



uOttawa

Ensemble-based computational enzyme design: methods and applications

Rojo Rakotoharisoa

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

© Rojo Rakotoharisoa, Ottawa, Canada, 2024

Abstract

Designing artificial enzymes to meet the increasing demand for industrial applications holds immense societal significance. However, despite successes in *de novo* enzyme design methodologies, artificial enzymes exhibit low catalytic efficiency compared to natural ones. In this thesis, we studied structural changes occurring during the evolution of a low-activity *de novo* Kemp eliminase HG3 ($k_{\text{cat}}/K_{\text{M}}$ 146 M⁻¹s⁻¹). We observed a rigidification of catalytic residues, a better active site pre-organization and a wider active site entrance. Using backbone ensembles from ensemble refinement (ER) of crystallographic data of HG3, our ensemble-based enzyme design method allowed a high prediction accuracy of HG4, a highly efficient enzyme we engineered ($k_{\text{cat}}/K_{\text{M}}$ 103,000 M⁻¹s⁻¹) containing key first and second-shell mutations found during evolution. Building upon these findings, we established a benchmark of our ensemble-based method using backbone ensembles from various methods obtained from a weakly active Kemp Eliminase, HG3. Our results showed that ensembles derived from PertMin, a method we previously developed, and ER accurately predicted the conformation of its optimized variant, HG4. This benchmark served as a guide for improving the catalytic efficiency of HG3 and KE70 ($k_{\text{cat}}/K_{\text{M}}$ 57 M⁻¹ s⁻¹). Designed sequences displayed catalytic efficiency that improved by 100–250-fold, comparable to screening thousands of variants from rounds of directed evolution. Finally, we benchmarked our ensemble-based design method for predicting to stabilize the multiple transition states of the reaction catalyzed by BH32.12, an evolved variant of BH32 catalyzing the Morita-Baylis-Hillman reaction. Our research advances the field of enzyme design by combining evolutionary insights with computational approaches, leading to the creation of highly efficient artificial enzymes for various applications.

Acknowledgements

First, I would like to thank my supervisor Dr. Roberto Chica for his precious advice and guidance. I would never be grateful enough for that. I would also like to thank him for believing in me from the beginning until the end and for his continuous motivation.

I also thank our collaborators from UCSF, Dr. James Fraser and Dr. Michael Thompson at UC Merced for crystallizing the HG-series of Kemp eliminases, and another thank you to Dr. Michael Thompson and Behnoush Seifinoferest at UC Merced for crystallizing the HG-designs. None of this great would have been achieved without them.

To write this thesis, I got help from many people including Dr. Aron Broom, Dr. Hang Pham, Claudèle Lemay-St-Denis, Niayesh Zarifi and Serena Hunt.

I also want to thank my current lab mates, Serena Hunt, Niayesh Zarifi, Dr. Hang Pham, Jack Miller, and Ilya Dementyev. who helped me through the years, by giving moral support, advice, scripts, and helping edit my thesis.

I also thank past Chica lab members, Dr. Aron Broom, Dr. Marina Cañellas, Dr. Matthew Eason, Dr. Anthony St-Jacques, Dr. Adam Damry, Dr. Jamey Davey, Dr. Jesús Bandas, who helped me when I started the Ph.D. and who also guided me on my studies. I want also to thank my dear friend Sandrine Legault, for her help on experimental work, and for moral support.

Finally, I thank my parents, my big brother, and my little brother for their unconditional love and support.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION	1
1.1. ENZYMES	1
1.1.1. <i>Enzyme reaction and mechanism</i>	2
1.1.1.1. The lock and key and induced fit models	3
1.1.1.2. Michaelis-Menten kinetics model.....	4
1.1.1.3. The Transition State theory.....	5
1.1.2. <i>Enzyme flexibility and function</i>	6
1.1.3. <i>Enzyme design methods and limitations</i>	8
1.1.3.1. Enzyme engineering.....	8
1.1.3.2. Laboratory directed evolution.....	9
1.2. COMPUTATIONAL ENZYME DESIGN	10
1.2.1. <i>Computational protein design</i>	11
1.2.1.1. Rotamer optimization.....	12
1.2.1.2. Scoring function.....	14
1.2.1.3. Sequence optimization and search algorithm	15
1.2.1.4. Multi-state computational design.....	16
1.2.2. <i>Methodologies to design de novo enzymes</i>	18
1.2.2.1. Theozyme placement	20
1.2.2.2. Active site repacking.....	21
1.3. SUCCESS IN COMPUTATIONAL ENZYME DESIGN.....	22
1.3.1. <i>Bond cleaving reactions</i>	24

1.3.1.1.	Single-step reaction: the Kemp elimination.....	24
1.3.1.2.	Multi-step reaction: retro-aldol condensation.....	27
1.3.2.	<i>Bond-forming reactions</i>	29
1.3.2.1.	Single-step reaction: Diels-Alder cycloaddition.....	30
1.3.2.2.	Multi-step reaction: the Morita Baylis Hillman reaction.....	31
1.4.	CHALLENGES	33
1.4.1.	<i>Prediction inaccuracy</i>	34
1.4.2.	<i>Presence of multiple conformational states</i>	35
1.4.3.	<i>Stabilizing multiple transition states</i>	36
1.5.	INCORPORATING BACKBONE FLEXIBILITY	37
1.6.	THESIS OBJECTIVES.....	38
 CHAPTER 2. ENSEMBLE-BASED ENZYME DESIGN CAN RECAPITULATE THE EFFECTS OF LABORATORY DIRECTED EVOLUTION IN SILICO		44
2.1.	STATEMENT OF CONTRIBUTION.....	44
2.2.	ABSTRACT	44
2.3.	INTRODUCTION	46
2.4.	RESULTS	47
2.4.1.	<i>HG series of Kemp eliminases</i>	47
2.4.2.	<i>Room-temperature crystal structures</i>	50
2.4.3.	<i>HG4, an efficient artificial enzyme</i>	56
2.4.4.	<i>Computational design of HG4</i>	57
2.5.	DISCUSSION	61
2.6.	METHODS	63

2.6.1.	<i>Protein expression and purification.....</i>	63
2.6.2.	<i>Steady-state kinetics.....</i>	64
2.6.3.	<i>Crystallization.....</i>	65
2.6.4.	<i>X-ray data collection and processing.....</i>	65
2.6.5.	<i>Structure determination</i>	66
2.6.6.	<i>Unconstrained molecular dynamics</i>	67
2.6.7.	<i>Computational enzyme design</i>	68
2.7.	SUPPLEMENTARY INFORMATION	70
 CHAPTER 3. IMPROVING THE ACCURACY OF DE NOVO ENZYME DESIGN		
USING A VARIETY OF BACKBONE ENSEMBLES.....		89
3.1.	STATEMENT OF CONTRIBUTION	89
3.2.	INTRODUCTION	89
3.3.	RESULTS	93
3.3.1.	<i>Hypothesis and general strategy.....</i>	93
3.3.2.	<i>Ensemble generation and characterization</i>	95
3.3.3.	<i>Computational Ensemble-based Design.....</i>	100
3.3.4.	<i>Recapitulation of the transition-state binding pose.....</i>	103
3.4.	DISCUSSION	105
3.4.1.	<i>Properties of the best ensembles.....</i>	105
3.4.2.	<i>Parameters used for the prediction.....</i>	106
3.4.3.	<i>Limitations</i>	107
3.4.4.	<i>Recommendations and perspectives</i>	109
3.5.	METHODS	110

3.6.	SUPPLEMENTARY INFORMATION	116
------	---------------------------------	-----

CHAPTER 4. HARNESSING THE CONFORMATIONAL ENSEMBLE TO DESIGN

EFFICIENT ARTIFICIAL ENZYMES..... 124

4.1.	STATEMENT OF CONTRIBUTION	124
4.2.	ABSTRACT	124
4.3.	INTRODUCTION	125
4.4.	RESULTS	127
4.4.1.	<i>Ensemble-based enzyme design</i>	127
4.4.2.	<i>Kinetic characterization of designs</i>	133
4.4.3.	<i>Crystal structures</i>	134
4.4.4.	<i>Enrichment of catalytically competent sub-states</i>	139
4.4.5.	<i>KE70 designs</i>	141
4.5.	DISCUSSION	143
4.6.	METHODS	146
4.7.	SUPPLEMENTARY INFORMATION	156

CHAPTER 5. ENSEMBLE-BASED COMPUTATIONAL ENZYME DESIGN FOR

STABILIZING THE TRANSITION STATES OF A MORE COMPLEX REACTION.. 178

5.1.	STATEMENT OF CONTRIBUTION	178
5.2.	INTRODUCTION	178
5.3.	RESULTS	183
5.3.1.	<i>Ensemble generation</i>	183
5.3.2.	<i>Design on BH32.12</i>	184
5.5.	METHODS	189

5.6.	SUPPLEMENTARY INFORMATION	191
CHAPTER 6.	DISCUSSION AND CONCLUSION	197
6.1.	SUMMARY	197
6.2.	PERSPECTIVE AND FUTURE DIRECTIONS	201
6.2.1.	<i>Advantages of the ensemble-based design approach.....</i>	<i>201</i>
6.2.2.	<i>Limitations</i>	<i>202</i>
6.2.3.	<i>The ensemble-based enzyme design method and other methods</i>	<i>204</i>
6.3.	FUTURE DIRECTIONS	205

LIST OF TABLES

Table 2.1. Kinetic parameters of Kemp eliminases	48
Table 3.1 List of enzymes used for the benchmark and a description of their crystal structures.	94
Table 3.2. Diversity and deviation.....	96
Table 3.3. HG3 designs best results.....	103
Table 4.1. Kinetic parameters of Kemp eliminases	132
Table 5.1. Crystal structures information and $R_{\text{work}}/R_{\text{free}}$ of crystal structures before and after ensemble refinement	184

LIST OF SUPPLEMENTARY TABLES

Supplementary Table 2.1. Crystallization conditions	72
Supplementary Table 2.2. Crystallographic data and refinement statistics for room-temperature structures (277 K)	73
Supplementary Table 2.3. Amino-acid positions optimized during computational design of HG4	74
Supplementary Table 2.4 Geometric definitions for generation of transition-state poses off the side chains of catalytic residues	75
Supplementary Table 2.5. Geometric constraints used to define catalytic contacts during HG4 computational design	76
Supplementary Table 2.6. Codon-optimized genes of HG-series Kemp eliminases	77
Supplementary Table 3.1 List of amino-acid positions optimized during the computational design procedure.....	116
Supplementary Table 3.2. AlphaFold2 ensembles scores	117
Supplementary Table 3.3. Ensemble refinement results.....	118

Supplementary Table 3.4. HG3 geometric constraints used to define catalytic contacts during computational design	118
Supplementary Table 3.5. Geometric definitions for generation of transition-state poses off the side chain of catalytic base	119
Supplementary Table 4.1. Ensemble refinement results.....	156
Supplementary Table 4.2. Design filtering – HG series	157
Supplementary Table 4.3. Angles defining the hydrogen bond between D127 and the transition state or transition-state analogue for various Kemp eliminase structures.....	158
Supplementary Table 4.4. Crystallization conditions	159
Supplementary Table 4.5. Crystallographic data and refinement statistics	160
Supplementary Table 4.6. Median pairwise backbone RMSDs for designed active-site residues	161
Supplementary Table 4.7. Design filtering – KE70 series.....	162
Supplementary Table 4.8. Amino-acid positions optimized during computational design of HG3 variants	163
Supplementary Table 4.9. Amino-acid positions optimized during computational design of KE70 variants.....	164
Supplementary Table 4.10. Geometric definitions for generation of transition-state poses off the side chain of catalytic base	165
Supplementary Table 4.11. Geometric constraints used to define catalytic contacts during computational design	166
Supplementary Table 4.12. Amino-acid sequences of Kemp eliminases from HG3 series	167
Supplementary Table 4.13. Amino-acid sequences of Kemp eliminases from KE70 series.....	168

