷ neutralized with explicit counterions (Na⁺, Cl⁻) using the Amber 20 leap module. A two-state geometry optimization approach was applied. The first stage involved the minimization of solvent molecules and ion positions while imposing positional restraints on solutes using a harmonic potential (force constant: 500 kcal mol⁻¹ Å⁻²). The second stage consisted of an unrestrained minimization of all atoms in the simulation cell. The systems were gradually heated in seven 50 ps steps, each increasing the temperature by 50 K, under constant-volume and periodic boundary conditions. Water molecules were treated using the SHAKE algorithm, and long-range electrostatic effects were modeled using the particle-mesh-Ewald method.⁵⁸ An 8 Å cut-off was applied to Lennard-Jones and electrostatic interactions. Decreasing harmonic restraints were applied to the protein during thermal equilibration (500, 210, 165, 125, 85, 45, 10 kcal mol⁻¹ Å⁻²), with temperature control achieved using the Langevin scheme. A timestep of 1 fs was maintained during these stages. Following the heating, equilibration of each system was conducted for 20 ns without restraints at a constant pressure of 1 atm and a 2 fs timestep. After equilibration in the NPT ensemble, production run MD simulations for each system in the NVT ensemble with periodic-boundary conditions were performed. These production trajectories were run for 2000 ns in three separate replicates.

Similar protocol was used to evaluate the conformation of 1A53-Core containing Models of 1A53-Core containing alternate Trp110 rotamers (Supplementary Figure 2.5) were prepared from the crystal structure by deleting the bound MES molecule and changing the rotamer in Pymol

(<http://www.pymol.org/>). Molecular dynamics simulations were performed as described above. Trp110 dihedral angles throughout the simulations were measured using the python interface of cpptraj package.⁵⁹

2.7.11. Bottleneck radius analysis

Caver3³⁵ was used to calculate substrate entry channels for representative snapshots of the most populated energy minima of HG3-Designed, HG3-Core, HG3-Evolved, and HG3-Shell. Parameters used for the entry channel search were 6 Å Shell depth, 5 Å Shell radius, clustering threshold value of 3.5 and a 0.9 Å minimum probe radius.

2.8. Supplementary information

Supplementary Table 2.1. Amino acid sequences of various Kemp eliminases

Enzyme	Sequence
HG3-Designed	MAEAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIQADFGMVWPENSMQWDATEPSQGNFNFAGADYLVNWAQQNGKL IGGGMLVWHSQPLPSWVSSITDKNTLTNVMKNHITLMTTRYKGKIRAWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIAFQTA RAADPNAKLYIMDYNLDSASYPKTQAIIVNRVKQWRAAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVSIIMLDVA GASPTDYVNVVNACLVQSCVGITVFGVADPDSWRASTTPLLFDGNFNPKPAYNAIVQDLQQGSIEGRGHHHHHH
HG3-Shell	MAEAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIKADFGMVWPENSMQWDATEPSQGNFNFAGADYLVNWAQQNGKL IGGGMLVWHSQPLPSWVSSITDKNTLTNVMKNHITLMTTRYKGKIRAWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIAFQTA RAADPNAKLYIMDYNLDSASYPKTQAIIVNRVKQWRAAGVPIDGIGSQMHLASAGQGAGVLQALPLLASAGTPEVSIIMLDVA GASPTDYVNVVNACLVQSCVGITVFGVADPDSWRASSSTPLLFDGNFNPKPAYNAIVQNLQQGSIEGRGHHHHHH
HG3-Core	MAEAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIQADFGMVWPENSMQWDATEPSQGNFNFAGADYLVNWAQQNGKL IGAGCLVWHSFPLPSWVSSITDKNTLTNVMKNHITLMTTRYKGKIRTWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIAFQTA RAADPNAKLYIMDYNLDSASYPKTQAIIVNRVKQWRAAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVSIIMLDVA GASPTDYVNVVNACLVQSCVGITVMGVADPDSAFASSTPLLFDGNFNPKPAYNAIVQDLQQGSIEGRGHHHHHH
HG3-Evolved	MAEAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIQADFGMVWPENSMQWDATEPSQGNFNFAGADYLVNWAQQNGKL IGAGCLVWHSFPLPSWVSSITDKNTLTNVMKNHITLMTTRYKGKIRTWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIAFQTA RAADPNAKLYIMDYNLDSASYPKTQAIIVNRVKQWRAAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVSIIMLDVA GASPTDYVNVVNACLVQSCVGITVMGVADPDSAFASSTPLLFDGNFNPKPAYNAIVQDLQQGSIEGRGHHHHHH
1A53-Designed	MPLYLKGLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITATIAAYKRKSPSGLDVERDPIEYSKFMEYAVGLA IATEEKYFNNGSYETLRKIASSVSIPLMWDFIVKESQIDDAYNLGADTVALIVKILTERELESLEYARSYGMPEAIVIND ENDLDIALRIGARFIEIASRDLETLEINKENQRKLISMIPSNVVKVAVWQGISERNEIEELRKLGVNAFGIGSSSLMRNPEKI KEFILGSIIEGRGHHHHHH
1A53-Shell	MPLYLKGLKDVVQLSLRRPSFHASRQRPIISLNERILEFNKRNITATIAAYKRKSPCGLDVERDPIEYSKFMEYAVGLA IATEEKYFNNGSYETLRKIASSVSIPLMWDFIVKESQIDDAYNLGADTVALIVKILTERELESLEYARSYGMPEAIVIND ENDLDIALRIGARFIEIASRDFETLEINKENQRKLISMIPSNVVKVAVWQGISERNEIEELRKLGVNAFGIGSSSLMRNPEKI KEFTVGSIEDRGHHHHHH
1A53-Core	MPLYLKGLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITATIAAYRRKSPSGLDVERDPIEYSKFMEYAVGLA VATEEKYFNNGSYETLRKIASSVSIPLMWDFIVKESQIDDAYNLGADTVALIVKILTERELESLEYARSYGMPEYIVIND ENDLDIALRIGARFIEICSRDLETLEINKENQRKLISMIPSNVVKVAVWGGISERNEIEELRKLGVNAFGIGSSSLRNPEKI KEFILGSIIEGRGHHHHHH
1A53-Evolved	MPLYLKGLKDVVQLSLRRPSFHASRQRPIISLNERILEFNKRNITATIAAYRRKSPCGLDVERDPIEYSKFMEYAVGLA VATEEKYFNNGSYETLRKIASSVSIPLMWDFIVKESQIDDAYNLGADTVALIVKILTERELESLEYARSYGMPEYIVIND ENDLDIALRIGARFIEICSRDFETLEINKENQRKLISMIPSNVVKVAVWGGISERNEIEELRKLGVNAFGIGSSSLRNPEKI KEFTVGSIEDRGHHHHHH
KE70-Designed	MTDLKASSLRALKMHLATSANDDDTDEKVIALCHQAKTPVGTTDAIYIYPRFIPIARKTLKEQGTPEIRIWTSTNFPHGN DDIDIALAETRAAIAYGADGVAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLLSVIIETGELKDEALIRKASEISIKA GADHIVTSTGKVAVGATPESARIMMEVIRDMGVEKTVGFI PAGGVRTAEDAQKYLAIADELFGADWADARHYAFGASASLL ASLLKALGHGDGKSASSYGSLEHHHHHH
KE70-Shell	MTDLKASSLRALKMHLATSANDDDTDENVIALCHQAKTPVGNTDAIYIYPRFIPIARKTLKEQGTPEIRIWTSTNFPHGN DDIDIALAETRAAIAYGADGVAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLLSVIIETGELKDEALIRKASEISIKA GADHIVTSTGKVAVGATPESARIMMEVIRDMGVEKTVGFI PAGGVRTAEDAQKYLAIADELFGADWADARHYAFGASASLL ASLLKALGHGDGKSASSYGSLEHHHHHH
KE70-Core	MTDLKASSLRALKMHLATSANDDDTDEKVIALCHQAKTPVGTTDAIFIYPRFIPIARKTLKEQGTPEIRICTSTNFPHGN DDIDIALAETRAAIAYGADSVAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLLAVIIETGELKDEALIRKASEISIKA GADNIVTSTGKVAVGATPESARIMMEVIRDMGVEKTVGFI PVGGVRTAEDAQKYLAIADELFGADWADARHYAFGASASLL ASLLKALGHGDGKSASSYGSLEHHHHHH
KE70-Evolved	MTDLKASSLRALKMHLATSANDDDTDENVIALCHQAKTPVGNTDAIFIYPRFIPIARKTLKEQGTPEIRICTSTNFPHGN DDIDIALAETRAAIAYGADSVAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLLAVIIETGELKDEALIRKASEISIKA GADNIVTSTGKVAVGATPESARIMMEVIRDMGVEKTVGFI PVGGVRTAEDAQKYLAIADELFGADWADARHYAFGASASLL ASLLKALGHGDGKSASSYGSLEHHHHHH

Supplementary Table 2.2. Active-site and distal mutations

Enzyme	Active-site mutations ^a	Distal mutations ^a
HG3	G82A M84C Q90F A125T F267M W275A R276F	V6I Q37K N47E S89N T105I T142N T208M T279S D300N
1A53	K53R I82V A157Y A180C Q211G M237L	R23H I48T S58C L184F I219L R238S L248V G253D
KE70	Y48F W72C G101S S138A H166N A204V	K29N T43N

^a Mutations are relative to the Designed variant.

Supplementary Table 2.3. Purification yields

Enzyme	Yield (mg/L culture)
HG3-Designed	3.6 ± 0.4
HG3-Core	13 ± 5
HG3-Shell	13 ± 2
HG3-Evolved	33 ± 4
1A53-Designed	5 ± 3
1A53-Core	6 ± 5
1A53-Shell	2 ± 1
1A53-Evolved	4 ± 1
KE70-Designed	30 ± 6
KE70-Core	38 ± 1
KE70-Shell	50 ± 30
KE70-Evolved	31 ± 2

All proteins were purified at least twice (mean ± s.d.)

Supplementary Table 2.4. Crystallization conditions

Enzyme ^a	6NT	Protein (mg mL ⁻¹)	Buffer	pH	Additive #1	Additive #2
1A53-Core	(-)	8.0	100 mM MES	6.0	40% PEG-400	5% PEG-3000
	(+)	7.0	10 mM MES	6.0	40% PEG-400	5% PEG-3000
KE70-Core	(-)	35.0	100 mM Tris	8.5	25% PEG-6000	–
	(+)	35.0	100 mM HEPES	8.5	20% PEG-4000	–
HG3-Shell	(-)	3.5	100 mM HEPES	7.0	0.5% Jeff ED-2001	1.1 M disodium malonate
	(+)	3.5	100 mM HEPES	7.0	1% PEG MME 2000	1.0 M disodium succinate

^a All proteins were crystallized at 20 °C.

Supplementary Table 2.5. Crystallographic data and refinement statistics

	1A53- Core	1A53- Core	KE70- Core	KE70- Core	HG3- Shell	HG3- Shell
6NT	(-)	(+)	(-)	(+)	(-)	(+)
PDB ID	8FMC	8FOQ	8FMD	8FOR	8FME	8FOS
Data collection^a						
Temperature (K)	100	100	100	100	100	100
Resolution (Å)	52.56– 2.36	48.48– 1.60	76.11– 2.20	39.29– 2.20	35.14– 1.44	49.05– 1.56
Space group	P 31 2 1	P 31 2 1	P 1 21 1	P 21 21 21	P 21 21 21	P 21 21 21
Cell params.						
a b c (Å)	60.6865 60.6865 121.762	61.003 61.003 121.977	62.915 53.992 81.455	52.789 155.642 190.588	50.018 67.381 72.606	48.908 66.569 72.518
α β γ (°)	90 90 120	90 90 120	90 110.9 90	90 90 90	90 90 90	90 90 90
Chains per asymm. unit	1	1	2	6	1	1
R_{pim}	0.044 (0.17)	0.019 (0.80)	0.028 (0.070)	0.089 (0.925)	0.030 (0.571)	0.043 (0.668)
CC_{1/2}	0.99 (0.93)	1.00 (0.43)	0.99 (0.98)	0.99 (0.49)	0.99 (0.576)	0.99 (0.36)
I/σI	11.9 (4.4)	18.7 (0.7)	19.0 (9.2)	6.8 (0.8)	15.3 (1.2)	12.9 (2.0)
Completeness (%)	99.96 (100.00)	98.24 (88.68)	99.30 (97.41)	99.87 (99.99)	94.33 (68.30)	81.29 (31.47)
Multiplicity						
Wilson B-factor (Å²)	35.48	24.91	23.73	36.24	15.09	17.77
# unique reflections	11210 (1099)	34632 (3048)	26012 (2517)	80717 (7923)	42595 (3034)	27957 (1063)
Refinement						
R work/free	0.1907/ 0.2314	0.1592/ 0.1859	0.1622/ 0.1943	0.2085/ 0.2363	0.1542/ 0.1659	0.1573/ 0.1817
No. atoms						
Protein	2026	2117	3777	11132	2424	2409
Ligand	17	17	0	72	0	25
Water	43	171	161	206	182	133
Averaged B-factors (Å²)						
Protein	40.05	32.08	30.45	46.76	17.87	18.35
Ligands	66.22	53.86	–	41.00	–	33.58
Water	34.05	42.48	29.62	35.86	23.87	26.14
RMSD						
bond lengths (Å)	0.003	0.012	0.002	0.002	0.008	0.005
bond angles (°)	0.527	1.179	0.520	0.467	0.927	0.729
Molprobrity statistics						
Ramachand. outliers (%)	0.00	0.00	0.00	0.00	0.00	0.00
Ramachand. allowed (%)	3.67	2.45	2.00	1.56	1.34	2.36
Ramachan. favored (%)	96.33	97.55	98.00	98.44	98.66	97.64
Rotamer outliers (%)	0.00	0.00	0.51	0.09	0.00	0.39
MolProbity clashscore	1.21	2.77	1.18	1.29	2.71	2.90

^a Highest resolution shell is shown in parentheses.

Supplementary Table 2.6. Ensemble refinement of Kemp eliminase structures with bound 6NBT

Enzyme	Original name	PDB ID	Resolution (Å)	Before ensemble refinement		After ensemble refinement		No. structures
				R _{work}	R _{free}	R _{work}	R _{free}	
HG3-Designed	HG3-K50Q	7K4U	1.30	0.1850	0.2110	0.1648	0.2034	63
HG3-Shell	–	8FOS	1.56	0.1573	0.1817	0.1308	0.1714	63
HG3-Core	HG4	5RGF	1.46	0.1320	0.1500	0.1216	0.1464	125
HG3-Evolved	HG3.17	5RGE	1.77	0.1500	0.1790	0.1322	0.1750	56
1A53-Designed	1A53-2	3NZ1	1.56	0.1720	0.2100	0.1412	0.1863	125
1A53-Core	–	8FOQ	1.60	0.1592	0.1859	0.1451	0.1815	72
KE70-Core	–	8FOR	2.20	0.2080	0.2360	0.1762	0.2424	42
KE70-Evolved	R6 10/11G	3Q2D ^a	2.19	0.1910	0.2330	0.1564	0.2136	56

^a Ensemble refinement of KE70-Evolved was performed starting from the structure that we re-refined to model missing residues (22-25 in chain A and 3-25 in chain B) into the available density and to flip 6NBT into a productive binding pose. R values for this structure are listed in the table.

Supplementary Table 2.7. 6-nitrobenzotriazole (6NBT) binding kinetics

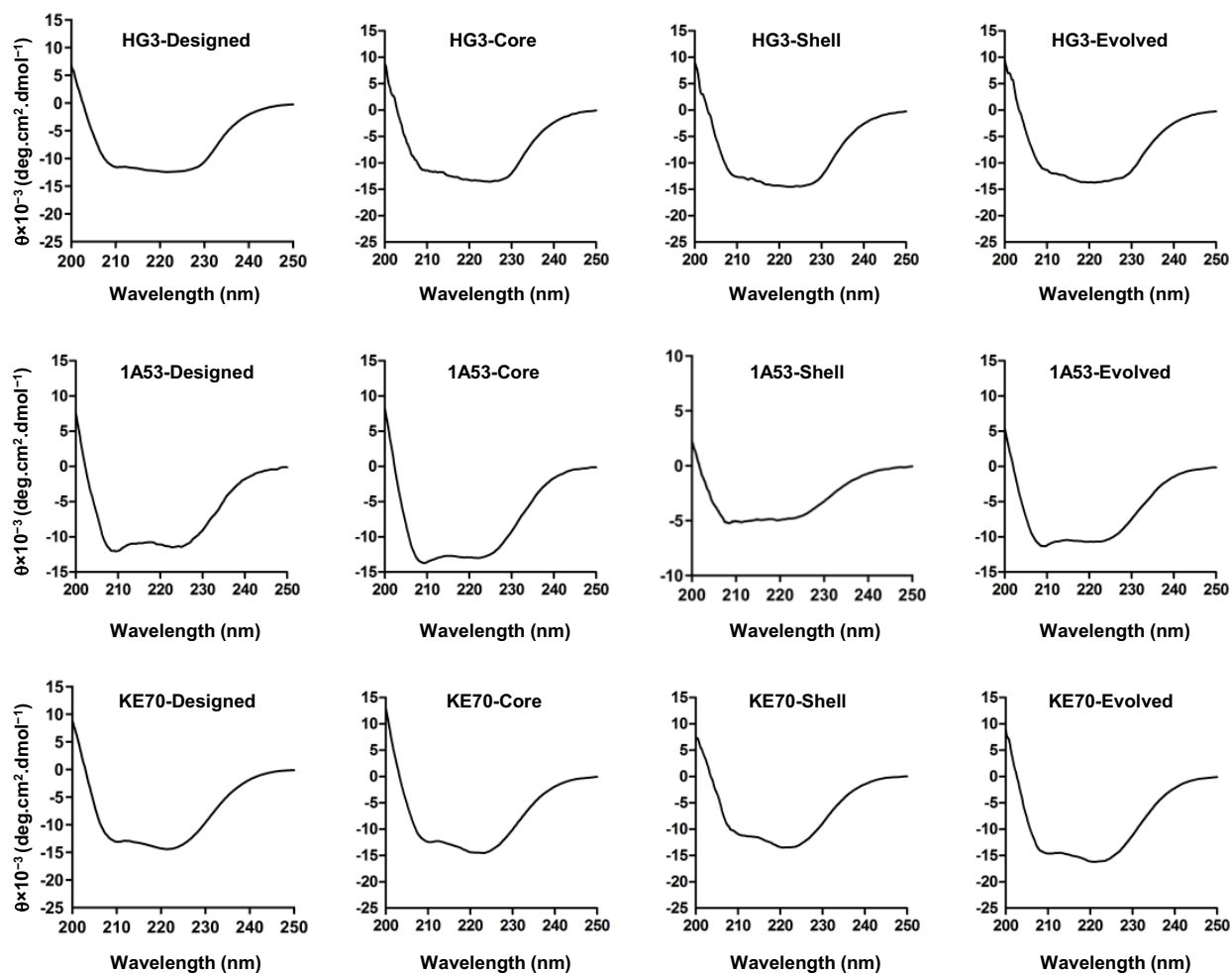
Constant	HG3-Core	HG3-Evolved	1A53-Core	1A53-Evolved
k_{on} 6NBT ($\mu\text{M}^{-1} \text{s}^{-1}$)	0.137 ± 0.005	0.179 ± 0.009	0.25 ± 0.02	0.091 ± 0.007
k_{off} 6NBT (s^{-1})	4.6 ± 0.3	4.5 ± 0.6	6.2 ± 0.4	7.3 ± 0.2
K_{D} 6NBT (μM)	34 ± 3	25 ± 5	25 ± 3	80 ± 8

Supplementary Table 2.8. DNA Sequences

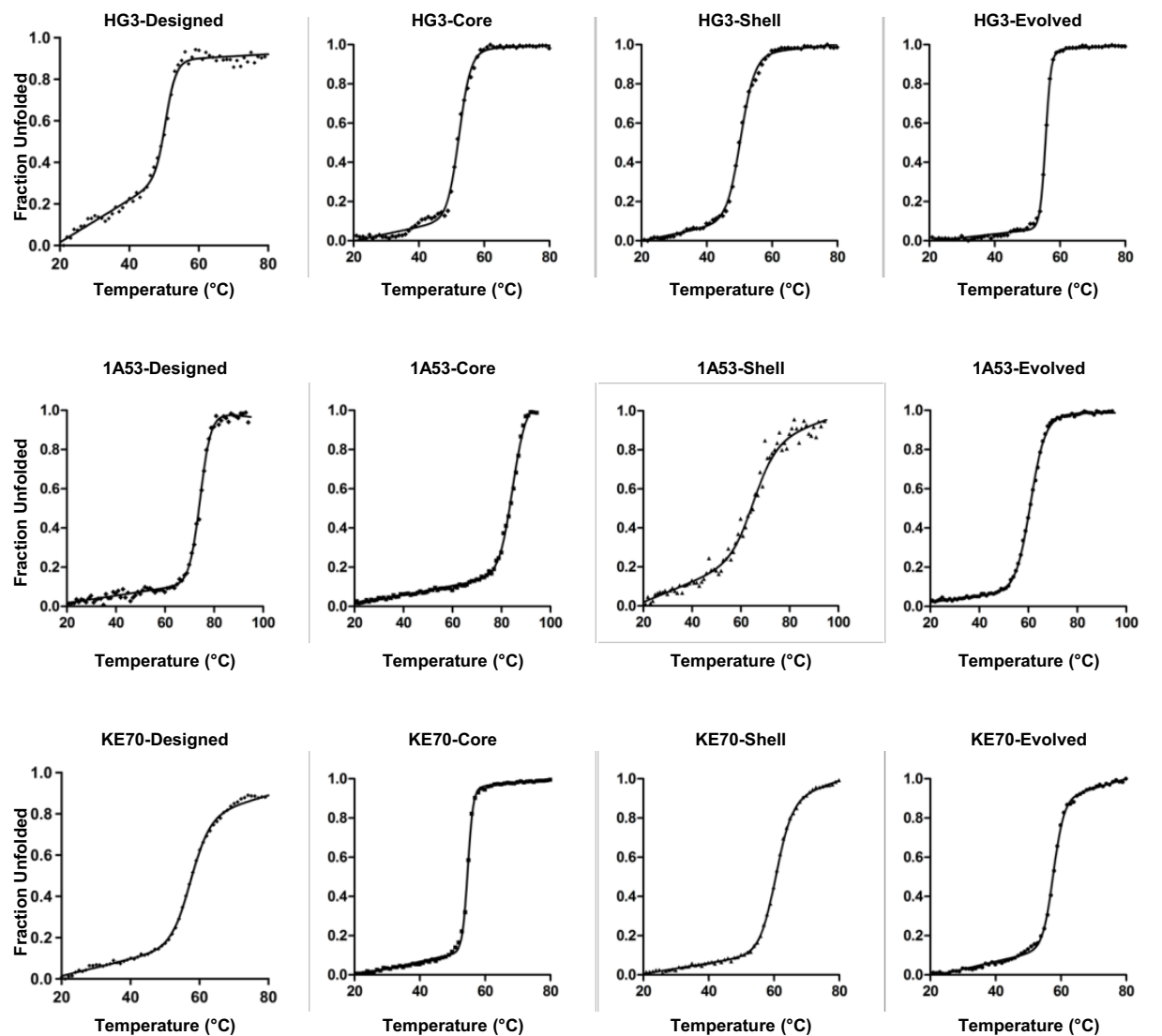
Enzyme	Sequence
HG3-Designed	GCGGAAGCGGCGCAAAGCGTAGACCAACTGATAAAGCACGCGGTAAAGGTCTACTTCGGGGTTGCTACGGATCAGAATCGGGCTGACGACCGGCAAGAAGCGCAGCAATAATTCAGGCTGACTTTGGTATGGTGTGGCCCGAAAATTCAATGCAGTGGGATGCCACCGAACCCTCACAAAGGAACTTCAACTTTGCCGGCGCCGACTACTTAGTAAACTGGGCACAACAAAATGGGAACTTATTGGTGGTGGTATGCTGGTATGGCATTCACAATTGCCGTCATGGGTGAGCAGCATTACCGACAAGAATACACTGACCAACGTTATGAAGAACCACATAA CTACCTTGATGACACGTTATAAGGGAAGATCCGAGCCTGGGACGTTGTTGGTGAGGCCTTAAATGAGGACGGATCTCTTCGCCA AACGGTTTTCTGAACGTGATTGGGGAAGACTACATTCCGATTGCGTTCCAACTGCACGTGCTGCGGACCCTAACGCAAAGTTG TACATTATGGATTACAACCTGGATTCCGCTTCTTACCCTAAGACACAAGCAATTGTCAATCGCGTGAAGCAGTGGCGGGCGGCCG GAGTACCCATCGATGGCATTGGCTCAGACACACCTTTCTGCTGGTCAGGGAGCAGGAGTTCTGCAAGCGCTTCCGCTGTAGC ATCCGACAGTACCCCTGAGGTGTCTATACTGATGTTGGACGTAGCAGGCGCGAGCCCCACGGACTATGTCAACGTGGTGAATGCA TGTCTGAATGTTCAAAGTTGCGTGGGATTACCGTGTTCGGAGTAGCTGACCCAGACTCTTGGAGAGCATCTACAACACCACTTT TATTTCGACGGAAATTTCAACCCTAAACCGGCTATAATGCGATAGTTCAAGATCTGCAGCAAGGCTCTATAGAGGGAAGAGGTCA TCACCACCACCATCACTAG
HG3-Shell	GCTGAGGCAGCCAGAGTATCGACCAACTGATTAAGCGCGAGGCAAGGTCTATTTTCGGAGTCGCTACTGACCAAAATCGGTTAA CGACAGGTAAAGAAGCGCAGCCATTATCAAAGCTGACTTCGGGATGGTGTGGCCTGAAGAGAGCATGCAGTGGGATGCTACAGAGCC TTCACAAGGCAACTTCAACTTCGACAGGAGCGGACTATTTAGTAAACTGGGCCAACAGAACGGTAAACTGATAGGTGGTGGTATG TTGGTCTGGCACAATCAATTACCTAGCTGGGTATCCTCTATCACCAGCAAGAATACTTTGATTAACTGAAGAACCACATTA CGACCCCTATGACTAGATACAAGGGAAGATTTCGAGCATGGGACGTAGTTGGGGAAGCGTTCAACGAGGATGGTTCCTTTCGCTCA GAATGTATTTCCTTAATGTTATCGGCGAGGACTACATACCCATTGCTTTCCAAACAGCACGTCGCCGCGGACCCGAATGCAAAATTA TACATCATGGACTACAACCTTAGACAGTGTAGCTACCCTAAACTCAAGCCATAGTCAACAGAGTCAAGCAATGGCGCGCTGCAG GCGTACCTATAGACGGAATTGGTTCACAAATGCACCTTAGTGCGGGACAAGGCGCGGGCGTTCTGCAAGGCCCTGCCGCTGTGGC GAGCGCCGGAACACCAGAAGTGTCCATTCTCATGCTTGATGTGGCCGAGCATCTCTACAGATTATGTCAATGTGGTGAATGCG TGTCTGAACGTCCAGAGTTGTGTGGGATCACGTTTTCGGGGTGGCTGACCTGACAGCTGGCGAGCCTCGTCCACTCCGCTTT TGTTCGATGGGAATTTTAACCCGAAGCCGGCATACAACGCGATTGTGCAAAACCTTCAGCAAGGCAGCATTGAGGGGCGGGGACA CCATCACCACCACCACTGA
HG3-Core	ATGGCGGAGGCGGCGCAGAGCGTGGACCAACTGATCAAGGCGCGTGGCAAGGTTTACTTTGGCGTGGCGACCGACCAGAATCGTC TGACCACCGGCAAGAAGCGCGGATCATTACGCGGACTTCGGCATGGTGTGGCCGGAGAACAGATGCAGTGGGATGCGACCGAA CCGAGCCAGGTAACCTTCAACTTTGCGGGCGCGGACTACCTGGTTAACTGGGCGCAGCAAAACGGCAAGCTGATCGGTGCGGGTT GCCTGGTGTGCGACAGCTTTCTGCCGAGCTGGGTTAGCAGCATTACCGATAAGAACACCCTGACCAACGTCGATGAAAAACCAT CACCACCTGATGACCCGTTATAAGGGTAAAATTCGTACGTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGCAGCCTGCGT CAGACCGTGTTCGTGAACGTTATCGGCGAGGACTACATCCCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAAAC TGTACATCATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACCCAGGCGATTGTGAACCGTGTAAACAATGGCGTGGCG GGGTGTGCCGATTGATGGTATTGGTAGCCAGACCCATCTGAGCGCGGGTCAGGGTGCAGGCGTTCTGCAAGCGCTGCCGCTGCTG GCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCGAGTCCGACCGATTACGTTAACGTGGTTAACG CGTGCTGAACGTGCAAGCTGCGTTGGTATTACCGTGATGGGTGTTGCGGACCCGGATAGCGCGTTTGCAGACACACCCCGCT GCTGTTGATGGCAACTTTAACC CGAAGCCGGCGTATAACGCGATTGTTCAGATCTGCAACAGGGTAGCATCGAGGGTCTGTTGCT CATCATCATCATCACTAA
HG3-Evolved	ATGGCGGAGGCGGCGCAGAGCATCGACCAACTGATTAAGGCGCGTGGCAAAGTGTACTTCGGTGTGCGACCGATCAGAACCCTG TGACCACCGGCAAGAAGCGCGGATCATTAAAGCGGACTTTGGTATGGTGTGGCCGGAGGAAAGCATGCAGTGGGATGCGACCGA ACCGAGCCAAAGGCAACTTCAACTTTGCGGGTGCGGACTATCTGGTTAACTGGGCGCAGCAAAACGGCAAGCTGATCGGTGCGGGC TGCCCTGGTGTGGCACAACCTTCTGCCGAGCTGGGTTAGCAGCATCACCATAAGAACACCCTGATTAACTGATGAAAAACCAACA TCACCACCTGATGACCCGTTACAAGGGCAAAATTCGTACTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGTAGCTAGCTGCG TCAGAACGTGTTTCTGAACGTTATCGGCGAGGACTACATCCCGATTGCGTTCAGACCGCGCGTGGCGGGGACCCGAACGCGAAA CTGTACATTATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACCCAGGCGATCGTGAACCGTGTAAACAGTGGCGTGGCG GGGGTGTGCCGATTGACGCTATTGGCAGCCAGATGCACCTGAGCGCGGGTCAAGGTGCGGGTGTTCGCAAGCGCTGCCGCTGCT GCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCGAGTCCGACCGATTACGTTAACGTGGTTAACG CGGTGCTGAACGTGCAAGCTGCGTTGGTATTACCGTGATGGGTGTTGCGGACCCGGATAGCGCGTTTGCAGACACACCCCGCT TGCTGTTGATGGCAACTTTAACC CGAAGCCGGCGTATAACGCGATTGTTCAGAACCTGCAGCAAGGTAGCATCGAAGGCCGTGGT CACCACCACCACCACCACTAA
1A53-Designed	CCGCGTTACCTTAAGGATGGCTCAAGGACGTAGTACAATTAAGCTTGGCTCGCCCTTCGTTCCGTGCTAGTCGCCAACGTCCCA TTATCTCTTTGAATGAACGCATCCTTGAATTTAATAAGCGCAATATCACAGCCATCATCGCGGCCCTACAAGCGTAAAAGCCCAAG CGGGTTGGACGTTGAGCGGATCCGATTGAATATAGCAAGTTTATGGAGCGTTATGCGGTAGGCCCTTGCCATTGCTACCGAAGAG AAGTATTTCAACGGTAGCTACGAGACGTTGCGCAAGATCGCCAGTAGCGTAAGCATCCCAATCTTGATGTGGGATTTTATTGTAA AAGAGTCACAAATCGATGACGCTATAACCTCGGTGCGGACACCGTCGCATTAATTGTCAAATCTTGACGGAACGCGAAGCTTGA GAGTCTTTTAGAGTACGCTCGCTCGTATGGCATGGAGCCTGCTATCGTCATTAACGATGAGAATGATCTCGACATTGCACTGCGT ATTGGAGCTCGCTTCATCGAGATCGCATCCCGTGACCTGGAGACGCTTGAGATCAACAAAGAGAATCAACGCAAACTTATCTCAA TGATCCCTTCGAACGTCGTAAAAGTCGCCTGGCAAGGCAATTAGTAGCGCAACGAAATTGAGGAGTTACGCAAGTTAGGAGTTAA CGCCTTTGGTATCGGTTCTAGCCTCATGCGTAATCCGGAAGATCAAGGAGTTTATCCTCGGAGTATCGAGGGTCTGGTGCAC CATCATCAACATCACTAA
1A53-Shell	CCCCGTACCTTAAGGATGGCTGAAGGACGTTGTACAATTATCCTTACGGCGGCCAGCTTCACGCGCTCACGCCAGCGCCCCA TAATCAGTCTGAATGAGCGGATCCTTGAATTTCAATAAGAGAAATATTACTGCTACCATCGCGGCCCTACAAGCGGAAGTCTCCGTG TGGTCTGGACGTTGAGCGGGATCCGATCGAATATTCTAAGTTTCATGGAGCGATACGAGTAGGCTTTCGAATTGCAACTGAAGAG AAATACTTCAACGGGAGTTACGAGACGCTTAGAAAAGATTGCCAGTTCAGTGTCTATCCCCATCCTTATGTGGGATTTTATCGTTA

	AGGAATCCCAAATAGATGACGCGTACAACCTGGGAGCAGACACTGTAGCGCTTATTGTGAAGATCCTTACCGAACGTGAACCTGA GTCTTACTTGAGTAGCTCGTAGCTACGGGATGGAGCCAGCCATAGTAATAAATGACGAAAACGACTTGACATAGCCCTGCGG ATCGGCGCGCTTTTCATTGAAATTGCAAGCCGTGACTTCGAGACGCTGGAGATAAACAAGGAAAATCAACGCAAGCTCATTAGCA TGATTCCAAAGTAATGTCGTGAAGGTAGCTGGCAGGGCATTTCGAAACGTAATGAGCTGGAAAGAGCTGCGTAAAGTTAGGCGTTAA TGCCTTCGGAATTGGTAGTTCTCTGATGTCTAATCTGAGAAAATCAAAGAGTTCATCGTAGGTTCCATCGAAGATAGAGGACAT CATCACCACCACCCTGA
1A53- Core	CCGCGTACTTAAAGGCTGGCTTAAAGATGTCGTCCAACCTTAGCTTGCGCGCTCCATCATTTTCGCGCTTCGCGTCAGCGTCCAA TTATTAGCCTCAACGAGCGTATTTTGAATTCAATAAGCGTAATATCACGGCAATCATTGCCGCGTACCGTCGTAAGAGCCCTAG CGGTCTGGACGTAGAGCGCGATCCCATCGAGTACTCAAATTCATGGAACGCTACGCAGTTGGACTGGCGGTAGCAACCGAGGAA AAGTACTTTAACGGAAGCTATGAGACCTGCGTAAAATTGCGTCATCTGTTAGCATCCCCATCTTAATGTGGGATTTTCATTGTGA AGGAATCCCAAATCGACGACGCCATAAACCTGGGTGCCGACACCGTGGCATTAAATCGTGAAAAATTTAACGGAGCGCGAGTTGGA GTCCCTGCTGGAGTATGCACGCTCGTATGGGATGGAACCTTATATTGTGATCAATGATGAGAACGACCTGGACATTGCACCTCCGC ATTGGCGCGCTTTTCATTGAGATCTGCTCGCGCGATCTGGAACGCTTGAGATCAATAAGGAAAACCAACGCAAGCTTATCTCAA TGATTCCGTCAAATGTTGTTAAGGTAGCATGGGTGGAATCTCTGAGCGTAATGAGATCGAAGAGTTACGCAAGCTCGGAGTTAA TGCCTTCGGAATTGGCTCAAGTCTTTTACGTAATCCAGAAAAGATCAAAGAGTTCATTCTGGGTTCGATCGAGGGACGCGGGCAT CACCATCACCACCATTGA
1A53- Evolved	CCTCGTACTTGAAGGGTTGGCTGAAGGATGTGGTACAACCTGTCCCTTCGTCGTCCATCCTTTACGCGAGCCGTCACGCCCCGA TTATCAGCTTGAACGAGCGCATCCTTGAGTTTAAACAAACGCAATATTACTGCGACCATTTGCGGCTTATCGCCGTAAAGTCGCTTG TGGTTTGGATGTAGAGCGCGATCCCATCGAGTACAGCAAGTTTCATGGAGCGTTATGCTGTAGGGCTGGCAGTAGCAACCGAGGAG AAGTACTTCAATGGATCGTATGAGACTCTGCGTAAGATCGCTCGTCCGTCTCGATCCCGATCTTGATGTGGGACTTCATTGTTTA AGGAAAGCCAAATCGACGATGCATACAATCTTGGAGCGGATACCGTCGCACTCATCGTGAAAGATTCTCACTGAACGCGAACTGGA AAGTTTACTTGAGTATGCCGCTTCGTATGGAATGGAGCCCTACATTGTTATCAATGACGAGAACGACCTCGATATCGCTTTGCGC ATTGGAGCCGCTTCATCGAGATCTGCTCCCGTGACTTTGAGACGCTCGAAATCAACAAGGAGAACCAGCGTAAGCTCATCTCTA TGATCCCTTCCAATGTTGTAAAGGTAGCATGGGTGGCATCTCAGAGCGCAATGAATTAGAGGAGTTGCGTAAGCTGGGAGTAA CGCATTTGGAATTGGATCGTCCCTTCTGTCTAACCCAGAGAAGATTAAGGAATTTATTGTAGGCTCAATTGAAGACCGCGGCCAC CATCATCATCATCTAA
KE70- Designed	ACCGATTTAAAGGCATCAAGCCTTCGCGCCTTGAAGTTAATGCATCTTGCCACGGCCAACGATGATGACACTGACGAGAAGGTTA TTGCACTTTGTCAACCAAGCCAAGACCCCTGTGGGCACTACGGACGCGATCTACATCTATCCCCGCTTTATTCGGATCGCGCGCAA AACACTGAAGGAGCAAGGGACACCGGAGATTTCGCAATTTGGACCAGACAAAATTTTCGCAATGGCAATGATGACATTGACATCGCG TTGGCCGAGACTCGCGCGGCAATTGCATACGGAGCAGACGGCGTCGCCGTAGTATTCCTTACCGCGCTCTGATGGCTGGTAATG AGCAAGTTGGCTTTGATTAGTAAAGCGCTGCAAGGAAGCTGTGCGGCGAGCCAACGTCCTGTTATCCGTATCATTCGAGACGGG CGAACTTAAGGACGAGGCGTTAATCCGTAAGCATCCGAGATCTCTATCAAAGCCGGCGCGGACCACATTGTTACGTCCACCGGC AAGGTGCGGTTGGTGCGACCCAGAGAGTGCCCGCATCATGATGGAAGTAATTCGCGACATGGGTGTGGAGAAAACCGTGGGAT TTATCCCGCGCGCGCGCTGCGTACTGCAGAAGATGCTCAAAGTATTTAGCAATCGCGGATGAGTTATTTGGCGCCGACTGGGC CGACGCGCTCACTATGCTTTTGGTGCTTCTTCACTGTGTAACCTCGCTGTTAAAGCGCTTGGTACCGGAGATGGGAAGAGTGCA TCGTCTTATGGCAGTTTAGAGCATCATCACCACCACCATTGA
KE70- Shell	ACGGATCTTAAAGCAAGTAGCCTTCGTGCGCTTAAATTAATGCACCTGGCCACAAGTGCCAACGATGACGATACGGATGAAAAATG TGATTGCTCTGTGTATCAAGCTAAGACTCCTGTGCGAAACACCGATGCGATATATATCTATCCTCGCTTTATTTCCTATTGCTCG GAAAACCTTGAAGAGCAAGGCACTCCGAAATCCGGATTTGGACCTCAACCAACTTTCACATGGAAACGACGACATCGATATC GCGTTGGCGGAGACACGTGCTGCGATCGCTTATGGTGCGGATGGCGTGGCAGTGGTGTTTCCCTACCGCGCACTGATGGCGGGTA ACGAGCAGTTCGGGTTGATTGGTTAAAGCCTGCAAGGAAGCTTGCGCGCGCAGCAACGTCGCTTGTCTGTGATTATAGAAAC TGGAGAGCTGAAAGATGAAGCTTTAATACGCAAAAGCATCTGAAATAAGTATTAAAGCAGGCGCGGATCATATTGTGACGTCCACG GGCAAAGTTGCCGTAGGAGCCACTCCAGAATCCGCGCAATAATGATGGAAGTCATCCGTGACATGGGAGTAGAAAAGACTGTTG GGTTCAATTCGGCGAGCGCGCTTCGTACAGCGGAAGATGCCGAGAAATACCTGGCTATAGCGGACGAATTTATTTGGAGCCGATTG GGCAGACGCTCGACATTATGCTTTTCGGCGCATCAGCGTCTTTACTGGCTAGCCTGCTGAAGGCCTTAGGACACGGTGATGGTAA TCAGCGTCATCCTATGGTAGTTTGAACATCATCATCATCATCTAA
KE70- Core	ACGGATTTAAAGGCGTCTCTTTACGCGCCCTCAAACCTGATGCATTTGGCCACCAGTGCTAATGACGATGACACCGATGAAAAGG TAATTGCACCTTGTCAACAGCGGAAAACGCGAGTTGGGACCACGGACGCGATTTTCATTTATCCGCGCTTCATCCGATTGACAG TAAGACGTTAAAGGAGCAAGGAACACCAGAGATCCGTATCTGTACGTGACAAAACCTTCCCTACCGGTAATGATGACATTGATATC GCACTGGCGGAGACCCGCGCGCGCATCGCCTACGGTGACAGATTCCGTGCGGCTGGTCTTCCCTTATCGCGCTCTTATGGCGGGTA ATGAGCAGGTAGGCTTTGACCTTGTTAAAGCGTGCAAGGAAGCTTGCGCGGCCGCCAACGTGTTGCTCGCTGTAATTATTGAAAC TGGCGAGCTGAAAGACGAGGCATTGATTTCGTAAGGCGAGCGAGATCAGCATCAAGGCCGCGCGGATAACATCGTTACGTCTACA GGAAAGGTTGCTGTTGGAGCCACGCCGGAAGCGCTCGCATTATGATGGAAGTTATCCGTGATATGGGTGTGGAAGAACCGGTTG GGTTCATCCCAAGTGGGCGGCGTGCGTACCGCTGAGGATGCGCAAAAAGTATCTTGCCATTGCGGATGAGCTCTTCGAGCGGATTG GGCTGATGACGCCACTACGCCTTTGGGGCTAGCGCAAGTCTGCTTGATCTCTGTTAAAGCTTTAGGTCATGGCGATGGCAAG TCCGCGTCTCGTATGGATCACTTGAGCACCATCATCATCATCTAA
KE70- Evolved	ACTGACCTGAAGGCGTCTTCGTTGCGTGCCCTGAAATTAATGCATCTTGCCACGTCAGCAAAATGACGACGATACCGATGAGAAGC TTATTGCGCTCTGCCACCAAGCCAAGACCGCTGTTGGTAATACCGACGCTATCTTTCATCTACCCGCGCTTATTTCCTATTGCGCG TAAGACCTTAAAGGAGCAGGGTACCCCCGAGATCCGCATTTGCACATCAACCAATTTTCCCATGGAAATGATGACATCGACATT GCCCTGGCCGAAACGCGTGCGGCCATTGCGTATGGCGCCGACAGCGTGGCGCTCGTTTTCCCGTACCGCGCCCTTATGGCGGCA ATGAGCAAGTTGGATTGCACTTGGTCAAGGCGTGTAAGGAAGCTTGTCGGCGCCGCAACGTCGCTTTTGGCAGTCATCATCGAGAC CGGTAGCTGAAAGATGAGGCATTAATCCGAAGGCTCGGATCAGCATCAAGGCCGAGCCGATAACATTGTGACATCATCACC GGCAAAGTGGCAGTAGGAGCAACTCCTGAAAGCGCTCGCATCATGATGGAAGTGATTCTGATATGGGTGTGAGAAAAACGGTAG GTTTTATCCCTGTAGGCGCGCTGCGCACAGCGGAAGATGCTCAGAAGTATTTAGCCATCGTGATGAATTGTTGGGGCCGATTG