Supplementary Table 5.1. Repacking results.....	191
Supplementary Table 5.2. Amino acid and positions included in the design procedure	192
Supplementary Table 5.3. Geometric definitions for generation of transition-state poses off the side chain of the nucleophile for each transition state	193
Supplementary Table 5.4. Geometric constraints used to define catalytic contacts during the design procedure for each transition state	194

LIST OF FIGURES

Figure 1.1. Enzyme reaction and mechanism.	3
Figure 1.2. Overview of protein dynamics timescale.	6
Figure 1.3. Example of a procedure for laboratory directed evolution.....	9
Figure 1.4. Computational protein design framework.	11
Figure 1.5. Computational enzyme design framework.	19
Figure 1.6. Examples of de novo enzyme.....	23
Figure 1.7. Kemp elimination reaction scheme.	24
Figure 1.8. The Retro-Aldo condensation reaction scheme.....	27
Figure 1.9. The Diels-Alder cycloaddition reaction scheme.	30
Figure 1.10. Morita Baylis Hillman reaction (MBH).	31
Figure 1.11. Challenges in computational enzyme design.	33
Figure 1.12. Ensemble-based enzyme design goal.	39
Figure 1.13. Thesis objectives.	39
Figure 2.1. HG series of Kemp eliminases.	49
Figure 2.2. Crystal structures of HG-series Kemp eliminases.....	51

Figure 2.3. Conformational heterogeneity.	54
Figure 2.4. Computational design of HG4 on various backbone templates.	59
Figure 3.1. Method for Kemp Eliminase benchmark.	93
Figure 3.2. Overlay of all the ensembles from HG3.....	97
Figure 3.3. Distribution of the deviation of ensembles from HG3_EV crystal structure.	99
Figure 3.4. Computational design on crystal structure.	100
Figure 3.5. Bar plot showing repacking results.	102
Figure 3.6. Best results obtained from HG3_DH ensembles.....	105
Figure 4.1. Ensemble-based computational enzyme design.	129
Figure 4.2. Ensemble-based design vs. directed evolution.	134
Figure 4.3. Crystal structures of HG-series Kemp eliminases.....	136
Figure 4.4. Ensemble-based design accurately predicts crystal structures of designed enzymes.	138
Figure 4.5. Enrichment of catalytically-competent sub-states in the conformational ensemble.	141
Figure 5.1. MBH reaction between 4-nitrobenzaldehyde and cyclohexenone.	180
Figure 5.2. Mutations introduced in variants from the evolutionary trajectory of BH32.	183
Figure 5.3. Repacking result of the individual transition state placement.	185
Figure 5.4. Energy diagram of the results obtained with the ensemble-based method.	186
Figure 5.5 Future work for designing efficient MBHases.	189
Figure 6.1. Summary of findings.	197
Figure 6.2. Future directions.	206

LIST OF SUPPLEMENTARY FIGURES

Supplementary Figure 2.1. Amino-acid sequences of HG-series Kemp eliminases	70
---	----

Supplementary Figure 2.2. HG2 is the direct precursor to HG3.	78
Supplementary Figure 2.3. Steady-state kinetics.	79
Supplementary Figure 2.4. 6NT-binding pocket.	80
Supplementary Figure 2.5. Difference density increases around the 83–84 peptide bond after refinement in the absence of minor conformer.	81
Supplementary Figure 2.6. Crystal structures of HG-series Kemp eliminases in the unbound state.	82
Supplementary Figure 2.7. Energy of HG4 design models generated on various backbone templates.	83
Supplementary Figure 2.8. Backbone ensembles.	84
Supplementary Figure 3.1. HG3 diversity distribution of active site positions.	120
Supplementary Figure 3.2. Best designs obtained from the design of HG3 family.	121
Supplementary Figure 4.1. The Kemp elimination reaction.	169
Supplementary Figure 4.2. Ideal geometry for catalytic base contact.	170
Supplementary Figure 4.3. Design filtering using molecular dynamics.	171
Supplementary Figure 4.4. Steady-state kinetics of HG-series Kemp eliminases.	172
Supplementary Figure 4.5. Design on HG3 crystal structure does not accurately predict crystal structures of designed enzymes.	173
Supplementary Figure 4.6. Steady-state kinetics of KE70-series Kemp eliminases.	174
Supplementary Figure 5.1. Ensembles from ensemble refinement (ER) of BH32 variants.	195

LIST OF ABBREVIATIONS

AF2	AlphaFold2
BR	Backrub
CPD	Computational Protein Design
DFT	Density Functional Theory
ER	Ensemble Refinement
MBH	Morita-Baylis-Hillman reaction
MD	Molecular Dynamics
MSD	Multi-State Design
MSA	Multi-State Analysis
PDB	Protein Data Base
PertMin	Coordinate Perturbation followed by energy Minimization
QM/MM	Quantum Mechanics/Molecular Mechanics
TS	Transition State
TSA	Transition State Analogue

Chapter 1. Introduction

1.1. Enzymes

Enzymes are fascinating biological machines capable of catalyzing essential reactions for the vital functions of their host organisms. Since the first enzyme was discovered in 1833 by the French chemist Anselme Payen,¹ enzymes became valuable tools for many industrial applications thanks to their exquisite catalytic properties. They find applications in various sectors of our daily lives, including food and beverage processing, technology, and cleaning products. They serve as analytical devices and assays in clinical, forensic, and environmental contexts.²

Synthetic chemistry, a branch of chemistry that focuses on design and synthesis of new chemical compounds, molecules, or materials, plays an important role in producing a broad spectrum of fine chemicals that are important to our daily lives. These encompass everything from petroleum-derived compounds to the molecules that constitute our food additives. Yet, the efficiency of synthetic chemistry falls short in comparison to their biological counterparts. Furthermore, using chemical catalysts presents challenges – they tend to be costly due to the need for precious metals, their reactions occur in extreme conditions, they are non-reusable, and they contribute to waste generation and environmental concerns due to the reliance on toxic reagents and solvents. Conversely, using enzymes offers a pathway toward sustainable processes, as their reactions use nature's abundant resources while occurring under mild conditions. Additionally, enzymes can reduce losses from competing reactions, distinguishing them from chemical catalysts, which are more predisposed to yield undesired by-products.

While choosing enzymes over chemical catalysts promises to provide greener chemistry for our society, the plethora of industrial applications and medical research still rely heavily on non-enzymatic catalysts, as the enzymes required to catalyze the desired reactions are often

unavailable. To circumvent this limitation, protein engineering techniques were used to enhance natural enzymes' existing activity, and attempts to introduce reactions not existing in nature have been made.³⁻⁹ Despite the tremendous scientific advances over the past decades in unravelling enzyme reaction mechanisms, significant aspects of enzyme catalysis remain obscure. Furthermore, designing artificial enzymes with activities comparable to those of natural enzymes is a great challenge. The objective of improving current enzyme design methodologies to engineer artificial enzymes tailored for specific reactions holds the potential to meet present industrial demands and deepen our understanding of enzyme catalysis.

1.1.1. Enzyme reaction and mechanism

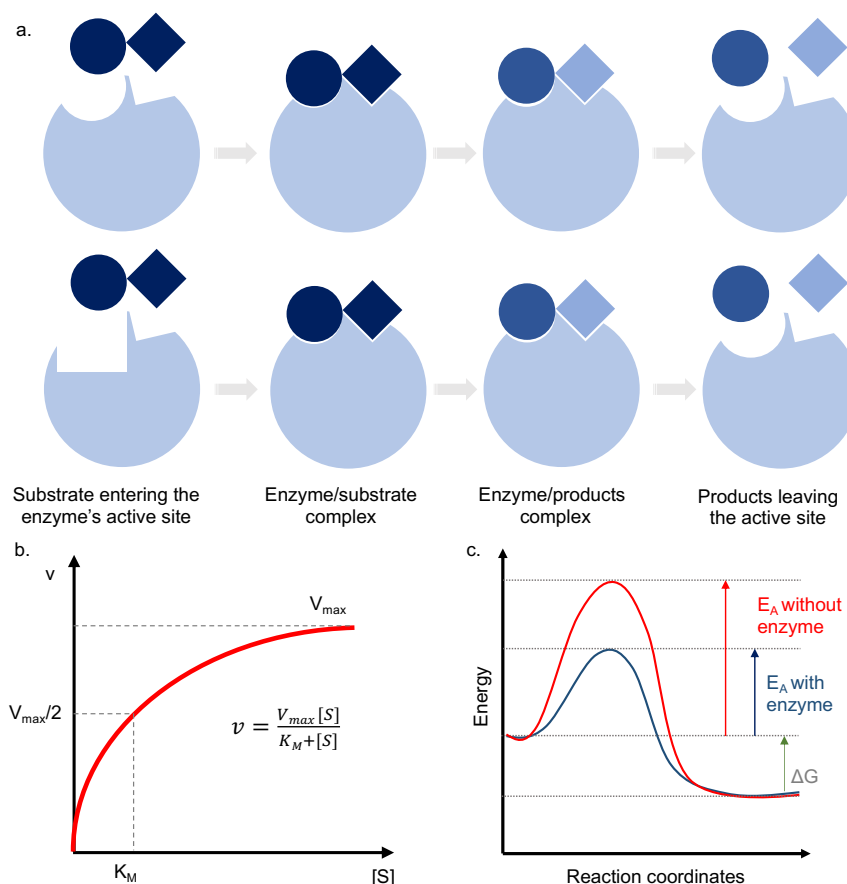


Figure 1.1. Enzyme reaction and mechanism. a) at the top is the lock and key theory from Fisher, the active site of the enzyme recognizes its substrate through shape complementarity. The bottom row is the induced fit model where the enzyme's active site changes shape until the substrate is fully bound. b) Michaelis-Menten model showing the relationship between the rate of an enzyme-catalysed reaction and substrate concentration. c) Theoretical enzyme reaction process, the red line corresponds to the reaction without enzyme and the blue line corresponds to the reaction with enzymes. The reaction has to overcome an energy barrier, to get the product. This barrier is decreased with the action of an enzyme.

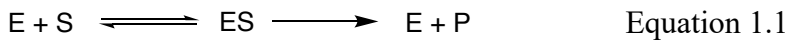
Honed through extensive evolution, enzymes exhibit a remarkable capability to accelerate catalytic reactions by up to 26 orders of magnitude greater than non-catalyzed reactions.¹⁰ Enzymes active sites offer an ideal microenvironment for catalysis by bringing substrates together in a geometry ideal for the reaction. These properties give the enzymes remarkable specificity, high regioselectivity, and high turnover rates when interacting with their substrates.^{2,10} These properties are linked to sophisticated mechanisms that were studied since Emil Fischer and decades later Daniel Koshland proposed the lock and key model or induced-fit model for enzymes' substrate recognition.

1.1.1.1. The lock and key and induced fit models

In 1894, Emil Fischer laid the foundation of the theory suggesting that enzymes identify their substrates through a lock and key model.¹¹ This model suggests that the active site of an enzyme possesses a complementary geometric shape (the lock) that corresponds to its substrates (the key) (Figure 1.1.a top row). In 1958, Daniel Koshland introduced the induced-fit model for substrate recognition.¹² This model suggests that an enzyme's active site is a dynamic entity that adapts until the substrate is fully bound (Figure 1.1.b bottom row). This notion enriches the lock and key model by introducing the idea that enzymes are flexible molecules rather than static, contrary to Fischer's original concept.

1.1.1.2. Michaelis-Menten kinetics model

In 1913, Leonor Michaelis and Maud Leonora Menten developed a mathematical model to describe enzyme kinetics. This model, called Michaelis-Menten enzyme kinetics is a fundamental concept in biochemistry describing the rate at which enzymes catalyze chemical reactions and the relationship between the rate and substrate concentration.¹³ In this model, the enzyme (E) and its substrate (S) bind reversibly to the enzyme-substrate complex (ES) and then catalyze the irreversible conversion of substrate into a product (P).



The reaction rate is dependent on the concentration of the enzyme-substrate complex (ES) and can be calculated using the following equation:

$$v = \frac{d[P]}{dt} = k_{cat}[ES] \quad \text{Equation 1.2}$$

where v is the initial rate of the reaction, $[P]$ is the concentration of the product, dt is the change in time, and k_{cat} is the turnover number, which represents the number of substrate molecules converted into products by one enzyme molecule in a unit of time. The enzyme-substrate complex concentration $[ES]$ depends on the enzyme's affinity for the substrate and the initial substrate concentration $[S]$. From the equation above, the rate of the enzymatic reaction can be obtained as follows:

$$v = \frac{k_{cat} [E][S]}{K_M + [S]} \quad \text{Equation 1.3}$$

where,

$$k_{cat} = \frac{V_{max}}{[E]_T} \quad \text{Equation 1.4}$$

$$v = \frac{V_{max} [S]}{K_M + [S]} \quad \text{Equation 1.5}$$

This equation describes the rate of the enzymatic reaction as a function of the substrate concentration, and it fits the plot shown in Figure 1.1.b. In this equation, K_M represents the Michaelis constant. It corresponds to the substrate concentration at which the reaction rate is half of V_{max} , providing an idea of the enzyme and substrate affinity.

When evaluating the efficiency of an enzyme, catalytic efficiency is often used, it is described as k_{cat}/K_M , and it tells the efficiency of an enzyme to convert substrates into products.

1.1.1.3. The Transition State theory

The Transition State theory states that the reaction between two elements reaches a reversible preequilibrium state between the reactant molecules and an activated complex near the saddle point of the potential energy surface.¹⁴ With enough energy provided, this high-energy state leads to product formation. The activation energy corresponds to the minimum energy required to overcome the activation barrier for a reaction to occur and is linked to the rate of the reaction. The reaction rate is exponentially proportional to the Gibbs energy of the highest-energy transition state as shown from the Arrhenius equation (Equation 1.6).

$$k = Ae^{\frac{-E_A}{RT}} \quad \text{Equation 1.6}$$

To achieve their high efficiency, enzymes lower a reaction's activation energy by tightly binding the unstable transition state with the enzyme, as proposed by Linus Pauling,^{15,3} (Figure 1.1.c). Wolfenden and coworkers added that the rate enhancement of enzymes is proportional to the enzyme's affinity for the transition state structure relative to the Michaelis complex during its

lifetime of 10^{-13} s.

1.1.2. Enzyme flexibility and function

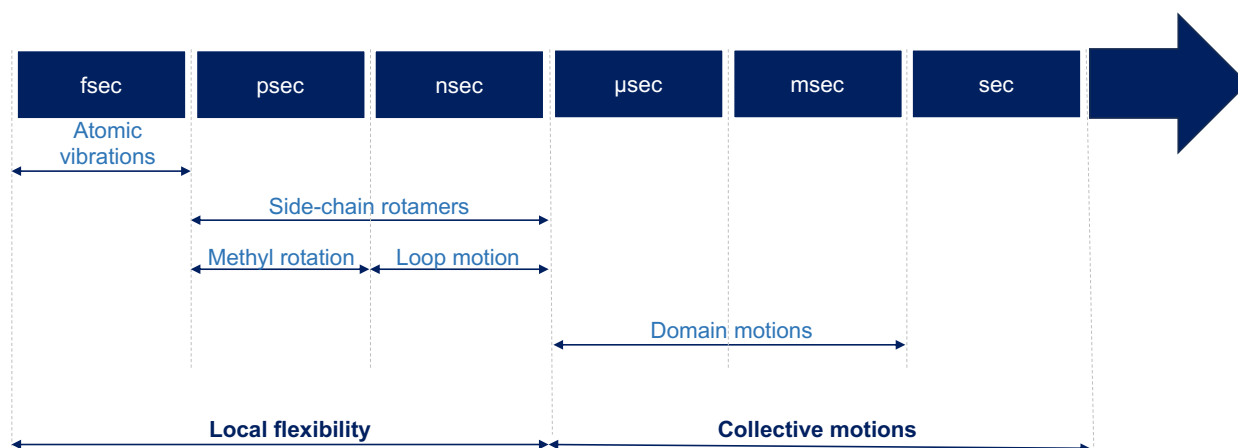


Figure 1.2. Overview of protein dynamics timescale.

Natural enzymes are dynamic entities whose flexibility is closely related to their functions. In the induced-fit model suggested by Koshland, an enzyme's active site changes conformation upon substrate binding, highlighting the link between enzyme flexibility and catalysis¹². Protein dynamics are defined as the time-dependent change in their atomic coordinates, and their functions are known to be governed by equilibrium effects. Enzyme motions can be described using an energy landscape, where multiple conformational substates fluctuate from local energy wells and are in thermal equilibrium influenced by solvent and ligand binding.¹⁶ These conformational ensembles are accessible at different timescales, ranging from bond vibrations and local fluctuations on a fast timescale to a slower conformational transition such as folding and unfolding on a longer timescale (Figure 1.2). Longer timescale dynamics are correlated to collective motions, including backbone conformational change such as the reversible *cis-trans* isomerization of prolyl peptide bonds or motions propagating from the active site to distant sites¹⁷. Faster motions on the picosecond to nanosecond time scale are correlated to sidechain fluctuations and loop motions,

and their highly populated conformational ensembles govern the entropy of a system. The femtosecond timescale motions are linked to bond vibrations corresponding to the breaking and bond formation or the transfer of protons, hydride ions and electrons. While substrate binding influences the conformational landscape of an enzyme, studies of free enzyme dynamics showed the pre-existence of conformational substates readily accessible for catalysis.^{17–19}

Many studies have shown the importance of protein dynamics in the enzyme function throughout its catalytic cycle, in substrate binding, chemical transformation and product release, stabilization and trapping of intermediates, or the orientation of the substrate in the binding cleft. The impressive catalytic power of these biological macromolecules has been attributed to their ability to fluctuate from one conformational state to another and to their highly preorganized active sites, in which the proper positioning of the catalytic machinery linked to protein dynamic allows a better stabilization of the transition state.²⁰ Other types of conformational motions observed with enzymes are the existence of opened and closed conformations, allowing entrance and tight binding of the substrate, and product release.

Studying the dynamic properties of the evolutionary trajectory of enzymes allowed a better understanding of their conformational flexibility. Catalytically inefficient enzymes can access multiple unproductive minor conformational states. Upon mutations introduced during evolution, the population of conformational states changed, gradually tuned to become a major productive state.^{21,22} These studies demonstrated the conformational plasticity of primordial enzymes and their promiscuous activity, while highly specialized enzymes' conformational states become narrower.

Regions of an enzyme remote from the active site have also contributed to enzyme catalysis. Mutations in distal positions can have a drastic impact on the overall catalysis. Some enzymes are

regulated by allostery, where they change their conformational ensemble upon binding an effector. The process of allosteric binding induces conformational redistribution, leading to a modulation in the active site conformation, thereby regulating enzyme catalysis.^{23–25} These conformational changes are related to coupled motions that could impede functions if interrupted.²³

1.1.3. Enzyme design methods and limitations

The finite repertoire of reactions catalyzed by natural enzymes limits their use for industrial applications. To address this limitation, studies have been conducted on re-engineering existing enzymes to repurpose their functions or enhance their efficiencies. However, this procedure needs the existence of an enzyme that has a modest catalytic activity. The challenge with this approach is the exhaustive initial work to scan existing enzyme databases.

1.1.3.1. Enzyme engineering

The ultimate objective of enzyme designers is to harness the potential of enzymes for a wide range of applications. To accomplish this, one effective strategy involves conducting a comprehensive exploration of enzyme databases to identify enzymes exhibiting the desired activity. Through billions of years of evolution, nature has provided enzymes with diverse catalytic functions, resulting in highly specialized enzymes. In the early stages, ancient enzymes needed to possess the ability to facilitate a broad spectrum of chemical transformations to ensure the survival of their host organisms. Such ancient features are found on enzymes that belong to mechanistically diverse superfamilies, and their structures are often conserved among them. As such, they often exhibit promiscuity, catalyzing numerous distinct reactions.^{26,27} Given the ability of promiscuous enzymes to catalyze diverse reactions, one strategy to redesign enzymes with enhanced

functionalities is to proceed with structure and sequence alignment to introduce key mutations in the enzyme active site.^{28–30} While sequence and structure alignment of homologous enzymes is a simple approach, it is limited by the mechanisms found within the superfamily they belong to. While many enzymes used in the industry today stem from this approach, it is not guaranteed that an enzyme with a desired promiscuous activity can be found.

1.1.3.2. Laboratory directed evolution

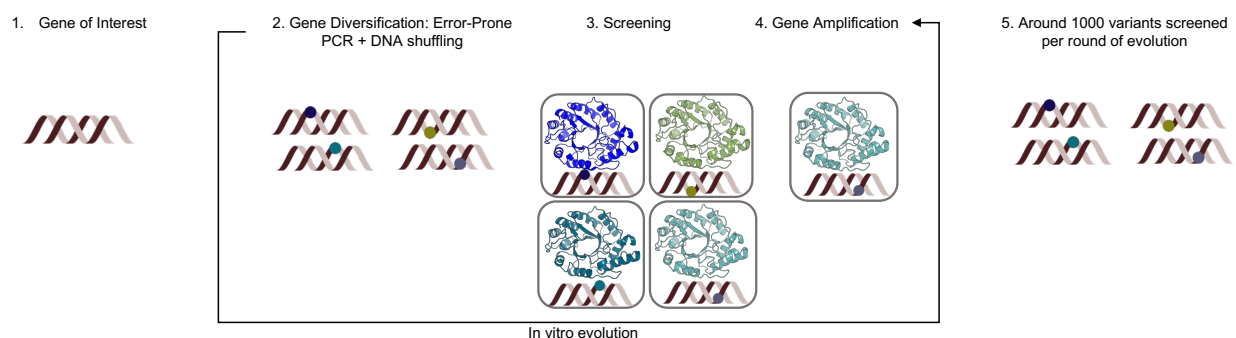


Figure 1.3. Example of a procedure for laboratory directed evolution. (1) a target objective is selected for example a low-activity enzyme, (2) gene diversification is proceeded using techniques such as Error-Prone PCR + DNA shuffling, (3) Screening methods are chosen for (4) gene amplification, (5) a typical directed evolution procedure will require around 1000 variants per round of evolution.