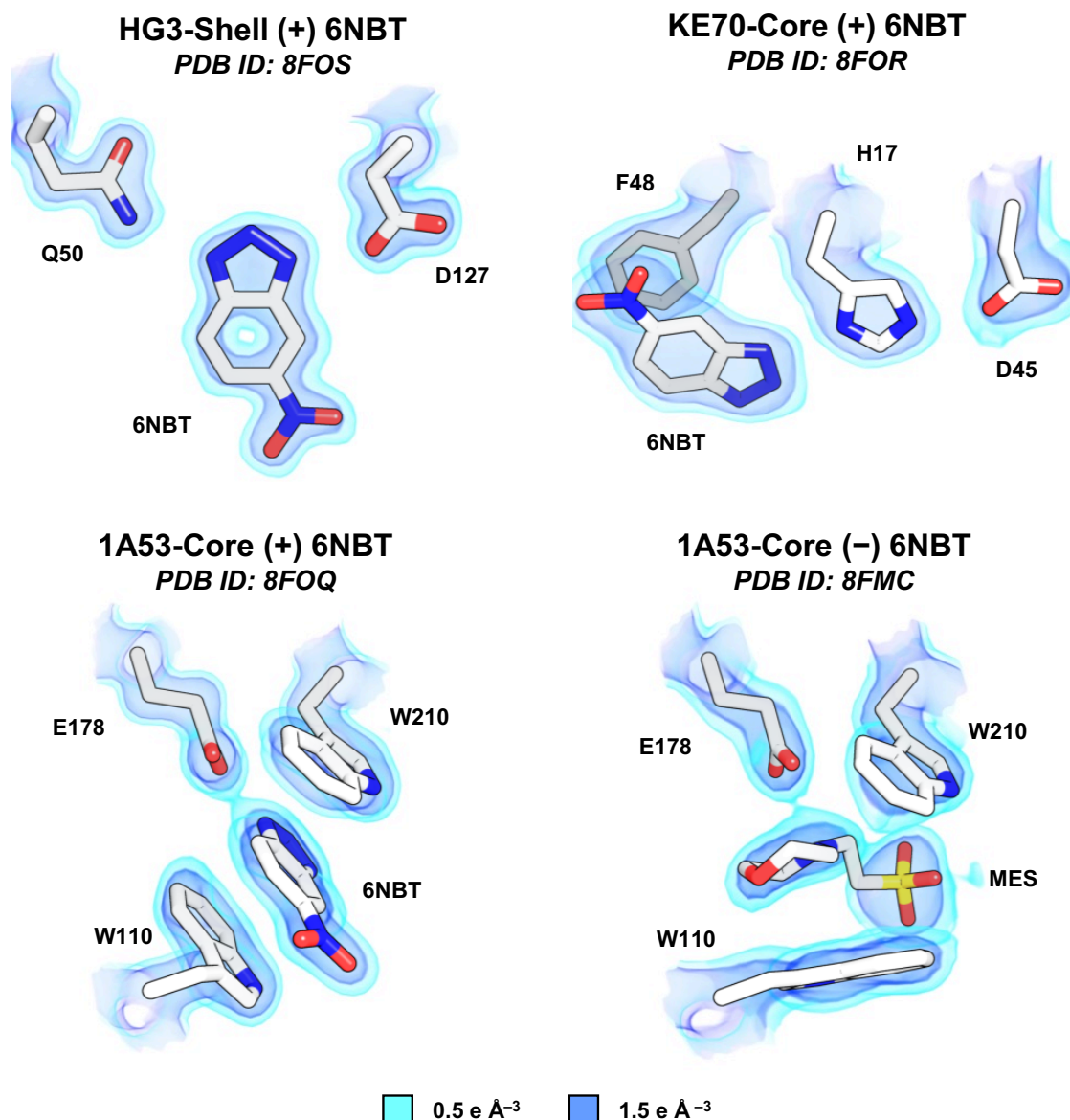
GGCTGACGCTCGCCATTACGCTTTCGGCGCTTCAGCATCCTTACTTGCGTCGCTCTTGAAGGCCTTGGGTCATGGGGATGGGAAG
AGCGCTTCCAGTTATGGCAGCCTGGAACATCACCACCACCATCACTGA



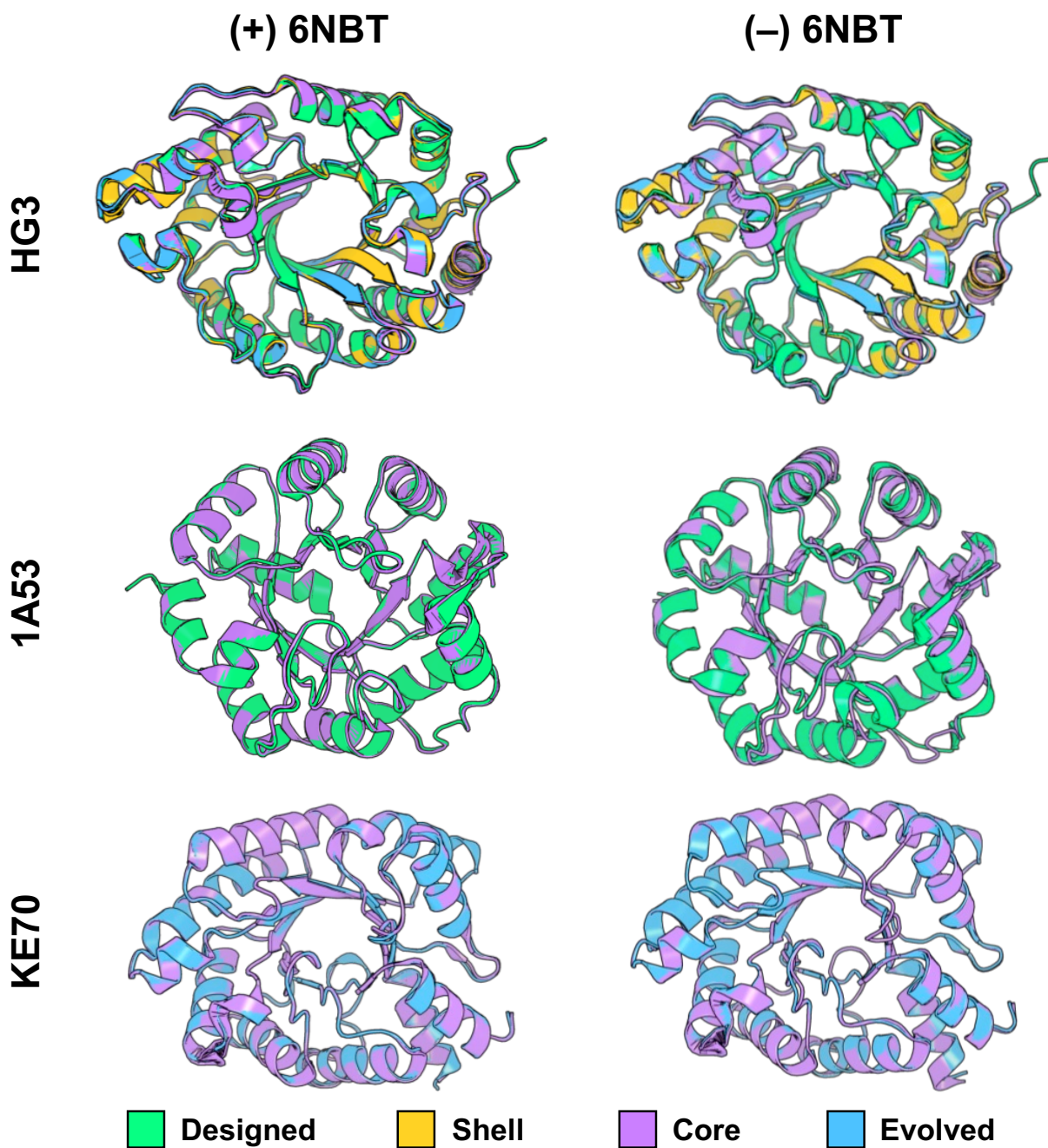
Supplementary Figure 2.1. Far-UV circular dichroism. All spectra were acquired at 20 °C in 10 mM sodium phosphate buffer pH 7 supplemented with 100 mM sodium chloride.



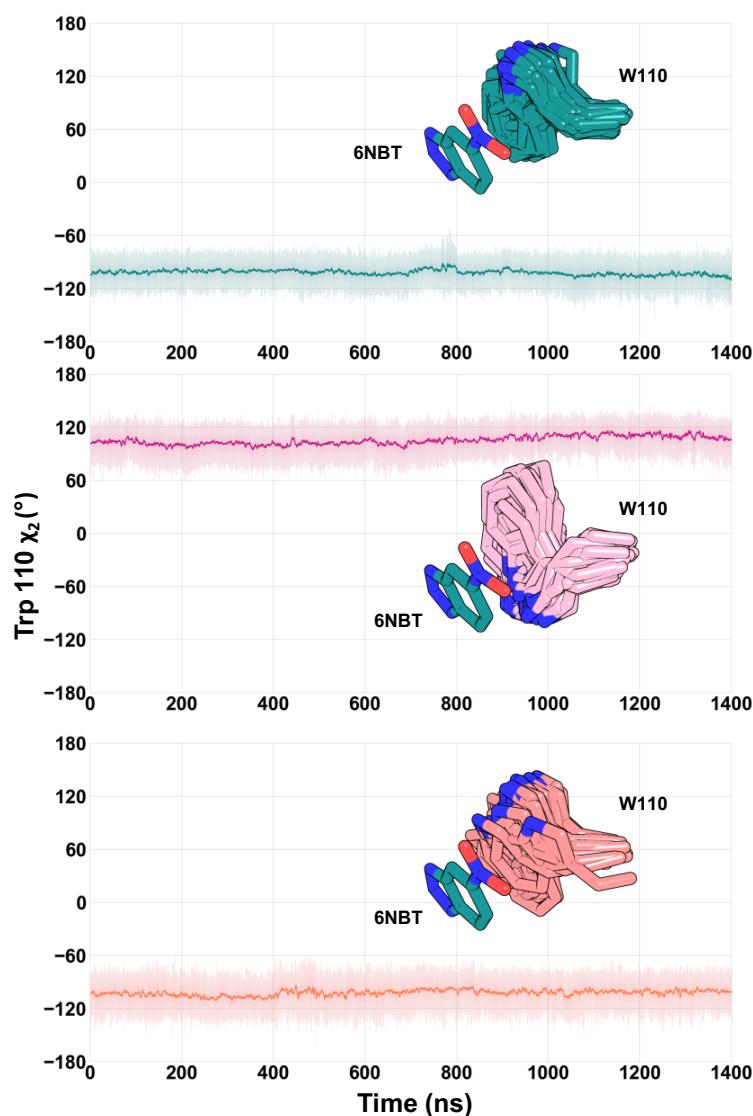
Supplementary Figure 2.2. Thermal denaturation of Kemp eliminases. Thermal denaturation was monitored by circular dichroism at 222 nm. Data were fit to a two-state unfolding model¹. T_m values are reported on Table 2.1.



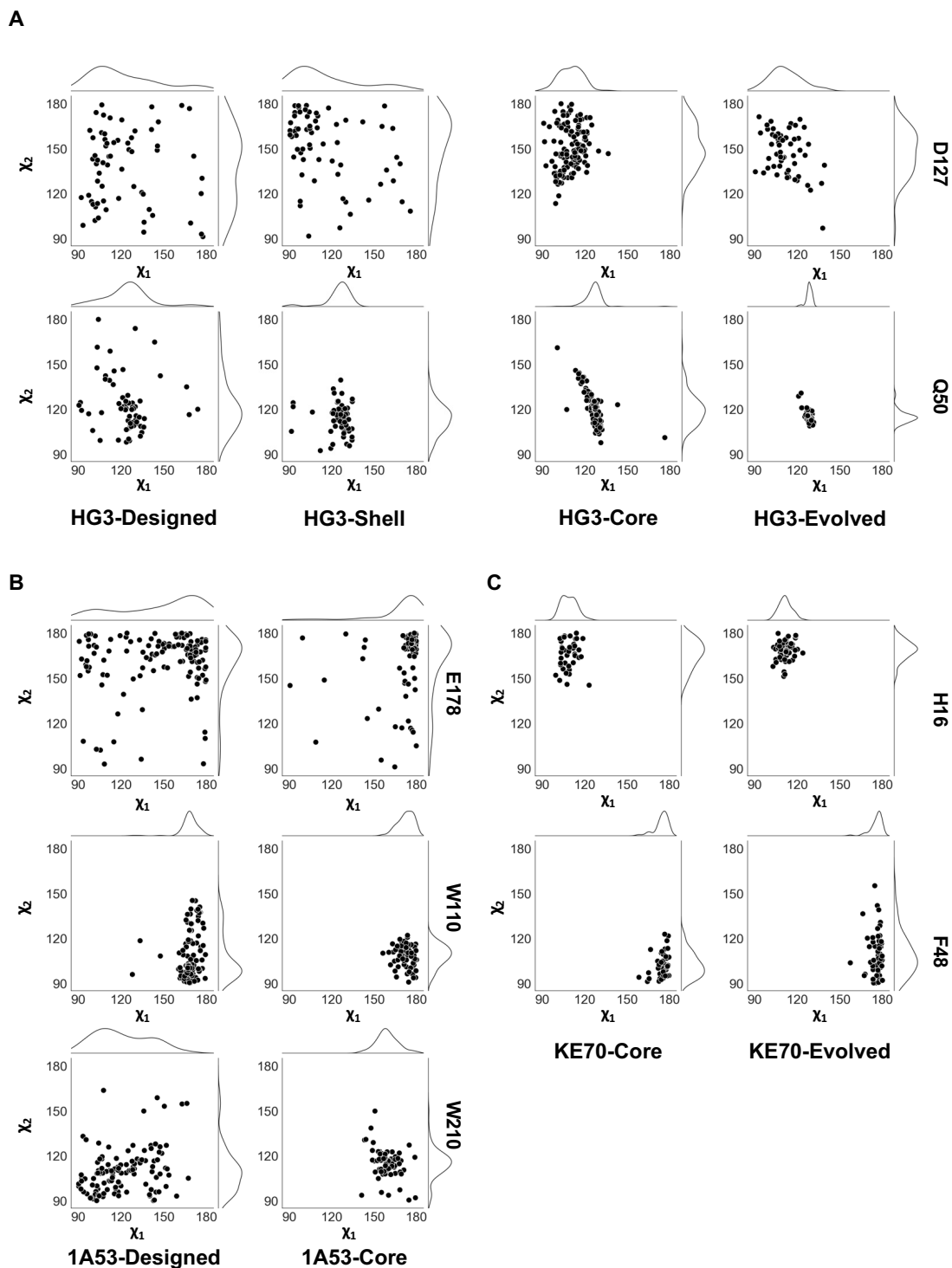
Supplementary Figure 2.3. Electron density of Kemp eliminase active sites. The structures of HG3-Shell, KE70-Core and 1A53-Core show clear density for the transition state analogue 6-nitrobenzotriazole (6NBT) in the active site. However, the structure of 1A53-Core obtained in the absence of transition-state analogue shows density for a 2-(*N*-morpholino)ethanesulfonic acid (MES) molecule from the crystallization buffer. Binding of MES is accompanied by a rotameric change to the W110 side chain, specifically a 120° rotation around χ_2 (from -107.5° to $+11.5^\circ$). The 2Fo-Fc map is shown in volume representation at two contour levels: 0.5 and 1.5 e Å⁻³ in light and dark blue, respectively.



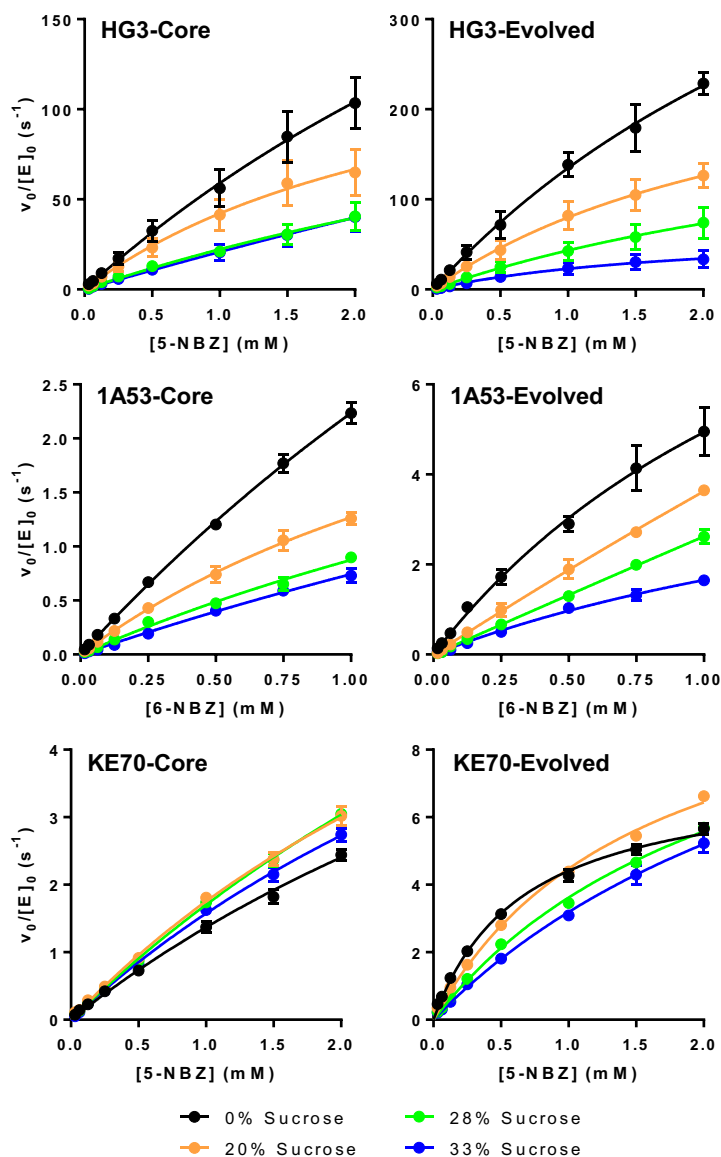
Supplementary Figure 2.4. Active-site and distal mutations do not cause substantial changes to the backbone structure of Kemp eliminases. Available crystal structures of Kemp eliminases in the presence and absence of bound transition-state analogue 6NBT are shown in cartoon representation.



Supplementary Figure 2.5. Evaluation of Trp110 preorganization in 1A53-Core. Microsecond-timescale molecular dynamics simulations of unbound 1A53-Core were conducted using three input structures: (top) the 8FOQ crystal structure with 6NBT deleted from the active site; (middle) the 8FMC crystal structure with MES deleted; and (bottom) the 8FMC crystal structure with MES deleted and the Trp110 side-chain conformation flipped to the t-100 rotamer ($\chi_1 \sim 170^\circ$, $\chi_2 \sim 250^\circ$) observed in unbound 1A53-Designed (PDB ID: 3NYZ) and in 6NBT-bound 1A53-Designed and 1A53-Core (PDB IDs: 3NZ1 and 8FOQ). Across all simulations, the Trp110 side-chain rotamer remains unchanged, and snapshots from trajectories (sticks) show that this residue does not occupy the binding site of 6NBT and instead adopts conformations that could π -stack with this ligand.



Supplementary Figure 2.6. Dihedral angles of catalytic residues in Kemp eliminase conformational ensembles. Ensemble refinement reveals decreased conformational heterogeneity of catalytic residues D127 and Q50 in HG3 enzymes upon introduction of active-site and distal mutations. Similarly, E178, W110, and W210 exhibit reduced heterogeneity when active-site mutations convert 1A53-Designed into 1A53-Core. In KE70-Core, the addition of distal mutations to generate KE70-Evolved reduces the conformational heterogeneity of catalytic residues H16 and F48.

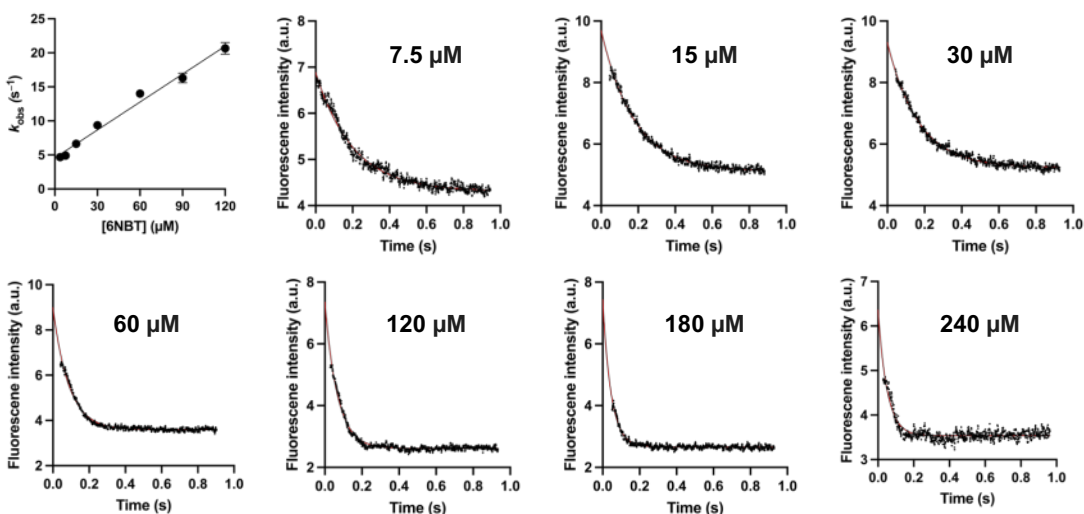


Supplementary Figure 2.7. Steady-state kinetics at various viscosities. Michaelis-Menten plots of normalized initial rates as a function of substrate concentration are shown. The effect of solvent viscosity on catalytic efficiency of KE70, 1A53, and HG3 variants was determined using sucrose as the viscogen at different concentrations (0, 20, 28 and 33 % w/v). Initial rates were determined and fitted to the linear portion of the Michaelis-Menten model ($v_0 = (k_{cat}/K_M) [E_0] [S]$), as saturation could not be achieved for most reactions in the presence of sucrose. The exception was KE70-Evolved, for which the standard Michaelis-Menten equation was applicable. Data represent the average of 6 individual replicates from 2 independent protein batches, with error bars reporting the SEM.

HG3-Core

$$k_{\text{on}} = 0.137 \pm 0.005 \mu\text{M}^{-1} \text{s}^{-1}$$

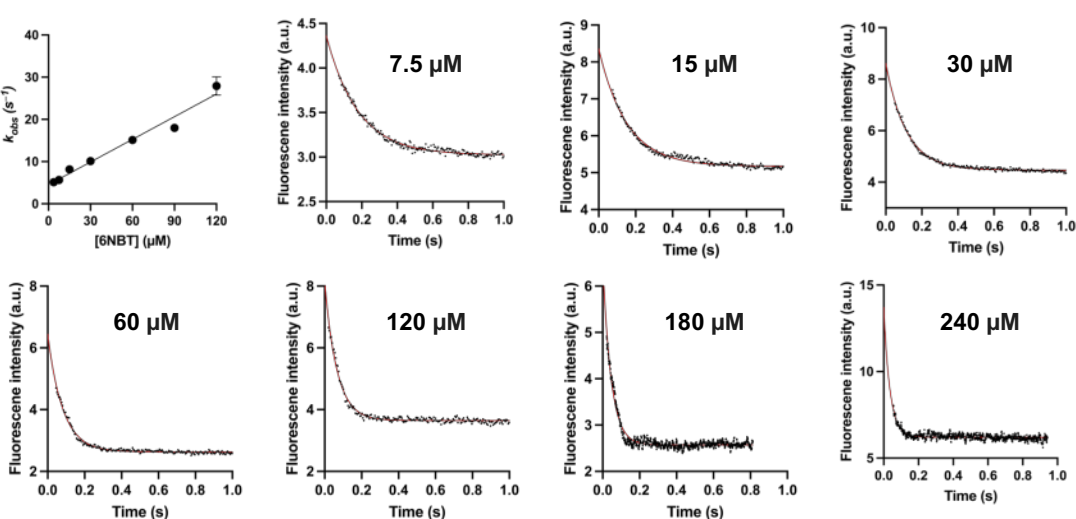
$$k_{\text{off}} = 4.56 \pm 0.34 \text{s}^{-1}$$



HG3-Evolved

$$k_{\text{on}} = 0.179 \pm 0.009 \mu\text{M}^{-1} \text{s}^{-1}$$

$$k_{\text{off}} = 4.51 \pm 0.61 \text{s}^{-1}$$

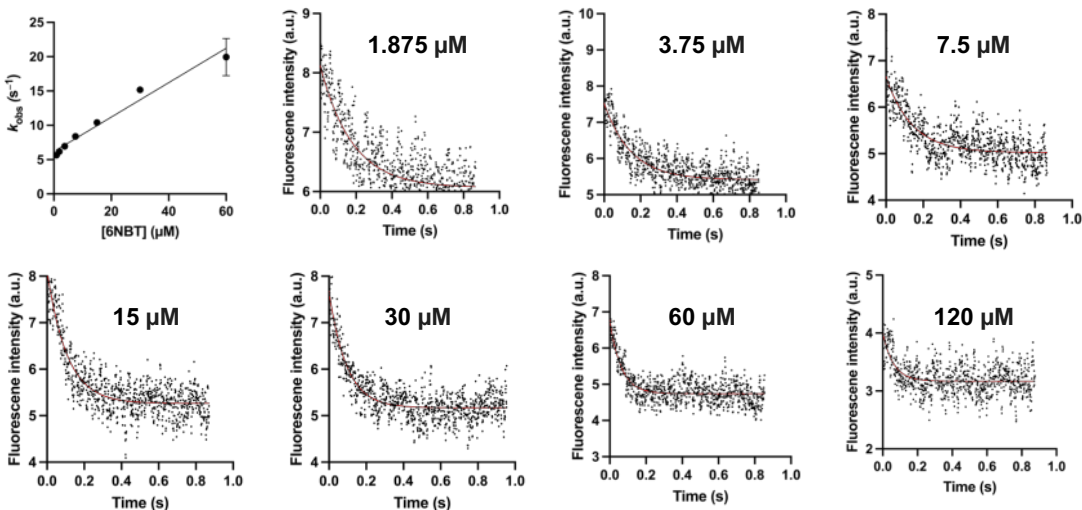


Supplementary Figure 2.8. Kinetics of 6NBT binding to HG3 enzymes measured by stopped-flow Trp fluorescence quenching experiments. Observed rate constants (k_{obs}) plotted against 6NBT concentration show a linear dependency, where the slope and y-intercept correspond to k_{on} and k_{off} , respectively. Data represent the average of at least three individual replicates from a single protein batch (mean $\pm$ SEM). k_{obs} values were extracted from individual fluorescence traces at various 6NBT concentrations using a one phase exponential decay function. A single representative fluorescence trace is shown for each 6NBT concentration.

1A53-Core

$$k_{\text{on}} = 0.25 \pm 0.017 \mu\text{M}^{-1} \text{s}^{-1}$$

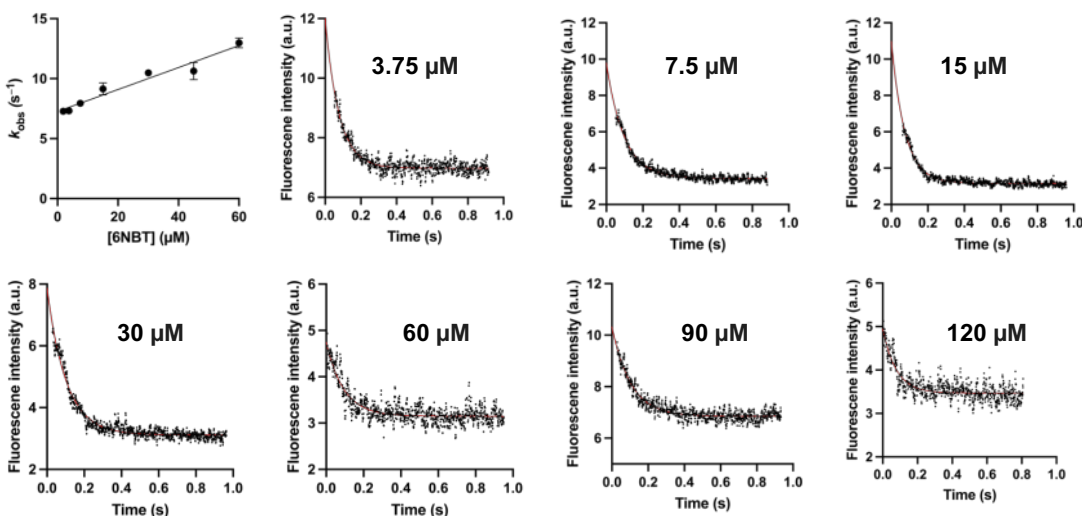
$$k_{\text{off}} = 6.18 \pm 0.39 \text{s}^{-1}$$



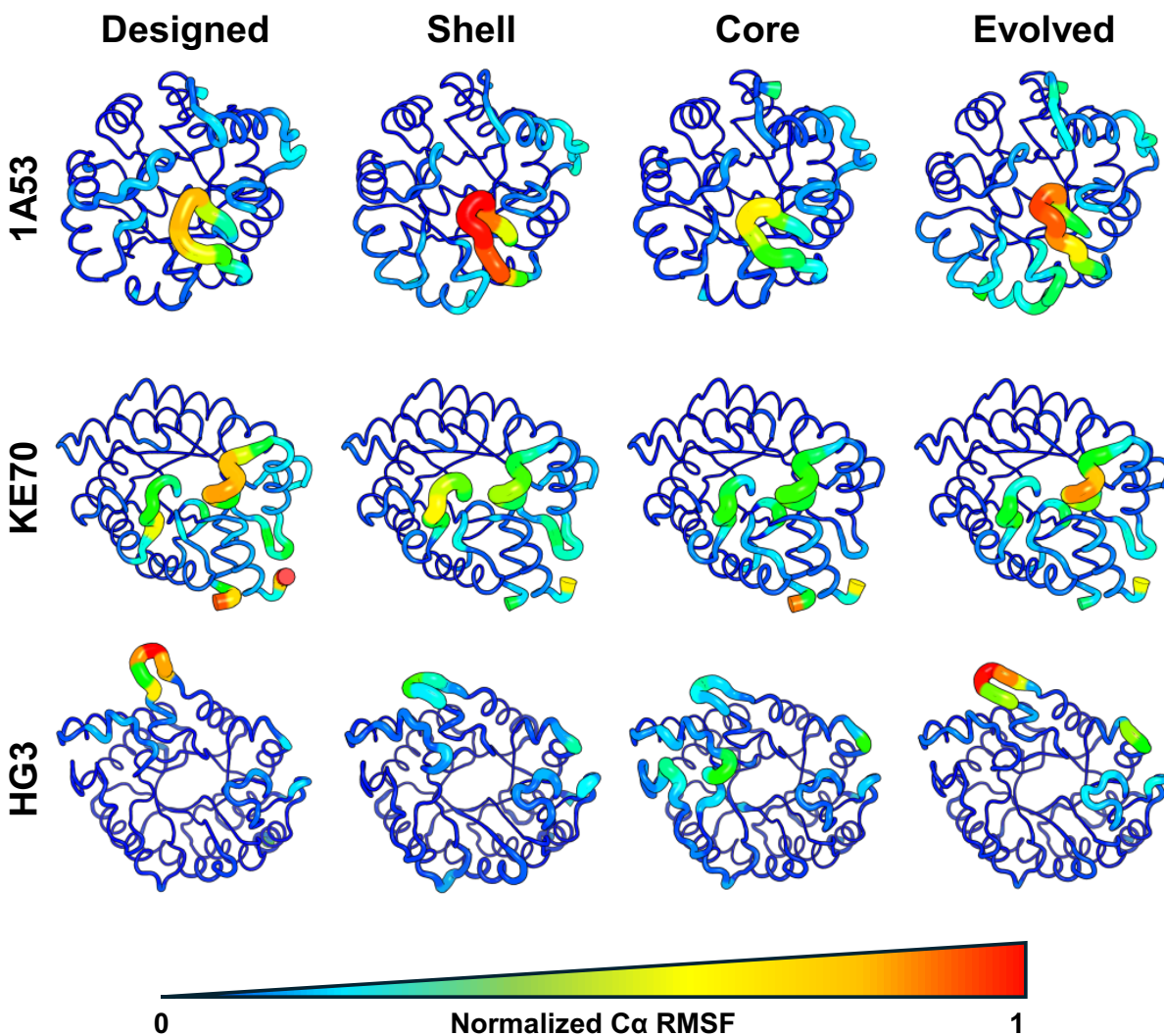
1A53-Evolved

$$k_{\text{on}} = 0.091 \pm 0.007 \mu\text{M}^{-1} \text{s}^{-1}$$

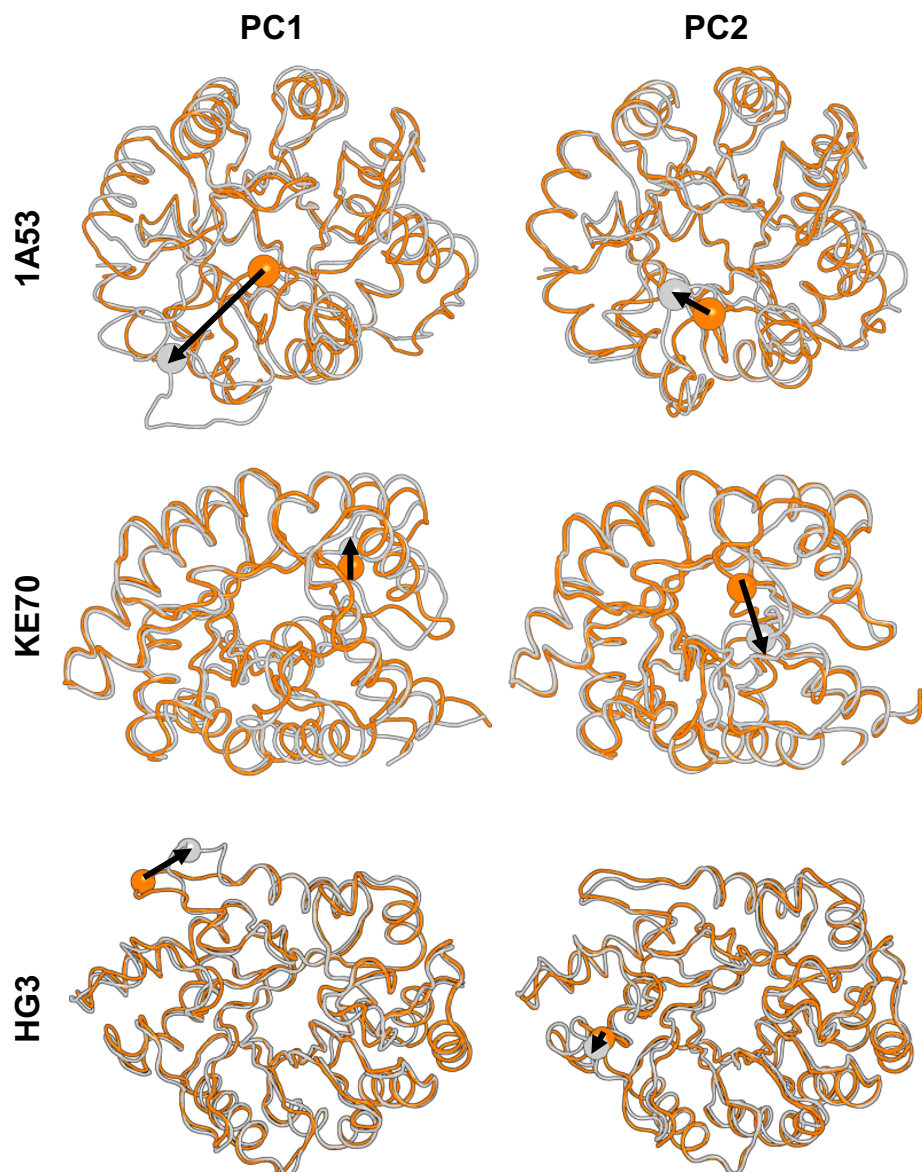
$$k_{\text{off}} = 7.26 \pm 0.23 \text{s}^{-1}$$



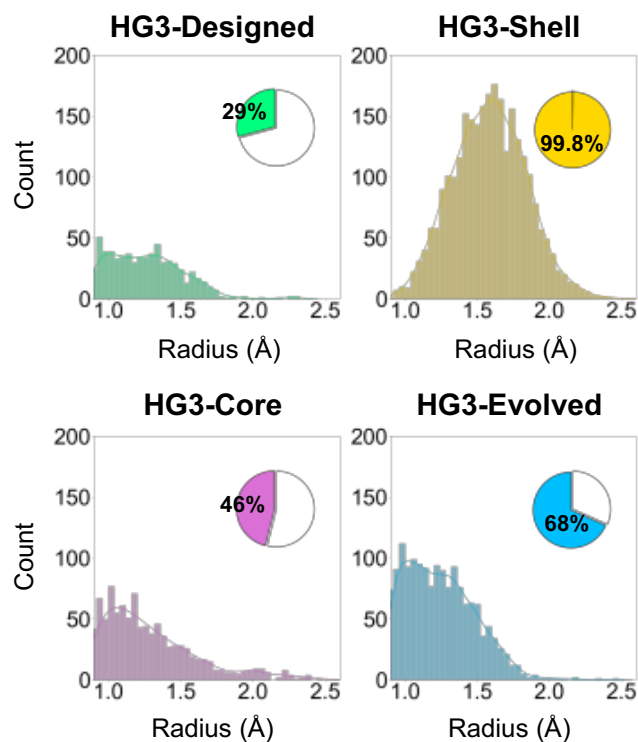
Supplementary Figure 2.9. Kinetics of 6NBT binding to 1A53 enzymes measured by stopped-flow Trp fluorescence quenching experiments. Observed rate constants (k_{obs}) plotted against 6NBT concentration show a linear dependency, where the slope and y-intercept correspond to k_{on} and k_{off} , respectively. Data represent the average of at least three individual replicates from a single protein batch (mean $\pm$ SEM). k_{obs} values were extracted from individual fluorescence traces at various 6NBT concentrations using a one phase exponential decay function. A single representative fluorescence trace is shown for each 6NBT concentration.



Supplementary Figure 2.10. Conformational flexibility of Kemp eliminases during molecular dynamics. Per-residue root mean square fluctuations (RMSF) mapped onto the Ca Backbone of HG3, 1A53 and KE70 Designed, Shell, Core and Evolved variants. Regions with lowest and highest RMSF values are colored in blue and red, respectively.



Supplementary Figure 2.11. Dominant conformational changes in each enzyme family revealed by principal component analysis. Enzyme structures (orange and white) illustrate the conformational changes contributing to the first two principal components (PC1 and PC2). Spheres represent the alpha carbon of residues located on structural elements undergoing these changes, with arrows indicating the direction and magnitude of movement corresponding to increasing PC values (Figure 2.5A). In 1A53 enzymes, PC1 corresponds to the displacement of an active-site loop (residues 55–58) from its position above the active site to approximately 8 Å away, which opens the active site. PC2 represents the movement of the same loop (residues 58–60) about 5 Å toward the active site. In KE70 variants, PC1 highlights a 3 Å movement of residues 66–67 located on a flexible loop far from the active site, whereas PC2 reveals a 6.5 Å displacement of an active-site loop (residues 21–23), increasing its separation from a second loop and opening the active site. In HG3 enzymes, PC1 reflects an 8 Å movement of the loop encompassing residues 58–60, far from the active site, while PC2 reveals a 2 Å shift in another distal loop (residues 137–145). These conformational changes occur away from the HG3 active site and do not directly impact its opening.