Laboratory directed evolution is a technique used to mimic natural evolution by artificial selection in a laboratory time scale, where individual genes are expressed in fast-growing microorganisms.^{31,32} This technique has enabled the design of enzymes catalyzing reactions not existing in nature, such as cyclopropanation, B-H insertion or S-H insertion, among other reactions.^{33–35} In a typical directed evolution process, an existing protein with a minimal desired function is selected as a starting point. By introducing mutations and applying selective pressure, proteins displaying the desired phenotypes are screened. The resulting proteins are then used as a starting point for another round of evolution, and beneficial mutations are recombined as needed. Mutations are introduced using error-prone PCR combined with random mutagenesis to target the

entire protein and saturated mutagenesis to target positions within the active site,³² (Figure 1.3). This technique allowed Frances Arnold to receive a Nobel Prize in 2018, demonstrating the effectiveness of directed evolution in introducing a non-natural reaction in weakly active enzyme. This method has been proven reliable and scalable for industrial purposes, where activity enhancement led to activities approaching those of the natural ones. However, its use is limited by the need for high throughput screening methods and for the target to exhibit minimal activity. As this method depends on a starting minimal activity of the target function, it can be combined with the first step of rational design. This approach has led to highly efficient artificial enzymes with activities not found in nature.^{4,36}

1.2. Computational enzyme design

Computational enzyme design aims to introduce catalytic activity into a protein scaffold devoid of the desired function. This differs from rational design, in which affinity for the substrate and a low activity are already present in the protein. Computational enzyme design is a subsection of computational protein design (CPD), where the goal is to introduce activity by altering amino-acid identities in the active site to convert the protein into an enzyme catalyzing the reaction of interest (Figure 1.5).

1.2.1. Computational protein design

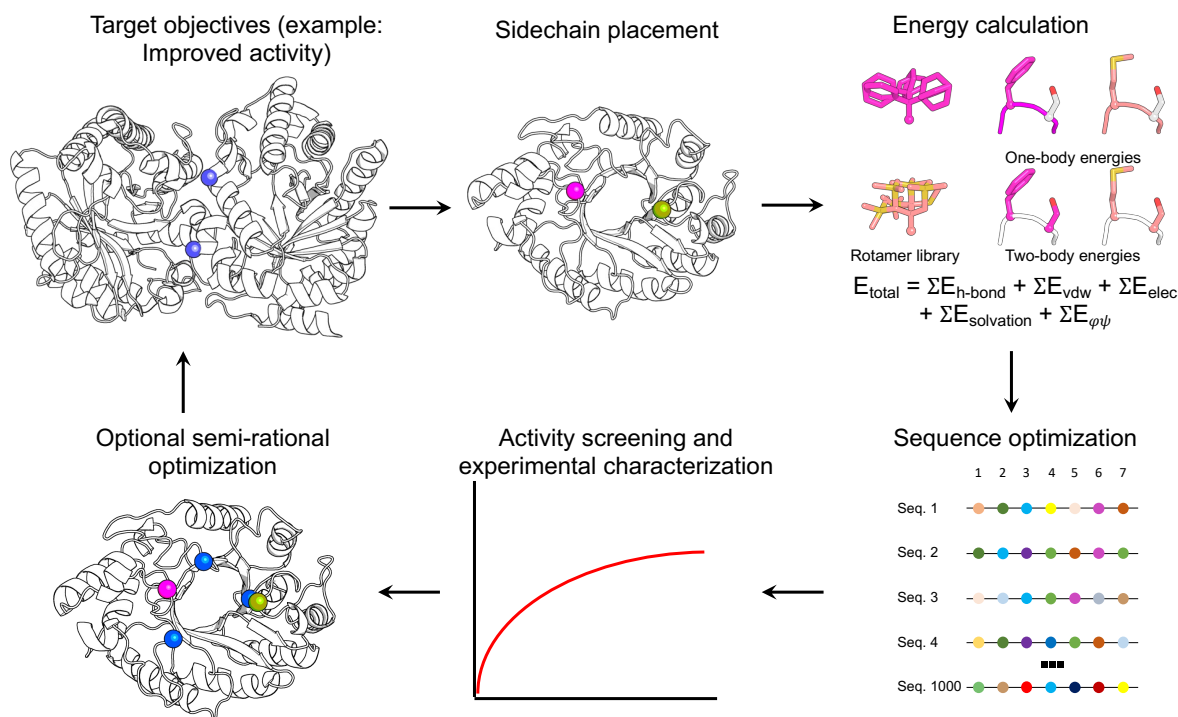


Figure 1.4. Computational protein design framework. (1) a target objective is chosen, for example improving the stability of a protein (2) positions are selected, and sidechain placement is conducted on a backbone template using a library of rotamers. (3) an energy calculation is performed to assess the energy interactions between rotamers and the backbone template (one-body energies) and pairs of rotamers (two-body energies) of each sequence. The computed energy typically corresponds to the sum of van der Waals (vdw), electrostatics (elec), hydrogen bonding (h-bond), and other types of interactions. (4) during the sequence optimization step, combination of rotamers and their interaction with the template are optimized using a search algorithm to find low energy sequences, the final sequences are then ranked based on their energies. (5) promising sequences are tested experimentally to confirm the design and finally (6) an optional semi-rational optimization might be conducted to improve the design.

Computational protein design (CPD) is defined as the computer-aided rational design of a protein, in which amino-acid sequences are identified to stabilize a three-dimensional protein scaffold for desired biophysical properties and functions such as improved thermostability, binding affinity or altered specificity.^{37–39} The main challenge of CPD is to solve the protein folding problem, as stated by Anfinsen Dogma, in which the three-dimensional structure of a protein is predicted based on its amino acid sequence.⁴⁰ To find the sequence with the desired function, an

astronomical sequence space has to be explored. For example, in a design requiring optimizing 100 positions with 20 possible amino acid types, the sequence space will be as large as 20^{100} sequences, where only a small fraction will fold into a stable protein. CPD's algorithms are optimized to search through this astronomical number of sequences and select the ones generating the lowest energy, which translates to the sequence exhibiting the desired property. During the procedure, defined positions in the protein scaffold will be replaced by other possible amino acid types to obtain the desired properties (Figure 1.4). The CPD procedure involves choosing a suitable protein scaffold, ideally a high-resolution crystal structure from the protein data bank, in which specific positions are mutated to a library of rotamers consisting of discrete stable conformations of amino acids. A list of user-defined amino acids is chosen for each position, and an energy calculation is performed where each amino acid interaction between the backbone template (one-body energies) and its immediate neighbouring amino acid sidechains (two-body energies) is calculated using a scoring function. A search algorithm is then selected to explore the large sequence space and the vast possibility of amino acid sequences to obtain optimal sequences. Finally, sequences are ranked according to their energy values, which reflect their stability and desired function.

1.2.1.1. Rotamer optimization

Rotamers refer to conformational isomers of a sidechain that undergo interconversion through single-bond rotation. Each rotamer can be represented by a collection of values assigned to dihedral angle degrees of freedom, representing various conformations. Typically, rotamers correspond to local energy minima or represent an average conformation within a specific range of the dihedral conformational space. The number of rotamers a residue can have depends on the

number of its χ angles. For example, glycine and alanine will always have one rotamer, while residues with more χ angles, like methionine, can have up to 18 rotamers.⁴¹

In computational protein design, rotamer libraries are employed to sample potential conformational states of residues. A rotamer library is comprised of a comprehensive set of rotamers for each residue type derived from statistical analysis. It encompasses information on the conformations and frequency of each rotamer. Variances around dihedral angle mean, and modes are often included to enhance sampling.

The definition of rotamers has been refined through statistical analysis of experimental conformational sidechains. Rotamers can be categorized by clustering conformations or dividing them into bins based on dihedral angles. Canonical χ angles can be divided into three bins centred on each staggered conformation (-60° , 60° , 180°). This approach allows for a more precise characterization of rotamers and facilitates their practical implementation.⁴¹

Rotamer libraries are used to predict a specific conformation of a given sidechain. This conformation depends on several factors, including the probability of the sidechain, its environment, and whether it is energetically favourable in its surroundings. The rotamer prediction can or may not depend on the backbone (in the case of the backbone-dependent rotamer library). Rotamer libraries exist as three distinct types, allowing their dependence on the local backbone conformations. They can be either backbone-independent, secondary-structure-dependent libraries or they can be backbone-dependent libraries. With the backbone-independent libraries, no reference is made to the local backbone conformation, and rotamers are calculated from all the available sidechain conformations of a particular residue type.⁴¹

While rotamer libraries play a crucial role in sampling sidechain flexibility in computational protein design (CPD), it is essential to acknowledge their limitations. Non-rotameric sidechains

have χ angles differing from the canonical angles by $\pm 20^\circ$.⁴² It is caused by different factors, such as the sidechains' immediate environment or steric interaction with the backbone atoms (N_i and C_i). Steric hindrance caused by two terminal methyl groups in one of the chemical conformations of n-pentane or syn-pentane interactions is a common contributor to non-rotameric conformations.

Typically, dihedral angle distributions are assumed to follow normal distributions, allowing the most prevalent rotamers to be described by either their means or modes. However, due to the steric interactions of χ_1 angle of some residues, such as valine, distributions tend to be skewed so that modes were selected instead of the means to represent the most common rotamers accurately.⁴³ Another challenge is sampling non-rotameric residues or residues with dihedral angles differing from canonical angles by $\pm 30^\circ$. To overcome this limitation, Dunbrack and Cohen have used a Bayesian function,⁴⁴ in which the probability of a given data is conditional to a prior distribution.

1.2.1.2. Scoring function

During the design procedure, an energy calculation is necessary to evaluate the approximate stability of a sequence in a backbone template. Scoring functions are then needed to reward good interactions and penalize poor interactions throughout the design process. During sequence optimization, sidechains of amino acids are threaded onto the backbone. The pairwise energy between the rotamer and the backbone and the energy between the rotamer and its surroundings are computed.

Four scoring functions have been developed: Knowledge-based, force field energy function, empirical and machine learning. While their usage depends on the application, the force field energy function can be used to assess both the interactions between the ligand and also the overall stability of the sequence.

Potential energy functions are calculated based on non-covalent interactions (van der Waals, electrostatics and hydrogen bonds, solvation energy terms) and covalent interactions (dihedral angles ϕ ψ).⁴⁵ Van der Waals potential is used to assess the overall sidechain packing specificity, and a Lennard-Jones 12-6 expression models it. Hydrogen bonds are represented by an angle-dependent, 12-10 hydrogen-bond potential. However, the non-physical hydrogen bond geometries that might arise can limit this potential energy. Electrostatic interactions play a role in assessing the specificity of the different interactions. These interactions are modelled by Coulomb's law. Internal bonding terms reflect the change in rotamer, and, finally, a solvation potential is sometimes required to model the hydrophobic effect. This potential includes polar and nonpolar atomic solvation parameters.⁴⁵

1.2.1.3. Sequence optimization and search algorithm

After careful definition of the target objective, the sequence space and the scoring function, the quality of a design procedure depends on the sequence optimization step and the choice of the search algorithm. One challenge of the CPD procedure is finding a stable sequence that will fold to the target objective, considering the immense space it needs to explore. During sequence optimization, the combination of rotamer interaction with the template and the interactions between each rotamer are optimized to minimize the target objective to find the Global Minimum Energy Conformation (GMEC).

A search algorithm typically selects sequences based on the precomputed energies to ensure conformational stability. Two types of search algorithms exist: deterministic algorithms and stochastic algorithms. Stochastic algorithms such as Monte Carlo (MC) simulated annealing,^{46,47} Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER),⁴⁸ or genetic

algorithms are used when exploring a significant sequence space is required. In such algorithms, random components are introduced. For example, in the Monte Carlo simulated annealing algorithm, a random amino acid rotamer is selected at a random position, and the sequence is kept based on the Metropolis criterion.⁴⁹ If a substitution lowers the energy of the system, it is accepted. Otherwise, acceptance or rejection of a substitution is based on Boltzmann's relationship between probability and energy differences at a given temperature. High temperatures allow large jumps between local energy minima in the energy landscape. With cooling, accessing a high-energy state decreases along with the Boltzmann probability of acceptance. Multiple independent runs can be performed to increase the probability of the sequence escaping local energy minima. Although these algorithms enable faster exploration of the sequence space, the random component of stochastic algorithms does not ensure that the GMEC can be found.

In addition to stochastic algorithms, deterministic algorithms, such as Dead-End Elimination (DEE), are an alternative for CPD. DEE is valuable because it guarantees the identification of the GMEC if it converges, thus ensuring design accuracy. However, it is important to note that it comes at the expense of a higher computational cost. With DEE, sequences incompatible with the GMEC will be systematically eliminated during optimization, ensuring that the GMEC is found.⁵⁰

1.2.1.4. Multi-state computational design

Multi-state design (MSD) is a technique used to introduce flexibility in the CPD procedure to improve prediction accuracy of a computational protein design problem. During MSD, multiple single-state design is used on an ensemble of backbone templates to generate sequences. A fitness energy function is then used to assess the stability of one sequence across multiple backbones.⁴⁸

CPD has demonstrated several successes in engineering novel specificity and functions to proteins, however, due to limitations in computational power, existing methodologies rely on several approximations. In traditional CPD, a single structure obtained from the Protein Data Bank is engineered for a specific application, in which selected positions are mutated to a list of user-defined amino acids. The nature of the designed amino acid and its conformation will then be dependent on the backbone conformation. However, enzymes are dynamic molecules, and their functions are closely related to their conformational flexibility. In that regard, approximating the protein conformation to a single state structure will introduce biases and lead to a higher proportion of false positive designs.^{51,52} To improve the accuracy of the computational protein design procedure, using a multi-state design approach is a better solution to approximate the intrinsic flexibility of the protein and then improve design prediction. In a study conducted by Davey et al.,⁵¹ using an ensemble of backbones through MSD algorithm improved the stability of a protein called Gβ1, while the design using a single crystal structure turned out to generate sequences with low prediction accuracy and with low stability. This methodology has a great advantage over introducing flexibility during the design calculation, which is computationally expensive. Its advantage lies in the simplicity of generating an ensemble of backbones before design and its implementation with current algorithms.⁴⁸ However, the quality of the ensemble generated will influence the final prediction. In the same work conducted by Davey et al.,⁵¹ ensembles were generated from molecular dynamics simulations temperatures with varying temperatures, from different parameters of perturbation and minimization of the input structure, from Rosetta backrub,⁵³ and ensembles generated by NMR spectroscopy. In this work, the ensemble with the highest prediction accuracy was the one with the most similarity to the crystal structure, highlighting that a subtle movement of the backbone increases prediction accuracy.⁵¹

As MSD proved to generate a higher prediction accuracy than using a single crystal structure by designing Gβ1 variants with improved stability, it highlighted the need to use an ensemble of backbones to sample different conformations of the protein in solution. In the case of computational enzyme design, a single crystal structure was used to introduce a catalytic activity in a protein scaffold devoid of the desired function.^{4-6,8} This procedure led to the design of artificial enzymes with a low activity, in which low prediction accuracy could contribute to weak binding of the transition state and leading to a high energy barrier preventing efficient catalysis.^{36,54,55} Using this information, a variation of the MSD or multi-state analysis (MSA) could be used in enzyme design, where multiple single-state design calculations are performed using an ensemble of backbone templates,⁵² and the energy of a sequence depends on its optimization in one backbone template. Using multiple conformations of the enzyme in solution should allow the sampling of a template that will better stabilize the TS during the design procedure.

1.2.2. Methodologies to design *de novo* enzymes

Enzymes achieve their high catalytic rate by stabilizing the transition state and lowering the energy barrier of the reaction. Current computational enzyme design methodologies use this concept to introduce a reaction in a protein scaffold devoid of the catalytic activity of interest. The approach consists of generating a computational transition state using Quantum Mechanics/Molecular Mechanics (QM/MM) techniques such as density functional theory and placing it inside an active site, providing a physical and chemical environment essential for stabilizing the transition state. To generate this theoretical transition state, Tantillo et al.,⁵⁶ conducted a study based on advances in NMR and crystallography to create what we call “theozyme” which is short for “theoretical enzyme.” The theozyme is a geometrical arrangement

of functional residues aiming to stabilize the transition state. QM/MM methods enable the generation of theozyme, providing accurate geometries and energies. Generally, quantum mechanics calculations (Hartree-Fock theory, density functional theory (DFT), or semi-empirical) generate a transition state. This step is followed by optimizing the transition state with functional residues, and each can interact non-covalently (Figure 1.4). Stabilization of the transition state consists of a stronger hydrogen bond, a specific orientation of functional groups to favour a strong hydrogen bond and a favourable electrostatic environment (as later confirmed by Warshel's work).

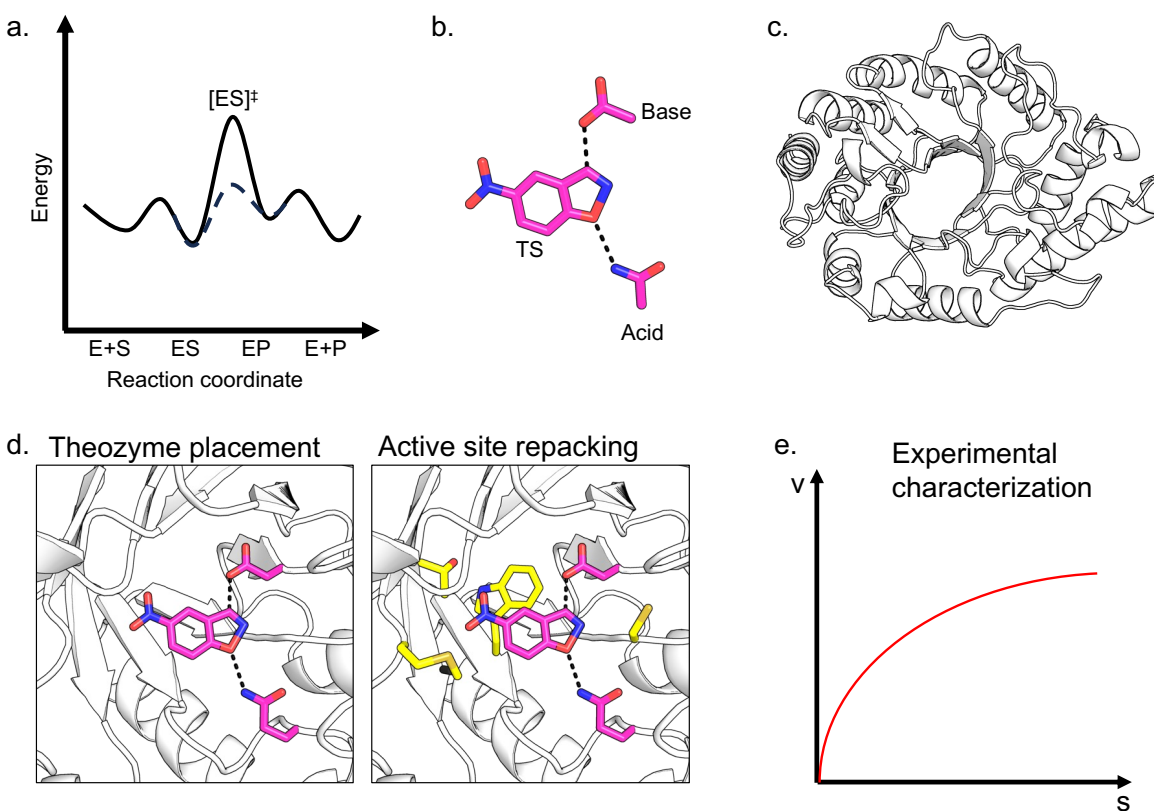


Figure 1.5. Computational enzyme design framework. a) The objective of computational enzyme design is to stabilize the transition state to increase the reaction rate. b) Theozyme calculation using QM/MM methods. Theozyme is the geometrical arrangement of functional groups for stabilizing the transition state. Here, a base is used to catalyze the reaction, and an acid stabilizes the transition state. c) A high-resolution crystal structure is selected to carve the active site for the theozyme. d) During theozyme placement, a combinatorial search is performed to place the theozyme (in magenta) inside the backbone template. In the active site repacking, positions in the active site are optimized to stabilize the transition state further. e) After sequence optimization, an experimental characterization is performed to confirm the activity of the designed enzymes.

1.2.2.1. Theozyme placement

Computational enzyme design is a technique used to design *de novo* enzymes using computational protein design (CPD) methodologies. As described in the previous section, CPD aims to optimize the sidechain placement from a rotamer library in a protein backbone template using a search algorithm.⁵⁷ The computational enzyme design procedure differs from the CPD protocol in that a cavity is created in the initial protein scaffold, where a transition state generated by a hybrid QM/MM (quantum mechanics/molecular mechanics) is placed inside (Figure 1.5). CPD methods enable the optimization and minimization of the transition state energy and allow rate acceleration, as described in Figure 1.5. The advantage of using computational methods is that they do not require sequence with an initial activity.

The first step of computational enzyme design is creating a theozyme. DFT is used to compute the optimal geometry of the interacting sidechains for transition-state stabilization. This theozyme is introduced inside a protein backbone template without the desired function. An active site is created by mutating all positions to Ala or Gly residues except for the catalytic residues (Figure 1.5.b, c, and d). The theozyme placement requires precise geometry between the catalytic sidechains and the transition state since a small deviation may change the enzyme-TS energy of interaction and may impact the overall k_{cat} of the reaction.³⁶ Two approaches have been devised to achieve this goal. One is called RosettaMatch,⁵⁸ developed by the Baker group, and phoenixMatch, developed by the Mayo group.⁵⁹

In RosettaMatch, the process involves constructing ligand poses while considering each catalytic contact simultaneously. By progressing from the protein's structural features towards the TS model from each catalytic side chain, TS configurations are identified that align with the

placement of each catalytic residue. For a given protein scaffold, a predetermined set of potential active site locations is established. These positions can encompass all available positions within the protein structure or may be limited to specific sites, such as cavities or areas where small molecules are known to bind.⁵⁸

In PhoenixMatch, ligand interactions are built from assigned catalytic residues, while an energetic benefit is included to favour required interactions.⁵⁹ In the first step, many substrate orientations are generated. Then, steric pruning and user-defined geometric criteria are used to reduce the number of substrate orientations. Next, pairwise energy interactions between sidechain/sidechain, sidechain/backbone, backbone/ligand, and sidechain/ligand interactions are calculated using a scoring function. Finally, a user-defined energy benefit is included to favour specific geometries between the ligand and the contacting residues. When the desired combination of sidechain geometry and TS is found, the other positions in the active site are optimized to stabilize the transition state further. The design calculations ensure functional groups are correctly positioned around the transition state. It has the advantage of explicitly placing key catalytic residues.

1.2.2.2. Active site repacking

After a desired ligand and sidechain conformation is found, all the positions in the active site are mutated to a list of user-defined amino acid types to stabilize the transition state further. Sequence optimization algorithms are then selected to optimize the combination of rotamers and their interaction with the backbone template. At the end of the procedure, a list of amino acid sequences stabilizing the TS binding pose, which is the position and orientation of this ligand in the protein pocket, and the specific intermolecular interactions it makes with the protein, is

generated and sequences are ranked according to their energies. A final list of sequences is selected for experimental characterization (Figure 1.5).