Supplementary Figure 2.12. Distal mutations widen active-site tunnels. Caver3 was used to compute tunnels to the active site for 1000 snapshots from the most populated energy cluster in PCA plots of HG3 enzymes (Figure 5a). Pie charts show the percentage of structures containing a tunnel with a minimum bottleneck radius of 0.9 Å, a threshold chosen because it matches the radius of small ions commonly bound in protein structures. The median bottleneck radius for Designed, Shell, Core and Evolved variants is 1.24 Å, 1.58 Å, 1.20 Å and 1.22 Å, respectively.

2.9. References

1. Rauwerdink, A. & Kazlauskas, R. J. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β -Hydrolase Fold Enzymes. *ACS Catal.* **5**, 6153–6176 (2015).
2. Du, S., Kretsch, R. C., Parres-Gold, J., Pieri, E., Cruzeiro, V. W. D., Zhu, M., Pinney, M. M., Yabukarski, F., Schwans, J. P., Martínez, T. J. & Herschlag, D. Conformational Ensembles Reveal the Origins of Serine Protease Catalysis. *Science*. **387**, eado5068 (2025).
3. Khersonsky, O. Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *Journal of Molecular Biology* **407**, 391-412 (2011).
4. Broom, A., Rakotoharisoa, R. V., Thompson, M. C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J. S. & Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 08 (2020).
5. Otten, R., Pádua, R. A. P., Bunzel, H. A., Nguyen, V., Pitsawong, W., Patterson, M., Sui, S., Perry, S. L., Cohen, A. E., Hilvert, D. & Kern, D. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).
6. Hong, N.-S. The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.* **9**, 3900 (2018).
7. Fröhlich, C., Bunzel, H. A., Buda, K., Mulholland, A. J., Van Der Kamp, M. W., Johnsen, P. J., Leiros, H.-K. S. & Tokuriki, N. *Epistasis Arises from Shifting the Rate-Limiting Step during Enzyme Evolution*. *Nature Catalysis* **7**, 499-509 (2024).
8. Peccati, F., Noey, E. L., Houk, K. N., Osuna, S. & Jiménez-Osés, G. Interplay Between Substrate Polarity and Protein Dynamics in Evolved Kemp Eliminases. *ChemCatChem* **16**, e202400444 (2024).
9. Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., Afriat-Jurnou, L., Buckle, A. M., Weik, M., Hollfelder, F., Tokuriki, N. & Jackson, C. J. The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).

10. Jiménez-Osés, G., Osuna, S., Gao, X., Sawaya, M. R., Gilson, L., Collier, S. J., Huisman, G. W., Yeates, T. O., Tang, Y. & Houk, K. N. The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **10**, 431–436 (2014).
11. Doustmohammadi, H., Sanchez, J., Ram Mahato, D. & Osuna, S. Evolution Enhances Kemp Eliminate Activity by Optimizing Oxyanion Stabilization and Conformational Flexibility. *Chem. – Eur. J.* **31**, e202403747 (2025).
12. Lee, J. & Goodey, N. M. Catalytic Contributions from Remote Regions of Enzyme Structure. *Chem. Rev.* **111**, 7595–7624 (2011).
13. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).
14. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comput. Mol. Sci.* **11**, (2021).
15. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
16. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. *De Novo* Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
17. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E. & Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
18. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).

19. Lassila, J. K. Conformational diversity and computational enzyme design. *Curr. Opin. Chem. Biol.* **14**, 676–682 (2010).
20. Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D. & Baker, D. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
21. Althoff, E. A., Wang, L., Jiang, L., Giger, L., Lassila, J. K., Wang, Z., Smith, M., Hari, S., Kast, P., Herschlag, D., Hilvert, D. & Baker, D. Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
22. Bjelic, S., Kipnis, Y., Wang, L., Pianowski, Z., Vorobiev, S., Su, M., Seetharaman, J., Xiao, R., Kornhaber, G., Hunt, J. F., Tong, L., Hilvert, D. & Baker, D. Exploration of Alternate Catalytic Mechanisms and Optimization Strategies for Retroaldolase Design. *J. Mol. Biol.* **426**, 256–271 (2014).
23. Khersonsky, O., Kiss, G., Röthlisberger, D., Dym, O., Albeck, S., Houk, K. N., Baker, D. & Tawfik, D. S. 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 (2012).
24. Khersonsky, O., Röthlisberger, D., Dym, O., Albeck, S., Jackson, C. J., Baker, D. & Tawfik, D. S. Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J. Mol. Biol.* **396**, 1025–1042.
25. Crawshaw, R., Crossley, A. E., Johannissen, L., Burke, A. J., Hay, S., Levy, C., Baker, D., Lovelock, S. L. & Green, A. P. Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
26. Blomberg, R., Kries, H., Pinkas, D. M., Mittl, P. R. E., Grütter, M. G., Privett, H. K., Mayo, S. L. & Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).

27. Bunzel, H. A., Kries, H., Marchetti, L., Zeymer, C., Mittl, P. R. E., Mulholland, A. J. & Hilvert, D. Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **141**, 11745–11748 (2019).
28. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N. & Hilvert, D. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
29. Obexer, R., Godina, A., Garrabou, X., Mittl, P. R. E., Baker, D., Griffiths, A. D. & Hilvert, D. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
30. Rakotoharisoa, R. V., Seifinoferest, B., Zarifi, N., Miller, J. D. M., Rodriguez, J. M., Thompson, M. C. & Chica, R. A. Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J Am Chem Soc* **146**, 10001–10013 (2024).
31. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci.* **92**, 452–456 (1995).
32. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
33. Gadda, G. & Sobrado, P. Kinetic Solvent Viscosity Effects as Probes for Studying the Mechanisms of Enzyme Action. *Biochemistry* **57**, 3445–3453 (2018).
34. Kries, H., Bloch, J. S., Bunzel, H. A., Pinkas, D. M. & Hilvert, D. Contribution of Oxyanion Stabilization to Kemp Eliminase Efficiency. *ACS Catal.* **10**, 4460–4464 (2020).
35. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sust, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J. & Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLOS Comput. Biol.* **8**, e1002708 (2012).

36. Bunzel, H. A., Anderson, J. L. R., Hilvert, D., Arcus, V. L., Van Der Kamp, M. W. & Mulholland, A. J. Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **13**, 1017–1022 (2021).
37. Yang, G., Miton, C. M. & Tokuriki, N. A mechanistic view of enzyme evolution. *Protein Sci.* **29**, 1724–1747 (2020).
38. Natarajan, S. K. & Sierks, M. R. Identification of Enzyme–Substrate and Enzyme–Product Complexes in the Catalytic Mechanism of Glucoamylase from *Aspergillus awamori*. *Biochemistry* **35**, 15269–15279 (1996).
39. Machado, T. F. G., Gloster, T. M. & Da Silva, R. G. Linear Eyring Plots Conceal a Change in the Rate-Limiting Step in an Enzyme Reaction. *Biochemistry* **57**, 6757–6761 (2018).
40. Yao, L., Li, Y., Wu, Y., Liu, A. & Yan, H. Product Release Is Rate-Limiting in the Activation of the Prodrug 5-Fluorocytosine by Yeast Cytosine Deaminase. *Biochemistry* **44**, 5940–5947 (2005).
41. Vaissier, V., Sharma, S. C., Schaettle, K., Zhang, T. & Head-Gordon, T. Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **8**, 219–227 (2018).
42. Wankowicz, S. & Fraser, J. Making sense of chaos: uncovering the mechanisms of conformational entropy. *ChemRxiv* (2024)
43. Lauko, A., Pellock, S. J., Anischanka, I., Sumida, K. H., Juergens, D., Ahern, W., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I. R., Jamieson, C., Kang, A., Brackenbrough, E., Bera, A. K., Sankaran, B., Houk, K. N. & Baker, D. Computational design of serine hydrolases. *bioRxiv* (2024).
44. Braun, M., Tripp, A., Chakatok, M., Kaltenbrunner, S., Totaro, M. G., Stoll, D., Bijelic, A., Elaily, W., Hoch, S. Y. Y., Aleotti, M., Hall, M. & Oberdorfer, G. Computational design of highly active *de novo* enzymes. *bioRxiv* (2024).
45. Crean, R. M., Gardner, J. M. & Kamerlin, S. C. L. Harnessing Conformational Plasticity to Generate Designer Enzymes. *J. Am. Chem. Soc.* **142**, 11324–11342 (2020).

46. Duran, C., Casadevall, G. & Osuna, S. Harnessing conformational dynamics in enzyme catalysis to achieve nature-like catalytic efficiencies: the shortest path map tool for computational enzyme redesign. *Faraday Discuss.* **252**, 306–322 (2024).
47. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel β -sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein science : a publication of the Protein Society* **8**, 841-853 (1999).
48. Swindells, J. F., Snyder, C. F. & Hardy, R. C. Viscosities of sucrose solutions at various temperatures : tables of recalculated values. (Washington : For sale by the Supt. of Docs., U.S. G.P.O., [1958], 1958).
49. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
50. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013).
51. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
52. Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877 (2019).
53. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
54. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).

55. Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migués, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q. & Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
56. Lee, F. S., Anderson, A. G. & Olafson, B. D. Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Eng. Des. Sel.* **36**, gzad013 (2023).
57. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
58. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
59. Roe, D. R. & Cheatham, T. E. I. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

CHAPTER 3. CUSTOMIZING THE STRUCTURE OF A *DE NOVO* TIM BARREL TO INTRODUCE ENZYMATIC ACTIVITY

Niayesh Zarifi,¹ Benjamin J. Smith,¹ Julian Beck,² Emily Freund,² Birte Höcker² & Roberto A. Chica¹

¹ Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

² Department of Biochemistry, University of Bayreuth, Bayreuth, Germany.

3.1. Foreword

Chapter 2 delved into how distal mutations impact enzyme catalysis, revealing that mutations far from the active site facilitate substrate binding and product release. We observed that Core variants containing only active-site mutations displayed catalytic efficiencies several orders of magnitude greater than the original designs. This suggests that it is possible to create efficient enzymes by focusing solely on the active site. Furthermore, we observed that distal mutations impact the conformational flexibility of surface loops, shifting the enzymes' conformational landscape towards more open states. Building upon these insights, Chapter 3 shifts focus towards designing structural fragments that can act as a lid and accommodate substrate binding and product release. Moreover, we aim to address another challenge in enzyme design, which is the limited availability of protein scaffolds that are well-suited to the reaction of interest. In this chapter, we developed a workflow to modify the structure of *de novo* TIM barrels to introduce Kemp eliminase activity. Using a combination of physics-based and machine learning tools, we customized the minimal structure of *de novo* TIM barrel scaffolds to create flexible lid fragments that enclose the active site. Our objective was to create a workflow where a minimal and functionless protein scaffold can be customized for any reaction of interest, providing a versatile platform for enzyme design.

3.2. Statement of contribution

Niayesh Zarifi and Benjamin J. Smith performed the computational enzyme design on F15C TIM barrel. Julian Beck and Emily Freund performed the computational enzyme design on CP1 and CP2 TIM barrels. Niayesh Zarifi and Benjamin J. Smith performed experimental characterization of the enzymes. Benjamin J. Smith performed molecular dynamics simulations and Niayesh Zarifi and Benjamin J. Smith analyzed the molecular dynamics simulations. Roberto A. Chica and Birte Höcker conceived the project. Niayesh Zarifi wrote the chapter.

3.3. Abstract

The ability to design enzymes from scratch for new-to-nature reactions is highly desired but challenging. Traditional enzyme design approaches often rely on repurposing natural scaffolds, which may not always be ideal given the native activities of these pre-existing scaffolds and the fact that there may not be existing scaffolds well-suited for the novel reaction of interest. In this study, we aimed to create a functional enzyme from a *de novo* functionless TIM barrel scaffold for the Kemp elimination reaction. We altered the structure of *de novo* inert TIM barrels to introduce a lid fragment that sequesters the active site from the solvent. This structural modification was achieved by using a combination of machine-learning tools, such as RFDiffusion for generating structural segments, ProteinMPNN for sequence design, and AlphaFold2 for structure validation. Among 20 experimentally characterized designs, one design, KempTIM 4, catalyzes the Kemp elimination reaction with a catalytic efficiency of $10 \text{ M}^{-1}\text{s}^{-1}$. Microsecond timescale molecular dynamics (MD) simulations revealed flexibility in the lid region and a loop adjacent to the active site of KempTIM 4. This flexibility may be crucial for facilitating substrate entry and product release, differentiating KempTIM 4 from the inactive variants, which contain more rigid lid fragments. This hypothesis is further supported by the observation that KempTIM 4 possesses wider tunnels leading to the active site, in contrast to the inactive designs. Our findings highlight the limitations of current machine-learning methods for enzyme design, particularly in achieving the precision required for efficient catalysis.

3.4. Introduction

Enzymes are the most efficient catalysts, exhibiting unparalleled rate acceleration, selectivity, and sustainability. These properties make them highly attractive for industrial applications in textiles, detergents, and pharmaceuticals. A long-standing goal in enzyme engineering has been to create tailor-made enzymes capable of catalyzing new-to-nature reactions. Over the past two decades, several *de novo* enzymes have been designed for various model organic transformations.^{1–5} The process of designing enzymes for novel reactions typically begins with generating an idealized active site, or “theozyme”, which consists of the transition state of the rate-limiting step and optimally oriented catalytic residues. This theozyme is then placed into a protein scaffold, and subsequently, residues in the active site are mutated to further stabilize the transition state (Figure 3.1).

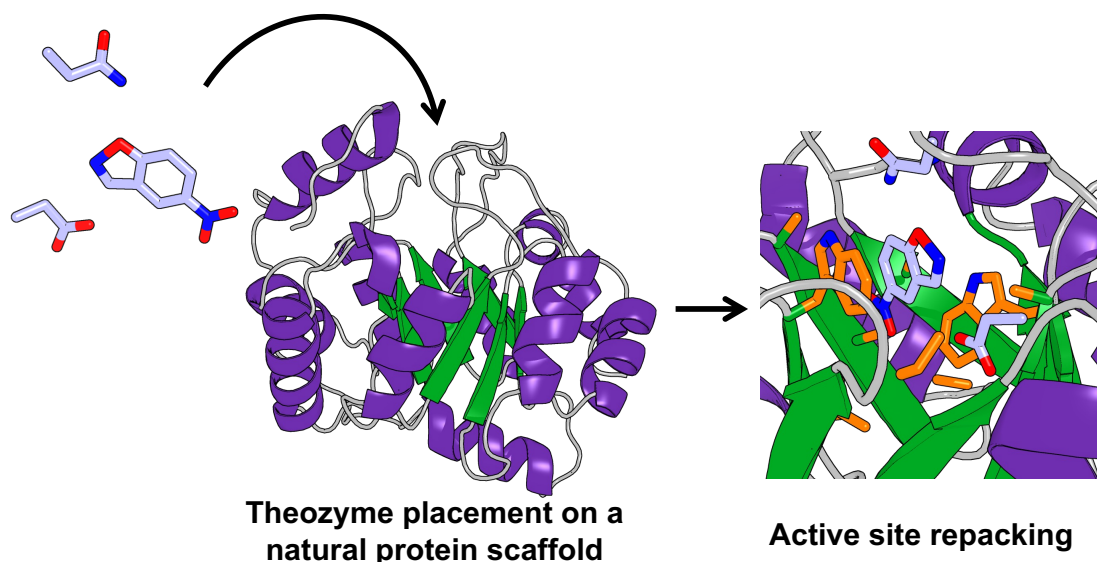


Figure 3.1. *De novo* enzyme design methodology. The theozyme comprising the transition state and catalytic residues is placed onto a natural protein scaffold. Subsequently, residues in the active site are mutated to further stabilize the transition state and improve packing.

Traditionally, the creation of *de novo* enzymes has been achieved by repurposing natural scaffolds to introduce novel catalytic activities. While this approach offers a diverse range of scaffolds, it has inherent limitations. Native protein scaffolds are highly evolved for their specific biological functions and are often only marginally stable to begin with.⁶ Introducing mutations to engineer new catalytic functions could further destabilize these scaffolds, as such mutations are predominantly destabilizing, potentially resulting in unstable proteins.⁷ Furthermore, there may not be a pre-existing scaffold well-suited to the new reaction of interest.

To address the challenges of using natural protein scaffolds, recent enzyme design efforts have shifted towards using *de novo* scaffolds. In this approach, a theozyme is either placed onto a newly designed scaffold or the scaffold itself is generated around the theozyme.^{8,9} *De novo* scaffolds offer advantages, such as a well-defined sequence-structure relationship based on the design principles used, and their idealized structures lack the complex features seen in natural proteins, such as flexible long loops. This strategy has already led to successful designs, including the design of a retroaldolase on a *de novo* β -barrel,⁸ the creation of a *de novo* luciferase,⁹ and the computational design of serine hydrolases.¹⁰ However, despite these achievements, this approach faces challenges, such as low experimental success rates due to false positives,⁹ difficulty in designing the flexibility required for many enzymatic reactions,¹¹ and a lack of sub-ångstrom precision required in the generation of efficient enzymes.¹²

In response to these challenges, we have developed a workflow to customize a simple inert scaffold to create a *de novo* enzyme for the model Kemp elimination reaction by using a combination of physics-based and machine learning tools. Specifically, we focused on artificial triosephosphate isomerase (TIM) barrel protein fold as our target scaffold. The TIM barrel, or $(\beta/\alpha)_8$ fold, is the most common enzyme fold, found in approximately 10% of all enzymes,

catalyzing a wide range of unrelated reactions.^{13,14,15} This fold is particularly well-suited for designing novel catalytic activities due to its structural diversity, frequent occurrence in nature, and presence of active sites with significant structural and functional flexibility. A notable advantage of TIM barrels is the mutability of their β/α loops, which are not critical for stability and can be modified to introduce catalytic residues or other structural segments.^{14,16} Additionally, residues in the eight β strands and adjacent loops are oriented towards the top of the barrel, where the active site is typically located, making the TIM barrel fold an ideal scaffold for introducing new catalytic functions.¹⁷ Early *de novo* enzyme design efforts frequently employed TIM barrels as the starting scaffold.¹⁻³ While these efforts successfully generated active designs, in some cases, long loops near the active site exhibited unexpected configurations in the crystal structures.^{17,18}

The widespread occurrence and functional diversity of the TIM barrel fold have made the *de novo* design of artificial TIM barrels from scratch highly desirable. In 2015, a *de novo* TIM barrel with four-fold symmetry, sTIM11, was successfully designed with a sequence distinct from natural TIM barrels.¹⁷ Artificial TIM barrels provide a unique opportunity to design novel functions due to their simple regular structures and minimalist loops that can be modified to create tailor-made catalysts. Here, we targeted three *de novo* TIM barrels. The four-fold symmetric *de novo* TIM barrel, F15C (PDB ID: 7MCD), was designed using a deep neural network sequence design model conditioned on local backbone structure and chemical environment.¹⁹ CP1 and CP2 were created by circular permutation of NT6-SB TIM barrel. NT6-SB is a *de novo* TIM barrel scaffold derived from the parent scaffold NT6 (PDB ID: 7OSV), created by introducing salt bridge clusters at the base of the barrel to enhance its stability and crystallizability.²⁰

De novo TIM barrels are structurally simple and functionally inert, typically containing the eight α -helices and β -strands characteristic of TIM barrels. This simplicity contrasts with natural

TIM barrels, which often contain additional helices and flexible loops that contribute to structural integrity, ligand binding, and catalysis (Figure 3.2). We hypothesized that the high-resolution structures of these idealized *de novo* TIM barrels could serve as “universal scaffolds” that are customizable for novel reactions through the addition of structural segments. This approach offers advantages over repurposing natural scaffolds, often marginally stable and encumbered by millions of years of evolutionary history. It also surpasses the generation of entirely *de novo* scaffolds, as the crystal structures provide higher accuracy than design models.

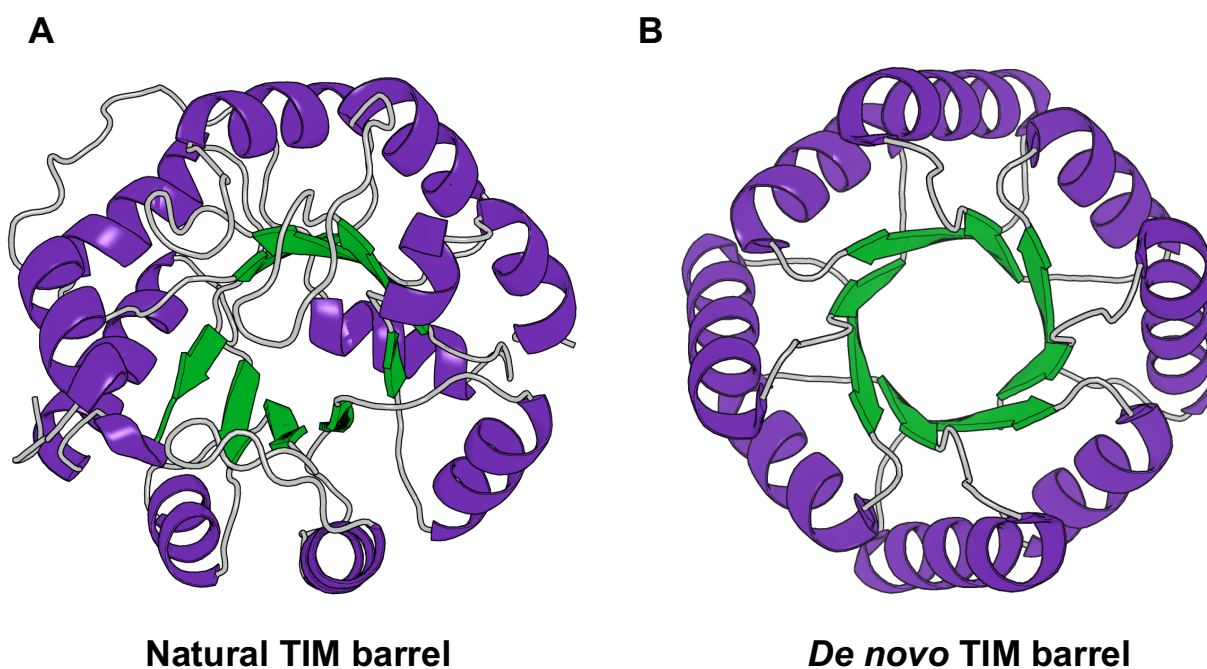


Figure 3.2. Natural versus *de novo* TIM barrels. (A) Natural TIM barrels are more structurally complex, (e.g., PDB ID: 1A53), being highly evolved for their native functions. (B) In contrast, *de novo* designed TIM barrels (e.g., PDB ID: 7MCD) are structurally simple and functionally inert.

Here, we present a framework to customize a simple, inert TIM barrel scaffold to create a *de novo* enzyme to catalyze Kemp elimination reaction. To do so, we combine physics-based methods of theozyme placement and rotamer optimization with machine learning methods, such as RFDiffusion, to introduce structural segments designed to act as a lid over the active-site and scaffold catalytic residues, ProteinMPNN, to design the sequence of the lid region, and AlphaFold2, to validate the folding of designs. We applied this framework to three *de novo* TIM barrels, and out of 20 designs tested experimentally, one exhibited Kemp elimination activity with a catalytic efficiency of $10 \text{ M}^{-1}\text{s}^{-1}$, comparable to previously designed Kemp eliminases. The active design, Kemp TIM 4, exhibited greater flexibility in both the lid region and a loop near the active site during molecular dynamics (MD) simulations compared to inactive designs. Additionally, it featured wider tunnels to the active site, which could play a role in facilitating substrate binding and product release.

3.5. Design strategy and *in silico* results

We targeted Kemp elimination reaction, a model proton transfer reaction and a *de novo* enzyme design benchmark. This reaction involves the deprotonation of a benzisoxazole by a catalytic base and delocalization of the resulting negative charge, which leads to the formation of a cyanophenolate through ring opening.^{21,22} The theozyme for this reaction includes the transition state, a catalytic base (Asp, Glu) to deprotonate the benzisoxazole, and a hydrogen bond donor (Asn, Gln) to stabilize the partial negative charge on the isoxazolic oxygen (Figure 3.3).¹

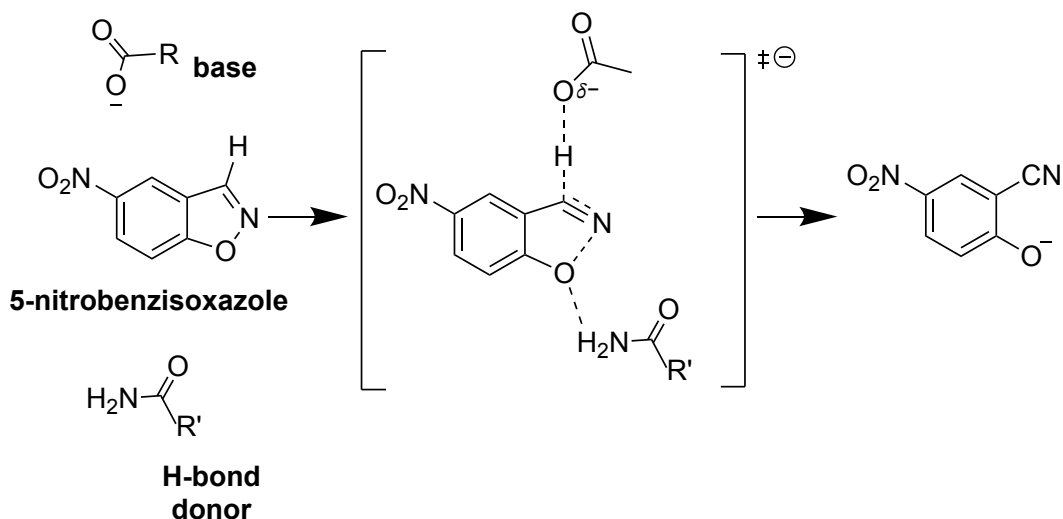


Figure 3.3. Kemp elimination of 5-nitrobenzisoxazole (5-NBZ). Kemp elimination reaction proceeds through a single transition state, involving the abstraction of hydrogen from the carbon of the substrate by a catalytic base. Additionally, a hydrogen bond donor stabilizes the negative charge that forms on the isoxazolic oxygen.

Efficient catalysis of the Kemp elimination reaction requires a hydrophobic environment in the active site, which could promote binding of the hydrophobic substrate, 5-nitrobenzisoxazole and increase the pK_a of the catalytic base through desolvation.² However, artificial TIM barrels do not naturally provide a deep hydrophobic pocket, and their β -barrel structure is overly exposed to solvent. Therefore, we postulated that introducing a lid structure over the barrel could shield the catalytic pocket from solvent, therefore lowering the active site dielectric constant. This, in turn, could enhance transition state stabilization by creating a more hydrophobic environment and increasing the basicity of the catalytic base.

3.5.1. Theozyme placement

Our design strategy focused on placing the catalytic Asp within the barrel to position the transition state with the isoxazolic oxygen pointing outward. This arrangement allows us to overlay the transition state along with the hydrogen-bond donor residue onto the transition state pose

generated from the catalytic base placement, positioning the H-bond donor residue above the barrel. This configuration enables the diffusion of the artificial TIM barrel's minimal loops to hold the H-bond donor in place and form a lid over the catalytic pocket. Sixteen positions within the barrel were mutated to Asp, resulting in four poses that successfully positioned the isoxazolic oxygen of the transition state pointing outward (Figure 3.4A–C, Supplementary Table 3.1). In separate calculations, we generated hydrogen bond donor poses that satisfied the geometric constraints of the theozyme. For this, we performed theozyme placement calculations by mutating the second residue of a tripeptide (residues 67–69 PDB ID: 7JLS) to either Asn or Gln (Figure 3.4D). Next, we superimposed the transition state poses from the hydrogen bond donor calculations onto those obtained from the catalytic base placement (Figure 3.4E). This process created the complete theozyme within the barrel, positioning the hydrogen bond donor floating above the opening of the barrel. This orientation allows the hydrogen bond donor to act as an anchor for generating a lid segment over the active site.

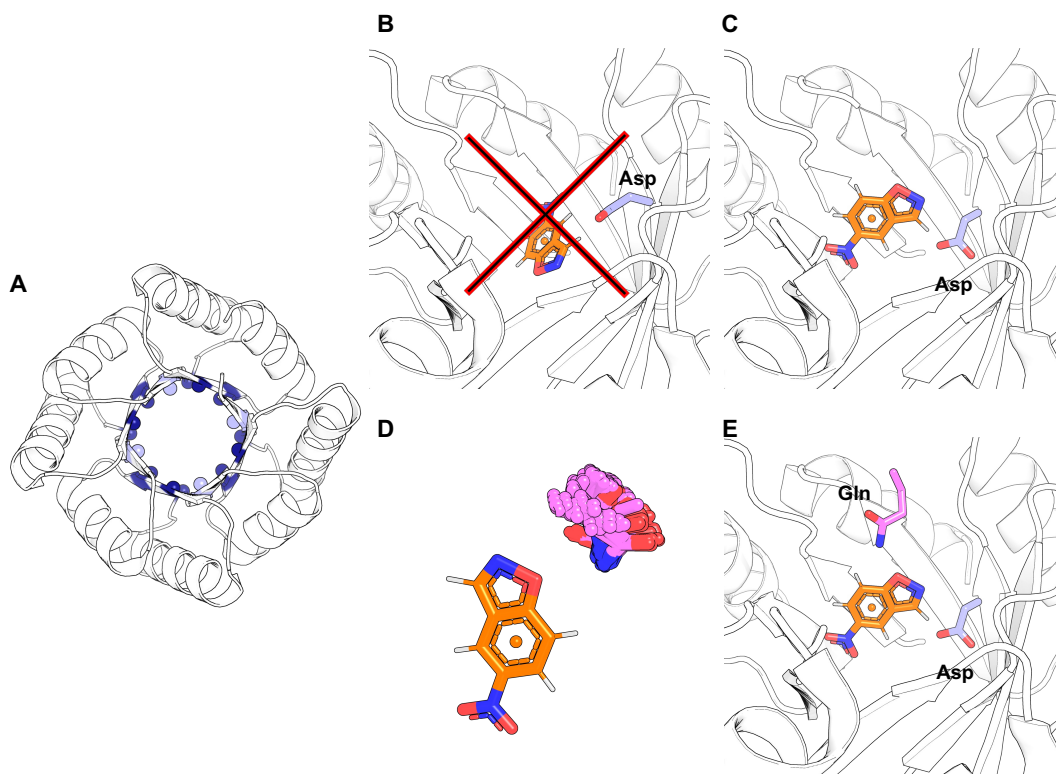


Figure 3.4. Theozyme placement. (A) Theozyme placement based on catalytic base was performed at each of the positions shown as blue spheres. Positions that resulted in transition state poses with isoxazolic oxygen pointing outward are highlighted in light blue. (B) A representative pose where the isoxazolic oxygen is oriented towards inside the barrel, which was subsequently eliminated. (C) A representative pose where the transition state is correctly positioned with the isoxazolic oxygen pointing outwards. (D) An ensemble of hydrogen bond donor poses (with Gln shown here) generated by theozyme placement calculations based on hydrogen bond constraints. (E) For successful transition state poses, the hydrogen bond donor (Asn or Gln) is placed in the optimal orientation for catalysis, positioned above the barrel.

3.5.2. Lid generation

After placing the theozyme, which includes the catalytic base inside the barrel, the transition state with the isoxazolic oxygen pointing outwards, and the hydrogen bond donor positioned floating above the barrel, we proceeded to the next step: diffusing loops near the hydrogen bond donor to form a lid over the active site and hold the hydrogen bond donor in place. Initially, we removed existing loop residues from the scaffold to create space for the new structural segment. We then used Rfdiffusion to generate a new structural segment made up of 10–30

residues, starting from the end of a β -strand and extending to the hydrogen bond donor, ultimately connecting to the start of the corresponding α -helix (Figure 3.5A). To create a diverse set of structural segments, we performed 100 diffusion trajectories for each input structure, resulting in a wide array of potential lid configurations. (Supplementary Figure 3.1). We then applied several filtering criteria to ensure the structural segments were compatible with our design goals. Structures were discarded if they (1) exhibited steric clash with the transition state, (2) displaced the Ca of the hydrogen bond donor by more than 8 Å away from the isoxazolic oxygen of the transition state, or (3) lacked well-defined secondary structure elements based on backbone geometry and hydrogen bonding patterns (Figure 3.5B). We hypothesized that long loops with poorly defined secondary structures would be challenging to design accurately. This might lead to excessive flexibility and unexpected conformational changes, potentially causing the hydrogen bond donor to deviate from its intended positions.

This filtering process eliminated approximately 95% of the generated structures. The remaining structures contained diverse secondary structure elements, including α -helices, β -turns and helix-turn-helix motifs. Since RFdiffusion generates only the backbone structure without the corresponding amino acid sequence, we used ProteinMPNN to design sequences for the filtered structural segments. Four sequences were generated for each remaining RFdiffusion output. Next, we validated the folding of the designed sequences using AlphaFold2 to predict their three-dimensional structures. We filtered AlphaFold2 predictions based on two criteria: (1) the backbone root-mean-square deviation (RMSD) between the AlphaFold2 model and the original RFdiffusion structure must be less than 3 Å, and (2) the average predicted local distance difference test (pLDDT) score for the five residues with the lowest pLDDT had to be greater than 45 (Figure

3.5C). This criterion guarantees that even the least confident regions of the predicted structure were modelled with reasonable accuracy.

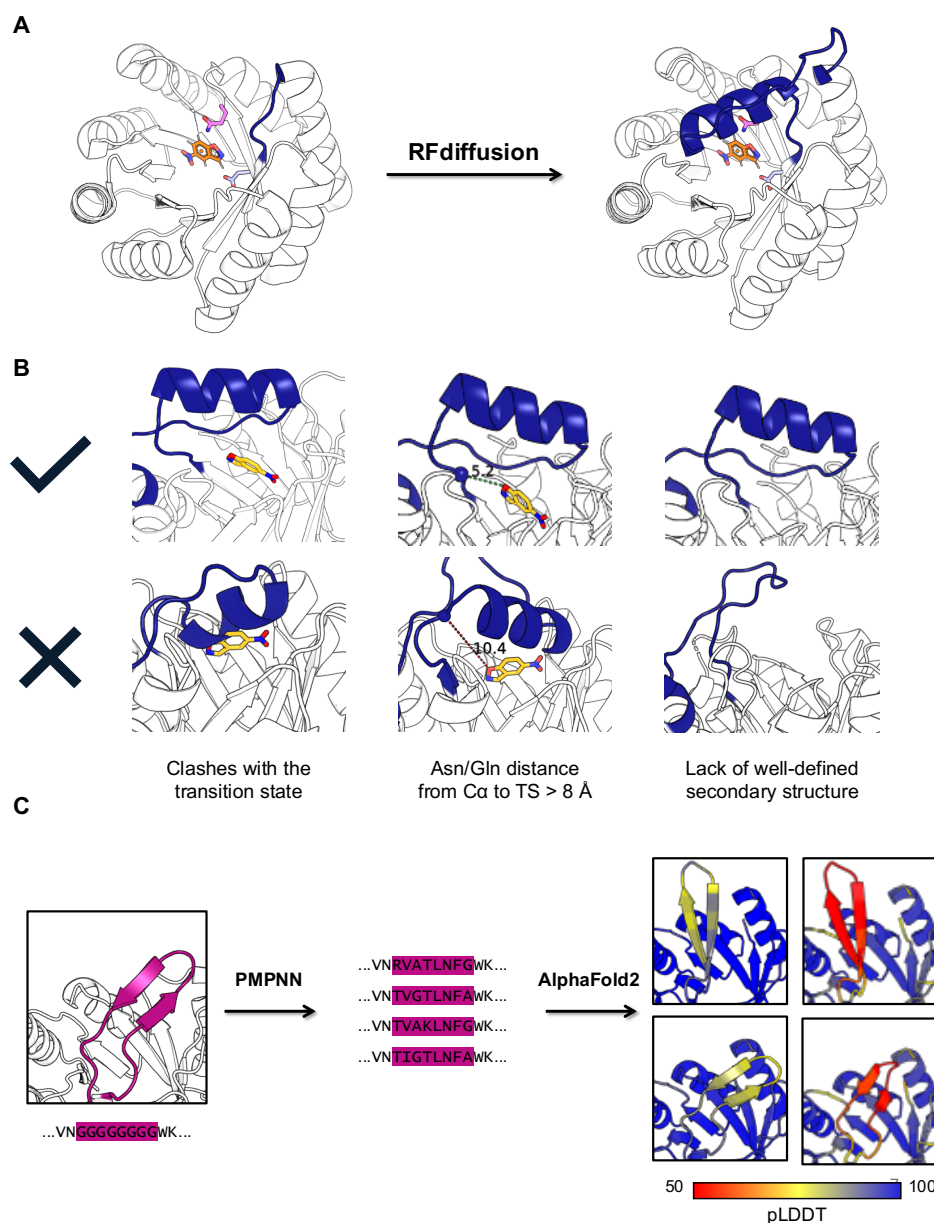


Figure 3.5. Lid generation. (A) Following theozyme placement, RFdiffusion was used to generate a structural motif from one of the β/α loops on the TIM barrel, designed to function as a lid over the active site. Starting from the end of a β -strand, RFdiffusion generated a structure to the hydrogen bond donor and then extended this segment to the corresponding α -helix. The generated structural motif is in blue, with the rest of the TIM barrel in white. (B) Generated structures were filtered based on several criteria. Structures were excluded if 1) the generated lid fragment clashed with the transition state (TS), 2) the C α of the H-bond donor was moved more than 8 Å away from the isoxazolic oxygen of the TS, or 3) they lacked a well-defined secondary structure. (C) ProteinMPNN (PMPNN) was used to design four sequences for each filtered structure. These sequences were validated by AlphaFold2 by measuring the RMSD between the predicted structure and original RFdiffusion structure, as well as the pLDDT score, AlphaFold's confidence metric. Dark pink indicates the diffused structure, while the structural models on the right are colour-coded by pLDDT confidence scores, with red and blue indicating lowest and highest pLDDT scores, respectively.

3.5.3. Active site repacking and filtering

Since AlphaFold2 models only predict the protein structure without including ligands, we needed to perform theozyme placement again for each AlphaFold2 model. During theozyme placement, residues near the transition state in the β -barrel and those on the lid structure oriented toward the barrel were mutated to glycine to create space for the transition state and catalytic residues. To account for backbone movements in the AlphaFold2 predictions, we allowed both Asn and Gln as potential hydrogen bond donors. Next, we performed active site repacking by designing the residues previously mutated to glycine. To create a hydrophobic pocket conducive to catalysis, we allowed hydrophobic and aromatic residues while also retaining wildtype lid residues designed by ProteinMPNN. We optimized the identity and rotameric configuration of residues in the β -barrel and lid structure to maximize favorable interactions and further stabilize the transition state. Following repacking, we applied a set of filtering parameters intended to reproduce mechanisms associated with enhanced catalysis.²³ First, we measured dihedral angles defining the hydrogen bond between the catalytic Asp and the transition state, as well as those between the hydrogen bond donor and the transition state. We retained only sequences with optimal catalytic dihedrals (Supplementary Table 3.6). Second, we measured the solvent-accessible surface area (SASA) of the transition state to assess the packing and shape-complementarity of the transition state in the active site and only kept sequences where $\geq 95\%$ of the transition state surface is buried. Third, we analyzed the preorganization of active site residues by optimizing their rotamers in the absence of the transition state. We then discarded sequences with more than three residues adopting different rotameric configurations in the presence and absence of the transition state, which we consider not preorganized for catalysis (Figure 3.6).

Fourth, we measured tunnels to the active site, eliminating sequences that lacked a tunnel with a minimum bottleneck radius of 0.9 Å. We also excluded sequences with more than 5 methionines in the active site to prevent excessive conformational flexibility in the active site. After filtering, the original set of 3472 designs based on TIM barrel F15C was reduced down to 16 designs. We performed similar calculations on CP1 and CP2 TIM barrels, resulting in the selection of 4 designs from these templates for experimental analysis. The 20 designs tested experimentally are based on four different lid structures—three from F15C and one from CP1 (Figure 3.7).