The David Baker group initiated the attempt to introduce the Kemp elimination reaction in different protein scaffolds. They have successfully designed artificial enzymes for unimolecular reactions, such as Kemp Eliminases,⁴ and Retro Aldolases,^{5,7} and they also successfully designed enzymes catalyzing bimolecular reactions, such as Diels-Alderase,⁶ and enzymes catalyzing the Morita Baylis Hillman reaction,⁸ (Figure 1.6). In the same objective to design *de novo* enzymes, the Stephen Mayo group designed artificial Kemp Eliminases using an iterative computational procedure,⁵⁸ (Figure 1.6).

1.3. Success in computational enzyme design

Using a “compute and built” approach, Bolon et al.,³ successfully introduced an esterase activity into the thiorodoxin scaffold from *Escherichia coli*. Their designs included the histidine-mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate (PNPA) into *p*-nitrophenol (PNP) and acetate. They proceeded to an active site scan to find positions accommodating the nucleophile histidine, followed by sequence optimization for stabilizing a high-energy state for rate acceleration,^{3,15}. Although their “protozymes” exhibited insufficient activity, this work paved the way for improving computational enzyme design methodologies to create *de novo* enzymes.

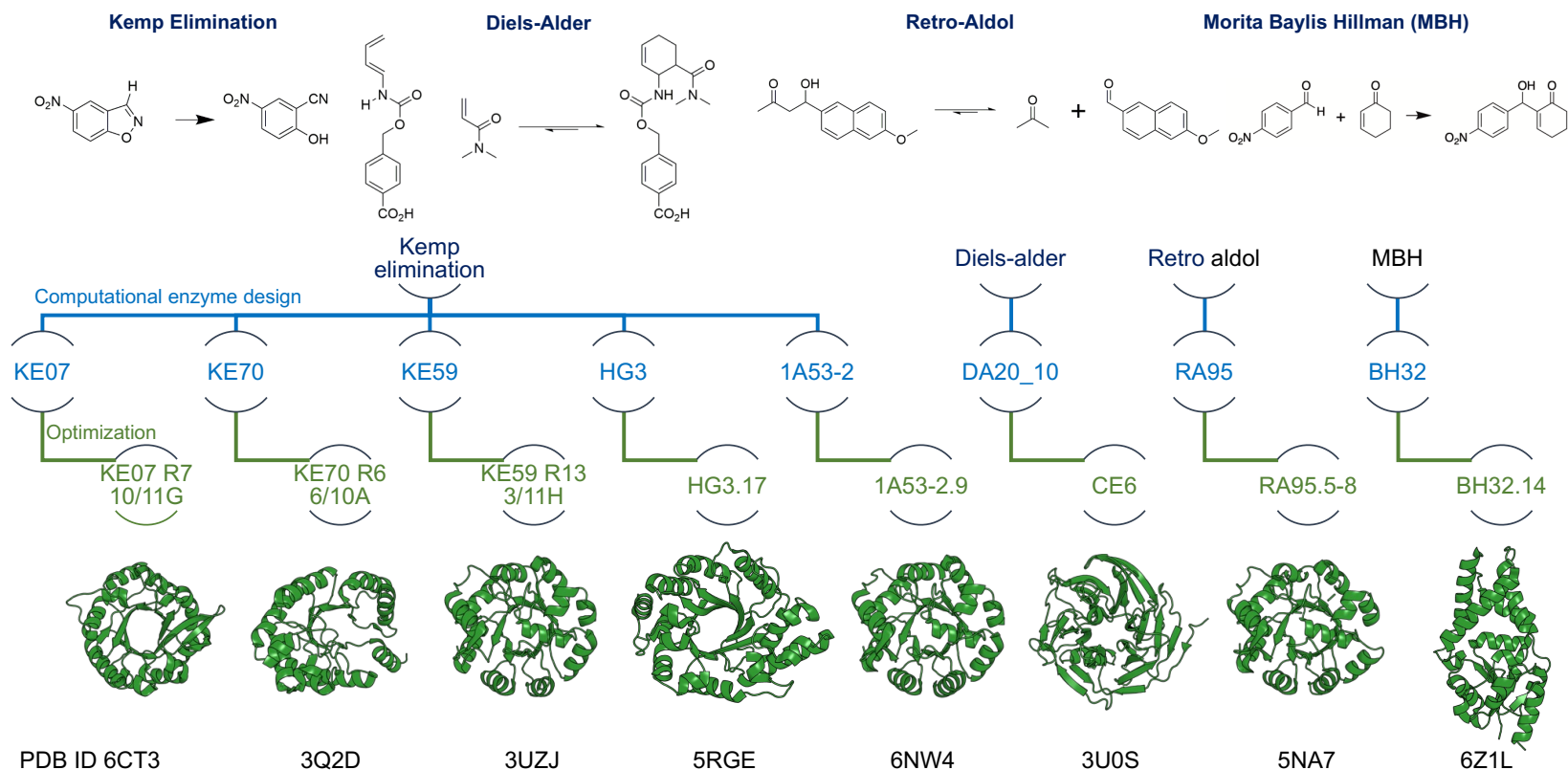


Figure 1.6. Examples of de novo enzyme. Only examples described in this chapter are shown here. At the top, four different reactions were successfully designed using the computational enzyme design framework developed by both the Baker group and the Mayo group. In the second row, a non-exhaustive list of artificially designed enzymes is shown in blue. In green, the respective variants optimized through either directed evolution or Foldit for DA20_10 are shown. The crystal structures of the best variants are shown in the last row in green.

1.3.1. Bond cleaving reactions

1.3.1.1. Single-step reaction: the Kemp elimination

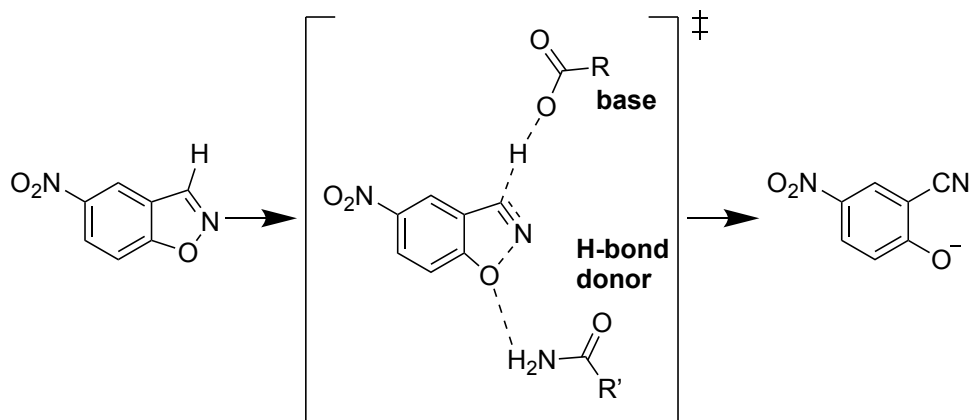


Figure 1.7. Kemp elimination reaction scheme. The reaction proceeds with the concerted base promoting E2 elimination and ring opening of 5-nitrobenzisoxazole leading to o-cyanophenolate. An H-bond donor is used to stabilize the developing negative charge on the phenolic oxygen from the transition state.

The Kemp elimination reaction provides an ideal model reaction for developing computational enzyme design algorithms. It is a well-defined model for catalysis of proton transfer from carbon and a rate-determining step in numerous enzymes. This model reaction is the concerted base promoting E2 elimination and ring opening of 5-nitrobenzisoxazole leading to the o-cyanophenolate product, in which an acid can be placed to further stabilize the charge dispersion on the oxygen at the phenol ring (Figure 1.7).

The David Baker group and the Stephen Mayo group introduced the Kemp elimination reaction into inert protein scaffolds. The Kemp elimination transition state coordinates were obtained from DFT calculations and were introduced into scaffolds with various folds. This approach led to eight active designs, with modest activities from two catalytic motifs. The catalytic activity of KE07,⁶¹ was improved by directed evolution using seven rounds of random mutagenesis and DNA shuffling. This led to a variant with catalytic efficiency improved by 200-fold compared to KE07. Mutations were found near the designed residues, while the catalytic

residues remained the same. The optimization process leads to better enzyme solubility and stability thanks to surface mutations, improvement of binding affinity, optimization of the electrostatic environment of the catalytic base, leading to the increase in pKa. ⁴ Following this work, KE70 has been improved by combining computational optimization and directed evolution. Rounds of optimization included the alternation of random mutagenesis and computational approaches such as increased sampling from β -strands perturbations using an ensemble generated by Rosetta Backrub, insertions in loops at the top of the active site, and loop redesign. This approach led to a 455-fold improvement in catalytic efficiency compared to the designed enzyme, in which random mutagenesis and computational optimization introduced an equal proportion of mutations. Interestingly, the two first rounds of directed evolution introduced mutations inside the active site, which they hypothesized could be predicted by computational methods. Throughout evolution, KE70 active site became more rigid. Similar to KE07, optimization of KE70 did not change the catalytic base initially designed. KE59 is another variant the David Baker lab designed and optimized through directed evolution. In the case of this designed enzyme, KE59 was less stable as its expression was low (< 2mg/L of culture). The evolvability of KE59 could be hampered by its low solubility and stability. Thus, their strategy to improve the activity included consensus mutations introduced through combinatorial incorporation from multiple sequence alignment with proteins homologous to Indole-3-glycerol phosphate synthase, which corresponds to the scaffold used for designing KE59. To improve its catalytic efficiency, mutations were introduced by combining random mutagenesis and introducing consensus mutations to increase the activity of KE59 variants. Their evolutionary strategy included using activated and less activated substrates as selective pressure. The best variant obtained by directed evolution showed an improved catalytic efficiency of 375-fold. The most active variant acquired 23 mutations, nine of which come from

consensus mutations and the remaining from random mutagenesis. Analysis of change in the active site environment by Brønsted slopes revealed that the best variants active site are better suited for stabilizing the developing negative charge in the transition state.

Another attempt to introduce the Kemp elimination reaction in an inert protein scaffold has been led by Privett et al.⁶⁰. In contrast to the work conducted by Röthlisberger et al.⁴, the design was focused on a single protein scaffold. Their initial design led to an inactive variant called HG1. Molecular dynamics simulations revealed high flexibility of the active site with high accessibility of water molecules inside. Instead of mutating the active site residues to more hydrophobic residues, they created a new pocket, which was more buried inside the barrel and designed the active variant HG2. This strategy occluded the active site from water molecules, increasing the catalytic base's reactivity. MD simulations on HG2 enabled the design of the HG3 variant with a mutation S265T. This variant exhibited a 4-fold improvement in catalytic efficiency compared to HG1. This work showed the importance of using MD simulations to guide designs⁶⁰. The specific case of HG1 revealed a high desolvation barrier highlighted by the presence of a high number of water molecules.^{4,54} This work has shown that higher catalytic activity is achieved by a more hydrophobic and deeper active site.⁶⁰ Desolvation of D127 designed as the catalytic base through an aprotic environment improves its reactivity by increasing its pKa.