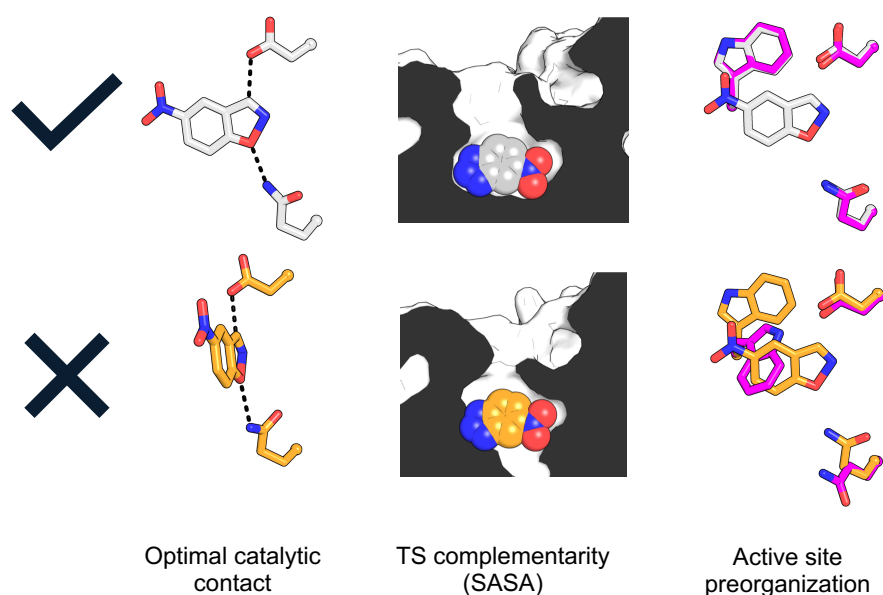


Figure 3.6. Filtering of sequences. Designed models were filtered using structural metrics to reproduce mechanisms known to contribute to efficient catalysis. Accepted designs met the following criteria: (1) optimal catalytic contact geometries between the transition state (TS) and catalytic residues as, defined by dihedral angles; (2) active sites with high structural complementarity to the TS, measured by TS solvent-accessible surface area (SASA); and (3) highly preorganized active sites, where most designed residues are predicted to adopt the same rotameric configuration in the presence (white/orange) and absence (magenta) of TS. Figure adapted with permission from Rakotoharisoa *et al.*²³ © 2024 American Chemical Society

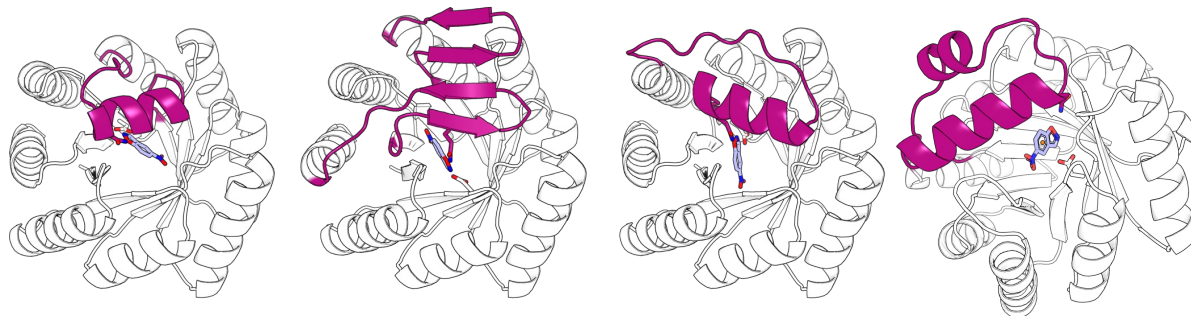


Figure 3.7. Structures of the experimentally characterized designs. The lid fragment generated by RFdiffusion is shown in pink. The first three structures on the left are based on TIM barrel F15C, while the fourth one is based on TIM barrel CP1.

3.6. Results

Next, we sought to experimentally characterize the designed Kemp eliminases. Genes encoding the sequences were expressed in *E. coli*, and the resulting proteins were purified by nickel affinity chromatography. All 20 designs expressed solubly. The original template, TIM barrel F15C exhibited α -helical circular dichroism (CD) spectrum characteristic of its respective secondary structure, with prominent negative peaks at 208 nm and 222 nm. The melting temperature (T_m) of F15C, was determined to be 88 °C, indicating high thermal stability (Figure 3.8). Five representative variants, encompassing all four designed lid structures, were analysed by CD. Design 1 and Design 13 showed CD spectra lacking distinct α -helical peaks, with no detectable melting temperature, suggesting partial misfolding or unfolding. The lid structure in these designs had an average pLDDT score of 60 for Design 1 and 87 for Design 13, as predicted by AlphaFold2. In contrast, Design 8 and Design 16, with lid pLDDT scores of 90 and 73, displayed CD spectra similar to the template TIM barrel, indicating proper folding. Therefore, there appeared to be no clear correlation between AlphaFold2 confidence scores and experimental folding behaviour. Additionally, the melting temperatures of Design 8 and Design 16 were lowered by 22 °C and 21

°C respectively, compared to the original template (Figure 3.8). Only one design, KempTIM 4 based on the TIM barrel CP1, displayed Kemp elimination activity, with a catalytic efficiency of $10 \text{ M}^{-1}\text{s}^{-1}$, comparable to previously designed Kemp eliminases.¹ Mutation of the catalytic aspartate to alanine abolished this activity (Figure 3.9).

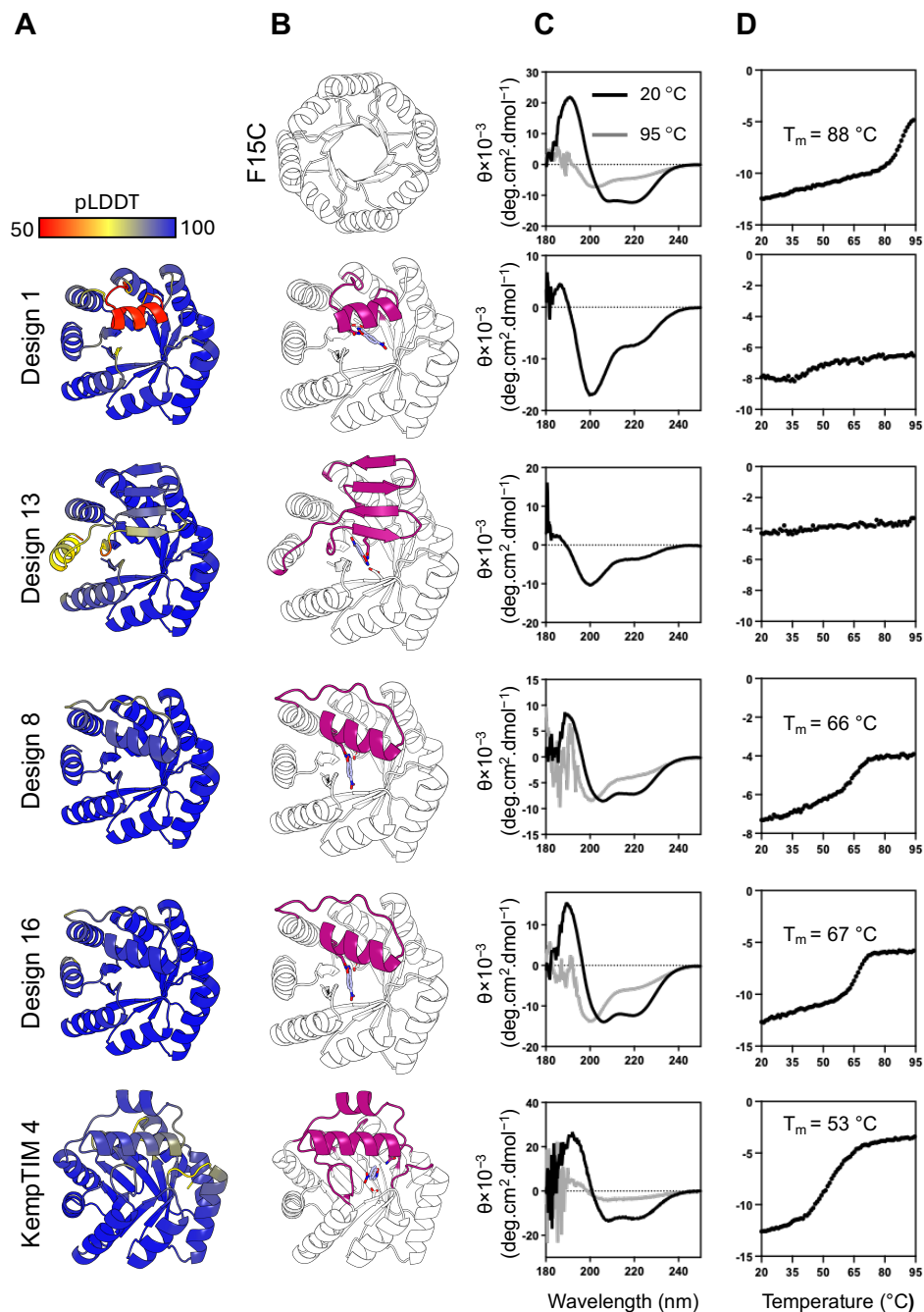


Figure 3.8. Circular dichroism (CD) and thermal melting results. (A) Structures of TIM barrel F15C, Designs 1, 13, 8, 16, and KempTIM 4, colored by pLDDT scores (50-100, with higher confidence in blue and lower in red). (B) Structures of the same TIM barrels with the designed lid fragment highlighted in pink. (C) CD spectra at 20 °C (Black) and 95 °C (Gray) for the original TIM barrel F15C, Designs 1, 13, 8, 16, and KempTIM 4. (D) Thermal unfolding curves, showing the melting temperatures (T_m) for TIM barrel F15C (~88°C), Design 8 (~66°C), Design 16 (~67°C), and KempTIM 4 (~53°C). No melting temperature was detected for Design 1 and Design 13.

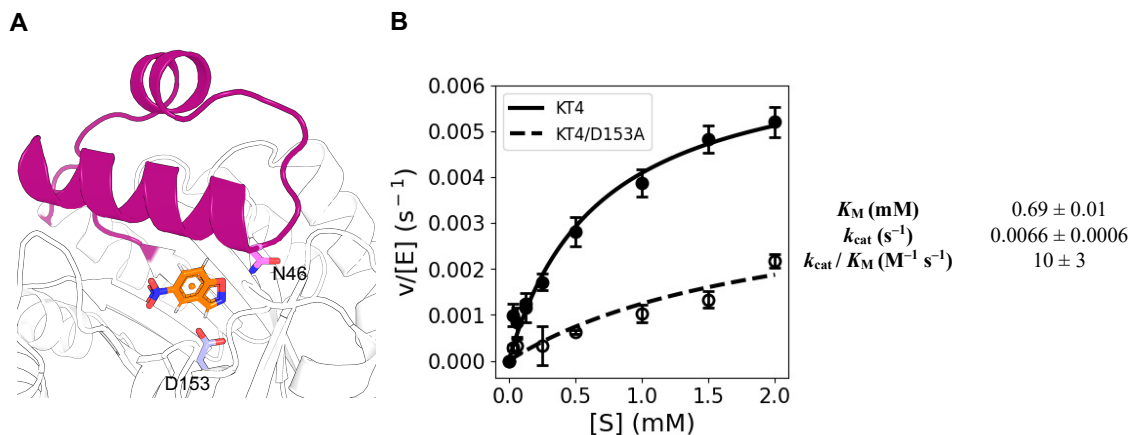


Figure 3.9. Kinetic characterization of KempTIM 4. (A) Design model of KempTIM 4, highlighting the catalytic Asp153 and hydrogen bond donor Asn46. The lid structure is shown in pink. (B) Michaelis-Menten graph of KempTIM 4 (KT4) and its mutant D153A. The kinetic parameters for KempTIM 4 are shown, with a k_{cat}/K_M value of $10 \text{ M}^{-1}\text{s}^{-1}$.

Next, we performed 1 μs molecular dynamics (MD) simulations of transition state analogue (TSA) bound and unbound design models to assess their conformational dynamics. To evaluate the flexibility of the design models, we calculated per-residue mean square fluctuations (RMSFs) for the folded but inactive designs, Design 8 and Design 16, and the active Kemp eliminase, KempTIM 4. Interestingly, in the two folded but inactive designs, Design 8 and Design 16, the lid structure exhibited low flexibility in the α -helical region, with an average RMSF value of $0.87 \text{ \AA}$ and $0.97 \text{ \AA}$, respectively. In contrast, the active KempTIM 4 displayed greater flexibility in the lid region, with an average RMSF of $1.56 \text{ \AA}$. Furthermore, loop 106–113, located near the lid and above the active site, showed significant flexibility in KempTIM 4, with an average RMSF of $3.6 \text{ \AA}$. We hypothesized that the rigidity of the lid region in the inactive designs could make the active site less accessible, impeding substrate entry and product release. To investigate this, we analyzed tunnels to the active site in the MD trajectories. In Design 8 and Design 16, 23% and 33% of structures, respectively, contain a tunnel, both with a median bottleneck radius of $1 \text{ \AA}$. In contrast, 95% of KempTIM 4 structures contain a tunnel to the active site with, a median bottleneck

radius of 2.1 Å (Figure 3.10). Therefore, the inactivity of Design 8 and Design 16 could stem from lacking tunnels to the active site, which could impede substrate entry.

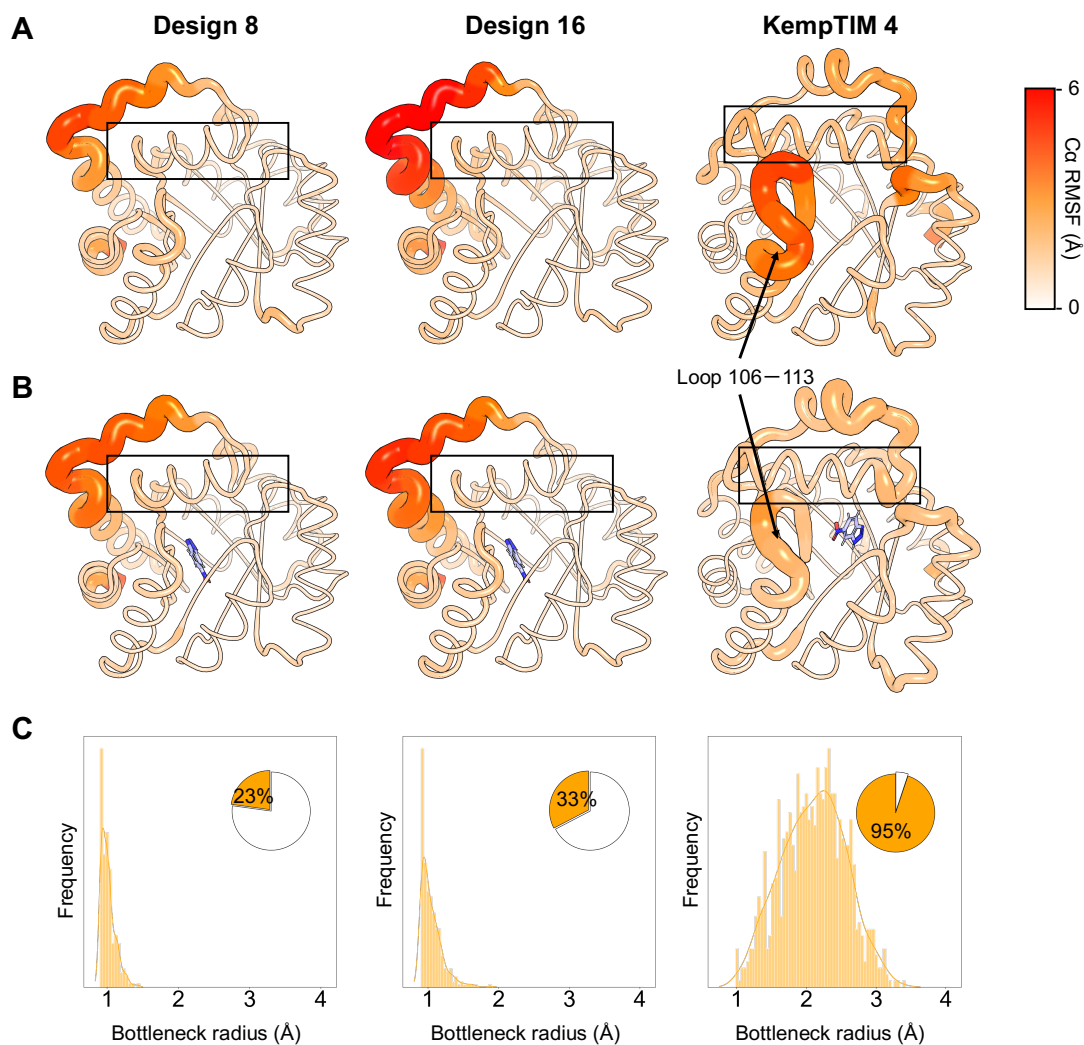


Figure 3.10. Conformational flexibility and active site accessibility of designed Kemp eliminases. Per-residue root mean square fluctuations (RMSF) mapped onto the Ca backbone of (A) unbound and (B) transition state analogue (TSA) bound Design 8, Design 16 and KempTIM 4. KempTIM 4 shows higher flexibility in both the lid region (highlighted by a box) and loop 106–113. (C) Distribution of tunnel bottleneck radii in Design 8, Design 16 and KempTIM 4. The percentage of structures containing a tunnel is shown in the pie charts.

3.7. Discussion

In this study, we aimed to transform a *de novo*, functionless TIM barrel into a Kemp eliminase by customizing its structure to introduce an additional lid fragment. Contrary to previous enzyme design methods, where natural scaffolds with native activity were used as the starting template, and active site structures were designed using physics-based methods,¹⁻⁴ our approach incorporates machine-learning techniques, such as RFdiffusion²⁴ for generating new structural segments, ProteinMPNN²⁵ for sequence design, and AlphaFold2²⁶ for structural validation. This approach allowed us to explore the potential of machine-learning methods for enzyme design. We have successfully demonstrated that a *de novo* TIM barrel, with an idealized structure consisting of eight α -helices and eight β -strands connected by short, minimal loops and devoid of any function, can be modified to catalyze a chemical reaction. The KempTIM 4 variant exhibited catalytic activity with efficiency comparable to previously designed Kemp eliminases, such as KE07.¹ This highlights the potential of scaffold customization to introduce novel functionality into otherwise inert proteins.

However, the low success rate—both in terms of the low catalytic efficiency of the active design and the fact that one out of 20 experimentally tested designs displayed Kemp elimination activity—reflects the inherent challenges in *de novo* enzyme design. Although a low success rate is common in enzyme design studies²⁷, this outcome underscores the approximations made during the design process. Methods such as RFdiffusion, ProteinMPNN, and AlphaFold2 introduce layers of uncertainty when predicting both the overall structure and the precise positioning of catalytic residues. These uncertainties likely contributed to misfolding observed in some designs, as evidenced by CD spectra of Design 1 and Design 13, which lacked clear α -helical peaks and

displayed no detectable melting temperature. Highly efficient catalysis requires sub-angstrom level of precision, which machine-learning methods, at least in their current state, fail to achieve.

Furthermore, MD simulations revealed that the active KempTIM 4 design exhibited greater conformational flexibility in both the designed lid region and loop 106–113 near the active site compared to the inactive designs. We hypothesize that this flexibility is critical for facilitating substrate entry and product release. In contrast, the lid structures in the inactive designs, Design 8 and Design 16, were more rigid and had lower RMSF values. This could have limited the accessibility to the active sites and contributed to their lack of activity. This theory is further supported by the tunnel analysis of MD trajectories. In the active KempTIM 4 design, 95% of structures contained a tunnel with a median bottleneck radius of 2.1 Å, similar to the most active Kemp eliminase, HG3.17.²⁸ In comparison, only 23% and 33% in Designs 8 and 16, respectively, had a tunnel, both with a median bottleneck radius of 1 Å. The overall rate of the reaction is determined by the rate-limiting step, which is not always the chemical transformation that is modeled with the theozyme and could be substrate entry,²⁹ product release,³⁰ or conformation exchange between sub-states.³¹ Therefore, limited flexibility in the lid fragment may restrict ligand binding and release, potentially leading to inactivity or significantly reduced activity. This highlights the need to improve current enzyme design methods by considering the complete catalytic cycle and not just the chemical step.

In this study, we chose to eliminate loop structural segments that lacked defined secondary structures, assuming that it would be challenging to accurately design loops without well-defined secondary structural elements and ensure that the hydrogen bond donor forms the necessary contacts. In retrospect, this decision may have limited the potential functionality of our designs.

Loops and their dynamic behaviour in TIM barrels have been linked to substrate recognition, binding, and specificity,^{32,33} as well as in product release^{33–35} and ligand-gated conformational changes.³⁶ Moreover, interactions between loops have been shown to provide the proper chemical environment for the reaction to occur.³⁷ By excluding flexible loops, we may have overlooked their potential in facilitating the reaction. A second limitation in our study is the placement of the hydrogen bond donor on the lid fragment. This fragment must alternate between an open conformation (allowing substrate entry) and a closed conformation (sequestering the active site from the solvent and extruding water). However, if the lid region is indeed sampling both states, this compromises the preorganization of the hydrogen bond donor, which is a key factor in efficient catalysis.³⁸

Overall, the results reported here contain several insights for *de novo* enzyme design. First, incorporating dynamics earlier in the design process could help address challenges of rigidity in inactive designs and enhance access to the active site. Second, machine learning protein design methods, while powerful, lack the preciseness that is necessary for designing highly efficient enzymes and should be used with caution. Finally, creating efficient enzymes likely requires a holistic approach where every possible state that the enzyme samples along its reaction coordinate is included in the design process.²⁸ In the future, we aim to customize *de novo* TIM barrels to catalyze other reactions, further exploring the utility of *de novo* TIM barrels as universal scaffolds for enzyme design.

3.8. Methods

3.8.1. Computational enzyme design

All calculations were performed using the Triad protein design software (Protabit, Pasadena, CA, USA). Rotamer optimization was performed using a Monte Carlo with simulated annealing search algorithm. The *de novo* TIM barrels F15C, CP1, and CP2 as well as individual templates from AlphaFold2 were prepared for computational design using the addH.py application within Triad, which optimized and renamed hydrogen atoms, assigned protonation state to His residues, flipped His/Asn/Gln side chains if necessary, and rotated hydroxyl, sulfhydryl, methyl, and amino groups in Ser/Thr, Cys, Met, and Lys, respectively. A backbone-independent conformer library described previously³⁹ was used for ligand placement, and a backbone-independent rotamer library (bbind02.May.e2)⁴⁰ with expansions of ± 1 standard deviation around $\chi 1$ and $\chi 2$ was used for active-site repacking. The Kemp elimination transition state was built using parameters described by Privett and co-workers.² For both theozyme placement and active-site repacking, a library of TS poses is generated by targeted ligand placement using the contacts specified in Supplementary Table 3.2. The TS pose energies were calculated using a modified Phoenix energy function⁴¹, which contains a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field⁴² with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal mol⁻¹ and an equilibrium donor–acceptor distance of 2.8 Å,⁴³ and an electrostatic energy term modeled using Coulomb’s law with a distance-dependent dielectric of 10. During the energy calculation step, energies for interactions between side-chains and the transition state were biased to favor interactions that satisfy contact geometries (Supplementary Table 3.4) by applying an energy benefit of –100 kcal mol⁻¹.

3.8.2. Lid generation

Motif scaffolding with RFdiffusion²⁴ was performed using the base model without applying auxiliary potentials. ProteinMPNN²⁵ was employed to design the entire structure generated by RFdiffusion, with the exception of the catalytic H-bond donor. Cys and Met were excluded from the sequence due to their side-chain flexibility, although no other amino acid biases were used. Four sequences were generated per structure, with a sampling temperature of 0.1. AlphaFold2²⁶ predictions were made using the monomer model, with a maximum template data of 2020-05-14 and using the full genetic database.

3.8.3. Active-site repacking

Following lid generation and theozyme placement, active-site repacking calculations were performed on the theozyme-containing structures predicted by AlphaFold2. In these calculations, the transition state was translated ± 0.4 Å in each Cartesian coordinate in 0.2 Å increments and rotated 10° about all three axes (origin at transition state geometric centre) in 5° increments for a total combinatorial rotation/translation search space of 15,625 possible poses. The identities of catalytic residues were fixed and allowed to sample all rotamers of that amino acid type. Other active-site residues were allowed to sample rotamers of hydrophobic and aromatic amino acids. Active-site residues on the lid were additionally allowed to sample rotamers of the amino acid introduced by ProteinMPNN. All residues within 3 Å of any of the designed positions were free to move during repacking. Gly was excluded from designed positions unless Ala caused a clash with the TS or catalytic residues, and Cys was excluded entirely to prevent the formation of disulfide bonds. For each template, 1000 repacked sequences were generated and ranked by energy. Energies for interactions between side-chains and the TS were biased as done for theozyme

placement (Supplementary Table 3.4). Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function mentioned above.

3.8.4. Library design and filtering

The 1000 energy-ranked sequences for each template generated by active-site repacking were used for library design, which was performed with the CLEARSS algorithm.⁴⁴ Each library of ~200 sequences was determined based on the Boltzmann weighting of both the frequency and stability (energy) of individual mutations introduced during repacking. The cleanSequences.py application within Triad was used to find the lowest energy conformation of each library sequence on its respective structure, generating “cleaned” sequences. In the “cleaning” procedure, the rotameric configurations of all active-site residues except the catalytic residues are optimized following the above-described protocol. To compare energies of “cleaned” sequences, we calculated their energy difference with a sequence containing Gly at all designed positions in the absence of the TS. The “cleaned” sequences were filtered by criteria described in Supplementary Table 3.6.

3.8.5. Protein expression and purification

Codon-optimized (*E. coli*) and histidine-tagged (C-terminus) genes for Kemp eliminases cloned into the pET-29b(+) vector via NdeI and XhoI were obtained from Twist Bioscience (Supplementary Table 3.5). Kemp eliminases were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth (LB) supplemented with 50 $\mu\text{g mL}^{-1}$ kanamycin for the designs or 50 $\mu\text{g mL}^{-1}$ ampicillin for F15C. Cultures were grown at 37 °C while shaking (250 rpm) until an optical density (OD) at 600 nm of 0.6–0.8 was reached. Protein expression was initiated with 1 mM isopropyl- β -D-1-thiogalactopyranoside (IPTG) and was carried out at 16 °C for 16 h while

shaking (250 rpm). Cells were then harvested by centrifugation at 4 °C and resuspended in 8 mL lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer pH 8.0) supplemented with 1 mg mL⁻¹ lyophilized lysozyme and 1 U mL⁻¹ benzonase nuclease (Merck Millipore). Cells were homogenized using an Avestin EmulsiFlex-B15 cell disruptor. Proteins were purified by immobilized metal affinity chromatography using Ni-NTA agarose (Qiagen) pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Contaminants were washed away using 10 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and then 20 mM imidazole in the same buffer. Proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0), and the buffer was exchanged into 50 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride using a PD-10 desalting column (Cytiva). The concentration of purified samples was determined by measuring the absorbance at 280 nm and applying Beer-Lambert's law. Extinction coefficients were predicted using the ExPASy ProtParam tool (<https://web.expasy.org/protparam>). The identity and purity of each protein was confirmed by SDS-PAGE.

3.8.6. Steady-state kinetics

Kinetic assays were performed in triplicate at 27 °C in 50 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride. Reactions (200 µL final volume) were initiated by addition of varying concentrations (0.03–2 mM) of 5-nitrobenzisoaxazole (AAblocks) dissolved in methanol to the enzyme in wells of a 96-well plate (Greiner Bio-One). Product formation was monitored spectrophotometrically at 380 nm ($\epsilon = 15,800 \text{ M}^{-1} \text{ cm}^{-1}$) using a Biotek Synergy H1 multimode plate reader (Agilent). Path lengths for each well were calculated ratiometrically using the difference in absorbance of 50 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride and 10% methanol at 900 and 975 nm (27 °C). Linear

phases of the kinetic traces were used to measure initial reaction rates. Data were fitted to the Michaelis-Menten equation to calculate individual k_{cat} and K_{M} parameters.

3.8.7. Circular dichroism and thermal denaturation assays

Circular dichroism measurements were performed with a Jasco J-815 spectrometer using 650- μL aliquots of each enzyme in a 1-mm path-length quartz cuvette (Jasco) at a concentration of 10 μM in 20 mM sodium phosphate buffer (pH 7). For structural characterization of protein folds, circular dichroism spectra were acquired from 180 to 250 nm, sampled every 0.2 nm at a rate of 10 nm min^{-1} . Five scans were acquired and averaged for each sample. For thermal denaturation assays, samples were heated at a rate of 1 $^{\circ}\text{C}$ per minute, and ellipticity at 222 nm was measured every 1 $^{\circ}\text{C}$. T_{m} values were determined by fitting a 2-term sigmoid function with baseline correction using nonlinear least-squares regression.⁴⁵ Thermal unfolding measurements were performed as a single replicate.

3.8.8. Molecular dynamics

The Amber 2020 software (<http://ambermd.org/>) with the ff19SB protein force field⁴⁶, gaff2 ligand force field⁴⁷, and OPC water force field⁴⁸ was used for all MD simulations. Long-range ($> 10 \text{ \AA}$) electrostatics were modeled using the particle mesh Ewald method.⁴⁹ Parameters for the transition state analogue 6-nitrobenzotriazole (6-NT) were generated using the Antechamber package.⁵⁰ All MD trajectories were run in triplicate and with a time step of 2 fs. Models of each design generated from ‘cleaning’ described above were prepared for MD. MD trajectories were run for both the bound and unbound structures, so the TS was converted to 6-NT or removed, respectively. Hydrogen atoms were added using Reduce,⁵¹ and the prepared structure was solvated with OPC water in a truncated octahedral box with periodic boundary conditions

where the distance between the protein surface and the box edges was set to 10 Å. The addions2 algorithm in Amber was used to place counterions to neutralize the system. The system was first minimized through the method of steepest descent using a force constant of 1000 kcal mol⁻¹ Å⁻². The system was heated to a temperature of 300 K for 240 ps, followed by equilibration under an NPT ensemble at 300 K and 1 bar for 10 ns and subsequently equilibration under an NVT ensemble at 300 K for another 10 ns. Constant temperature and pressure were achieved using the Nose-Hoover thermostat⁵² and Berendsen barostat⁵³, respectively. After removing positional restraints on protein heavy atoms imposed for heating and equilibration, a 1 μs production run was performed under Parrinello–Rahman pressure coupling.⁵⁴ For analyzing bottleneck radii, 1000 snapshots separated by 1 ns each were extracted from the production trajectory. A minimum cutoff of 0.9 Å was used to identify tunnels in the unbound MD trajectories using CAVER 3.0.⁵⁵

3.9. Supplementary information

Supplementary Table 3.1. Positions in the TIM barrel F15C tested for Asp placement

Position	Amino acid in F15C
4^a	Ile
6	Asn
29	Ala
31	Ile
50	Ile
52	Asn
75	Ala
77	Ile
96	Ile
98	Aan
121	Ala
123	Ile
142	Ile
144	Asn
167	Ala
169	Ile

^a Positions in bold successfully generated transition state pose with isoxazolic oxygen pointing outward

Supplementary Table 3.2. Geometric definitions for generation of transition-state poses off the side chain of catalytic base

Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
Asp	Distance	OD1 or OD2	H3			1.0, 1.2, 1.5
	Angle	CG	OD1 or OD2	H3		112, 117, 122
	Angle	OD1 or OD2	H3	C3		159, 164, 169, 174, 179
	Torsion	CB	CG	OD1 or OD2	H3	0, 5, 10, 170, 175, 180
	Torsion	CG	OD1 or OD2	H3	C3	170, 175, 180, 185, 190
	Torsion	OD1 or OD2	H3	C3	N2	0, 5, 170, 175, 180

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residue.

^b Distance measurements given in Å, all others are in degrees.

Supplementary Table 3.3. Geometric definitions for generation of transition-state poses off the side chain of hydrogen bond donor

Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
Asn	Distance	2HD2 or 1HD2	O1			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	ND2	2HD2 or 1HD2	O1		145, 151, 157
	Angle	2HD2 or 1HD2	O1	N2		120, 130, 140
	Torsion	CG	ND2	2HD2 or 1HD2	O1	120, 130, 140
	Torsion	ND2	2HD2 or 1HD2	O1	N2	-121, -116, -111
	Torsion	2HD2 or 1HD2	O1	N2	C3	160, 180, 200
Gln	Distance	2HE2 or 1HE2	O1			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	NE2	2HE2 or 1HE2	O1		145, 151, 157
	Angle	2HE2 or 1HE2	O1	N2		120, 130, 140
	Torsion	CD	NE2	2HE2 or 1HE2	O1	120, 130, 140
	Torsion	NE2	2HE2 or 1HE2	O1	N2	-121, -116, -111
	Torsion	2HE2 or 1HE2	O1	N2	C3	160, 180, 200

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residue.

^b Distance measurements given in Å, all others are in degrees.

Supplementary Table 3.4. Geometric constraints used to define catalytic contacts during computational design

Contact	Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max ^b
Catalytic base	Asp	Distance	OD1 or OD2	H3			1.0	1.6
		Angle	CG	OD1 or OD2	H3		109	131
		Angle	OD1 or OD2	H3	C3		159	180
		Torsion	CB	CG	OD1 or OD2	H3	−21	21
H-bond donor	Gln	Distance	1HE2 or 2HE2	O1			1.2	2.3
		Angle	NE2	1HE2 or 2HE2	O1		145	157
		Angle	1HE2 or 2HE2	O1	N2		120	140
		Torsion	1HE2 or 2HE2	O1	N2	C3	160	200
	Asn	Distance	1HD2 or 2HD2	O1			1.2	2.3
		Angle	ND2	1HD2 or 2HD2	O1		145	157
		Angle	1HD2 or 2HD2	O1	N2		120	140
		Torsion	1HD2 or 2HD2	O1	N2	C3	160	200

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residue.

^b Distance measurements given in Å, all others are in degrees.

Supplementary Table 3.5. Amino acid sequences of Kemp eliminase designs

Enzyme	Sequence
F15C	MDILIVNPDDFEKGVVEVKELKRHGAKIIAYISKSAEELKKA EKAGADILVNPDDFEKGVVEVKELKRHGAKIIA YISKSAEELKKA EKAGADILVNPDDFEKGVVEVKELKRHGAKIIAYISKSAEELKKA EKAGADILVNPDDFEKGV VEVKELKRHGAKIIAYISKSAEELKKA EKAGHHHHHH
Design 1	MDILIVGPDDFEKGVVEVKELKRHGAKIIAYASKSAEELKKA EKAGADILAVAPDDFEKGVVEVKELKRHGAKII AYASKSAEELKKA EKAGADILMVAPDDFEKGVVEVKELKRHGAKIIAYVGGSVKKVIEANQKGT SRGATVEELK KA EKAGADILDVVPDDFEKGVVEVKELKRHGAKIIAYASKSAEELKKA EKAGWHHHHHH
Design 2	MDILGVAPDDFEKGVVEVKELKRHGAKIIIGYMSKSAEELKKA EKAGADILLVAPDDFEKGVVEVKELKRHGAKI IAYMSKSAEELKKA EKAGADILMVFPDDFEKGVVEVKELKRHGAKIIAYLSKSAEELKKA EKAGADILDVANAE QEKIASKFLGRKTKVKIEENDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGWHHHHHH
Design 3	MDILMVMPDDFEKGVVEVKELKRHGAKIIAYASKSAEELKKA EKAGADILMVMPDDFEKGVVEVKELKRHGA KIIAYASKSAEELKKA EKAGADILDVAPDDFEKGVVEVKELKRHGAKIIAYAAAAGRIETPGETHRSSEEVVEELKK AEKAGADILAVLPDDFEKGVVEVKELKRHGAKIIAYYAGGQTFEIGGSKNKAAPLGVEELKKA EKAGWHHHHHH H
Design 4	MDILAVAPDDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVVEVKELKRHGAKII GYASKSAEELKKA EKAGADILFVMPDDFEKGVVEVKELKRHGAKIIAYMAANIETPGKTIKRSSEEVVEELKKA EKAGADILAVAPDDFEKGVVEVKELKRHGAKIIAYGAGGKSQTLGGSTMKLAKLTVEELKKA EKAGWHHHHHH H
Design 5	MDILAVAPDDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVVEVKELKRHGAKII GYMSKSAEELKKA EKAGADILFVAPDDFEKGVVEVKELKRHGAKIIAYMAANIETPGKTIKRSSEEVVEELKKA EKAGADILAVAPDDFEKGVVEVKELKRHGAKIIAYGAGGKSQTLGGSTMKLAKLTVEELKKA EKAGWHHHHHH H
Design 7	MDILAVAPDDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVVEVKELKRHGAKII GYMSKSAEELKKA EKAGADILYVAPDDFEKGVVEVKELKRHGAKIIAYLASGVIETPGEKIVRSSEEVVEELKKA EKAGADILAVAPDDFEKGVVEVKELKRHGAKIIAYGAGGKTQTVDGATFKLAPLGVEELKKA EKAGWHHHHHH H
Design 8	MDILGVAPDDFEKGVVEVKELKRHGAKIIIGYMSKSAEELKKA EKAGADILLVAPDDFEKGVVEVKELKRHGAKI IAYMSKSAEELKKA EKAGADILMVFPDDFEKGVVEVKELKRHGAKIIAYLSKSAEELKKA EKAGADILDVANAE QEKIASKFMGRKTKVKIEENDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGWHHHHHH
Design 9	MDILAVAPDDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVVEVKELKRHGAKII GYMSKSAEELKKA EKAGADILFVMPDDFEKGVVEVKELKRHGAKIIAYLAANIETPGKTIKRSSEEVVEELKKA EKAGADILAVAPDDFEKGVVEVKELKRHGAKIIAYGAGGKSQTLGGSTMKLAKLTVEELKKA EKAGWHHHHHH H
Design 10	MDILAVMPDDFEKGVVEVKELKRHGAKIIAYASKSAEELKKA EKAGADILMVMPDDFEKGVVEVKELKRHGAK IIAYASKSAEELKKA EKAGADILDVAPDDFEKGVVEVKELKRHGAKIIAYAAAAGRIETPGETHRSSEEVVEELKKA EKAGADILAVLPDDFEKGVVEVKELKRHGAKIIAYYAGGQTFEIGGSKNKAAPLGVEELKKA EKAGWHHHHHH H
Design 11	MDILGVAPDDFEKGVVEVKELKRHGAKIIIGYMSKSAEELKKA EKAGADILMVAPDDFEKGVVEVKELKRHGAK IIAYMSKSAEELKKA EKAGADILAVFPDDFEKGVVEVKELKRHGAKIIAYLSKSAEELKKA EKAGADILDVANAE QEKIASKLMGRKTKVKIEENDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGWHHHHHH
Design 12	MDILGVAPDDFEKGVVEVKELKRHGAKIIIGYMSKSAEELKKA EKAGADILLVAPDDFEKGVVEVKELKRHGAKI IAYMSKSAEELKKA EKAGADILMVFPDDFEKGVVEVKELKRHGAKIIAYLSKSAEELKKA EKAGADILDVANAE QEKIASKLLGRKTKVKIEENDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGWHHHHHH
Design 13	MDILAVAPDDFEKGVVEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVVEVKELKRHGAKII GYMSKSAEELKKA EKAGADILMVAPDDFEKGVVEVKELKRHGAKIIAYMAANIETPGKTIKRSSEEVVEELKK AEKAGADILAVAPDDFEKGVVEVKELKRHGAKIIAYGAGGKSQTVGGSTMKLAKLTVEELKKA EKAGWHHHHH HH

Design 14	MDILAVAPDDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVEEVKELKRHGAKII GYMSKSAEELKKA EKAGADILFVAPDDFEKGVEEVKELKRHGAKIIAYLASGVIETPGEKIVRSSEEVVEELKKA EKAGADILAVAPDDFEKGVEEVKELKRHGAKIIAYGAGGKTQTVDGATHKLAPLGEELKKA EKAGWHHHHH H
Design 15	MDILMVAPDDFEKGVEEVKELKRHGAKIIAYASKSAEELKKA EKAGADILAVVPDDFEKGVEEVKELKRHGAKI IAYGSKSAEELKKA EKAGADILAVAPDDFEKGVEEVKELKRHGAKIIAYAAAGRIETPGETIIRSSEEVVEELKKA KAGADILDVVPDDFEKGVEEVKELKRHGAKIIAYMAGGQTQEMGGSKLKMAPLGEELKKA EKAGWHHHHH H
Design 16	MDILAVAPDDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKA EKAGADILDVVPDDFEKGVEEVKELKRHGAKII GYMSKSAEELKKA EKAGADILYVAPDDFEKGVEEVKELKRHGAKIIAYLASGVIETPGEKIVRSSEEVVEELKKA EKAGADILAVAPDDFEKGVEEVKELKRHGAKIIAYGAGGKTQTVDGATHKLAPLGEELKKA EKAGWHHHHH H
CP1	MDKDEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKEIAYRSDD WRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWK QVEILRRLGAKEIAYRSDDWRDLQEALKKGADILIVDATGLEHHHHHH
KempTIM 1	MEIAYASDDWRDLQEALKKGADILDVNVNPEEAWKQVEILRRLGAKRILYASLRWRDLQEALKKGADILAVPL VDK DEAWKQVEILRRLGAKEIAYGSDDWRDLQEALKKGADILGVASNTPRVRRLVLRNLETLYSREEAERIVEEKIKL NDPE EAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADILVAADESEESKDEAWKQVEILRRLEHHHHHH
KempTIM 2	MEIAYASDDWRDLQEALKKGADILDVNVNPEEAWKQ VEILRRLGAKRIAYASLRWRDLQEALKKGADILAVPLVD KDEAWKQVEILRRLGAKEIAYGSDDWRDLQEALKKGAD ILGVASNTPRVRRLVLRNLETLYSREEAERIVEEKIKLNDP EEAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADIL LVVADDESEESKDEAWKQVEILRRLEHHHHHH
KempTIM 3	MRIAYASLASRWRDLQEALKKGADILMVALARDAASVR ALAEGAAGNGDPRVAELLALLRPTDKDEAWKQVEILR RLGAKEIAYMSDDWRDLQEALKKGADILMVLASSALSPE EAWKQVEILRRLGAKRIAYGSGDWRDLQEALKKGADIL DVLGAKGKDEAWKQVEILRRLGAKEIAYASFDWRDLQE ALKKGADILMVFAVLDAERALKQVEILRRLEHHHHHH
KempTIM 4	MRIAYMSLASRWRDLQEALKKGADILMVALARDAASVR ALAEGAAGNGDPRVAELLALLRPTDKDEAWKQVEILR RLGAKEIAYMSDDWRDLQEALKKGADILMVLASSALSPE EAWKQVEILRRLGAKRIAYGSGDWRDLQEALKKGADIL DVLGAKGKDEAWKQVEILRRLGAKEIAYASFDWRDLQE ALKKGADILVVFAVLDAERALKQVEILRRLEHHHHHH

Supplementary Table 3.6. Criteria for design filtering: Designs were excluded if they failed to meet all the criteria listed below.

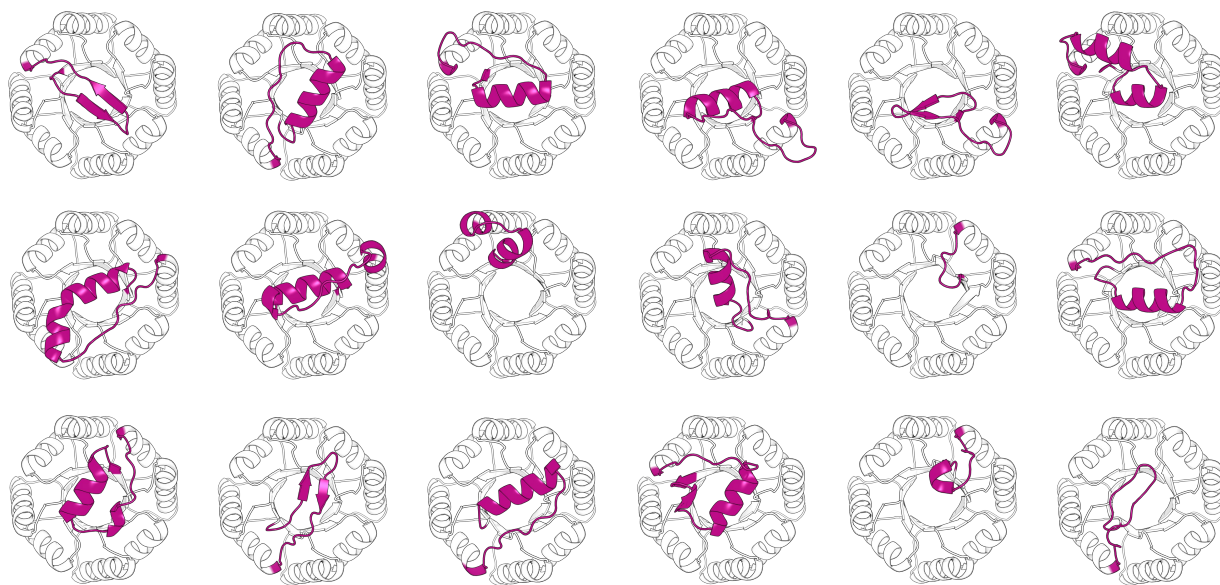
Criterion	Description
Optimal catalytic contacts	Neither hydrogen bond donor or base dihedral angle are between -130° and -50° or 50° and 130°
Energy	All energy values < 0
TS complementarity	Solvent-accessible surface area (SASA) between $80 \text{ \AA}^2$ and $120 \text{ \AA}^2$
Number of Gly and Met	Less than 4 Gly or 6 Met in the positions designed during repacking
Preorganized active site	Less than 3 residues not preorganized, having different rotamers in the presence and absence of the transition state
Tunnels to the active site	Contains a tunnel $> 0.9 \text{ \AA}$, as identified by CAVER

Supplementary Table 3.7. Design filtering – F15C template

Design	base dihedral	H-bond donor dihedral	Energy (kcal mol ⁻¹)	SASA (Å ²)	# Met	# Preorganized residues (on 27)	Bottleneck radius (Å)	Lid structure average pLDDT
Design 1	-24.7	38.8	-94.0	89.9	1	26	0.96	60.1
Design 2	-1.0	28.8	-115.5	86.9	3	26	0.96	90.7
Design 3	-26.5	-155.0	-121.7	94.4	4	27	1.12	75.5
Design 4	-28.2	160.0	-111.7	94.7	3	27	0.91	84.3
Design 5	-28.2	160.0	-115.1	88.6	3	28	0.91	85.5
Design 7	-17.8	162.1	-131.4	88.7	1	27	0.92	81.5
Design 8	-1.0	28.8	-111.8	86.9	4	26	0.94	90.3
Design 9	-28.2	160.0	-111.6	93.2	3	27	0.91	83.9
Design 10	-26.5	-155	-114.9	94.9	3	27	0.92	76.0
Design 11	-0.7	44.2	-108.9	99.6	4	25	1.08	92.8
Design 12	-13.3	29.6	-107.4	91.3	3	26	0.96	92.5
Design 13	-28.2	160.0	-111.4	110.1	4	26	0.91	87.4
Design 14	-17.8	162.1	-136.3	95.2	1	27	0.92	74.9
Design 15	-11.3	-147.1	-99.7	99.9	4	26	0.92	64.8
Design 16	-13.8	149.6	-116.9	84.6	1	27	0.95	73.0

Supplementary Table 3.8. Design filtering – CP1 and CP2 templates

Design	base dihedral	H-bond donor dihedral	Energy (kcal mol ⁻¹)	SASA (Å ²)	# Met	# Preorganized residues (on 16)	Bottleneck radius (Å)
KempTIM 1	-42.5	25.7	-39.4	174.4	1	11	0.92
KempTIM 2	-7.3	25.7	-36.3	188.5	1	10	0.92
KempTIM 3	27.6	-156.2	-10.4	107.5	4	12	1.53
KempTIM 4	27.6	156.2	-2.3	106.6	4	13	1.53



Supplementary Figure 3.1. Structural segments generated by RFdiffusion.

3.10. References

1. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
2. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).
3. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. *De Novo* Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
4. Bjelic, S., Nivón, L. G., Çelebi-Ölçüm, N., Kiss, G., Rosewall, C. F., Lovick, H. M., Ingalls, E. L., Gallaher, J. L., Seetharaman, J., Lew, S., Montelione, G. T., Hunt, J. F., Michael, F. E., Houk, K. N. & Baker, D. Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita–Baylis–Hillman Reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).
5. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E. & Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
6. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Bioinforma.* **46**, 105–109 (2002).
7. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
8. Kipnis, Y., Chaib, A. O., Vorobieva, A. A., Cai, G., Reggiano, G., Basanta, B., Kumar, E., Mittl, P. R. E., Hilvert, D. & Baker, D. Design and optimization of enzymatic activity in a *de novo* β -barrel scaffold. *Protein Sci.* **31**, e4405 (2022).

9. Yeh, A. H.-W., Norn, C., Kipnis, Y., Tischler, D., Pellock, S. J., Evans, D., Ma, P., Lee, G. R., Zhang, J. Z., Anishchenko, I., Coventry, B., Cao, L., Dauparas, J., Halabiya, S., DeWitt, M., Carter, L., Houk, K. N. & Baker, D. *De novo* design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
10. Lauko, A., Pellock, S. J., Anischanka, I., Sumida, K. H., Juergens, D., Ahern, W., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I. R., Jamieson, C., Kang, A., Brackenbrough, E., Bera, A. K., Sankaran, B., Houk, K. N. & Baker, D. Computational design of serine hydrolases. *Preprint at bioRxiv* (2024).
11. Nestl, B. M. & Hauer, B. Engineering of Flexible Loops in Enzymes. *ACS Catal.* **4**, 3201–3211 (2014).
12. Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by *de novo* protein design. *Nat. Biotechnol.* **42**, 203–215 (2024).
13. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).
14. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the ($\beta\alpha$)₈-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).
15. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
16. Romero-Romero, S., Kordes, S., Michel, F. & Höcker, B. Evolution, folding, and design of TIM barrels and related proteins. *Curr. Opin. Struct. Biol.* **68**, 94–104 (2021).
17. Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Höcker, B. & Baker, D. *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
18. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N. & Hilvert, D. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
19. Anand, N., Eguchi, R., Mathews, I. I., Perez, C. P., Derry, A., Altman, R. B. & Huang, P.-S. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).

20. Kordes, S., Romero-Romero, S., Lutz, L. & Höcker, B. A newly introduced salt bridge cluster improves structural and biophysical properties of *de novo* TIM barrels. *Protein Sci.* **31**, 513–527 (2022).
21. Casey, M. L., Kemp, D. S., Paul, K. G. & Cox, D. D. Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **38**, 2294–2301 (1973).
22. Kemp, D. S. & Casey, M. L. Physical organic chemistry of benzisoxazoles. II. Linearity of the Broensted free energy relation for the base-catalyzed decomposition of benzisoxazoles. *J. Am. Chem. Soc.* **95**, 6670–6680 (1973).
23. Rakotoharisoa, R. V., Seifinoferest, B., Zarifi, N., Miller, J. D. M., Rodriguez, J. M., Thompson, M. C. & Chica, R. A. Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J Am Chem Soc* **146**, 10001–10013 (2024).
24. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., ... Baker, D. *De novo* design of protein structure and function with RFdiffusion. *Nature*, **620**, 1089–1100 (2023).
25. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., De Haas, R J, Bethes, N, Leung, P J Y, Huddy, T F, Pellock, S, Tischer, D, Chan, F, Koepnick, B, Nguyen, H, Kang, A, Sankaran, B, Bera, A K, King, N P, & Baker, David. Robust deep learning based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
26. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
27. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Computational Enzyme Design: Successes, Challenges, and Future Directions. in *Modern Biocatalysis* (eds. Williams, G. & Hall, M.) 88–116 (Royal Society of Chemistry, Cambridge, 2018). doi:10.1039/9781788010450-00088.

28. Broom, A., Rakotoharisoa, R. V., Thompson, M. C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J. S. & Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
29. Wang, L. H., Tsai, A. L. & Hsu, P. Y. Substrate binding is the rate-limiting step in thromboxane synthase catalysis. *J. Biol. Chem.* **276**, 14737–14743 (2001).
30. Cleland, W. W. What limits the rate of an enzyme-catalyzed reaction. *Acc. Chem. Res.* **8**, 145–151 (1975).
31. Kerns, S. J., Agafonov, R. V., Cho, Y.-J., Pontiggia, F., Otten, R., Pachov, D. V., Kutter, S., Phung, L. A., Murphy, P. N., Thai, V., Alber, T., Hagan, M. F. & Kern, D. The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124–131 (2015).
32. Schlee, S., Dietrich, S., Kur  on, T., Delaney, P., Goodey, N. M. & Sterner, R. Kinetic Mechanism of Indole-3-glycerol Phosphate Synthase. *Biochemistry* **52**, 132–142 (2013).
33. Romero-Rivera, A., Corbella, M., Parracino, A., Patrick, W. M. & Kamerlin, S. C. L. Complex Loop Dynamics Underpin Activity, Specificity, and Evolvability in the ($\beta\alpha$)₈ Barrel Enzymes of Histidine and Tryptophan Biosynthesis. *JACS Au* **2**, 943–960 (2022).
34. Zhao, Y., She, N., Zhang, X., Wang, C. & Mo, Y. Product release mechanism and the complete enzyme catalysis cycle in yeast cytosine deaminase (yCD): A computational study. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1865**, 1020–1029 (2017).
35. Yao, L., Yan, H. & Cukier, R. I. A Molecular Dynamics Study of the Ligand Release Path in Yeast Cytosine Deaminase. *Biophys. J.* **92**, 2301–2310 (2007).
36. Richard, J. P., Zhai, X. & Malabanan, M. M. Reflections on the catalytic power of a TIM-barrel. *Bioorganic Chem.* **57**, 206–212 (2014).
37. Wang, Y., Berlow, R. B. & Loria, J. P. Role of Loop–Loop Interactions in Coordinating Motions and Enzymatic Function in Triosephosphate Isomerase. *Biochemistry* **48**, 4548–4556 (2009).

38. Villà, J. & Warshel, A. Energetics and Dynamics of Enzymatic Reactions. *J. Phys. Chem. B* **105**, 7887–7907 (2001).
39. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci.* **103**, 16710–16715 (2006).
40. Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
41. Tang, Y. T. & Marshall, G. R. PHOENIX: A Scoring Function for Affinity Prediction Derived Using High-Resolution Crystal Structures and Calorimetry Measurements. *J. Chem. Inf. Model.* **51**, 214–228 (2011).
42. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
43. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10172–10177 (1997).
44. Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc. Natl. Acad. Sci.* **107**, 19838–19843 (2010).
45. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel β -sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein science : a publication of the Protein Society*, **8**, 841–853 (1999).
46. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O. & Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).
47. He, X., Man, V. H., Yang, W., Lee, T.-S. & Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J. Chem. Phys.* **153**, 114502 (2020).

48. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
49. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
50. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).
51. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation¹. *J. Mol. Biol.* **285**, 1735–1747 (1999).
52. Evans, D. J. & Holian, B. L. The Nose–Hoover thermostat. *J. Chem. Phys.* **83**, 4069–4074 (1985).
53. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
54. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
55. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J. & Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLOS Comput. Biol.* **8**, e1002708 (2012).

CHAPTER 4. *DE NOVO* DESIGN OF MICHAELASES USING A MULTISTATE ENZYME DESIGN APPROACH

Niayesh Zarifi¹, Marc Garcia-Borràs² & Roberto A. Chica¹

¹ Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

² Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Girona, Spain

4.1. Foreword

In Chapter 2, we studied the role of active-site versus distal mutations on enzyme catalysis. We observed that active-site mutations alone are sufficient to form an ordered active site for efficient catalysis, while distal mutations further enhance activity by facilitating substrate binding and product release. In Chapter 3, we developed a workflow to customize the structure of a *de novo*, inert TIM barrel scaffold to introduce Kemp elimination activity. Using machine learning tools, we added a lid fragment to enclose the active site. However, the generated lid regions did not exhibit the flexibility required for rapid substrate binding and product release. The design that showed Kemp elimination activity catalyzes this reaction with modest efficiency. The enzymes discussed in Chapter 2 and designed in Chapter 3 catalyze the Kemp elimination reaction, a model proton abstraction reaction that involves a single transition state. However, industrially relevant reactions are more complex and often involve multiple reaction steps, transition states, and intermediates. Designing enzymes for such reactions requires more sophisticated approaches. In Chapter 4, we aim to apply a multistate enzyme design approach to design an enzyme to catalyze a carbon-carbon bond-forming reaction. Given the complexity of designing enzymes for a multistep reaction, we opted to use pre-existing TIM barrels as template scaffolds to avoid additional complications from designing lid structures.

4.2. Statement of contribution

Niayesh Zarifi performed the DFT calculations, computational enzyme design, experimental characterization of enzymes, and molecular dynamics simulations and analysis. Niayesh Zarifi, Roberto A. Chica, and Marc Garcia-Borràs conceived the project. Niayesh Zarifi wrote the chapter.

4.3. Abstract

Creating *de novo* enzymes to catalyze complex multistep reactions is a challenging endeavor with significant potential for industrial applications. Traditional single-state design approaches, which focus on stabilizing a single or composite transition state, often fall short for multistep reactions as efficient catalysis requires a low energy barrier across all reaction steps. Here, we applied a multistate enzyme design approach to develop an enzyme to catalyze the Michael addition reaction, a C—C bond-forming reaction. We performed density functional theory (DFT) calculations to characterize the reaction mechanism and generate theozymes for each step and then designed sequences capable of stabilizing two critical transition states. We experimentally characterized 12 designed enzymes. All variants catalyze the Michael addition reaction with lower K_M^{app} values than the starting scaffold, indicating improved substrate binding. However, these designs exhibited reduced apparent turnover numbers, likely due to disproportionate stabilization of the targeted steps relative to other steps within the catalytic cycle. Our study underscores the potential of multistate enzyme design in expanding the biocatalytic repertoire for complex chemical processes and highlights the importance of considering enzyme dynamics and the entire catalytic cycle in enzyme design.

4.4. Introduction

De novo design of enzymes to catalyze complex, multistep reactions useful for industrial biocatalysis remains a significant challenge. Previous computational enzyme design efforts often focused on stabilizing a single^{1,2} or a composite^{3,4} transition state, which is inadequate for designing highly efficient enzymes to catalyze reactions involving multiple steps. To ensure efficient catalysis, an enzyme must maintain low activation energy barriers throughout all reaction steps.⁵ Solely focusing on stabilizing the computationally predicted rate-determining step could inadvertently destabilize other steps in the reaction pathway and lead to poor catalytic performance. Furthermore, accurate identification of the rate-limiting step requires experimental characterization since it may differ from computational predictions.^{6,7} To address the shortcomings of single-state enzyme design, a more sophisticated approach is needed, one that considers the entire reaction pathway. Previously, multistate protein design has been employed to design sequences compatible with multiple chemical or conformational states.^{8–10} In this study, we expand this concept to enzyme design. We hypothesize that multistate enzyme design, which involves simultaneous stabilization of multiple transition states along the reaction pathway, is necessary to design efficient enzymes for useful, multistep reactions.

One such reaction is the Michael addition reaction, a conjugate addition of, also known as a ‘Michael donor,’ to an α,β -unsaturated carbonyl, the ‘Michael acceptor’. In the case of carbon-based nucleophiles, this results in the formation of a carbon-carbon (C–C) bond in the ‘Michael adduct.’^{11,12} Discovered by Arthur Michael in 1887,¹³ the Michael addition reaction has garnered attention in organic synthesis due to its ability to form C–C bonds, which is essential for synthesizing organic molecules. This reaction is widely used in the synthesis of polymers and

pharmaceuticals, including the anti-inflammatory drug ibuprofen.¹⁴ Despite its utility, naturally occurring enzymes that can catalyze the Michael addition reaction are rare. Those that do exist often exploit the catalytic promiscuity of enzymes, meaning they were not evolutionary optimized to catalyze this reaction but can perform it under certain conditions.^{15,16} Therefore, design of enzymes that can catalyze the Michael addition reaction with high efficiency could significantly expand the repertoire of biocatalysts available for synthetic chemistry, facilitating more sustainable and efficient chemical processes.

Here, we aimed to design an enzyme for the Michael addition reaction. We first performed density functional theory (DFT) calculations to characterize the reaction pathway and create theozymes for each step and compute their corresponding activation energy barriers. Next, using a multistate enzyme design approach, we focused on stabilizing key transition states. Experimental characterization of 12 designs revealed that all variants exhibited Michaelase activity. The designs displayed improved apparent K_M for the Michael acceptor compared to the original scaffold. However, they showed diminished apparent turnovers. The low catalytic efficiency of the designed enzymes could arise from over-stabilizing the targeted steps relative to other stages within the catalytic cycle. This study represents the first application of multistate enzyme design. Future efforts could focus on optimizing the catalytic efficiency through additional rounds of design and refinement.

4.5. Characterization of the reaction mechanism

We targeted the Michael addition reaction between (*E*)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) and ethyl cyanoacetate (**2a**) to form the product 3*S*-(*E*)-4-(4-methoxyphenyl)but-3-en-2-one (**3a**) (Figure 4.1). Previously, Hilvert and colleagues investigated the catalytic promiscuity of the engineered retro-aldolase RA95.5-8 in promoting the Michael addition of various nucleophiles to the α,β -unsaturated ketone **1a**.^{17,18} RA95.5-8 catalyzes this reaction with an apparent catalytic efficiency $k_{\text{cat}}^{\text{app}}/K_{\text{M}}^{\text{app}}$ of $53 \text{ M}^{-1}\text{s}^{-1}$. To enhance this activity, site-saturation mutagenesis was performed on nine positions within the enzyme's active site, resulting in the identification of a double mutant variant RA95.5-8 T53L/K210H, which exhibited a 3-fold increase in catalytic efficiency. The catalytic mechanism, similar to the retro-aldol reaction of methodol, involves the formation of an iminium ion intermediate between the catalytic lysine and the substrate (Figure 4.1B).^{19,20} We targeted the Michael addition reaction between **1a** and **2a** because previously designed and evolved retro-aldolase enzymes could serve as suitable scaffolds, featuring a pocket for the naphthyl ring of methodol that could accommodate the aromatic substituent of **1a** and a low- $\text{p}K_{\text{a}}$ lysine essential for catalysis. Moreover, the reaction progress could be monitored spectroscopically by measuring the reduction in absorbance of the conjugated double bond in **1a**. To our knowledge, no *de novo* enzyme has been previously designed for this reaction.

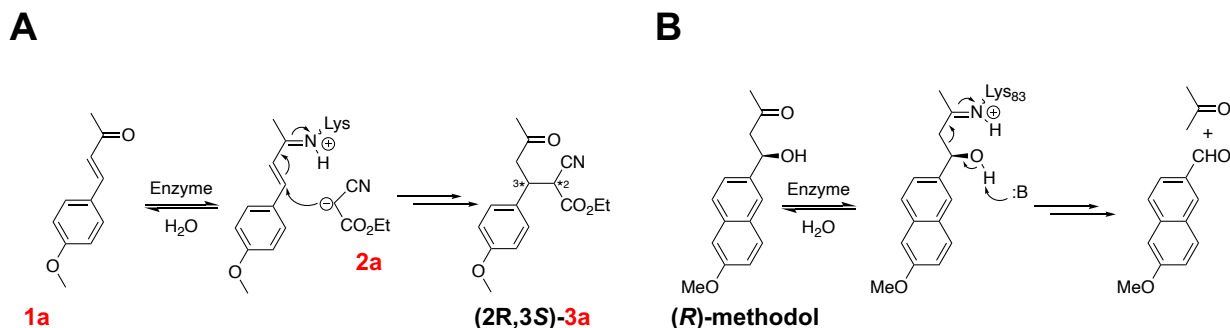


Figure 4.1. Michael addition and retro-aldol reactions catalyzed by RA95.5-8. In both (A) Michael addition of (*E*)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) and ethyl cyanoacetate (**2a**), and (B) retro-aldol reaction of (*R*)-methodol, a Schiff base intermediate is formed by reaction with the catalytic lysine 83.

To characterize the mechanism of Michael addition reaction between **1a** and **2a**, we performed density functional theory (DFT) calculations using Gaussian 16C.01 software package. Geometry optimization and single-point energy calculations were performed with the B3LYP-D3(BJ) theoretical method and the 6-31G(d) basis set. In calculating the transition states, we followed a procedure consistent with the transition state theory:

1. Proposing a transition state geometry: we began by proposing a transition state geometry informed by our understanding of the retro-aldol reaction mechanism.
2. Optimization of the transition state: The proposed transition state was optimized by traversing the potential energy surface and calculating energy gradients and the Hessian matrix for each geometry along the pathway. The transition state was considered optimized once the energy gradient vanished, and the Hessian matrix exhibited a single imaginary frequency.
3. Intrinsic Reaction Coordinate (IRC) Calculation: We carried out IRC calculations to ensure that the transition state connects the intended reactants and products.

4. Optimization of Reactant and Product Geometries: The geometries of the reactants and products were further optimized to accurately reflect their stable forms.

For these calculations, we used a simple, truncated model that included key catalytic residues such as lysine and tyrosine. The mechanism begins with an acid-base reaction in which the hydroxyl hydrogen of tyrosine is abstracted by the carbonyl oxygen of the substrate's aldol group. This process occurs concomitantly with the formation of an N–C bond between the ϵ -amino group of lysine and the carbonyl carbon of the substrate and deprotonation of the lysine residue by the same tyrosine.²⁰

The mechanism for the Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) and ethyl cyanoacetate (**2a**), as determined through DFT calculations, is illustrated in Figure 4.2.

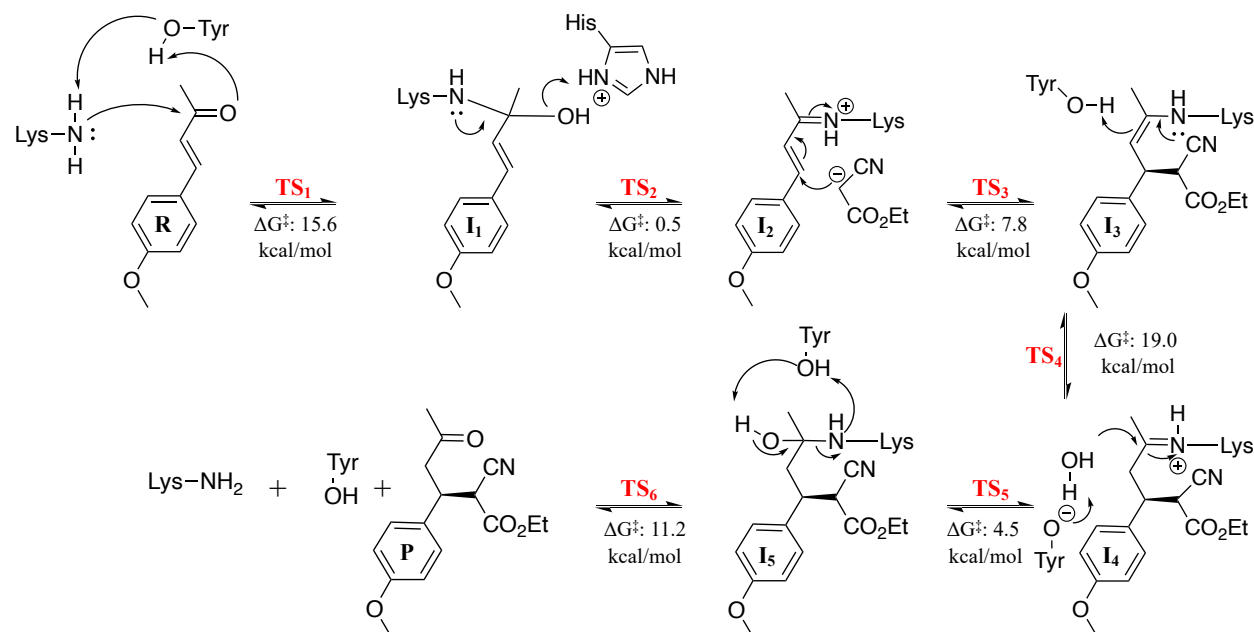


Figure 4.2. Reaction mechanism for Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) and ethyl cyanoacetate (**2a**).

It is important to note that the proposed reaction mechanism is a hypothesis that serves as the foundation for the entire design strategy. The design calculations are based on the assumption that the reaction proceeds via this mechanism. While this approach facilitates the rational design of enzymes tailored to stabilize the transition states, several limitations must be acknowledged. Firstly, the accuracy of the mechanism relies heavily on the computational models and assumptions used, such as the representation of the active site environment and the choice of density functional theory (DFT) methods. These models may not fully capture the dynamic and heterogeneous nature of enzyme active sites. Additionally, the modeling of charged intermediates and transition states presents inherent challenges, particularly with respect to solvation effects. The simplified treatment of solvation within an enzyme's active site may lead to uncertainties in the calculated energy barriers. Furthermore, there is a possibility that the actual reaction mechanism deviates from the proposed one due to unknown or overlooked catalytic factors. Despite these limitations, this mechanism represents our working theory and serves as the basis for the subsequent analysis. In the following sections, we will delve into each step of the reaction in detail, followed by the design calculations aimed at stabilizing the key transition states.

For the nucleophilic attack to occur, the nucleophile ethyl cyanoacetate must be deprotonated. We characterized this deprotonation process using either an aspartate or a histidine residue as the catalytic base. In the reaction with histidine the activation energy is $13.3 \text{ kcal mol}^{-1}$ and a reaction free energy of $12.3 \text{ kcal mol}^{-1}$. The reaction catalyzed by aspartate is more favourable with lower activation and reaction free energies of $4.7 \text{ kcal mol}^{-1}$ and $1.4 \text{ kcal mol}^{-1}$, respectively. The energy diagrams and structures of the transition states for both scenarios are depicted in Figure 4.3.

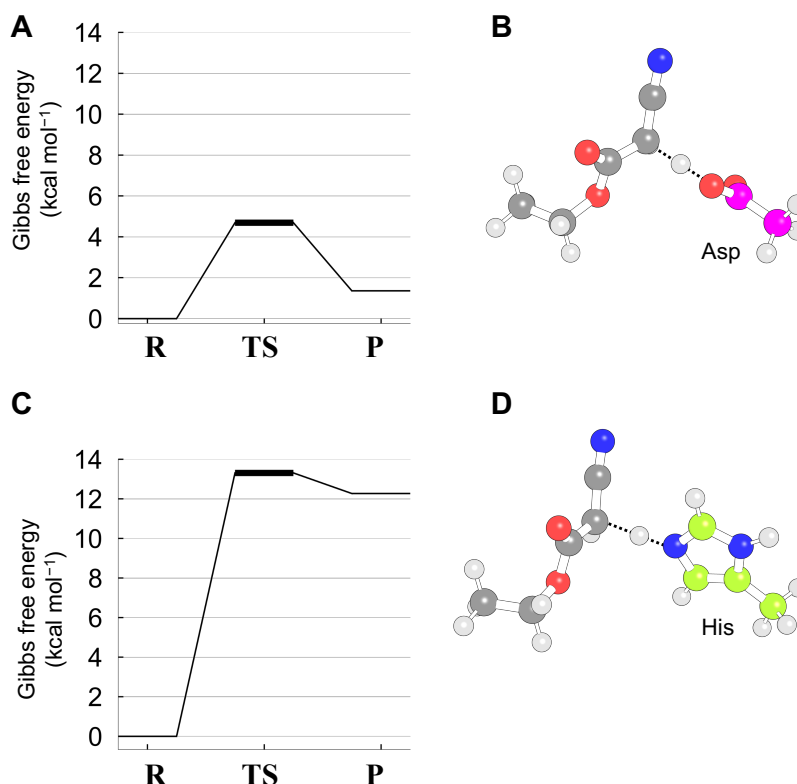


Figure 4.3. Deprotonation of ethyl cyanoacetate. Energy diagram for the deprotonation of ethyl cyanoacetate by an aspartate residue, showing an activation energy of 4.7 kcal mol⁻¹ and a reaction free energy of 1.4 kcal mol⁻¹. (B) Optimized structure of the transition state (TS) with aspartate (Asp) depicted in magenta. (C) Energy diagram for the deprotonation of ethyl cyanoacetate by a histidine residue, showing an activation energy of 13.3 kcal mol⁻¹ and a reaction free energy of 12.3 kcal mol⁻¹. (D) Optimized structure of the transition state (TS) with histidine (His) depicted in light green. Partial bonds are depicted in black dashed lines.

The first step of the reaction with an activation free energy of 15.6 kcal mol⁻¹ involves the concerted abstraction of hydroxyl hydrogen from a tyrosine residue, accompanied by the formation of an N—C bond between **1a** and lysine. Simultaneously, the same tyrosine deprotonates the lysine residue. The energy diagram for this step, along with the optimized structures of the reactants (R), transition state (TS₁), and intermediate (I₁), is presented in Figure 4.4. The overall ΔG of the reaction is 7.0 kcal mol⁻¹.

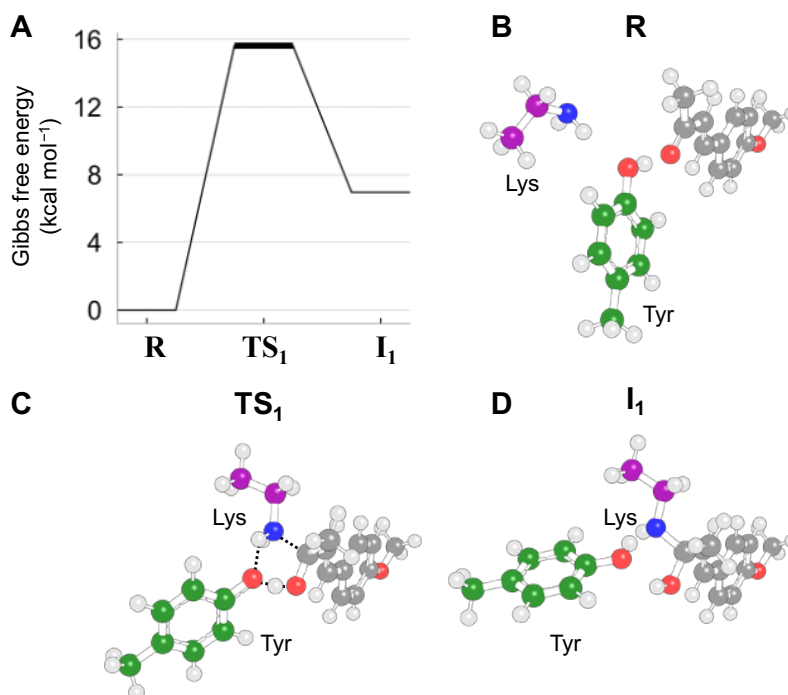


Figure 4.4. Step 1 of the reaction. (A) Energy diagram for step 1 of the reaction, showing an activation energy of 15.6 kcal mol⁻¹. Optimized structures of (B) the reactants (R), (C) Transition state 1 (TS₁), and (D) Intermediate 1 (I₁) structures. The substrate (E)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) is represented in gray, the lysine residue in purple, and the tyrosine residue in green. Partial bonds are depicted in black dashed lines.

The second step involves the elimination of a water molecule facilitated by a histidine residue, leading to the formation of a Schiff base. As depicted in the energy diagram shown in Figure 4.5A, this exergonic reaction occurs with an activation free energy of 0.5 kcal mol⁻¹. The resulting intermediate, I₂, is 52.0 kcal mol⁻¹ more stable than I₁.

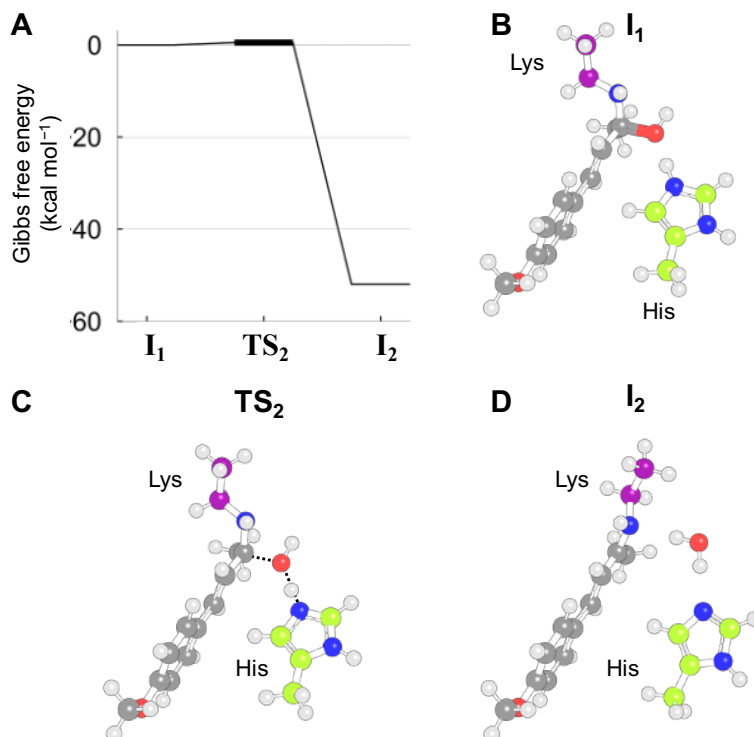


Figure 4.5. Step 2 of the reaction. (A) Energy diagram for step 2, with an activation energy of 0.5 kcal mol⁻¹. Optimized structures of (B) Intermediate 1 (I₁), (C) Transition state 2 (TS₂), (D) Intermediate 2 (I₂). In these structures substrate (1a) atoms are depicted in gray, the lysine residue in purple, and the histidine residue in light green. Partial bonds are depicted in black dashed lines.

In the third step, the deprotonated ethyl cyanoacetate acts as a nucleophile, attacking the double bond to form the C—C bond. This step has an activation energy of 7.8 kcal mol⁻¹ and a reaction free energy of -5.2 kcal mol⁻¹. The energy diagram for this step, along with corresponding structures, is depicted in Figure 4.6.

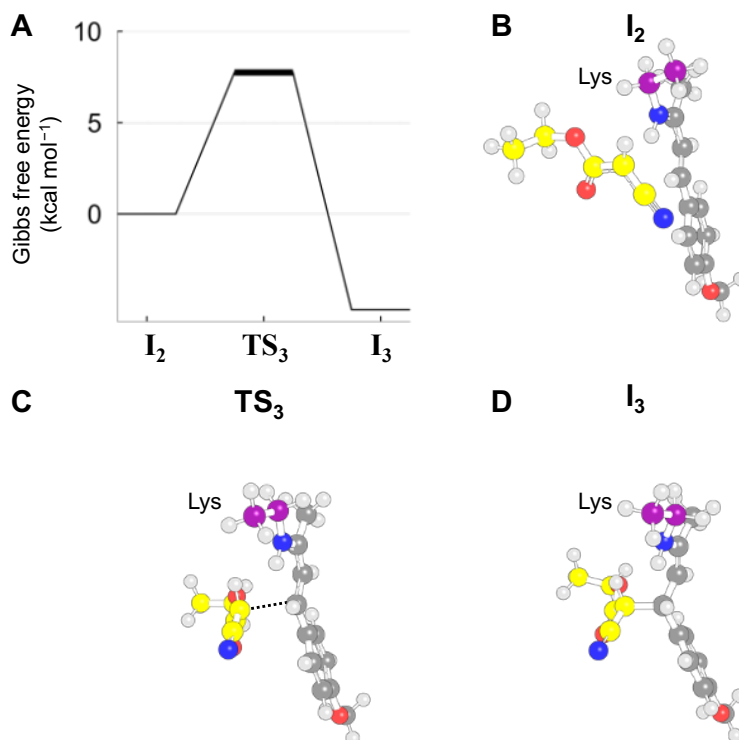


Figure 4.6. Step 3 of the reaction. (A) Energy diagram for step 3, with an activation energy of 7.8 kcal mol⁻¹. Optimized structures of (B) Intermediate 2 (I₂), (C) Transition state 3 (TS₃), and (D) Intermediate 3 (I₃). In these structures, the substrate 1a is depicted in gray, the lysine residue in purple and the substrate ethyl cyanoacetate (2a) in yellow. Partial bonds are depicted in black dashed lines.

In Step 4, the double bond is protonated by a tyrosine residue, with an activation free energy of 19.0 kcal mol⁻¹, making this the predicted rate-limiting step of the overall reaction. The reaction free energy is 18.8 kcal mol⁻¹ (Figure 4.7)

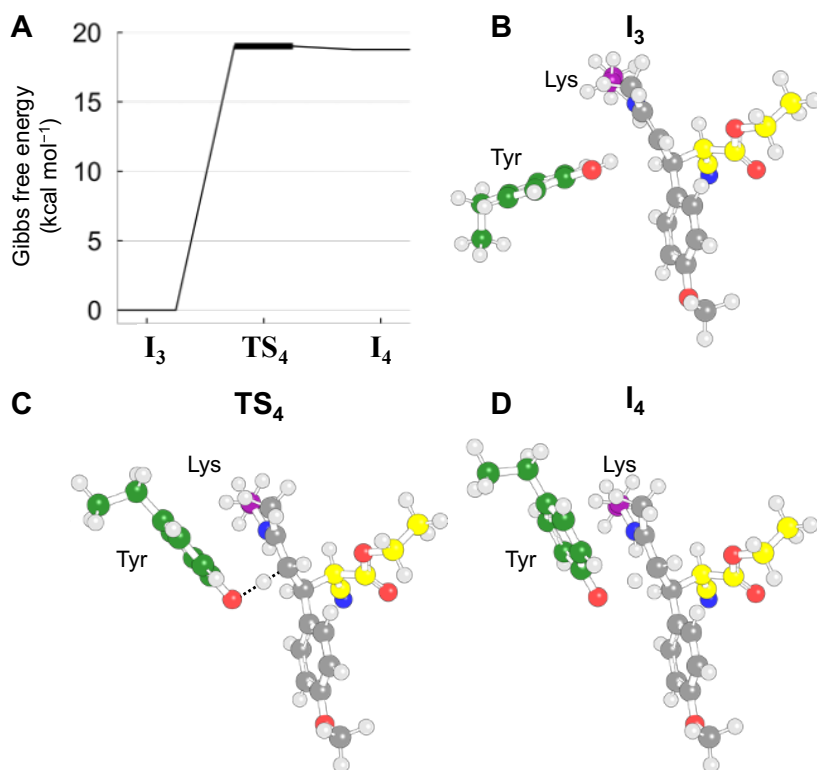


Figure 4.7. Step 4 of the reaction. (A) Energy diagram for step 4, with an activation energy of 19.01 kcal mol⁻¹. Optimized structures of (B) Intermediate 3 (I₃), (C) Transition state 4 (TS₄) and (D) Intermediate 4 (I₄). In these structures, the substrate 1a is depicted in gray, the lysine residue in purple, the tyrosine residue in green, and the substrate 2a in yellow. Partial bonds are depicted in black dashed lines.

The fifth step involves the nucleophilic attack by a water molecule on the sp² carbon of the Schiff base, facilitated by a tyrosine residue to form a carbinolamine. This step proceeds with an activation energy of 4.5 kcal mol⁻¹ and is exergonic, with a free energy change of -14.8 kcal mol⁻¹ (Figure 4.8).

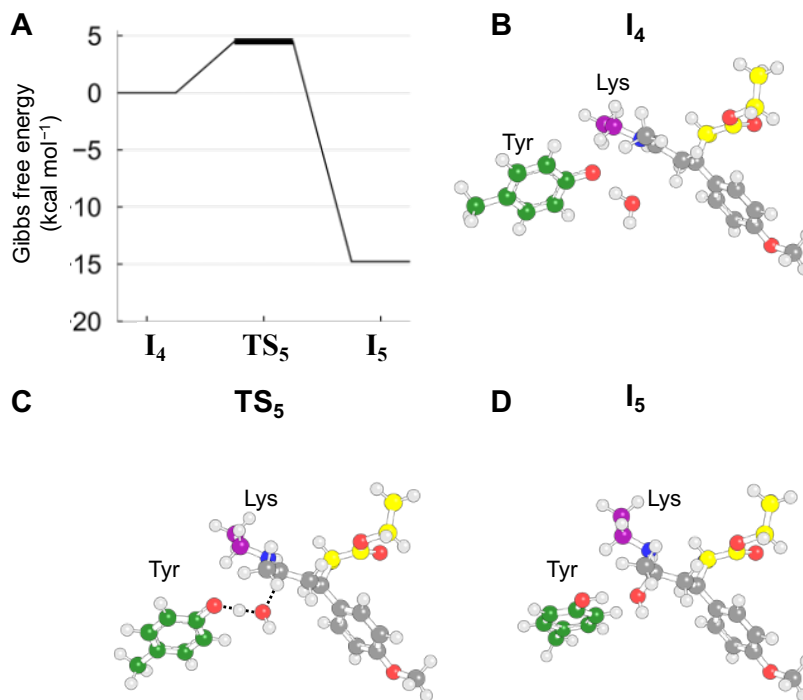


Figure 4.8. Step 5 of the reaction. (A) Energy diagram for step 5, with an activation energy of 4.5 kcal mol⁻¹. Optimized structures of (B) Intermediate 4 (I₄), (C) Transition state 5 (TS₅) and (D) Intermediate 5 (I₅). In these structures, the substrate 1a is depicted in gray, the lysine residue in purple, the tyrosine residue in green, and the substrate 2a in yellow. Partial bonds are depicted in black dashed lines.