HG3 has been subjected to 17 rounds of directed evolution, in which the catalytic activity of the best variant improved by around 600-fold compared to the starting point. Like KE07 and KE59, mutations accumulated in the active site during the first rounds of evolution, while last rounds of evolution introduced mutations outside of the active site. This study revealed two important structural improvements upon the introduction of beneficial mutations. The crystal

structure of the highest active variant shows a better active site shape complementarity with the ligand, and alignment of the catalytic base Asp127 with the ligand was optimized. Finally, an h-bond donor stabilizing the developing charge on the phenolic oxygen of the transition state has emerged. Together, these improvements contribute to an increase in catalytic efficiency.³⁶

Another Kemp eliminase designed by Privett et al.⁶⁰ is 1A53-2. It was initially designed to catalyze the proton abstraction of 5-nitrobenzoxisazole using their iterative computational approach. This designed enzyme was optimized by screening a less reactive substrate, 6-nitrobenzisoxazole. The best variant 1A53-2.9 catalytic efficiency improved by 8900-fold compared to the initial design catalyzing the reaction with 6-nitrobenzisoxazole with a modest activity ($k_{\text{cat}}/K_{\text{M}} \sim 3.9 \text{ M}^{-1} \text{ s}^{-1}$). The loose binding site is increasingly rigidified throughout evolution to allow a more tightly bound transition state, decreasing the conformational degree of freedom of the transition state.⁶²

1.3.1.2. Multi-step reaction: retro-aldol condensation

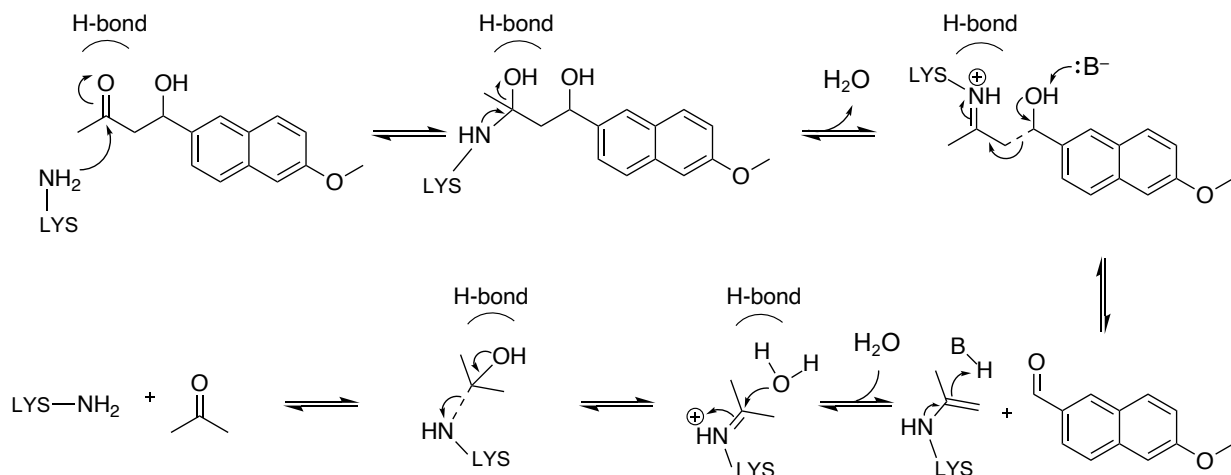


Figure 1.8. The Retro-Aldo condensation reaction scheme. The reaction is triggered by a nucleophilic attack of lysine on the ketone of the substrate, forming a carbinolamine intermediate, eliminating a water molecule and form the iminium species. A carbon-carbon bond cleavage is initiated by the deprotonation of the β -alcohol, with the

iminium acting as an electron sink. Lastly, the enamine tautomerizes to an imine that is then hydrolyzed to release the covalently bound product and free the enzyme.

The first design of a model reaction with multiple transition states was conducted by Jiang et al.⁵, in which a Retro-Aldol condensation was introduced. The retro aldol condensation is a complex multi-step reaction encompassing various transition states. In this study, the reaction corresponds to the carbon-carbon bond breaking of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (Figure 1.8). A composite transition state was employed to accommodate the diverse structures of these transition states. This approach incorporated four distinct motifs into scaffolds, including TIM-barrel and Jelly roll folds. Designs in TIM-barrel folds turned out to lead to active retro aldolases. They have more enclosed active site pockets, which may restrict substrate entrance and product release.⁵

The reaction utilizes a Schiff base from a lysine, which acts as an electron sink during C–C bond cleavage. The composite transition state was used to generate four distinct motifs. Among them, one motif successfully enabled the design of artificial retro aldolases, where a water molecule was explicitly modelled to facilitate proton transfers involving the β -alcohol and the hydroxyl group of the carbinolamine reaction intermediate. Their activities reached up to four orders of magnitude over the uncatalyzed reaction, although their rate enhancement was far from naturally occurring enzymes. Althoff et al.⁷ tested the robustness of their procedure by improving the active site preorganization of their designed enzymes by repacking and minimizing the designed sidechains with and without transition state. They created a bigger active site to accommodate the carbinolamine alcohol formation. Their designs led to sixty-five active designs on fourteen protein backbones.^{5,7} Among the active enzymes, the modest rate enhancement activity of RA95 was improved by 4,400-fold through directed evolution. During evolution, a new lysine has emerged and is hypothesized to activate the Schiff base-forming lysine by decreasing its pKa through a

better electrostatics environment. Crystal structure analysis of the evolved variant showed that the lysine initially designed to act as a Schiff base is has lost its function over the appearance of a new lysine during evolution. In addition to this new active site, a long loop close to the active site change in conformation might contribute to the increased activity, maybe because of displacement of the loop to let the substrate get in the active site and the product to be released.⁶³

A second retro aldolase RA110 has been evolved through directed evolution. Evolution has changed the binding mode and has introduced an H-bond donor binding the β -carbonyl group of the inhibitor used for crystallography. A water molecule and an adjacent tryptophan coordinate an additional H-bond with the inhibitor. Directed evolution has remodelled the backbone and the loops to enable the emergence of a new surface groove to accommodate the ligand. Consistent with other studies, RA110 evolution highlighted the importance of accounting for the backbone flexibility to increase the accuracy of the prediction.⁶⁴

Retro aldolases include multiple transition states, and a more accurate prediction will imply the implicit modelling of the transition state along the reaction coordinate.

1.3.2. Bond-forming reactions

Chemical reactions involving carbon-carbon bond formation are important reactions for many industrial applications as they allow the synthesis of complex molecules from simple building blocks. These reactions are more challenging to design as they require precise orientation of the substrates in the active site.

1.3.2.1. Single-step reaction: Diels-Alder cycloaddition

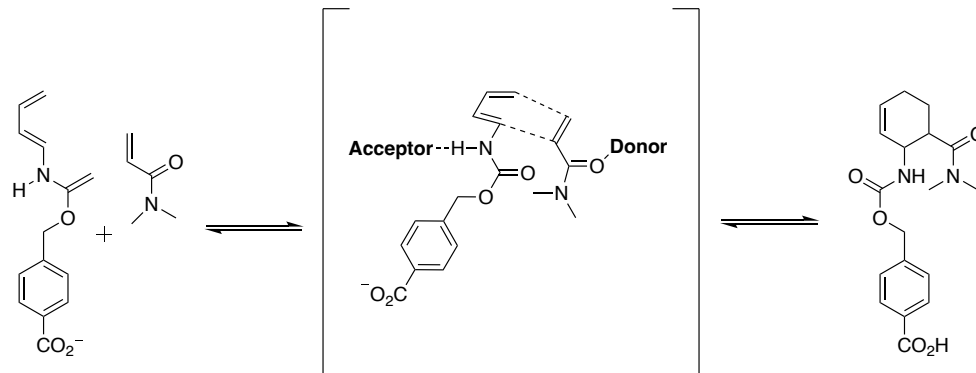


Figure 1.9. The Diels-Alder cycloaddition reaction scheme. The reaction proceeds to the [4 + 2] cycloaddition of 4-carboxybenzyl-*trans*-1,3-butadiene-1-carbamate on *N,N*-dimethylacrylamide. In this reaction scheme, an acceptor is used to activate the 4-carboxybenzyl-*trans*-1,3-butadiene-1-carbamate and a donor is used to activate *N,N*-dimethylacrylamide.

An interesting model reaction is the Diels-Alder bimolecular reaction successfully introduced in an inert protein scaffold by Siegel et al. ⁶. This artificial Diels-Alderase catalyzed the [4 + 2] cycloaddition of 4-carboxybenzyl-*trans*-1,3-butadiene-1-carbamate (diene) on *N,N*-dimethylacrylamide (dienophile) (Figure 1.9). A composite transition state where a QM-derived *endo-re* transition state, the two substrates bound in a transition-state-like orientation and conformation, and the *3R,4S endo* product was generated. Their approach aimed to decrease the energy gap between this reaction's HOMO (diene) and LUMO (dienophile) states using several theozymes. A large ensemble of minimal active sites using QM/MM was generated by scanning the whole PDB database, where a library of 207 stable protein scaffolds allowed the selection of 84 designs for experimental characterization. Their designs spanned over beta-propeller folds. Out of 50 designs, two had Diels-Alderase activity. To afford for higher activity, computational methodologies should improve the precision of the substrates' orientations relative to one another and to the catalytic residues during the design procedure, allowing for higher prediction accuracy.

1.3.2.2. Multi-step reaction: the Morita Baylis Hillman reaction

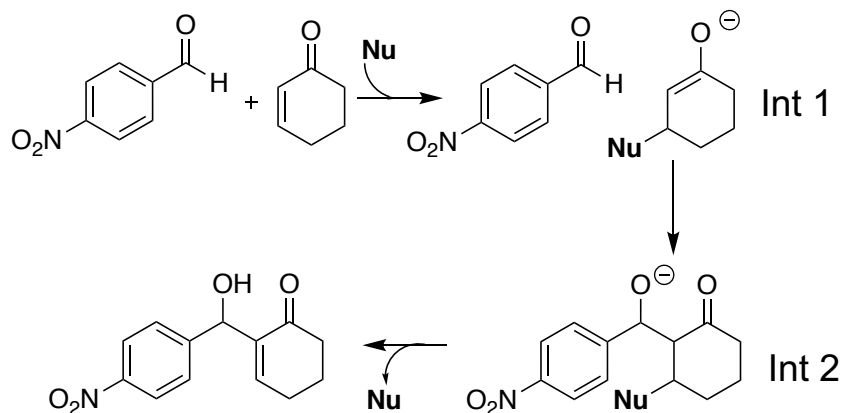


Figure 1.10. Morita Baylis Hillman reaction (MBH). Reaction between 4-nitrobenzaldehyde and cyclohexenone. First, an activated nucleophile attacks the β -carbon of the α,β -unsaturated carbonyl compound, which leads to the formation of a carbon-carbon bond between the α -carbon of the Michael acceptor and the nucleophile. This attack triggers the addition of the carbon-carbon bond between the carbon at the α position of cyclohexenone and the electrophilic carbon from 4-nitrobenzaldehyde.