The reaction of the sixth step is the reverse of step 1. The tyrosine is deprotonated by the nitrogen of the carbinolamine, the N—C bond is cleaved, and the tyrosine residue subsequently deprotonates the hydroxyl group. This step has an activation energy of 11.2 kcal mol⁻¹ (Figure 4.9).

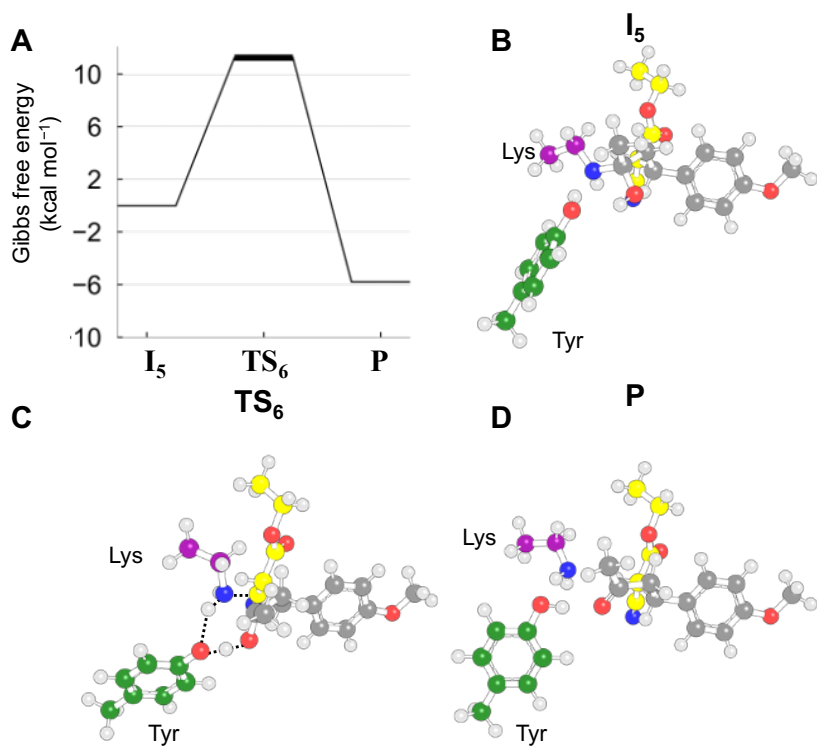


Figure 4.9. Step 6 of the reaction. (A) Energy diagram for step 6, with an activation energy of 11.2 kcal mol⁻¹. Optimized structures of (B) Intermediate 5 (I₅), (C) Transition state 6 (TS₆) and (D) Product (P). In these structures, the substrate 1a is depicted in gray, the lysine residue in purple, the tyrosine residue in green, and the substrate 2a in yellow. Partial bonds are depicted in black dashed lines.

4.6. Computational design procedure

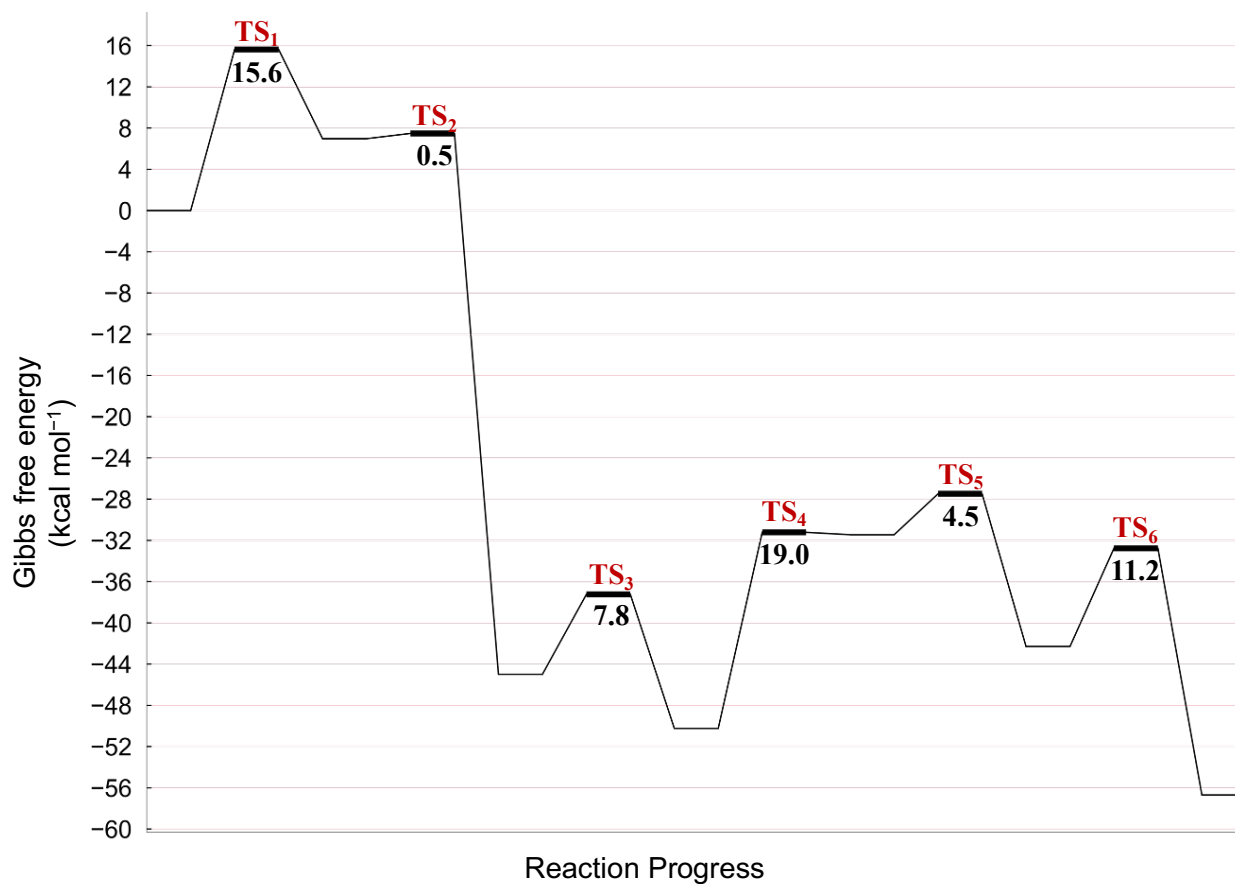


Figure 4.10. Energy profile for Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (1a) and ethyl cyanoacetate (2a).

Once we have constructed the reaction pathway and identified all the theozymes along with their corresponding free activation energies (Figure 4.10), the next step is selecting appropriate scaffolds capable of accommodating these theozymes. In designing an enzyme for a multistep reaction, it is crucial to account for backbone flexibility due to the conformational changes the enzyme undergoes during different reaction steps. This flexibility is essential for the enzyme to accommodate various intermediates and transition states effectively.

To address this, we selected scaffolds that are not only structurally compatible but also catalytically relevant. Specifically, we chose to use scaffolds from retro-aldolase enzymes that contain a catalytic lysine in their active sites. Additionally, we included class I aldolases in our selection, as these enzymes also contain a catalytic lysine that forms a Schiff base with the substrate fructose 1,6-bisphosphate. A total of 21 scaffolds were selected from the Protein Data Bank (PDB) based on their structural and catalytic properties. To account for backbone flexibility, we performed ensemble refinement on these scaffolds. The ensemble refinement resulted in a set of 1,613 templates to be used for theozyme placement (Supplementary Table 4.1).

The three transition states with the highest free activation energies are TS₁, TS₄, and TS₆. TS₁, with a barrier of 15.6 kcal mol⁻¹, is characterized by the formation of an N—C bond between lysine and the substrate, coupled with the deprotonation of a tyrosine residue, which subsequently deprotonates the lysine. TS₄, exhibiting the highest barrier at 19.0 kcal mol⁻¹, involves the protonation of the double bond by tyrosine and the formation of a Schiff base. TS₆, with a barrier of 11.2 kcal mol⁻¹, is characterized by the cleavage of the N—C bond, the deprotonation of tyrosine by nitrogen of the lysine, and the subsequent deprotonation of the hydroxyl group by the tyrosine. Given that TS₁ and TS₆ are catalytically reverse of each other and share structural similarities, with TS₆ representing a later stage where ethyl cyanoacetate has already attacked, making it a larger theozyme, we decided to focus our efforts on stabilizing TS₄ and TS₆.

After identifying candidate scaffolds and selecting the transition states to be stabilized, the next step is theozyme placement. During this process, ligand poses are generated from rotamers of a single catalytic contact, in this case, lysine, followed by testing of tyrosine rotamers to identify poses that satisfy all necessary contacts. Additionally, ligand poses are generated by allowing rigid

body rotations of rotatable bonds. Both TS₄ and TS₆ contain the catalytic lysine and a tyrosine residue, although the specific tyrosine differs between these theozymes. We introduced each transition state individually into each backbone template using triad phoenixMatch.py²¹ app as outlined in the materials and methods section. Once both theozymes were placed onto all backbone templates, the next task was to identify templates from the same sequence that could accommodate both transition states in such a way that they maintain similar positions and orientations. Given that the enzyme is unlikely to undergo major conformational changes, such as significant loop movements, the templates selected must also exhibit similar loop orientations.

To identify backbone templates capable of accommodating two transition states with minimal conformational deviation, we employed an in-house matching algorithm specifically developed for this purpose. This algorithm calculates the distance between the two transition states within templates derived from a single backbone as well as the backbone root mean square deviation (RMSD). We then normalized the ligand distances and RMSD values, combining them into a single score, which was the sum of the normalized ligand distances and backbone RMSD values (Figure 4.11). We selected motifs with a combined score below 1, which yielded 58 pairs. Given that each transition state involves a different tyrosine residue, we then added the tyrosine from TS₄ to TS₆ motif and vice versa within each template. This additional step further refined our selection, narrowing down the choices to five pairs.

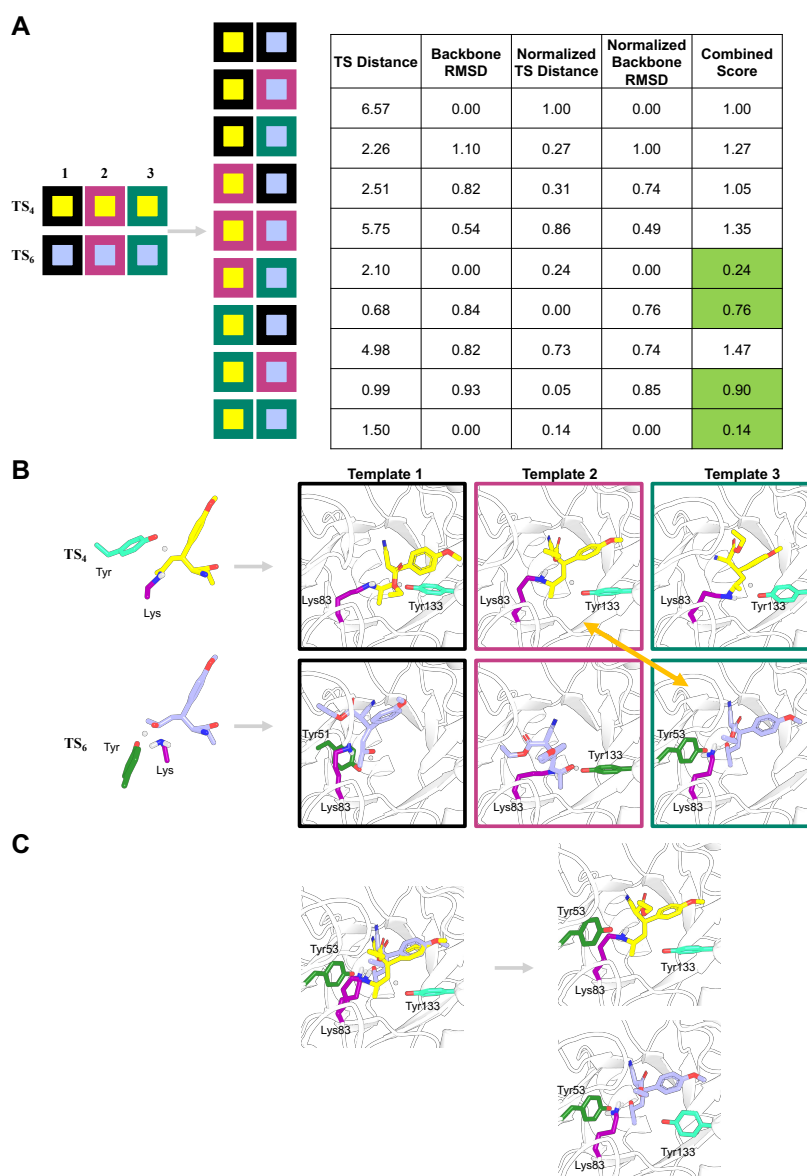


Figure 4.11. Schematic of the matching algorithm. Representation of matching algorithm used to identify backbone templates capable of accommodating TS₄ and TS₆ with minimal conformational deviation. (A) Each theozyme, TS₄ (yellow square) and TS₆ (blue square), is individually placed on all backbone templates, represented by frames in black, pink and green. For every possible pair of motifs within a template derived from the same sequence, the distance between two transition states and backbone root mean square deviation (RMSD) are calculated. These values are normalized and combined into a single score, representing the sum of the normalized ligand distances and backbone RMSD values. Templates with a combined score below 1, highlighted in green, are selected for the next step. (B) TS₄ (yellow) and TS₆ (light blue) are placed on all backbone templates. The catalytic lysine is shown in purple. Tyrosines for TS₄ and TS₆ are shown in cyan and dark green, respectively. The yellow arrow indicates a match between TS₄ on template 2 and TS₆ on template 3. (C) Overlay of both theozymes placed on two different backbone templates. Tyrosine 53 is added to TS₄ motif and Tyrosine 133 is added to TS₆ motif.

Following theozyme placement of TS₄ and TS₆, the next step is to identify sequences capable of stabilizing both transition states on their respective backbone templates. To achieve this, we employed a multistate sequence design (MSD) algorithm.²² MSD expands single-state design (SSD) by optimizing a sequence across multiple target backbone templates. It does so by calculating an energy combination function that evaluates the overall fitness of a sequence across multiple backbone structures. Similar to SSD, the energy for each sequence is determined by summing pairwise rotamer energies and the rotamer/template energies for each backbone. Next, a fitness score is calculated using an energy combination function and this score is optimized using a Monte Carlo sequence optimization algorithm to find optimal sequences for stabilizing multiple backbone templates. We extended the same principle to design sequences that stabilize multiple transition states on different backbone templates (Figure 4.12).

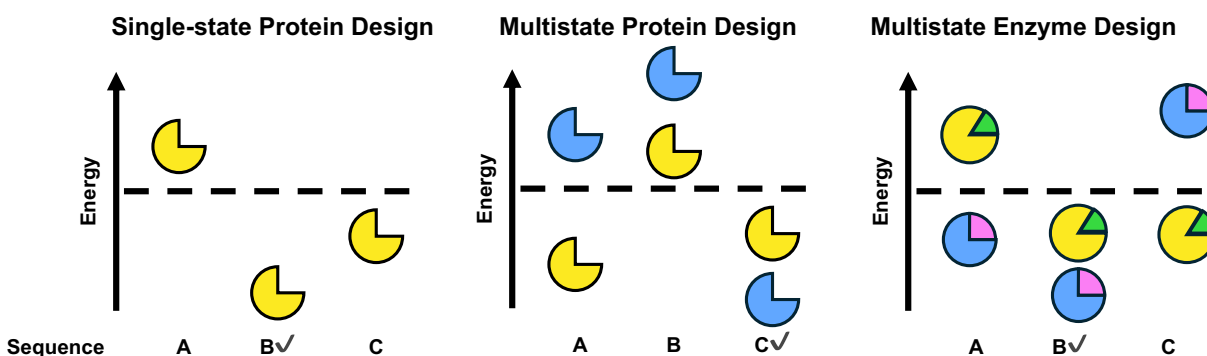


Figure 4.12. Schematic representation of protein and enzyme design strategies. In single-state protein design, sequences are optimized to stabilize a single backbone conformation (yellow) in this schematic, sequence B represents the lowest energy. In multistate protein design, sequences are optimized to stabilize multiple backbone conformations (yellow and blue) Here, sequence C represents the lowest energy sequence across both conformations. In multistate enzyme design, sequences are optimized to stabilize multiple chemical (green and pink) and conformational (yellow and blue) states. Here, sequence B represents the lowest energy for both transition states on two different backbone templates.

We first performed a single-state design calculation focused solely on stabilizing TS₄ as a control, yielding sequences with the lowest energy of $-240 \text{ kcal mol}^{-1}$. Using this energy as reference, we performed MSD on the five candidate pairs using a combination function that minimizes the energy difference between TS₄ and TS₆, while also imposing penalties if the energy for each state exceeded -200 , -220 , and $-250 \text{ kcal mol}^{-1}$ in three separate calculations. From 1,000 sequences generated per protocol, only one pair, based on the retro-aldolase RA95.5-8F I133F scaffold exhibited favorable energies, which was selected for next step.

Following sequence design, we applied a series of filters aimed at reproducing mechanisms known to enhance enzyme catalysis. First, we eliminated sequences with an energy more than $100 \text{ kcal mol}^{-1}$ above the lowest energy in each calculation. Second, we evaluated active-site complementarity to TS₄ and TS₆ by calculating the solvent-accessible surface area (SASA) of each ligand in the computational models generated for 1,000 sequences designed in each protocol. Only designs where $\geq 75\%$ of the TS surface is buried were kept. Finally, we filtered based on the predicted lysine pK_a, keeping only sequences with a pK_a below 9. This filtering process narrowed the original list of 4000 designs down to 12, which we experimentally characterized (Table 4.1). For the single-state design SSD, these filters were only applied to TS₄.

Table 4.1. Filtering parameters for the 12 experimentally tested designs

Design	TS ₄ Energy	TS ₆ Energy	ΔE	#mutations	Lys pK _a TS ₄	Lys pK _a TS ₆	TS ₄ SASA	TS ₆ SASA
MSD_200_163	-203.8	-202.2	1.6	13	8.5	8.3	76.9	79.4
MSD_220_116	-216.2	-216.9	0.7	14	8.55	8.3	76.9	76.3
MSD_220_164	-215.3	-216.6	1.2	14	8.5	8.3	77.2	76.6
MSD_220_266	-213.6	-215.9	2.3	14	8.6	8.7	75.0	75.2
MSD_220_33	-219.8	-218.9	0.9	14	8.5	8.3	77.3	76.4
MSD_220_376	-212.2	-214.7	2.5	14	8.6	8.7	75.0	75.9
MSD_250_13	-234.6	-244.3	9.6	14	8.6	8.5	75.0	76.5
SSD_201	-225.9	-234.9	9.0	14	8.6	8.8	75.4	67.8
SSD_370	-220.4	-256.5	36.1	14	8.6	8.6	76.5	75.1
SSD_656	-213.9	-244.9	30.9	14	8.5	8.5	75.5	77.4
SSD_811	-210.1	-232.7	22.5	13	8.6	8.6	75.6	73.7
SSD_897	-207.5	-85.0	122.5	13	8.6	8.7	75.5	69.2

4.7. Results

We next sought to experimentally characterize the designed Michaelases. Genes encoding the sequences were expressed in *E. coli*, and the resulting proteins were purified by nickel affinity chromatography. All twelve designs catalyzed the Michael addition reaction, though their catalytic efficiencies were lower than the starting scaffold RA95.5-8F used for design ($k_{\text{cat}}^{\text{app}} / K_{\text{M}}^{\text{app}} = 46.5 \text{ M}^{-1} \text{ s}^{-1}$) (Table 4.2, Figure 4.13). Interestingly, all but one design exhibited lower apparent K_{M} values for the Michael acceptor compared to the starting scaffold, indicating tighter substrate binding. However, all designs exhibited reduced apparent turnover numbers with the highest $k_{\text{cat}}^{\text{app}}$ one order of magnitude lower than the starting scaffold (Figure 4.14).

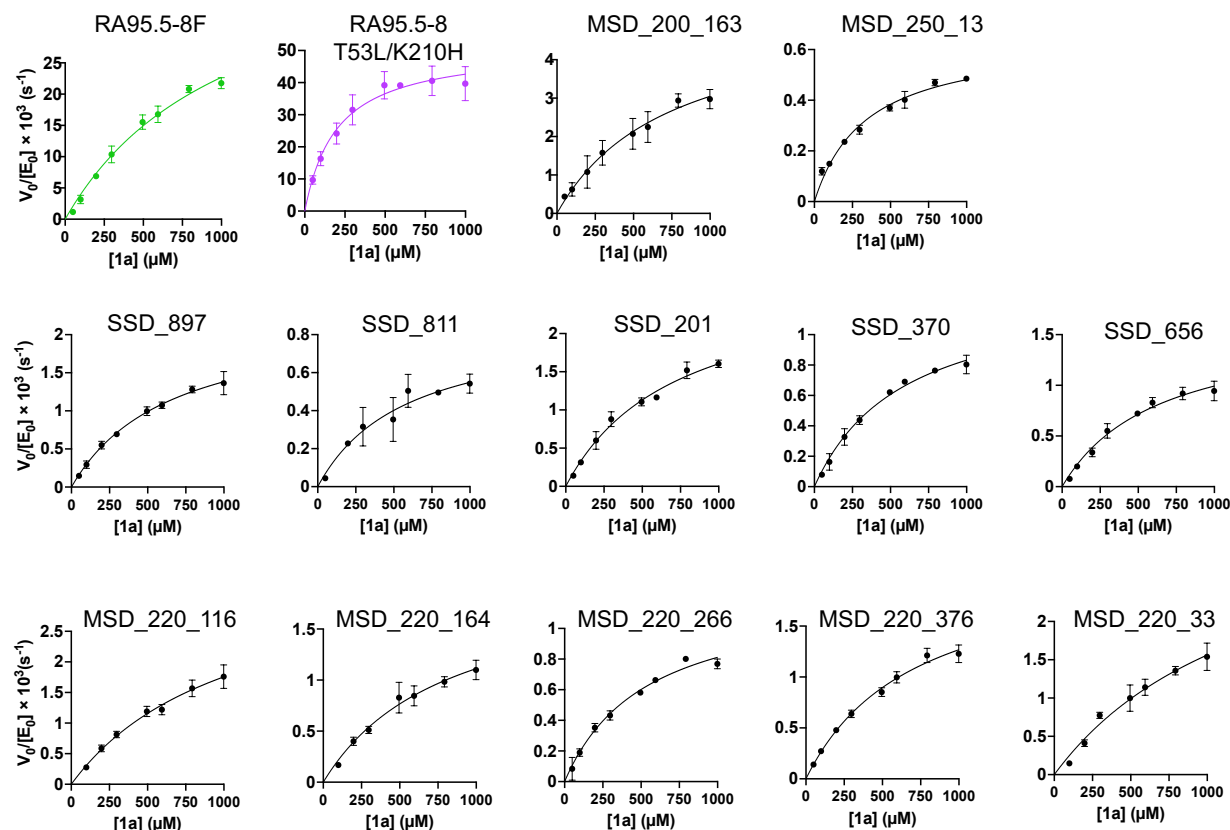


Figure 4.13. Michaelis-Menten graphs of designed Michaelases. Michael addition reaction between (E)-4-(4-methoxyphenyl)but-3-en-2-one (1a) and ethyl cyanoacetate (2a) was monitored for each Michaelase variant. Data fitted at constant concentration of 2a (5 mM) and various concentrations of 1a.

Table 4.2. Apparent steady-state parameters of Michaelase designs

Enzyme	K_M^{app} (μM)	k_{cat}^{app} (s^{-1})	k_{cat}^{app}/K_M^{app} ($\text{M}^{-1} \text{s}^{-1}$)
MSD_200_163	737	0.005	7.2
MSD_250_13	341	0.0006	1.9
MSD_220_266	537	0.0012	2.3
MSD_220_376	707	0.002	3.5
MSD_220_33	1359	0.0036	2.7
MSD_220_116	1068	0.0036	3.4
MSD_220_164	821	0.002	2.5
SSD_897	655	0.002	3.1
SSD_811	581	0.0008	1.5
SSD_201	724	0.003	3.8
SSD_370	593	0.0013	2.2
SSD_656	632	0.0016	2.5
RA95.5-8F (Template)	1074	0.05	46
RA95.5-8 T53L/K210H (evolved Michaelase) ¹⁷	205	0.05	244

Reaction conditions: 25 mM Hepes buffer, 100 mM NaCl, pH=7.5, 29 °C, 3% acetonitrile, **2a** = 5 mM.

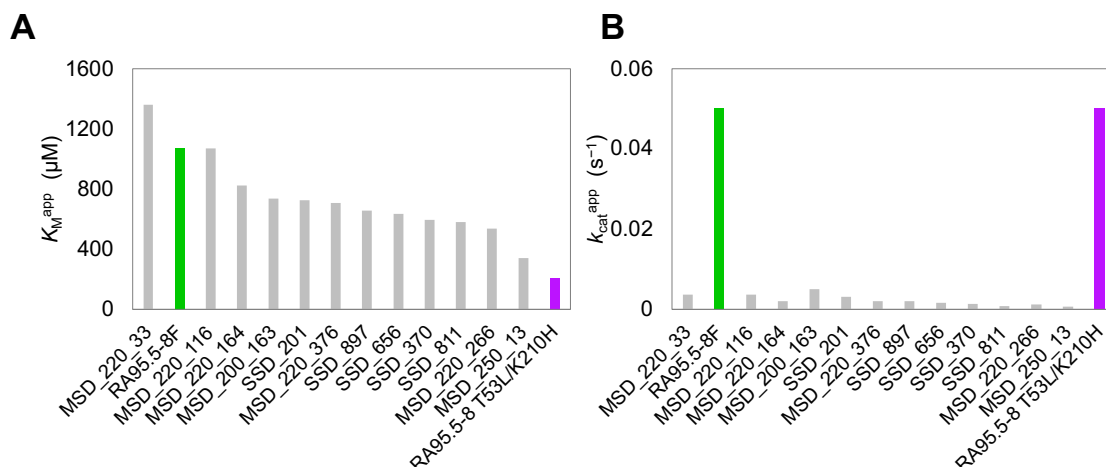


Figure 4.14. Comparison of kinetic parameters for Michaelase variants. Apparent (A) Michaelis constant, K_M^{app} values and (B) turnover number, $k_{\text{cat}}^{\text{app}}$ for the designed Michaelase variants, the starting scaffold RA95.5-8F (green), and the evolved variant RA95.5-8 T53L/K210H (purple).

Despite the multistate design (MSD) approach aimed at stabilizing both TS_4 and TS_6 , no significant differences were observed between the catalytic activities of single-state (SSD) and multistate designed Michaelases. This was somewhat expected, as the sequences generated by SSD, while only minimizing the energy of TS_4 , still retained favourable energies for TS_6 (Table 4.1).

To explore the origins of the reduced turnover rates in the designed Michaelases, we modelled other transition states, including the deprotonation of ethyl cyanoacetate (**2a**), on the backbone ensembles of the starting scaffold RA95.5-8F, the evolved Michaelase RA95.5-8 T53L/K210H, the most active design MSD_200_163 and the design with the lowest K_M^{app} for **1a**, MSD_250_13. These theozymes were placed on all members of each ensemble using the phoenixMatch theozyme placement algorithm. A similar matching algorithm, as previously described, was employed to identify templates that aligned the theozymes in comparable orientations. The energies of the resulting designs were then calculated to construct an energy

diagram for each enzyme. For MSD_200_163 and MSD_250_13, we used the ensembles of their starting scaffolds RA95.5-8F II33F. The evolved Michaelase RA95.5-8 T53L/K210H harbours the mutation K210H, suggesting that this histidine residue could catalyze the nucleophile deprotonation. We successfully modelled the deprotonation of **2a** with Asp212 on the starting scaffold RA95.5-8F and His210 on the evolved Michaelase RA95.5-8 T53L/K210H. However, modelling the deprotonation of **2a** on MSD_200_163 and MSD_250_13 using Asp212 resulted in steric clashes with Gln58. Therefore, we hypothesize that a contributing factor to the lower turnover rates might be the lack of explicit design of a catalytic residue for the deprotonation of the nucleophile **2a** (Figure 4.15).

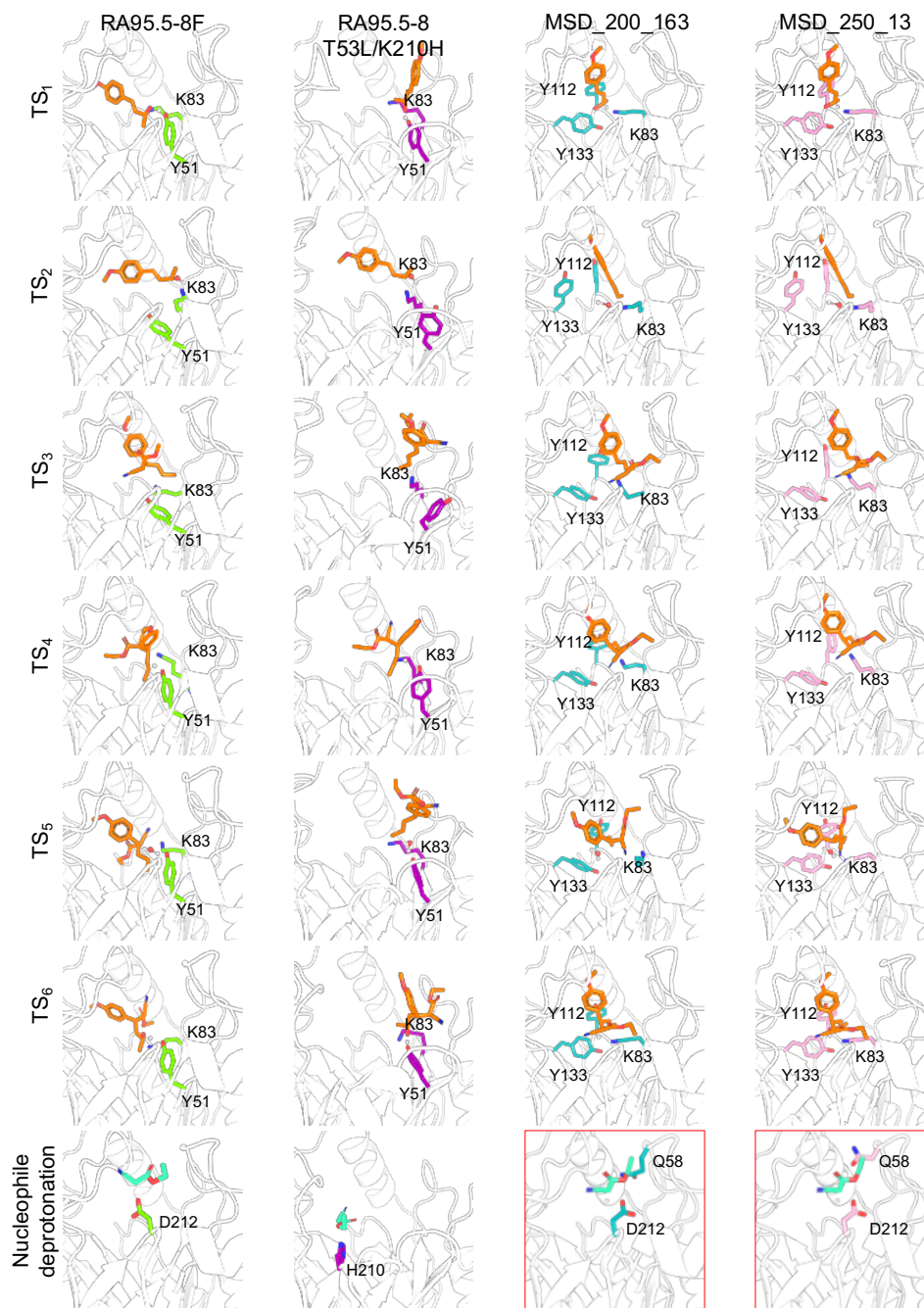


Figure 4.15. Structural models of transition states. Transition states 1–6 and transition state of deprotonation of ethyl cyanoacetate (**2a**) is modeled across ensembles of four Michaelase scaffolds. RA95.5-8F, RA95.5-8 T53L/K210H, MSD_200_13 and MSD_250_13. The catalytic lysine, Lys83, is conserved in all variants, while the catalytic tyrosine differs. Tyr51 in RA95.5-8F and RA95.5-8 T53L/K210H and Tyr112 or Tyr133 in MSD_200_13 and MSD_250_13. The deprotonation of **2a** is modeled using Asp212 in RA95.5-8F and His210 in RA95.5-8 T53L/K210H. However, modeling the deprotonation step in the designed variants MSD_200_13 and MSD_250_13 leads to steric clashes with Gln58, as highlighted by red squares.

Furthermore, the energy profiles of steps 1–6 transition state models on these four enzymes revealed that MSD_200_163 and MSD_250_13 exhibit lower energies for TS₄ and TS₆ relative to the starting scaffold, indicating effective stabilization of these targeted transition states. However, this stabilization comes at the cost of higher energy barriers for other states, such as TS₃ and TS₅ in MSD_250_13 (Figure 4.16A). We also computed the ΔE —the energy difference between the least and most stable transition states. For MSD_200_163 and MSD_250_13, the ΔE values are 216 and 212 kcal mol⁻¹, respectively, whereas the higher activity variants, RA95.5-8F and RA95.5-8 T53L/K210H, display smaller ΔE values of 94 and 82 kcal mol⁻¹. The imbalance in transition state stabilization in the designed enzymes may cause the enzyme to get trapped in a less dynamic conformation, hindering its ability to progress through the catalytic cycle efficiently.

Moreover, we performed 200 ns molecular dynamics (MD) simulations in triplicate to evaluate the conformational dynamics of these enzymes. The evolved Michaelase exhibited greater loop distances between two loops located over the active site, with a median loop distance of 27 Å, compared to 14 Å for MSD_200_163, 17 Å for MSD_250_13, and 14 Å for RA95.5-8F (Figure 4.16B). This increased loop distance in the evolved variant could facilitate substrate binding and product release, potentially explaining its superior activity compared to the designed variants.

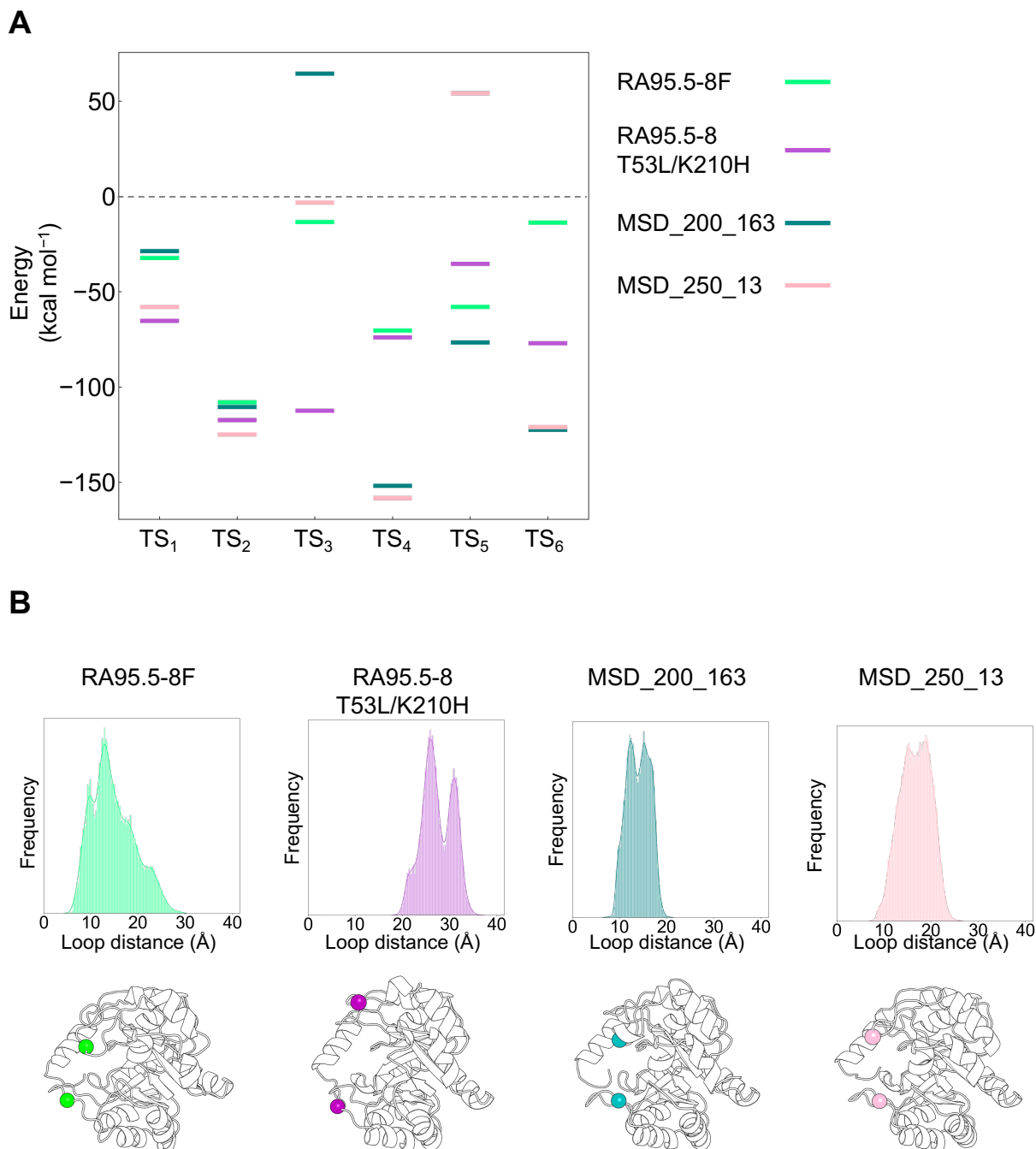


Figure 4.16. Energy diagram and loop distances in Michaelase variants. (A) Energy diagram of different transition state models on the backbone ensembles of RA95.5-8F (green), RA95.5-8 T53L/K210H (purple), MSD_200_13 (teal), and MSD_250_13 (light pink). (B) Distribution of loop distances in the MD trajectories of the four scaffolds, alongside representative structures. The median inter-loop distance is 14 Å for RA95.5-8F, 27 Å for RA95.5-8 T53L/K210H, 14 Å for MSD_200_163, and 17 Å for MSD_250_13. The loop distances are calculated by measuring the Ca distance between residues 58 and 188.

Circular dichroism analysis revealed that MSD_200_163 and MSD_250_13 displayed similar spectra to the starting scaffold RA95.5-8F, suggesting proper folding (Figure 4.17A). The melting temperature (T_m) of MSD_250_13 was identical to that of RA95.5-8F at 72 °C, indicating retained thermal stability. In contrast, MSD_200_163 showed a decreased T_m of 56 °C (Figure 4.17B). All kinetic assays were performed at 29 °C, well below the T_m of both designs, ensuring their structural integrity was maintained during the experiments. The irregular shapes of the thermal unfolding graphs may result from protein precipitation during melting, which could be mitigated by optimizing buffer conditions, such as adjusting pH and ionic strength, to enhance protein solubility and stability.

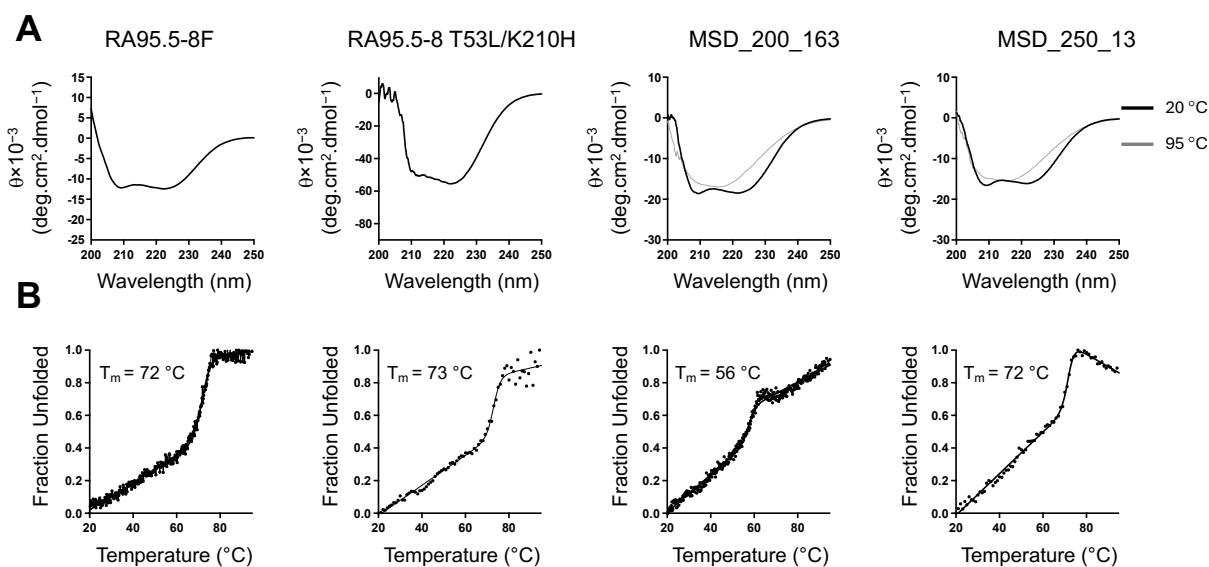


Figure 4.17. Circular dichroism (CD) spectra and thermal melting graphs. (A) CD spectra and (B) thermal unfolding graphs of the starting scaffold RA95.5-8F, the evolved Michaelase RA95.5-8F T53L/K210H, MSD_200_163 and MSD_250_13.

4.8. Discussion

This chapter represents the first application of multistate enzyme design to generate a series of enzymes capable of catalyzing the Michael addition reaction. All 12 designed enzymes successfully catalyze the Michael addition reaction, demonstrating the feasibility of the multistate approach for enzyme design. We did not observe a significant difference between the catalytic activities of designs obtained by single-state design (SSD) where sequences were optimized to stabilize only TS₄ versus multistate design where sequences were optimized to stabilize both TS₄ and TS₆, and also minimize their energy difference. We hypothesize that this is because SSD generated sequences that had favourable energy values for TS₆ as well.

The experimentally characterized designs were able to lower the apparent Michaelis constant of the Michael acceptor compared to the starting scaffold RA95.5-8. However, these enzymes exhibited significantly reduced apparent turnover numbers $k_{\text{cat}}^{\text{app}}$. This suggests that multistate enzyme design is able to create novel enzymes with active sites that are able to productively bind the substrate, but this alone does not guarantee highly efficient catalysis. This pattern has been observed in other designed and engineered enzymes, where the modest catalytic efficiencies of these enzymes are mostly due to lower k_{cat} values, which are 1–6 orders of magnitude lower than natural enzymes, and not the Michaelis constants, which fall within the range of natural enzymes.^{23,24}

Our investigation into the origins of the diminished turnover rates revealed that the designed enzymes stabilize TS₄ and TS₆ significantly more than other transition states. While this targeted stabilization successfully lowered the energy barriers for these steps, it might have inadvertently destabilized other transition states. This suggests that the designed enzymes might

get trapped in less dynamic states, hindering the necessary conformational changes for efficient catalysis. This implies that to design enzymes with higher catalytic efficiencies, the entire catalytic cycle must be considered during design calculations—including the free enzyme, the substrate-bound, and the product-bound enzyme, along with transition-state bound states. This can help prevent the over-stabilization of the substrate or product-bound states, which might otherwise make these steps rate-limiting. Furthermore, the lack of a catalytic base to deprotonate the nucleophile ethyl cyanoacetate could result in lower turnovers, as the pK_a of ethyl cyanoacetate is 8.98.

An efficient enzyme must maintain low energy barrier across all reaction steps. We hypothesized that flattening the energy landscape—reducing the difference in energy between various reaction steps—would result in higher catalytic efficiency. To test this, we measured ΔE , the energy difference between the least and most stable transition state models in different enzymes. The most active Michaelase, RA95.5-8 T53L/K210H, displayed the lowest ΔE , followed by RA95.5-8F, indicating flatter energy profiles. In contrast, the designs MSD_200_163 and MSD_250_13 exhibited higher ΔE values, indicating greater energy variation between transition states, which may account for their lower catalytic efficiencies.

The analysis of loop distances during MD simulations revealed that the evolved Michaelase RA95.5-8 T53L/K210H has higher loop distances compared to the starting scaffold and designs MSD_200_163 and MSD_250_13. This increased loop distance could play a role in facilitating substrate binding and product release. Therefore, it is important to consider enzyme dynamics early in the design process. The overall rate of enzymatic catalysis involves not just the chemical transformations modeled by theozymes but also binding of the substrates, release of products and conformational exchange in between steps. Neglecting these aspects during design can result in

enzymes that, while structurally compatible with the transition states, are trapped in less active conformations. This could be addressed by explicitly designing both the open and closed states of the enzyme.

Another strategy to improve the enzymes is to construct more sophisticated theozymes with additional residues that not only stabilize the transition states but also interact with and orient the catalytic residues. While *de novo* enzymes are typically created using simple theozymes containing only catalytic residues, directed evolution often leads to the emergence of more complex active sites. A notable example is the retro-aldolase RA95, which was originally designed with a catalytic lysine in a hydrophobic environment.³ Multiple rounds of directed evolution resulted in the emergence of a catalytic tetrad with extensive hydrogen bond network.²⁵ Using this tetrad to design retro-aldolases led to designed enzymes with higher catalytic efficiencies than previously designed retro-aldolases. Therefore, by constructing more sophisticated theozymes with additional stabilizing interactions, we may be able to create enzymes with greater catalytic efficiency.

Furthermore, *de novo* enzyme design is an iterative process, often requiring multiple rounds of optimization to enhance the catalytic efficiency of designs.²⁷ Therefore, a second round of design where we target other transition states to stabilize, model loop movements between different states and design open and closed conformations could optimize the activity of the designs generated in the first round.

4.9. Methods

4.9.1. DFT calculations

Quantum mechanical calculations were conducted using density functional theory (DFT) with the B3LYP functional and the 6-31G(d,p) basis set to locate transition structures. These structures were confirmed by vibrational frequency analysis to ensure they correspond to transition states. The solvent effect was modelled using the Polarizable Continuum Model (PCM) with diethyl ether as the solvent with a dielectric constant of 4.33 to mimic enzyme active site. Critical functional groups were included in the models: a lysine residue was present in all transition states, histidine was included in TS₂ to facilitate elimination of water, a tyrosine was incorporated in TS₁, TS₄, TS₅ and TS₆. Deprotonation of ethyl cyanoacetate was modelled using either an aspartate or histidine residue. Additionally, Intrinsic Reaction Coordinate (IRC) calculations were performed to trace the minimum energy pathway connecting the transition state to the reactant and products. Finally, the geometries of both reactants and products of each step were optimized. All calculations were carried out using Gaussian16.²⁸

4.9.2. Ensemble refinement

Time-averaged ensembles were generated with phenix.ensemble_refinement implemented in PHENIX v.1.15.2-3472. In ensemble refinement,²⁹ local atomic fluctuations are sampled using molecular dynamics simulations accelerated and restrained by electron density to produce ensemble models fitted to diffraction data. Briefly, input crystal structures were edited to remove low-occupancy conformers and assign an occupancy of 1.0 to the remaining conformer. Following addition of riding hydrogens, parallel ensemble refinement simulations were performed using various combinations of the parameters p_{TLS} (0.6, 0.8, 0.9, 1.0), τ_x (0.5, 1.5, 2.0) and $w_{\text{x-ray}}$ (2.5,

5.0, 10.0), where p_{TLS} describes the percentage of atoms included in a translation-libration-screw (TLS) model use to remove the effects of global disorder, τ_x is the simulation time-step and $w_{\text{x-ray}}$ is the coupled tbath offset, which controls the extent to which the electron density contributes to the simulation force field such that the simulation runs at a target temperature of 300 K. The ensemble generated from each crystal structure that displayed the lowest R_{free} value was selected for design and analysis (Supplementary Table 4.1).

4.9.3. Ligand placement

All calculations were performed with the Triad protein design software (Protabit, Pasadena, CA, USA) using a Monte Carlo with simulated annealing search algorithm for rotamer optimization. WT lysine in each template was used as the catalytic lysine and tyrosine was allowed at several positions in the active site (Supplementary Table 4.2). Except for catalytic residues, all the positions in the active site were mutated to glycine during theozyme placement to avoid steric clashes with designed catalytic residues. The 2002 Dunbrack backbone-independent rotamer library³⁰ with expansions of ± 1 standard deviation around χ_1 and χ_2 was used to provide side-chain conformations. A library of TS poses was generated in the active site by targeted ligand placement³¹ using the contact geometries listed in Supplementary Table 4.3 for TS₄ and Supplementary Table 4.4 for TS₆. TS pose energies were calculated using a modified version of the Phoenix energy function³² consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field³³ with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well-depth of 8.0 kcal mol⁻¹ and an equilibrium donor-acceptor distance of 2.8 Å,³⁴ and an electrostatic energy term modelled using Coulomb's law with a distance-dependent dielectric of 10. During the energy calculation step, each TS–side-chain interaction energies were biased to

favour interactions that satisfy contact geometries (Supplementary Table 4.5 and Supplementary Table 4.6) by applying an energy benefit of $-100 \text{ kcal mol}^{-1}$ as described by Lassila *et al.*³¹ A penalty of 500 kcal/mol was applied to poses that did not satisfy the biases.

4.9.4. Matching

Following ligand placement, all possible combinations of TS₄ and TS₆ poses generated from templates with identical sequences were aligned using Pymol,³⁵ and the distances between four atoms (O29, C26, C19 and O39) were calculated and averaged. Additionally, the backbone RMSD between the two templates was calculated. These values were normalized, and a combined score—representing the sum of the average ligand distance and backbone RMSD—was computed. Poses with a combined score ≤ 1 were retained for the next step. Subsequently, the tyrosine from TS₄ was incorporated into the TS₆ motif, and vice versa. Any motifs with steric clashes were eliminated.

4.9.5. Single-state design

Single state design calculations were performed to stabilize the TS₄ bound structures. The identities and rotamers of catalytic residues were fixed. Residues mutated to Gly in the theozyme placement step were allowed to sample rotamers of selected amino acids (Supplementary Table 4.9) to form a mostly hydrophobic active site. Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. While only the energy of TS₄ motif was stabilized during these calculations, sequences were also optimized on TS₆ motif. After sequence design, we calculated energy difference between each designed structure and the structure in the absence of the TS where all designed residues were converted to Gly. This energy subtraction was done to allow comparison of transition-state binding energies across the two backbone templates.

4.9.6. Multistate design

Multistate design calculations were performed for both TS₄ and TS₆ motifs using an energy combination function to determine the overall fitness score for a sequence across both transition states. The single-state design yielded TS₄ energies ranging from $-239 \text{ kcal mol}^{-1}$ to $-203 \text{ kcal mol}^{-1}$. Therefore, we conducted three sets of multistate design calculations where the combination function minimizes the energy difference between TS₄ and TS₆ while applying penalties if the energy of either state exceeds that of the corresponding all-glycine structure by -250 , -220 and $-200 \text{ kcal mol}^{-1}$. These calculations were performed using the `multi_design.py` application with the Triad software suite.

4.9.7. Analysis and filtering

For each design protocol—single-state design and the three multistate designs—1,000 sequences were generated. To reduce the number of sequences to test experimentally, we discarded those with energies exceeding $100 \text{ kcal mol}^{-1}$ above the lowest energy in each calculation. Additionally, the catalytic lysine pK_a for both TS₄ and TS₆ was using propka3,³⁶ and sequences with a predicted lysine pK_a greater than 9 were eliminated. Finally, the solvent-accessible surface area (SASA) of TS₄ and TS₆ was calculated using Pymol³⁵, and only sequences where $\geq 75\%$ of the TS surface is buried were retained.

4.9.8. Protein expression and purification

Codon-optimized and his-tagged (C-terminus) genes for Michaelase designs (Supplementary Table 4.7) cloned into the pET-29b(+) vector via NdeI and XhoI were obtained from Twist Bioscience. Enzymes were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using

lysogeny broth (LB) supplemented with 100 $\mu\text{g mL}^{-1}$ kanamycin. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of 0.3–0.7, and protein expression was initiated with 1 mM isopropyl- β -D-1-thiogalactopyranoside. Following incubation at 16 °C for 16 h with shaking (250 rpm), cells were harvested by centrifugation, resuspended in 8 mL of lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer pH 8.0) supplemented with 1.25 g mL^{-1} of lyophilized lysozyme and 1 U mL^{-1} of benzonase nuclease (Merck Millipore), and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography using Ni-NTA agarose (Qiagen) pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Columns were washed twice, first with 10 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0), and then with the same buffer containing 20 mM imidazole. Bound proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and exchanged into 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride using Econo-Pac 10DG desalting prepacked gravity-flow columns (Bio-Rad). Purified samples were quantified by measuring the absorbance at 280 nm and applying Beer–Lambert’s law using calculated extinction coefficients obtained from the ExPASy ProtParam tool (<https://web.expasy.org/protparam>).

4.9.9. Steady-state kinetics

All assays were carried out at 29 °C in 25 mM Hepes buffer (pH 7.5) supplemented with 100 mM sodium chloride. Michael addition reaction between (*E*)-4-(4-methoxyphenyl)but-3-en-2-one (**1a**) and ethyl cyanoacetate (**2a**) was monitored spectroscopically at 370 nm ($\Delta\epsilon = 1858.5 \text{ M}^{-1} \text{ cm}^{-1}$) using a SpectraMax384Plus plate reader (Molecular Devices). 200- μL reactions were performed in triplicate using varying concentrations of freshly prepared **1a** (A2B Chem) dissolved in acetonitrile (3% final concentrations) and 5 mM **2a** (A2B Chem). Reactions were initiated by

the addition of **2a** to the mixture of 2–4 μM enzyme and **1a**. Linear phases of the kinetic traces were used to measure initial reaction rates. Steady-state kinetic parameters were derived by fitting the experimental data to the Michaelis-Menten equation $v_0/[E_0] = k_{\text{cat}} \cdot [S]/(K_M + [S])$.

4.9.10. Circular dichroism and thermal denaturation assays

Circular dichroism measurements were performed with a Jasco J-815 spectrometer using 650- μL aliquots of each Michaelase in a 1-mm path-length quartz cuvette (Jasco) at a concentration of 10 μM in 20 mM sodium phosphate buffer (pH 7). For structural characterization of protein folds, circular dichroism spectra were acquired from 200 to 250 nm, sampled every 0.2 nm at a rate of 10 nm min^{-1} . Five scans were acquired and averaged for each sample. For thermal denaturation assays, samples were heated at a rate of 0.5 $^{\circ}\text{C}$ per minute, and ellipticity at 222 nm was measured every 0.2 $^{\circ}\text{C}$. T_m values were determined by fitting a 2-term sigmoid function with baseline correction using nonlinear least-squares regression³⁷. Thermal unfolding measurements were performed as a single replicate.

4.9.11. Molecular dynamics

The Amber 2020 software (<http://ambermd.org/>) with the ff19SB protein force field,³⁸ gaff2 ligand force field,³⁹ and OPC water force field⁴⁰ was used for all MD simulations. Long-range ($> 10 \text{ \AA}$) electrostatics were modeled using the particle mesh Ewald method.⁴¹ All MD trajectories were run in triplicate and with a time step of 2 fs. Hydrogen atoms were added using Reduce⁴², and the prepared structure was solvated with OPC water in a truncated octahedral box with periodic boundary conditions where the distance between the protein surface and the box edges was set to 10 $\text{\AA}$. The additions2 algorithm in amber was used to place counterions to neutralize the system. The system was first minimized through the method of steepest descent using a force

constant of $1000 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$. The system was heated to a temperature of 300 K for 240 ps, followed by equilibration under an NPT ensemble at 300 K and 1 bar for 10 ns and subsequently equilibration under an NVT ensemble at 300 K for another 10 ns. Constant temperature and pressure were achieved using the Nose-Hoover thermostat⁴³ and Berendsen barostat⁴⁴, respectively. After removing positional restraints on protein heavy atoms imposed for heating and equilibration, a 200 ns production run was performed under Parrinello–Rahman pressure coupling.⁴⁵

4.10. Supplementary information

Supplementary Table 4.1. Ensemble refinement of template scaffolds used in design of Michaelases

PDB ID	Start Rf/Rw	Final Rf/Rw	No. of templates
1A53	0.159/0.212	0.138/0.109	98
1EUA	0.243/0.201	0.228/0.157	63
1F7B	0.210/0.167	0.162/0.120	56
1HL2	0.210/0.183	0.178/0.139	56
1JCJ	0.168/0.142	0.162/0.141	100
1JCL	0.169/0.143	0.158/0.138	167
1P1X	None	0.154/0.141	100
1VHC	0.256/0.219	0.209/0.170	42
1W3I	0.211/0.176	0.178/0.140	63
3TC6	0.198/0.179	0.189/0.138	42
3O6Y	0.291/0.233	0.252/0.182	42
3TC7	0.193/0.158	0.167/0.135	42
3UD6	0.258/0.227	0.208/0.153	42
4A2R	0.165/0.138	0.169/0.144	72
4A2S	0.162/0.132	0.179/0.151	72
4LNY	0.222/0.191	0.231/0.158	56
4OU1	0.203/0.175	0.198/0.171	84
5EKY	0.144/0.121	0.140/0.125	100
5EL1	0.144/0.117	0.139/0.118	100
6TFA	0.279/0.229	0.273/0.188	50
6TF8 (Chain B)	0.247/0.200	0.222/0.154	72
6TF8 (Chain C)	0.247/0.200	0.222/0.154	72

Supplementary Table 4.2. Lysine positions and tyrosine candidate positions in each template scaffold

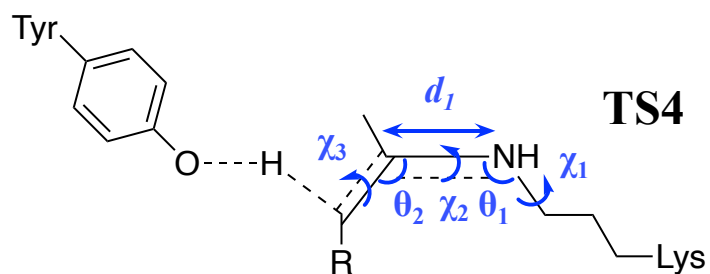
PDB ID	Lysine position	Tyrosine candidates
1A53	83	51,53,110,112,133,159,180,210
1EUA	133	20,45,47,73,94,135,161
1F7B	164	43,137,167,187,188,206
1HL2	165	43,137,167,187,188,206
1JCJ	167	18,20,47,73,102,139,201,236
1JCL	167	18,20,47,73,102,139,201,236
1P1X	167	18,20,47,73,102,139,201,236
1VHC	133	20,45,47,73,94,135,161
1W3I	155	43,130,132,157,179,196
3TC6	81	51,108,159,180,210,231,233
3O6Y	159	51,81,108,180,210,231,233
3TC7	210	51,110,133,159,180,210
3UD6	159	51,81,108,180,210,231,233
4A2R	83	51,53,110,112,133,159,180,210
4A2S	83	51,53,110,112,133,159,180,210
4LNY	210	51,110,133,159,180
4OU1	210	51,110,133,159,180
5EKY	167	18,20,47,73,102,139,201,236
5EL1	167	18,20,47,73,102,139,201,236
6TFA	83	51,53,110,112,133,159,180,210
6TF8 (Chain B)	83	51,53,110,112,133,159,180,210
6TF8 (Chain C)	83	51,53,110,112,133,159,180,210

Supplementary Table 4.3. Geometric definitions for generation of TS4 poses off the lysine side chain

Contact	Symbol	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
Lys	d_1	Distance	NZ	C10			1.1 1.3 1.5
	θ_1	Angle	CE	NZ	C10		121 126 131
	θ_2	Angle	NZ	C10	C11		117 122 127
	χ_1	Torsion	CD	CE	NZ	C10	0 60 120 180 240 300
	χ_2	Torsion	CE	NZ	C10	C11	150 155 160 165 170 175 180 185 190
	χ_3	Torsion	NZ	C10	C11	C17	8 13 18 23 28 33 38

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residues.

^b Distance measurements given in Å all others in degrees.

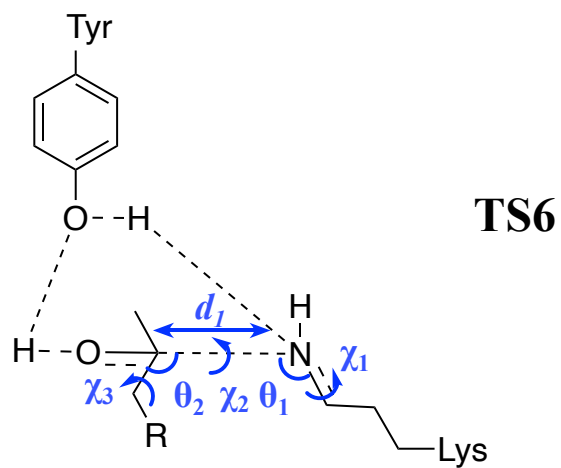


Supplementary Table 4.4. Geometric definitions for generation of TS₆ poses off the lysine side chain

Contact	Symbol	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
Lys	d_1	Distance	NZ	C10			1.4 1.6 1.8
	θ_1	Angle	CE	NZ	C10		117 122 127
	θ_2	Angle	NZ	C10	C11		105 110 115
	χ_1	Torsion	CD	CE	NZ	C10	0 60 120 180 240 300
	χ_2	Torsion	CE	NZ	C10	C11	283 288 293 298 303
	χ_3	Torsion	NZ	C10	C11	C17	266 271 276 281 286

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residues.

^b Distance measurements given in Å all others in degrees.

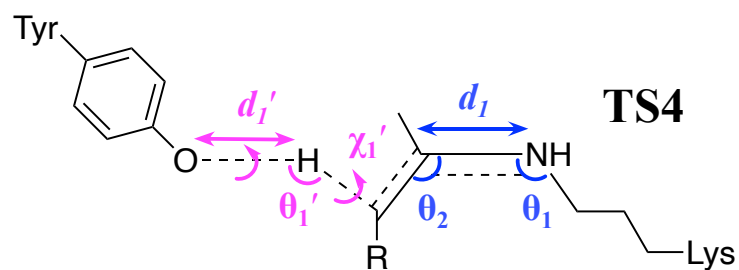


Supplementary Table 4.5. Geometric constraints used to define TS₄ catalytic contacts

Contact	Symbol	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max ^b
Lys	d_I	Distance	NZ	C10			1.1	1.5
	θ_1	Angle	CE	NZ	C10		100	140
	θ_2	Angle	NZ	C10	C11		100	140
Tyr	d_I'	Distance	OH	H64			0.9	2.0
	θ_1'	Angle	OH	H64	C11		150	180
	χ_1'	Torsion	OH	H64	C11	C10	-30	30

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residues.

^b Distance measurements given in Å all others in degrees.

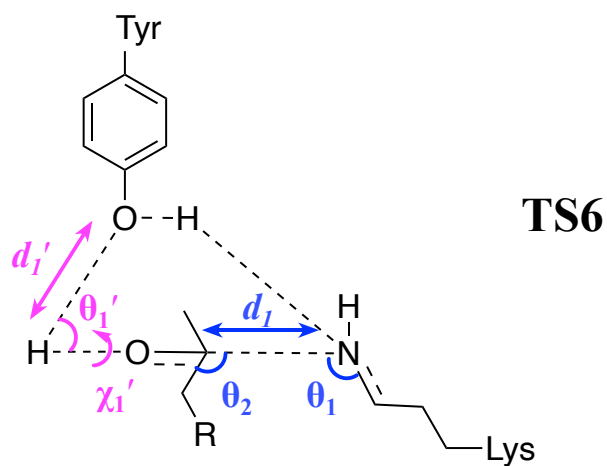


Supplementary Table 4.6. Geometric constraints used to define TS₆ catalytic contacts

Contact	Symbol	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max ^b
Lys	d_I	Distance	NZ	C10			1.4	2.0
	θ_1	Angle	CE	NZ	C10		100	140
	θ_2	Angle	NZ	C10	C11		90	130
Tyr	d_I'	Distance	OH	H65			0.9	2.5
	θ_1'	Angle	OH	H65	O64		150	180
	χ_1'	Torsion	OH	H65	O64	C10	0	60

^a Atoms in bold are from the transition state. All other atoms are from the catalytic residues.

^b Distance measurements given in Å all others in degrees.



Supplementary Table 4.7. Amino-acid sequences of Michaelase designs

Enzyme	Sequence
MSD_220_116	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_220_33	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_220_164	MPRYLKGHLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_220_266	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IAYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_220_376	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IAHYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_250_13	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIIIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
SSD_201	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IAMYARKGPVGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIIIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
SSD_811	MPRYLKGHLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPVGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIIIND ENDLDIALRIGARFITIFSNFETGEINKENQRKLISMIPSNVVKVPLDFFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
SSD_656	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPLGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIIIND ENDLDIALRIGARFITIFSVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
SSD_370	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPNGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIIIND ENDLDIALRIGARFITIHVNFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
SSD_897	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IAMYARKGPGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIYSINFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH
MSD_200_163	MPRYLKGQLEDVVQLSLRRPSVHASRQRPIISLNERILEFNKRNITAI IALYARKGPQGLDVERDP I EYAKYMEPYAVGLS IKTEEKYNDGSYEMLRKIASSVSIPIILMQDYIVKESQIDDAYNLGADTVNLYVEILTERELESLEYARGYGMEPLIQIND ENDLDIALRIGARFITIYSINFETGEINKENQRKLISMIPSNVVKVPWLDFEFEPNEIEELRKLGVNAFMISSSLMRNPEKI KELIEGSLEHHHHHH

Supplementary Table 4.8. DNA sequences of Michaelase designs

Enzyme	Sequence
MSD_220_116	CCGCGCTATCTGAAGGGGCAGCTGGAAGACGTTGTGCAATTATCACTGCGCCGCCGAGTGTCCATGCGTCTCGTCAGCGG CCTATAATCTCACTTAACGAGCGGATACTGGAGTTC AATAAGCGGAACATCACGGCTATTATCGCGCTGTACGCCGAAAG GGCCCTAACGGTTTAGACGTTGAACGCGACCCCATCGAGTACGCAAAATACATGGAGCCTTATGCGGTGGGGCTTTCATC AAAAC TGAGGAAAAGTATAATGACGGCAGCTACGAGATGTTGCGAAAGATCGCGTCATCAGTTAGCATTCCAATATTGATG CAAGACTATATAGTCAAGGAGTCTCAGATCGATGATGCCTACAACCTTGGGGCTGATACCGTGAACCTTATACGTAGAGATC CTGACAGAGCGTGAGCTGGAGTCCCTGTTAGAGTACGCGGTGGGTACGGTATGGAGCCTTAAATCAACATCAATGATGAG AACGACCTTGACATTGCACTTCGGATTGGCGCCAGATTTATCACTATCTTTAGCGTGAATTCGAGACAGGAGAGATAAAC AAGGAGAACCAACGGAAGTAAATTTCAATGATCCCAAGCAATGTAGTCAAGGTGCCGTGGTTAGATTTCTTTGAGCCGAAC GAAATCGAAGAGTTACGTAAGCTCGGCGTTAACGCTTTATGATCAGCAGCTCACTGATGAGAAACCCAGAGAAGATTAA GAGTTAATAGAAGGGTCGCTTGAGCATCACCATCACCACCACTGA
MSD_220_33	CCTAGATACTTAAAGGGTATGTTAGAGGATGTAGTTCAGCTTTCCTTACGTGCGCCGTCAGTGCATGCCTCTCGTCAACGT CCGATCATCAGCTTGAACGAGCGAATTCTCGAATTC AACAAGCGGAACATCACGGCAATTATTGCCTTATATGCGAGAAAG GGTCCGAACGGCCTGGACGTCGAGCGTGACCCCATAGAATACGCGAAATATATGGAGCCTTACGCCGTTGGACTGTCAATT AAGACCGAGGAAAAGTACAATGACGGCAGTTACGAGATGTTACGGAAGATAGCCTCAAGTGTAGCATCCCGATATTAATG CAAGACTACATCGTCAAGGAAAAGTCAGATTGACGACGCGTATAAATTTGGGTGCTGATACAGTCAACTTGTACGTTGAGATC CTGACCGAAGAGAGATTAGAAAGCCTTCTCGAATACGACGCGGTTACGGGATGGAACCATTGATACAGATCAACGACGAA AACGACCTTGACATAGCACTGAGAATTGGTGCCCGATTATCACAACTCTTTAGCGTTAATTTTGAAACCGGAGAGATTAA AAGGAAAATCAGCGAAAACCTTATCTCTATGATCCCGTCTAACGTGGTTAAAGTCCCGTGGTTGATTTCTTTGAGCCGAAC GAAATTGAGGAACCTCAGAAAGCTTGGAGTGAATGCGTTTATGATTTCAGCAGCTTGATGCGAAACCCGAAAAGATCAAA GAAC TGATAGAGGGGAGTTTAGAGCATCACCACCATCACCATTAA
MSD_220_164	CCGCGCTATCTGAAGGGGCATTTAGAGGATGTTGTGCAAGTGTCTCTGAGACGTCCATCAGTCCACGCTAGCCGACAGCGA CCTATTATTAGTTTAAATGAGCGCATTTTAGAATTCAACAAACGTAACATAACGGCAATTATTGCTTTTATATGCCCGTAAG GGCCCGAACGGGTTGAGATGTCGAGCGGGACCCATCGAGTATGCGAAGTACATGGAGCCTTACGAGTTGGGGCTTTCGATT AAAACGGAAGAGAAGTACAACGACGGCTCCTACGAGATGTTACGCAAGATTGCATCTAGCGTATCAATCCGATTTTGATG CAGGACTACATCGTAAAAGAGTCTCAAATCGATGATGCATACAACCTGGGCGCAGACACAGTCAACTTATACGTTGAAATA TTGACAGAGAGAGAGTTGGAAGTCTTTTAGAGTACGACGCTGGCTATGGCATGGAGCCCTCATTAACATTAACGACGAG AATGACCTTGATATCGCCCTCAGAATTGGAGCTCGTTTACATCACTATCTTCAGTGTGAATTTCCGAAACCGGAGAGATCAAC AAGGAGAATCAAAGAAAATTAATTTCAATGATCCCTCTAACGTTGTTAAGGTACCTTGGCTCGACTTCTTTGAGCCAAAT GAGATCGAAGAGCTCCGCAAGCTGGGAGTAAATGCGTTTATGATCAGCTCAAGCTTAATGCGGAACCCGAGAAGATTAA GAGTTGATCGAAGGAAGTCTTGAACACCATCATCACCACCACTGA
MSD_220_266	CCGCGGTACCTTAAAGGCATGTTAGAGGATGTGGTGAATTGAGTCTGCGACGCCCGTCAGTACACGCGTCTAGACAACGC CCCATAATCAGTTTAAATGAGCGCATTTTAGAATTCAATAAGCGGAACATCACCGCAATTATAGCGCAATACGCTAGAAAG GGTCCGAACGGGCTCGACGTTGAGAGAGACCCATTGAGTACGCTAAGTACATGGAGCCGTATGCGGTGGGATTTGTCAATC AAGACCGAAGAAAAGTACAATGATGGCAGCTACGAGATGCTTCGCAAAATCGCCAGCTCAGTGAGCATTTCCAATCTTAATG ATGACTACATTTGTAAGGAGTTCGAGATTGATGATGCTTACAACCTTAGGGGCGGACACAGTTAATTTGTACGTGGAGATC CTGACTGAACGCGAGTTAGAATCGCTTCTGGAGTACGCCAGAGGTTACGGCATGGAGCCACTGATCCAGATCAATGACGAG AACGACCTGGATATCGCGTTACGATCGGTGCTCGGTTTATCACCATATTTTCTGTAACTTTGAGACTGGTGAGATTAA AAAGAGAACCAACGTAAGCTGATCTCTATGATTCCTTCTAATGTGCTTAAAGTCCCGTGGCTGGATTCTTTGAGCCAAAC GAGATTGAAGAGCTTCGAAAGCTTGGTGTCAATGCTTTTATGATATCTTCAAGTCTTATGCGTAAATCCCGAGAAGATCAAA GAGTTGATAGAGGGTCACTGGAGCATCACCACCACCACTGA
MSD_220_376	CCACGTTATCTTAAAGGCATGCTGGAAGACGTCGTTCAATTATCTTTACGGCGGCCGTCTGTGACGCGCTCGCGACAACGT CCTATAATTTCACTCAACGAGCGTATCCTTGAGTTTAAACAAACGGAATATCACGGCTATCATTGCGCACTACGCCGTAAG GGCCCGAACGGACTCGACGTGGAGCGTGATCCTATCGAGTATGCTAAATACATGGAGCCATACGAGTTGGACTGTCAATC AAGACTGAGGAGAAGTATAACGATGGGAGCTACGAAATGTTACGTAAGATCGCGTCTAGTGTGAGCATTCCTATTTTGATG ATGGAATATATCGTCAAGGAATCAAAATAGACGATGCCTACAACCTTAGCGCTGACACGGTTAATTTATACGTGGAGATA TTGACGGAGCGCAACTGGAAGGTTTGCTTGAATATGCTCGTGGTTATGGAATGGAACCTCTTATACAAATTAACGACGAA AACGACCTGGACATAGCCCTCAGAATCGGGGCCGATTTCATACAATATTTAGCGTCAATTTTGAGACAGGGGAGATTAA AAGAAAACCGAGCTAAGCTGATAAGTATGATACCTTCAAACGTTGTGAAGGTTCCCTGGCTCGATTTCTTCGAACCAAT GAGATTGAGGAGCTTAGAAAATTAGGGGTTAACGCGTTTATGATCTCGTCATCACTTATGCGCAACCCGAAAAGATTAA GAGCTGATCGAGGGTTCTCTTGAGCACCACCACCATCACCACCTGA
MSD_250_13	CCTCGGTATCTTAAAGGGTATGTTAGAGGACGTCGTTCAATTAAAGCTTGCAGCGGCCAAGCGTTTATGCATCTCGTCAGCGG CCAATCATCAGTCTGAATGAGCGGATCCTTGAGTTC AACAAGCGCAACATAACGGCTATCATTTGCGCTTTATGCCCGTAAG GGCCCAACGGCTTAGACGTAGAGCGTGATCCCATGAGTATGCTAAGTACATGGAGCCTTATGCAGTCGGTCTCAGTATT AAAAC TGAGGAAAAGTACAATGATGGTAGTTATGAAATGCTTCGCAAGATTGCGTCCAGTGTGTCCATCCCTATATTAATG CAAGACTACATAGTCAAGGAGAGCCAAATTGACGACGCTACAACCTGGGCGCAGATACTGTTGCCCTTATGTGGAGATC CTGACAGAAAGAGAGCTTGAGTCATTATTAGAGTACGCGCGGGTACGGAATGGAACCGCTTATCATCATTAATGATGAG AATGACTTGGATATTGCTCTCCGGATCGGTGCACGCTTCATAACCATATTCTCTATCAACTTCGAGACCGGTGAGATAAAC AAGGAAAACCAACGCAAACTGATCAGCATGATTCCAGTAACGTGGTTAAGGTACCCATGCTCGACTTCTTTGAGCCGAAC

	GAAATAGAGGAGTTGCGTAAGTTAGGCGTTAACGCTTTTATGATAAGCTCATCCCTTATGCGCAACCCGGAGAAAATCAAA GAACTTATAGAAGGTTTCCTTAGAACACCATCACCACCATCATTAA
SSD_201	CCTCGATACTTGAAGGGTATGTTAGAGGACGTCGTACAACCTAGTCTCCGCCGTCCTCAAGTGACACGCCAGCCGCCAACGT CCGATCATTCTCTCAACGAGCGAATACTCGAATTCACAAGAGAAACATCACTGCAATTATTGCTATGTATGCTCGGAAG GGCCAGTGGGATTAGACGTCGAGAGAGATCCGATAGAGTACGCAAAGTACATGGAGCCCTACGCCGTCGGCTTGTCTATA AAGACCGAGGAAAAGTATAACGACGGATCTTATGAGATGTTGCGCAAGATCGCCAGCTCCGTGAGCATTCCAATCCTGATG ATGGACTACATCGTAAAGGAATCGCAAATTGACGATGCCACAATTTGGGCGCGGACACGGTTGCGTTGTACGTGGAGATT CTTACTGAGAGAGAACTGGAGAGCCTGTTAGAGTACGCAAGAGGGTACGGCATGGAACCTCTGATCATCATTAAATGACGAG AATGATTTAGACATCGCACTTCGAATTGGAGCAGGATTTATTACAATATTCTCTATAAACTTTGAGACCGCGGCAAAATCAAC AAGGAGAACCAGCGAAAAGCTGATATCCATGATCCCTCTAATGTTGTTAAGGTACCGATGCTGGACTTCTTCGAGCCAAAC GAGATCGAGGAACCTCGGAAGCTCGGTGTGAACGATTCATGATAAGCAGTTCACTGATGCGGAACCCTGAGAAGATCAAG GAGCTGATAGAAGGTAGTTTAGAACATCACCATCATCATCACTGA
SSD_811	CCCCGCTATTTGAAGGGTCACCTTGAAGACGTTGTCCAATTGTCCCTTCGTGCTCCGTCTGTGCATGCTTCCCGTCAGAGA CCAATTATTTCTCTGAATGAACGTATCTTAGAATTCACAAGCGTAACATCACCGCGATCATAGCTCTTTACGCGCGAAAG GGGCGGTTGGGCTGGACGTCGAGCGAGACCCAATTGAATATGCGAAATACATGGAGCCGTACGCAGTCGGTCTTTCCATT AAGACCGAAGAGAAATACAATGATGGCTCGTACGAAATGTTGCGGAAGATAGCATCCTCTGTTAGCATCCCCATCTCATG CAAGACTATATAGTCAAGGAGAGTCAAATCGACGACGCTTACAACCTTAGGGGCTGACACCGTGGCGTTATATGTAGAAATC TTGACGGAGAGAAATTGGAGAGCCTGTTAGAGTACGCGCGTGGATACGGAATGGAACCCCTTGATCATTATCAACGACGAA AACGACCTGGACATTGCTTTGCGCATTTGGCGCTAGATTTCATCAGCATATTTAGCCAGAACTTCGAGACCGGAGAGATCAAC AAAGAAAATCAACGCAAGCTGATCAGCATGATCCCGTCAACGTTGTCAAGGTCCCGCTGCTGGACTTCTTCGAGCCCAAT GAGATCGAGGAGCTTCGCAAGCTTTGGGGTGAATGCTTTTCATGATTAGCAGTTCGCTTATGAGAAACCCCGAAAAGATCAAG GAGTTGATCGAGGGTTCATTGGAGCATCACCATCATCATTAAT
SSD_656	CCCCGTTACCTTAAGGGCATGTTAGAAGACGTGGTACAGCTGAGTTTACGCCGGCCATCTGTTACGCAAGTCGCCAACGA CCGATAATCTCCCTGAATGAGAGAATCTTGAATTCAATAAGCGGAATATCACCGCGATTATAGCGCTGTATGCACGCAAG GGACCTTTAGGCTTAGACGTAGAACGTGATCCGATTGAATACGCGAAGTATATGGAGCCTTACGCGGTGCGGCTTAGTATT AAGACGGAAGAGAAATACAACGATGGCTCCTACGAGATGCTTCGGAAGATCGCGAGTTCTGTGTCTATACCTATTCTCATG CAAGATTACATAGTAAAGGAATCCCAAATAGATGACGCTTACAACCTTGAGAGCTGATACCGTTGCCCTCTACGTCGAGATT TTGACTGAAAGAGAATTAGAGAGTTTGTAGAGTACGCGCGAGGATATGGCATGGAGCCATTAATTATCATCAACGACGAG AATGACCTTGACATAGCGTTGAGAATCGGCGCGCGCTTTATTACAATATTCTCAATTAACCTTCGAGACTGGCGAAATTAAC AAGGAAAACCGCGGAACTCATATCCATGATACCTTCCAACGTCGTCAAGGTACCGATGCTTGACTTCTTCGAACCGAAC GAGATCGAAGAGCTGCGGAAGTTAGGAGTCAACGCTTTTCATGATTTCGTCCAGCCTTATGCGGAATCCGAGAGAAAATAAA GAGCTTATAGAGGGATCTCTGGAGCATCATCACCACCACCACTAA
SSD_370	CCAAGATACCTGAAAGGCCAGCTCGAAGACGTGGTCCAACCTGTCCCTTCGCCGCCCTTCCGTGACGCGAGCCGCCAGCGG CCGATAATTTCAATTAAATGAGCGCATCCTGGAGTTTAAACAAGCGTAACATCACCGCGATTATAGCTTTGTATGCGCGCAAG GGCCCGAACGGTTTAGACGTCGAGCGCGACCCATCGAATATGCAAAGTACATGGAACCGTACGCAGTGGGACTGTCAATT AAGACCGAAGAGAAATATAATGACGGCTCTTATGAGATGTTACGCAAGATTGCTTCCAGCGTGAGTATCCCAATTTTAAATG CAAGATTACATTGTGAAGGAGAGTCAAATAGATGATGCCATACAATCTTGGCGCCGACACCGTAGCGCTGTACGTGCAAAAT CTTACTGAACGGGAGTTGGAGAGTCTGCTCGAATATGCACGCGGGTATGGTATGGAGCCACTCATTATTATTAACGATGAG AATGATCTTGACATAGCATGCGGATCGGCGCTAGATTTATCACTATCCATAGCGTCAACTTCGAGACGGGCGAGATTAAT AAAGAGAATCAACGCAAGCTTATCTCTATGATACCATCAAATGTGGTGAAGGTGCCTTGTTGGACTTCTTCGAGCCTAAC GAAATCGAAGAGTTGCGTAAGTTAGGCGTGAATGCATTGATTAGTTCTTCTTAATGCGTAACCCAGAAAAGATCAAG GAGTTAATCGAGGGCTCCCTGGAACACCACCATCACCATCACTGA
SSD_897	CCCCGATACCTTAAGGGAATGTTGGAAGATGTAGTACAGCTCAGTTTGCGTAGACCGTCGGTACACGCGAGTCGACAGCGC CCGATTATCAGCCTGAACGAGAGAATCCTGGAGTTCAATAAGCGGAATATAACTGCGATTATAGCGATGTACGCTCGCAAG GGGCCGATCGGCCTTGATGTCGAACGCGACCCCATAGAATACGCGAAGTACATGGAACCGTATGCTGTAGGGTTATCCATC AAGACAGAGGAAAAGTACAATGATGGCTCTTACGAGATGCTCGGGAAGATAGCCAGCTCAGTCTCTATTCCAAATCCTGATG ATGGACTACATTGTAAAGGAGTCACAAATTGATGACGCCATACAATTTAGGTGCAGATACTGTAATCTGTACGTGGAAATC CTGACAGAAAGAGAGTTGGAGTCACTTCTGGAGTACGCGCGAGGGTACGGCATGGAACCTTTAATCCAATCAACGACGAG AATGACCTGGACATTGCACTCCGGATCGGGGCCGCTTTTATCACTATCTATAGCATAAAATTCGAGACTGGGGAGATTAAAC AAGGAGAACCAGCGTAAGCTGATTAGCATGATTCCCTCAAACGTCGTTAAGGTACCAATGTTAGACTTCTTCGAACCGAAC GAGATAGAGGAACCTCCGCAAATTTGGGCGTTAATGCCTTTATGATCTCCTCTTCACTGATGCGTAACCCAGAAAAGATCAAG GAATTAATAGAAGGATCCTTAGAACACCATCACCACCATCACTGA
MSD_200_163	CCGAGATATTTGAAGGGACAACCTCGAAGACGTTGTGCAGCTGAGTTTGCGTCGTCCATCAGTACATGCCAGTCGTCAAAGA CCAATTATATCCTTAAACGAGCGTATTTTAGAGTTCAACAAGCGGAACATTACAGCTATAATAGCACTTTACGCCCGGAAG GGGCCGACGGGACTTGATGTAGAGCGCGATCCGATAGAGTACGCGAAGTACATGGAGCCGTACGCAGTTGGACTCTCAATT AAGACGGAAGAAAAGTACATGATGGGTCTTACGAGATGCTGAGAAAGATAGCGTCAAGCGTCAGTATACCGATTCTGATG CAAGATTACATTGTGAAGGAGAGTCAGATCGACGACGCTTACAATTTAGGAGCGGACACCGTGAATCTCTACGTCGAAATC TTGACTGAGCGGGAGCTGGAGAGCCTGTTGGAGTATGCAACGGATACGGTATGGAACCGTTGATCCAAATTAACGACGAG AACGATCTGGACATTGCTTTACGCATCGGTGCCCGGTTTATTACAATCTACTCGGTCAATTTTGAGACTGGGGAGATTAAAC AAAGAAAATCAGCGTAAGCTCATTAGCATGATACCGTCAAACGTTGGTAAAGGTCCCTTGTTGGACTTCTTCGAGCCAAAT

GAGATAGAGGAATTACGTAAGTTGGGTGTTAACGCCTTTATGATTAGTTCCTCACTTATGCGTAATCCAGAAAAGATAAAG
GAGTCATCGAGGGAAGCCTGGAGCACCATCATCACCATCACTAG

Supplementary Table 4.9. Amino-acid positions optimized during computational design of Michaelases

Design objective	Designed residues	Allowed amino acids
Catalytic lysine	83	K
TS ₄ tyrosine	112	Y
TS ₆ tyrosine	133	Y
Transition-state packing	8 51 53 89 131 159 180 182 210	H, T, N, Q, A, V, I, L, M, F, Y, W
Loop positions	57 59 60 84 85 86 87 88 90 111 183 184 185 186 187	H, T, N, Q, A, V, I, L, M, F, Y, W, S, G
Floated positions	56 58	WT

4.11. References

1. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
2. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).
3. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. *De Novo* Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
4. Bjelic, S., Nivón, L. G., Çelebi-Ölçüm, N., Kiss, G., Rosewall, C. F., Lovick, H. M., Ingalls, E. L., Gallaher, J. L., Seetharaman, J., Lew, S., Montelione, G. T., Hunt, J. F., Michael, F. E., Houk, K. N. & Baker, D. Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita–Baylis–Hillman Reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).
5. Villali, J. & Kern, D. Choreographing an enzyme’s dance. *Curr. Opin. Chem. Biol.* **14**, 636–643 (2010).
6. Kumar, N. & Gupta, P. DFT Struggles to Predict the Energy Landscape for Iron Pyridine Diimine-Catalyzed [2 + 2] Cycloaddition of Alkenes: Insights into the Problem and Alternative Solutions. *J. Phys. Chem. A* **128**, 4114–4127 (2024).
7. Price, K. E., Broadwater, S. J., Jung, H. M. & McQuade, D. T. Baylis–Hillman Mechanism: A New Interpretation in Aprotic Solvents. *Org. Lett.* **7**, 147–150 (2005).
8. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A Generic Program for Multistate Protein Design. *PLoS ONE* **6**, e20937 (2011).
9. Sevy, A. M., Jacobs, T. M., Crowe, J. E. & Meiler, J. Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLOS Comput. Biol.* **11**, e1004300 (2015).

10. Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Sci.* **21**, 1241–1252 (2012).
11. Zixuan, L. Michael addition reaction and its examples. *Appl. Comput. Eng.* **NA**, NA-NA (2023).
12. Mather, B. D., Viswanathan, K., Miller, K. M. & Long, T. E. Michael addition reactions in macromolecular design for emerging technologies. *Prog. Polym. Sci.* **31**, 487–531 (2006).
13. Michael. Über die Addition von Natriumacetessig-und Natriummalonsäureäthern zu den Methylene der aromatischen α - und β -ungesättigten Säuren. *J. Prakt. Chem* **35**, 349–356 (1887).
14. Adams, S. S. & Nichol森, J. S. The synthesis of ibuprofen. *J. Chem. Educ.* **43**, 581–584 (1966).
15. Miao, Y., Rahimi, M., Geertsema, E. M. & Poelarends, G. J. Recent developments in enzyme promiscuity for carbon–carbon bond-forming reactions. *Curr. Opin. Chem. Biol.* **25**, 115–123 (2015).
16. Geertsema, E. & Poelarends, G. Enzymatic carbon-carbon bond-forming Michael-type additions. *Biocatal. Org. Synth.* **2**, 313–334 (2014).
17. Garrabou, X., Beck, T. & Hilvert, D. A Promiscuous *De Novo* Retro-Aldolase Catalyzes Asymmetric Michael Additions via Schiff Base Intermediates. *Angew. Chem. Int. Ed.* **54**, 5609–5612 (2015).
18. Garrabou, X., Macdonald, D. S., Wicky, B. I. M. & Hilvert, D. Stereodivergent Evolution of Artificial Enzymes for the Michael Reaction. *Angew. Chem. Int. Ed.* **57**, 5288–5291 (2018).
19. De Raffe, D., Martí, S. & Moliner, V. QM/MM Theoretical Studies of a *De Novo* Retro-Aldolase Design. *ACS Catal.* **9**, 2482–2492 (2019).
20. De Raffe, D., Martí, S. & Moliner, V. Understanding the Directed Evolution of *De Novo* Retro-Aldolases from QM/MM Studies. *ACS Catal.* **10**, 7871–7883 (2020).
21. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci.* **103**, 16710–16715 (2006).
22. Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.* **31**, 904–916 (2010).

23. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Computational Enzyme Design: Successes, Challenges, and Future Directions. in *Modern Biocatalysis* (eds. Williams, G. & Hall, M.) 88–116 (Royal Society of Chemistry, Cambridge, 2018). doi:10.1039/9781788010450-00088.
24. Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S. & Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).
25. Obexer, R., Godina, A., Garrabou, X., Mittl, P. R. E., Baker, D., Griffiths, A. D. & Hilvert, D. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
26. Braun, M., Tripp, A., Chakatok, M., Kaltenbrunner, S., Totaro, M. G., Stoll, D., Bijelic, A., Elaily, W., Hoch, S. Y. Y., Aleotti, M., Hall, M. & Oberdorfer, G. Computational design of highly active *de novo* enzymes. Preprint at <https://doi.org/10.1101/2024.08.02.606416> (2024).
27. Lauko, A., Pellock, S. J., Anischanka, I., Sumida, K. H., Juergens, D., Ahern, W., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I. R., Jamieson, C., Kang, A., Brackenbrough, E., Bera, A. K., Sankaran, B., Houk, K. N. & Baker, D. Computational design of serine hydrolases. Preprint at <https://doi.org/10.1101/2024.08.29.610411> (2024).
28. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Robb, M. A., Cheeseman, J. R. & Scalmani, G. Gaussian 16, Revision C.01.
29. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).
30. Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
31. Lassila, J. K., Baker, D. & Herschlag, D. Origins of catalysis by computationally designed retroaldolase enzymes. *Proc. Natl. Acad. Sci.* **107**, 4937–4942 (2010).

32. Tang, Y. T. & Marshall, G. R. PHOENIX: A Scoring Function for Affinity Prediction Derived Using High-Resolution Crystal Structures and Calorimetry Measurements. *J. Chem. Inf. Model.* **51**, 214–228 (2011).
33. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
34. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10172–10177 (1997).
35. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. (2015).
36. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
37. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel β -sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state.
38. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O. & Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).
39. He, X., Man, V. H., Yang, W., Lee, T.-S. & Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J. Chem. Phys.* **153**, 114502 (2020).
40. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
41. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

42. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation¹. *J. Mol. Biol.* **285**, 1735–1747 (1999).
43. Evans, D. J. & Holian, B. L. The Nose–Hoover thermostat. *J. Chem. Phys.* **83**, 4069–4074 (1985).
44. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
45. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

CHAPTER 5. DISCUSSION & CONCLUSIONS

5.1. Summary of Findings

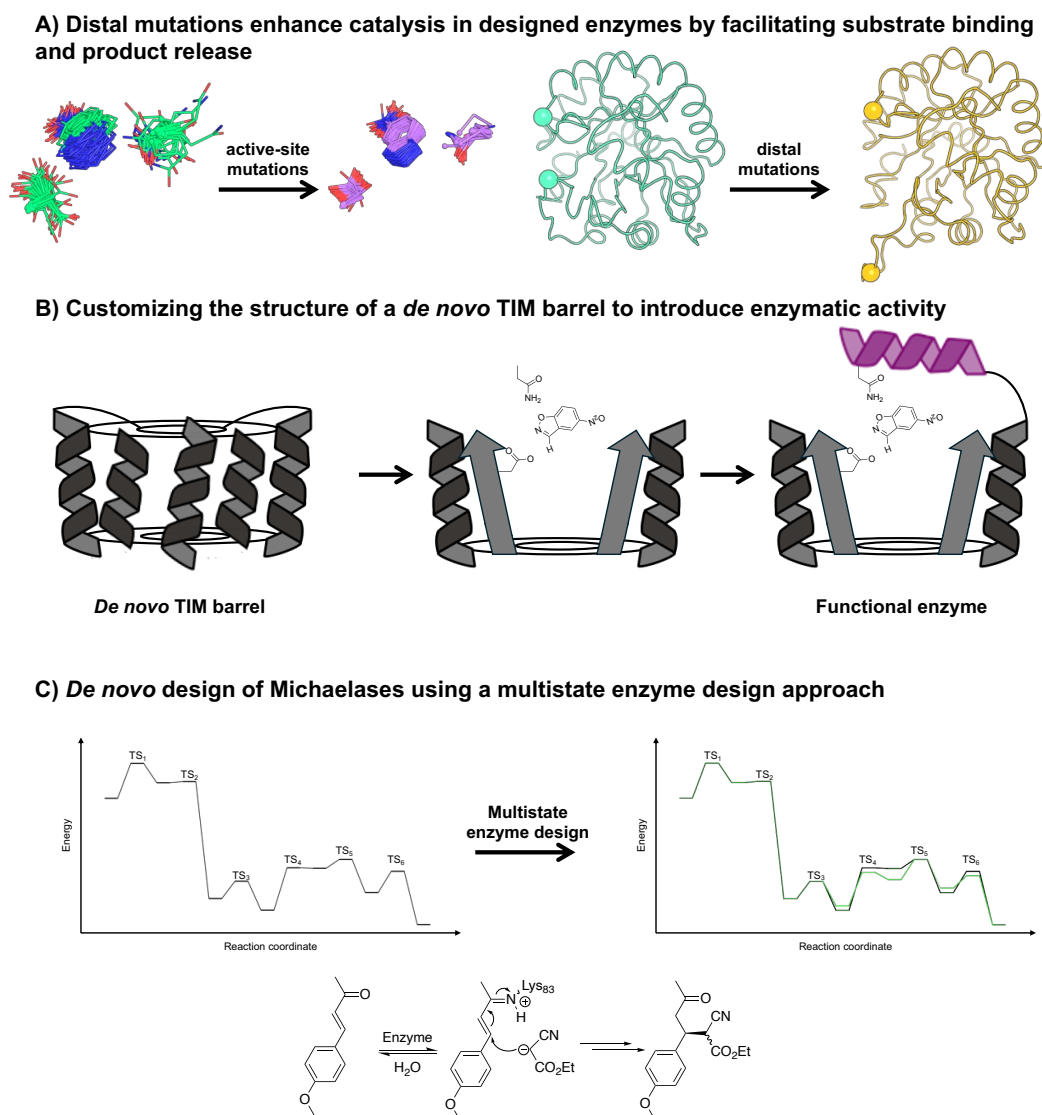


Figure 5.1. Summary of findings. (A) In Chapter 2, we observed that active-site mutations shift the conformational ensemble towards more productive substrates and create active sites ordered for efficient catalysis. Distal mutations reorganize surface loops and widen tunnels to the active site and in doing so, facilitate substrate binding and product release. (B) In Chapter 3, we developed a workflow to customize the structure of a simple *de novo* TIM barrel to introduce Kemp elimination activity. We used machine learning tools to generate additional structural segments that act as a lid and could accommodate substrate binding and product release. (C) In Chapter 4, we employed multistate design to design enzymes for a carbon-carbon bond forming reaction by stabilizing multiple conformational and chemical states.

The ability to create enzymes to catalyze any reaction on demand holds significant potential in synthetic chemistry and industrial biocatalysis. Over the past two decades, numerous enzymes have been designed for various model organic transformations.¹⁻⁵ These designed enzymes typically display catalytic efficiencies orders of magnitude lower than their natural counterparts, suggesting that current enzyme design methods must be improved to enable the design of highly efficient enzymes.⁶ In this thesis, our objective is to address current challenges facing enzyme design.

Directed evolution has been used to increase the activities of designed enzymes by up to several orders of magnitude.⁷⁻¹⁴ Mutations found during evolution appear both within the active site and at distal positions. The role of distal mutations in enhancing the catalytic efficiency of enzymes is poorly understood. Furthermore, enzyme design algorithms primarily focus on constructing the active site and distal sites are not designed as it is very challenging to identify distal residues that contribute to catalysis. Considering the well-documented evolutionary trajectories of *de novo* designed and *in vitro* evolved enzymes and the fact that mutations were selected by evolution specifically for their ability to enhance catalytic activity, these enzymes allow us to study the effects of active-site and distal mutations without complications of the evolutionary history of natural enzymes. **In Chapter 2**, we created variants of Evolved enzymes containing either active-site or distal mutations, referred to as Core and Shell variants, respectively. Kinetic analyses, X-ray crystallography, and molecular dynamics simulations revealed that active-site mutations are sufficient to form ordered active sites for efficient catalysis. The Core variants, containing only active-site mutations, displayed catalytic efficiencies several orders of magnitude greater than the original design. However, distal mutations, enhance catalytic activity by facilitating substrate entry and product release through reorganizing surface loops and widening

tunnels to the active site, (Figure 5.1A) and their addition on Designed enzymes had relatively small impacts on catalytic efficiency. Given that that active-site mutations alone can create highly efficient active sites, designing efficient enzymes should be achievable by focusing solely on the active site. However, previous enzyme design efforts using single-state design strategies with fixed backbone approximations have fallen short of generating optimal active sites.¹⁻⁴ By contrast, ensemble-base design approaches that incorporate backbone flexibility in the design procedure, have enabled creation of enzymes with efficiencies comparable to those obtained through multiple rounds of directed evolution.^{15,16}

In **Chapter 3**, we tackle the challenge of the availability of a protein scaffold well-suited to the reaction of interest. We hypothesized that a minimal *de novo* TIM barrel could serve as a blank template which can be customized for any desired reaction. Using a combination of physics-based and machine-learning tools, we developed a workflow to tailor a functionally inert TIM barrel with short, minimal loops to introduce enzymatic activity. We altered the structure of *de novo* TIM barrels to introduce a lid fragment that shields the active site from the solvent (Figure 5.1B). We employed a combination of machine-learning tools such as RFdiffusion for generating structural segments, ProteinMPNN for sequence design, and AlphaFold2 for structure validation. We experimentally characterized 20 designs. One design, KempTIM 4, catalyzes the Kemp elimination reaction with a catalytic efficiency of $10 \text{ M}^{-1}\text{s}^{-1}$. Molecular dynamics analyses revealed that KempTIM 4 exhibits greater flexibility in the lid region and a loop near the active site compared to the inactive variants. Furthermore, in contrast to inactive designs, KempTIM 4 contains wider tunnels to the active site. We hypothesize that this flexibility is required for substrate binding and product release. In future, we aim to apply the same workflow to other reactions, testing its generalizability.

Kemp elimination is a concerted reaction with a single transition state. Designing enzymes for multistep reactions is more complicated as stabilizing a single transition state is insufficient to create efficient enzymes for multistep reactions. An efficient enzyme must ensure low energy barriers throughout the reaction pathway. In **Chapter 4**, we present the first instance of multistate enzyme design, where we designed enzymes for a Michael addition reaction involving two substrates and six chemical steps. We designed Michaelases by stabilizing two chemical steps and employing an ensemble-based design approach (Figure 5.1C). Kinetic analysis revealed that all 12 designs catalyze the Michael addition reaction. However, they display lower catalytic activity compared to the starting scaffold. The diminished activity was due to lower turnover, and the designed variants displayed improved K_M^{app} for the Michael acceptor, implying enhanced substrate binding. We hypothesize that the lower turnovers might be due to disproportionate stabilization of the two targeted steps compared to other steps in the reaction. In the future, we aim to improve these designs through additional rounds of design and refinement.

5.2. Discussion

Enzymes catalyze virtually all chemical reactions in nature. The goal of enzyme design is to create new enzymes with functions that extend beyond those of natural enzymes. The ability to design enzymes for specific reactions holds significant potential in synthetic chemistry and industrial biocatalysis. However, currently, even for simple model reactions, the success rate is low, and the few active designed enzymes display low catalytic efficiencies, necessitating directed evolution to enhance their activities. In Chapter 2, we showed that active-site mutations are the main contributors to increased activities and are sufficient to form an ordered active site for efficient catalysis. To design these efficient active sites, traditional single-state design approaches using a fixed backbone template are inadequate, and backbone flexibility must be incorporated. Ensemble-based design approaches, where an ensemble of backbone templates is used, can create these highly efficient active sites.^{15,16} To achieve optimal catalytic efficiency rivaling those of natural enzymes, the incorporation of distal mutations that fine-tune the enzyme's conformational ensemble and facilitate substrate binding and product release is essential. However, it is challenging to determine *a priori* which positions to target. Various approaches have been developed to identify distal positions that contribute to catalysis. One such method is the Shortest Path Map (SPM), which constructs a graph based on correlated movements of residues and inter-residue distances from long time-scale molecular dynamics simulations. A map identifying pairs of residues with higher contributions to the enzyme's hypothetical allosteric communication pathway is generated. The algorithm then finds the shortest path from the first to the last residue.^{17–19} SPM has been successfully applied to a range of unrelated enzymes to predict distal mutation hotspots targeted by directed evolution. For instance, the RA95 series of retro-aldolases,²⁰ Monoamine Oxidase from *Aspergillus niger* (MAO-N)²¹ and Tryptophan synthase (TrpS) from

Pyrococcus furiosus.²² Another method of identifying distal mutation hotspots is the NMR-guided selection of candidate positions, which relies on chemical shift perturbations.²³ When a transition state analogue binds, residues requiring significant reorganization to increase the population of a rotamer to accommodate the TSA exhibit substantial shifts in their chemical shifts. These large perturbations can be used to identify distal mutation hotspots. This approach was used in the generation of Kemp eliminases from Myoglobin by creating saturation mutagenesis libraries of all positions with high chemical shift perturbation. All mutants catalyzed Kemp elimination reaction with a 2- to 71-fold increase in catalytic efficiency. Nine out of 19 productive mutations were located far from the active site.²³ In another study, an algorithm that integrates bioinformatics and structure-based approaches was used to identify distal hotspots capable of influencing the enzyme's conformational dynamics and the population of catalytically productive substrates. This method was applied to lactococcal multidrug resistance regulator (LmrR), where two variants with mutations 11 Å away from the active site increased catalytic efficiency towards hydrazone formation reaction. Their combination resulted in a 66% higher turnover.²⁴ While these methods are effective at predicting distal mutation hotspots, they do not identify the nature of mutations. For that, techniques such as site-saturation mutagenesis and ancestral sequence reconstruction could be employed.

In Chapter 3, we applied machine learning methods to customize a functionally inert TIM barrel to introduce Kemp eliminase activity. Our results highlight the limitations of current machine learning methodologies, particularly regarding the precision required for efficient catalysis. While AlphaFold2 predictions have achieved an atomic level of accuracy, this is not sufficient for optimal catalysis. Highly efficient catalysis requires a sub-angstrom level of precision to effectively discriminate the transition state from the ground state, as even slight

angstrom-level deviations in side-chain positioning can severely compromise catalytic function.²⁵ Furthermore, an efficient enzyme must exhibit conformational flexibility to efficiently exchange between conformational substates required for binding of the substrates, chemical transformations, and release of the product. Current machine learning tools have achieved a high success rate in designing rigid, static structures. RFdiffusion can generate highly thermostable protein folds that are dissimilar to the ones in the PDB, with diverse topologies, including alpha, beta, and mixed alpha-beta folds.²⁶ ProteinMPNN has a high native sequence recovery and has been successfully used in enhancing protein expression, thermostability and function.²⁷ The success of machine learning methods in generating stable, rigid structures is logical, as the training data consists of static structures. Incorporating dynamics into these models requires extensive high-resolution data related to protein dynamics to train models, which is currently not available.²⁸

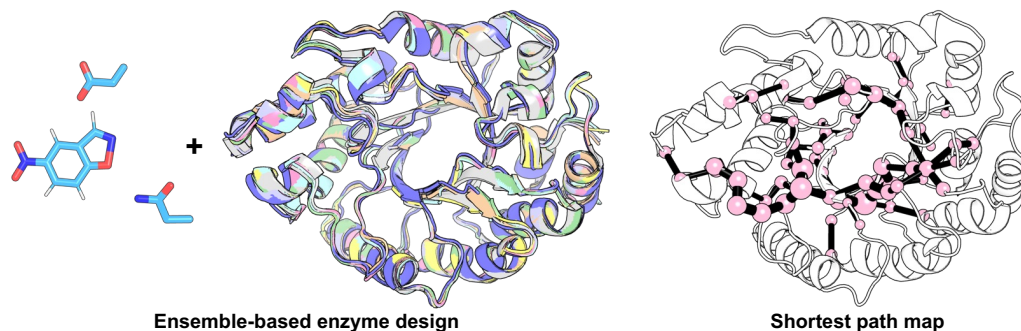
In Chapter 4, we applied multistate enzyme design to create Michaelases by stabilizing multiple chemical and conformational states. Although all designed variants exhibited Michaelase activity, their reduced turnovers implied that disproportionate stabilization of specific steps could be detrimental to catalytic efficiency. Therefore, we hypothesize that to design efficient enzymes for multistep reactions, the entire catalytic cycle needs to be incorporated in design calculations, including the free enzyme, the Michaelis complex, the enzyme-bound transition state for each chemical step, and the product-bound state. In designing enzymes for multistep reactions, a balance needs to be maintained in designing a highly ordered active site for efficient catalysis and maintaining conformational plasticity to sample different rotamers to accommodate and stabilize different transition states. Another aspect of multistate enzyme design that could be improved is by using more sophisticated theozymes. Typically, theozymes are generated by using truncated residues thought to be essential for catalysis. Previous enzyme design efforts utilized such simple

theozymes. However, directed evolution of designed enzymes often resulted in the development of complex active sites containing residues that interact with and orient catalytic residues.^{11,25} Using more sophisticated theozymes could lead to the generation of enzymes with higher activities.

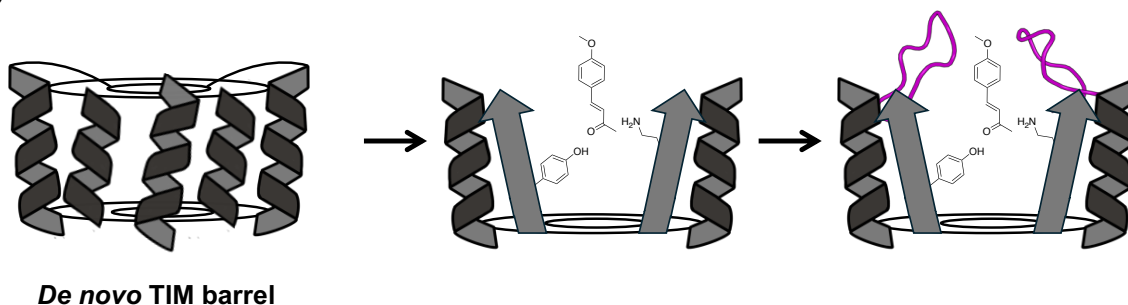
This thesis contributes to our understanding of enzyme catalysis and design. We have shown that active-site and distal mutations exert distinct effects on catalysis. Active-site mutations can create ordered active sites, and efficient enzymes can be constructed by focusing solely on the active site. On the other hand, distal mutations contribute to enhanced catalysis by fine-tuning enzyme dynamics and facilitating substrate binding and product release. Furthermore, the customization of a *de novo* TIM barrel scaffold for the Kemp elimination reaction underscores the potential of minimal, functionless scaffolds as blank templates customizable for various reactions, providing an alternative to repurposing natural enzymes with complexities arising from their evolutionary history. Finally, utilization of a multistate approach to design enzymes for the Michael addition reaction further pushes the boundaries of enzyme design. Our results suggest that it is possible to design sequences capable of stabilizing multiple chemical and conformational states; however, to achieve efficient catalysis, every step of the entire reaction pathway should be included in design calculations. Collectively, this work advances our understanding of enzymes, their design, evolution and catalysis. We hope this work lays the groundwork for creating highly efficient enzymes tailored for industrial and synthetic chemistry applications.

5.3. Future Directions

A) Prediction of distal mutations



B) Customize *de novo* TIM barrels for various reactions



C) Include the entire reaction coordinate in design calculations & construct more sophisticated theozymes

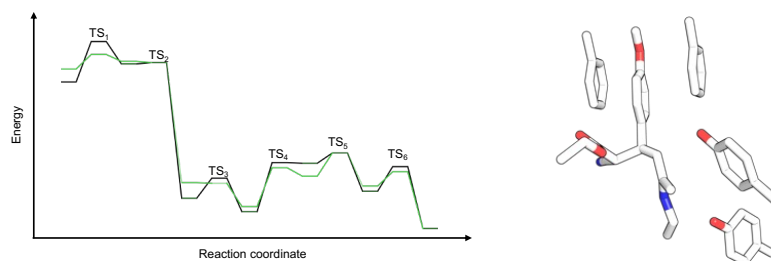


Figure 5.2. Future directions. In future we aim to (A) design efficient active sites using ensemble-based enzyme design and incorporate distal residues via methods such as shortest path map to create highly efficient enzymes, (B) customize *de novo* TIM barrels for diverse reactions, building on the insights from Chapter 3, by evaluating lid flexibility and accounting for substrate entry and product release, and (C) enhance Michaelase enzymes by including the entire reaction pathway in design calculations and creating more sophisticated theozymes with additional residues to provide further stabilization.

In future work, we aim to improve current design methods to incorporate the prediction of distal mutations. We hypothesize that using an ensemble-based design approach, in combination

with methods such as shortest path map (SPM), will enable the creation of highly efficient active sites ordered for catalysis while also incorporating distal mutations that facilitate substrate binding and product release. Our design strategy will begin by using an ensemble to generate active site structures optimized for catalysis. Next, we will use SPM to identify potential distal mutation hotspots. To determine the nature of these mutations, we will leverage methods such as ProteinMPNN,²⁷ Funclib,²⁹ or multiple sequence alignment.³⁰

Furthermore, we aim to apply the framework developed in Chapter 3 to other enzymatic reactions. We hypothesize that the *de novo* TIM barrels used in our work have the potential to be customized for a variety of desired reactions. To achieve this, we will integrate the sampling of protein dynamics early in the design process to generate flexible lid structures capable of accommodating substrate binding and product release. Furthermore, we will take advantage of the latest iterations of machine-learning tools for this purpose.^{31–33}

The Michaelases designed in Chapter 4 also present an opportunity for further refinement. By targeting additional transition states for stabilization and remodelling surface loops using diffusion methods to accommodate various intermediates, we aim to improve the rate acceleration of these enzymes. Previous studies have shown that iterative rounds of design and refinement can significantly improve the catalytic efficiency of designed enzymes.^{2,34} Moreover, our designs may benefit from the development of more sophisticated theozymes, wherein additional residues, such as those that can stabilize transition states through interactions like π -stacking with aromatic rings, are incorporated. Another strategy we plan to explore is negative design, where we destabilize the ground state relative to the transition states to improve catalytic efficiency. We will employ a holistic approach where every conformational and chemical state sampled by the enzyme is included in design calculations.

Ultimately, the goal of enzyme design is to create highly efficient enzymes from scratch. Building on the work discussed in this thesis, we aim to move closer to achieving this goal by integrating the methods and insights gained here. Our proposed approach includes: first, computing sophisticated theozymes that incorporate additional stabilizing residues alongside catalytic residues; second, generating ensembles of inert *de novo* TIM barrels and sampling the entire reaction coordinate in the design calculations. Including the free enzyme, the substrate-bound enzyme, the transition states for each chemical step, and the product bound-enzyme; third, utilizing machine learning approaches to design lid structures that sequester the active site from solvent and provide packing to ligands, and verifying the flexibility of lid structures by molecular dynamics simulations. Finally, we will use shortest path map to identify distal residues linked to catalysis and apply sequence alignment frequencies or ProteinMPNN to introduce distal mutations.

5.4. A paradigm shift in enzyme design

Protein design, which refers to design of amino acid sequences that fold to a desired three-dimensional structure, has traditionally been approached as an optimization problem. Physics-based methods, which rely on potential energy functions to navigate the protein fitness landscape, have long been used to identify minimum energy conformations. However, in recent years, the field of protein design has transitioned to a new way of thinking and is shifting away from physics-based approaches. The rapid improvements in computing power, coupled with advancements in software and hardware systems, combined with the accumulation of vast experimental data on protein sequences³⁵ and structures,³⁶ have paved the way for emergence of data-driven approaches to protein structure generation models,^{26,37,38} protein sequence design tools,^{27,39,40} and protein structure prediction algorithms.^{41–43} Deep learning methods have revolutionized the field of protein design, and this field is transitioning away from traditional physics-based approaches.²⁸ Machine learning tools are now offering alternative ways to design protein structure, sequence and function. Reflecting these advancements, part of the 2024 Nobel Prize in Chemistry was awarded to pioneers in computational protein design and protein structure prediction. Despite the power of these methods in generating stable, well-folded, and static structures, their application in enzyme design is hindered by limitations, including the absence of sub-angstrom level of precision and the inability to account for conformational flexibility in designs.

Despite these limitations, machine learning tools are evolving rapidly. During our work on customizing a *de novo* TIM barrel for the Kemp elimination reaction, as described in Chapter 2, an all-atom version of RFDiffusion, RFDiffusion AA, was developed, which can generate protein structures able to bind small molecules.³¹ Similarly, LigandMPNN, a deep learning neural network model, was developed to design protein sequences capable of interacting with non-protein atoms

and small molecules. Experimental characterization demonstrated that LigandMPNN can generate small molecule and DNA-binding proteins with high affinity and specificity.³² Additionally, the new version of AlphaFold2, AlphaFold3, is updated to predict structures of proteins alongside small molecules.³³ While these recent developments are exciting and hold great promise, their application in enzyme design should be approached with caution. Although these methods can generate proteins capable of binding small molecule ligands,³¹ efficient catalysis requires more than just binding. It demands precise positioning of catalytic residues, stabilization of transition states, and conformational flexibility to sample various substates, which remain challenging for these methods to achieve.

Two recent studies have used RFdiffusion to generate *de novo* templates capable of accommodating catalytic motifs. Lauko *et al.*³⁴ applied RFdiffusion to design protein structures that scaffold the active site of serine hydrolases, comprising a catalytic triad and oxyanion hole. After iterative rounds of design, refinement, and testing over 300 designs, only two exhibited multiple turnover activity. While the catalytic efficiencies of these designs are higher than those of previously designed serine hydrolases, the success rate remains low as hundreds of designs had to be tested. Furthermore, their activities are orders of magnitude lower than natural hydrolases. Similarly, Braun *et al.*⁴⁴ used RFdiffusion to design enzymes for the retro-aldol reaction by incorporating the highly optimized catalytic tetrad of evolved retro-aldolases, obtained after multiple rounds of directed evolution. They generated alpha-helical proteins able to accommodate the catalytic motif. Out of 35 experimentally tested designs, 31 displayed retro-aldolase activity, with higher catalytic efficiencies than previously designed retro-aldolases. However, their activities were still lower than those of evolved retro-aldolases, primarily due to lower k_{cat} values.

The most active design displayed a k_{cat} of $3.1 \times 10^{-2} \text{ s}^{-1}$, while the evolved retro-aldolase RA95.5-8F exhibits a 300-fold higher k_{cat} of 10 s^{-1} .

While these examples showcase the potential of deep learning approaches in enzyme design, they suffer from low success rates, and the catalytic activity of the successful designs remains orders of magnitude lower than that of native enzymes. To achieve optimal catalytic efficiency, enzyme dynamics needs to be considered, as static models do not fully capture the flexibility needed for efficient catalysis. One powerful method for sampling dynamics is physics-based molecular dynamics (MD) simulation. However, comprehensive sampling of the conformational ensemble is only feasible for small systems and is limited by high computational costs. Recent advancements in deep learning algorithms have given rise to enhanced sampling methods.⁴⁵ For instance, Tao *et al.* developed a deep learning sampling method capable of predicting large conformational changes in shorter simulation times compared with conventional MD.⁴⁶ Incorporating deep learning-assisted MD simulations could be an effective method to sample the conformational dynamics required for more accurate enzyme design.

Overall, we anticipate that data-driven approaches will continue to advance the field of enzyme design. While there are many challenges remaining ahead, it is conceivable that an automated enzyme design approach could allow for creation of protein sequences with desired catalytic functions from scratch. However, we believe that the reliance on directed evolution to optimize the catalytic efficiency of designed enzymes will persist. Nonetheless, the future of enzyme design holds great promise, with the potential for groundbreaking innovations on the horizon.

5.5. References

1. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
2. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).
3. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. *De Novo* Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
4. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E. & Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
5. Bjelic, S., Nivón, L. G., Çelebi-Ölçüm, N., Kiss, G., Rosewall, C. F., Lovick, H. M., Ingalls, E. L., Gallaher, J. L., Seetharaman, J., Lew, S., Montelione, G. T., Hunt, J. F., Michael, F. E., Houk, K. N. & Baker, D. Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita–Baylis–Hillman Reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).
6. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Computational Enzyme Design: Successes, Challenges, and Future Directions. in *Modern Biocatalysis* (eds. Williams, G. & Hall, M.) 88–116 (Royal Society of Chemistry, Cambridge, 2018). doi:10.1039/9781788010450-00088.
7. Khersonsky, O., Röthlisberger, D., Wollacott, A. M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K. N., Baker, D., & Tawfik, D. S. Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *Journal of molecular biology*, **407**, 391–412 (2011).

8. Khersonsky, O., Kiss, G., Röthlisberger, D., Dym, O., Albeck, S., Houk, K. N., Baker, D., & Tawfik, D. S. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci.*, **109**, 10358–10363 (2012).
9. Blomberg, R., Kries, H., Pinkas, D. M., Mittl, P. R. E., Grütter, M. G., Privett, H. K., Mayo, S. L. & Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
10. Althoff, E. A., Wang, L., Jiang, L., Giger, L., Lassila, J. K., Wang, Z., Smith, M., Hari, S., Kast, P., Herschlag, D., Hilvert, D. & Baker, D. Robust design and optimization of retroaldol enzymes: Robust Design and Optimization of Retroaldol Enzymes. *Protein Sci.* **21**, 717–726 (2012).
11. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N. & Hilvert, D. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
12. Bunzel, H. A., Kries, H., Marchetti, L., Zeymer, C., Mittl, P. R. E., Mulholland, A. J. & Hilvert, D. Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **141**, 11745–11748 (2019).
13. Obexer, R., Godina, A., Garrabou, X., Mittl, P. R. E., Baker, D., Griffiths, A. D. & Hilvert, D. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
14. Crawshaw, R., Crossley, A. E., Johannissen, L., Burke, A. J., Hay, S., Levy, C., Baker, D., Lovelock, S. L. & Green, A. P. Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
15. Broom, A., Rakotoharisoa, R. V., Thompson, M. C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J. S. & Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).

16. Rakotoharisoa, R. V., Seifinoferest, B., Zarifi, N., Miller, J. D. M., Rodriguez, J. M., Thompson, M. C. & Chica, R. A. Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J Am Chem Soc* **146**, 10001–10013 (2024).
17. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comput. Mol. Sci.* **11**, (2021).
18. Duran, C., Casadevall, G. & Osuna, S. Harnessing conformational dynamics in enzyme catalysis to achieve nature-like catalytic efficiencies: the shortest path map tool for computational enzyme redesign. *Faraday Discuss.* **252**, 306–322 (2024).
19. Casadevall, G., Casadevall, J., Duran, C. & Osuna, S. The shortest path method (SPM) webserver for computational enzyme design. *Protein Eng. Des. Sel.* **37**, gzae005 (2024).
20. Maria-Solano, M. A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J. & Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun.* **54**, 6622–6634 (2018).
21. Curado-Carballada, C., Feixas, F., Iglesias-Fernandez, J. & Osuna, S. Hidden Conformations in *Aspergillus niger* Monoamine Oxidase are Key for Catalytic Efficiency. *Angew Chem Int Ed* **58**, 3097–3101.
22. Maria-Solano, M. A., Kinateter, T., Iglesias-Fernández, J., Sterner, R. & Osuna, S. *In Silico* Identification and Experimental Validation of Distal Activity-Enhancing Mutations in Tryptophan Synthase. *ACS Catal.* **11**, 13733–13743 (2021).
23. Bhattacharya, S., Margheritis, E. G., Takahashi, K., Kulesha, A., D'Souza, A., Kim, I., Yoon, J. H., Tame, J. R. H., Volkov, A. N., Makhlynets, O. V. & Korendovych, I. V. NMR-guided directed evolution. *Nature* **610**, 389–393 (2022).
24. Casilli, F., Canyelles-Niño, M., Roelfes, G., & Alonso-Cotchico, L. (2024). Computation-guided engineering of distal mutations in an artificial enzyme. *Faraday discussions*, **252**, 262–278 (2024).

25. Lovelock, S. L., Crawshaw, R., Basler, S., Levy, C., Baker, D., Hilvert, D. & Green, A. P. The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).
26. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., ... Baker, D. *De novo* design of protein structure and function with RFdiffusion. *Nature*, **620**, 1089–1100 (2023).
27. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., De Haas, R. J., Bethes, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., & Baker, David. Robust deep learning based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
28. Liu, H. & Chen, Q. Computational protein design with data-driven approaches: Recent developments and perspectives. *WIREs Comput. Mol. Sci.* **13**, e1646 (2023).
29. Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D. L., Prilusky, J., Amengual-Rigo, P., Guallar, V., Tawfik, D. S. & Fleishman, S. J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178-186 (2018).
30. Kazan, I. C., Sharma, P., Rahman, M. I., Bobkov, A., Fromme, R., Ghirlanda, G. & Ozkan, S. B. Design of novel cyanovirin-N variants by modulation of binding dynamics through distal mutations. *eLife* **11**, e67474 (2022).
31. Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Kang, A., Brackenbrough, E., Bera, A. K., Baek, M., ... Baker, D. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, **384** eadl2528 (2024)

32. Dauparas, J., Lee, G. R., Pecoraro, R., An, L., Anishchenko, I., Glasscock, C. & Baker, D. Atomic context-conditioned protein sequence design using LigandMPNN. *Preprint at bioRxiv* (2023).
33. Abramson, J. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
34. Lauko, A., Pellock, S. J., Anischanka, I., Sumida, K. H., Juergens, D., Ahern, W., Shida, A., Hunt, A., Kalvet, I., Norn, C., Humphreys, I. R., Jamieson, C., Kang, A., Brackenbrough, E., Bera, A. K., Sankaran, B., Houk, K. N. & Baker, D. Computational design of serine hydrolases. *Preprint at bioRxiv* (2024).
35. The UniProt Consortium *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
36. wwPDB consortium *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
37. Wu, K. E., Yang, K. K., Van Den Berg, R., Alamdari, S., Zou, J. Y., Lu, A. X. & Amini, A. P. Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2024).
38. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A. & Kim, P. M. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst.* **11**, 402-411.e4 (2020).
39. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
40. Anand, N., Eguchi, R., Mathews, I. I., Perez, C. P., Derry, A., Altman, R. B. & Huang, P.-S. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).
41. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., ... Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876 (2021).

42. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. (2023).
43. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
44. Braun, M., Tripp, A., Chakatok, M., Kaltenbrunner, S., Totaro, M. G., Stoll, D., Bijelic, A., Elaily, W., Hoch, S. Y. Y., Aleotti, M., Hall, M. & Oberdorfer, G. Computational design of highly active *de novo* enzymes. *Preprint at bioRxiv* (2024).
45. Zhou, J. & Huang, M. Navigating the landscape of enzyme design: from molecular simulations to machine learning. *Chem. Soc. Rev.* **53**, 8202–8239 (2024).
46. Tian, H., Jiang, X., Xiao, S., La Force, H., Larson, E. C. & Tao, P. LAST: Latent Space-Assisted Adaptive Sampling for Protein Trajectories. *J. Chem. Inf. Model.* **63**, 67–75 (2023).