Designing enzymes capable of catalyzing the Morita Baylis Hillman (MBH) reaction is highly desirable as it generates highly functionalized products that can be used for subsequent synthetic steps. Bjelic et al., 2013 introduced the MBH reaction between 4-nitrobenzaldehyde and 2-cyclohexenone (Figure 1.10). Similar to the retro-aldolases, a composite transition state was generated by overlay of QM- and MD-derived transition states and reaction intermediates. Two motifs were generated for this reaction, in which either histidine stabilized by an aspartate or glutamate, or cysteine activated by a lysine were used as nucleophiles. Pairs of backbone amide and/or H-bond donor amino acids were used to stabilize the enolate formed from 2-cyclohexenone, aromatic amino acids to stabilize 4-nitrobenzaldehyde via π -stacking, and a water molecule was used to facilitate the final deprotonation. This procedure enabled the successful design of *de novo* enzymes catalyzing the MBH reaction. However, the activity was low and catalytic efficiency was optimized by testing point mutants based on visual inspections and automated scanning in Rosetta. Experimental analysis showed the importance of a precise orientation of catalytic residues for

substrate binding. In the case of MBH, it is a more challenging procedure as the two substrates' specificity must be considered. A knockout of the catalytic residue abolished activity, showing the importance of the designed nucleophile residue. In addition, one- μ s MD simulation revealed a helix shift, which has a consequence for sidechain positioning, revealing the importance of considering backbone motions in the design procedure.⁸

BH32 has been optimized through directed evolution,⁶⁵ in which the activity increased by 1000-fold for the best variant BH32.14. The backbone scaffold has been extensively remodelled throughout the evolution, and 23 positions were mutated. Analyzing the mutations introduced during evolution revealed that the initially designed hydrogen bond donor Gln128 was abandoned as it was mutated to a proline. Structural analysis and density functional theory (DFT) of one of the best variants of the evolutionary trajectory of BH32 showed that Arg124 is well placed for a bidentate hydrogen bonding with the developing charges on both the oxyanion on the first intermediate and the second intermediate,⁶⁵ allowing for better stabilization of the multiple transition states, and higher rate acceleration.

While the prospect of introducing catalytic reactions catalyzed by enzymes that do not exist in nature is undeniably exciting, the design procedure presents many challenges. This is highlighted by the need to use directed evolution to improve their catalytic efficiencies.

1.4. Challenges

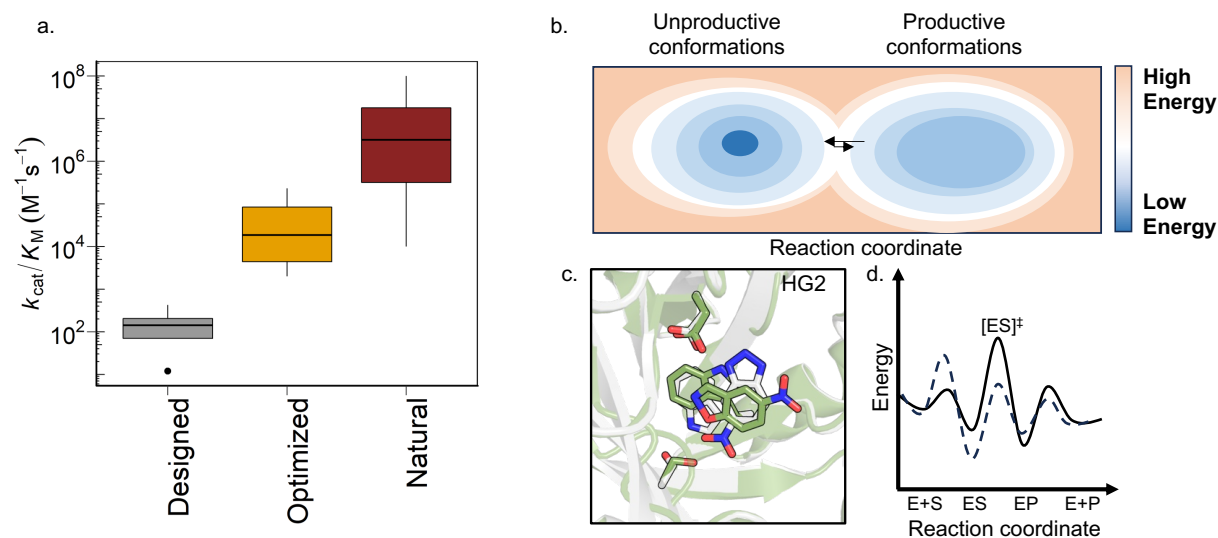


Figure 1.11. Challenges in computational enzyme design. a) Designed enzymes have catalytic efficiency orders of magnitude lower than natural enzymes, and they need to be optimized by directed evolution to increase their activity. b) A theoretical energy landscape of a low-activity enzyme, which can sample unproductive and productive conformations in a similar proportion. c) Examples of de novo-designed enzyme active sites. c) Crystal structures of HG2 (PDB ID: 3NYD) coloured in grey are superimposed with its designed model in forest green. For each enzyme, designed catalytic residues and their interacting ligands are shown in stick representations. Here, the crystal structure does not interact with the transition state analogue as predicted by the model. d) A theoretical model showing the limitation of the transition state stabilization, which can result to an overstabilization of the enzyme-substrate complex, leading to a higher energy barrier for active site or product release, decreasing catalytic efficiency.

As mentioned in previous sections, creating enzymatic systems from scratch for reactions with various complexity holds tremendous potential for computational methods in enzyme engineering and design.^{58,59} Although these methods enabled the successful design of various model reactions,^{3–9,60} their catalytic efficiencies are lower than the natural ones. Their catalytic rate enhancements range from k_{cat}/k_{uncat} 10^3 to 10^6 , which falls below natural enzymes' catalytic rate enhancement (typically ranges from 10^{11} to 10^{26}).¹⁰ To boost their modest activity, these enzymes are subjected to rounds of directed evolution to reach catalytic efficiencies comparable to those of naturally occurring enzymes.^{4,36,54,55,63,64} Catalytic rates fold improvements of these optimized enzymes compared to the designed ones ranged between 200-fold to roughly 8000-fold,

highlighting the need to improve the computational enzyme design methodology.

A combination of computational methods, such as molecular dynamics simulation techniques and analysis of crystallographic structures of the designed enzymes and their improved variants, revealed many pitfalls regarding the initial design methodology. Among them is the poor prediction of the binding poses confirmed by high RMSD between the designed model and its crystal structure. Secondly, introducing destabilizing mutations during the design induces conformational heterogeneity, highlighted by high flexibility of side chains or loops in critical regions.

1.4.1. Prediction inaccuracy

Various methods were employed to decipher the reasons for low catalytic efficiency. Catalytic residues initially predicted to interact with the ligand adopt different conformations, leading to the loss of the predicted binding poses as observed in the case of KE07, KE59 and HG2 models superposed with their crystal structures (example Figure 1.11). Another active design, 1A53-2, was designed using the same computational approach, with a catalytic efficiency of $62 \text{ M}^{-1} \text{ s}^{-1}$. As observed with HG2, the ligand has flipped in its crystal structure compared to the designed model, highlighting the inaccuracy of the design procedure.⁶⁰ The difference in the rotamer states between the predicted model and the crystal structure may result from unanticipated interactions with the catalytic residues with other residues in its vicinity. For example, in the case of KE07, improvement in catalytic efficiency was likely due to a shift in the catalytic residue Glu101 pKa (from <4.9 to 5.9). This change was hypothesized to result from a nearby mutation Ile7Asp, which holds the Lys222 and prevents it from interacting with Glu101.⁴ In the case of the first-generation design of HG3, HG1 did not show any activity toward the

substrate. When subjected to MD simulations with explicitly modelled water molecules, the trajectory samples conformational states that significantly differ from the static structure obtained through computational design processes.⁶⁷

1.4.2. Presence of multiple conformational states

Studying the conformational states that are sampled during MD simulations low activity enzymes revealed that conformational heterogeneity contributes to their low efficiency.^{55,60} Conformational heterogeneity can be described as the ability of an enzyme to access multiple conformations (Figure 1.11.b). Molecular dynamics studies and crystal structure analysis of the designed enzymes show higher flexibility of residues within the active site than their evolved counterparts. This higher flexibility might lead to the loss of critical interactions.⁶⁰

Analysis of the variants from the evolutionary trajectory of KE07 reveals that the improvement in catalytic activity during the early rounds of evolution was likely due to decreased activation energy (E_a). However, this decrease in E_a was offset by high entropy. Computational analysis showed that the catalytic base Glu101 adopted different conformations from the first to the fourth round of directed evolution. Structural analysis of the variants showed the presence of multiple active site configurations in the less active variants, consistent with high collision frequency. During directed evolution, the conformational heterogeneity increasingly disappeared, while stabilizing a single active site configuration. These observations come from residues increasingly enriching rotamers' active configuration during evolution. In this study, the authors highlighted the importance of substitutions outside the active site that improve catalytic activity by stabilizing the most catalytically active configuration.⁶¹ These observations were corroborated by structural analysis of the best variant from the evolutionary trajectory of KE59. In the case of

this enzyme, the catalytically active conformation of the catalytic base E230 was stabilized thanks to neighbouring mutations. It was reflected by the absence of multiple conformations of this residue in the crystal structure.⁵⁵

In the case of HG2, the second generation of the computationally designed enzymes from the HG-series, its crystal structure shows a binding pose, where the position of the transition state analogue used for crystallization differs from the position of the transition state of the designed model. This observation supports results on MD simulations on HG2 crystal structure showing the emergence of two distinct active site conformations. The transition state analogue was flipped, and the initial hydrogen bond donor S265 was abandoned leading to an unproductive pose. The two possible binding poses observed from the crystal structure of HG2 may contribute to its low activity ($k_{\text{cat}}/K_{\text{M}}$ 122 M⁻¹ s⁻¹). MD simulations of HG2, using its crystal structure only bound with the productive pose of the TSA as input structure, showed that the high degree of liberty of S265 renders this residue to be too flexible during the simulation and contribute to a high structural heterogeneity of the active site. A mutation of serine to a threonine, a residue that has less possible rotamer configurations compared to a serine would allow a better packing of the active site.

1.4.3. Stabilizing multiple transition states

In complex reactions with multiple transition states, such as retro-aldolase and MBH reactions, a composite transition state was used. Several shortcomings can arise when the explicit consideration of all the transition states of the reaction is disregarded. First, the inaccurate substrate alignment with the catalytic residue can lead to a suboptimal positioning of active site residues relative to the catalytic residues, as observed in the case of BH32, where key-designed residues adopt different conformational states, especially the designed Gln128 predicted to stabilize the

oxyanion hole. Or in the case of RA95, the low reactivity of the imine-forming lysine contributes to its low k_{cat} . Second, as depicted in Figure 1.11.d, stabilization of the transition state due to overpacking and overstabilization of intermediates may lead to a higher energy barrier for the reaction to overcome and a slower reaction rate.

One major limitation of traditional methods used to design the enzymes enumerated in this study is the single-state design approximation, where a single crystal structure was used during the design procedure (Figure 1.5.b). However, enzymes are dynamic adopting multiple conformations in solution. Using a single backbone during the design procedure might lead to weak transition state binding and limit the accurate positioning of catalytic residues. Using multiple backbone would approximate the flexibility of the enzyme in solution and allow the sampling of active conformational states and address the limitation introduced by the single backbone approximation.

4

1.5. Incorporating backbone flexibility

Introducing backbone flexibility during theozyme placement and active site repacking will ensure a precise positioning of catalytic residues, which is crucial to achieve high activity.⁶⁸ Most of the previous computational enzyme design methodologies have primarily focused on the chemical aspects of the reaction, in which only the nature of active site residues were optimized. However, crucial stages such as substrate binding, product release and the alteration of conformational states throughout catalysis have been disregarded. The observed low prediction accuracy resulted in a higher number of false positive designs and significantly reduced the overall success rate of the computational enzyme design process.

Consequently, efforts have to be made to incorporate enzyme flexibility in the design to improve the effectiveness of the design methodology. To address the observed limitations due to the use of a static structure during the design procedure, backbone flexibility needs to be introduced in the methodology to account for the dynamic behaviour of enzymes, allowing a more precise positioning of catalytic residues. To simulate the effect of backbone flexibility of the designed enzyme KE70, ensembles were generated using a method called backrub developed by the Baker lab,⁶⁹ and CPD was conducted on multiple templates. However, only 2 to 4-fold improvement was achieved using this method,⁵⁴ and directed evolution was necessary to improve the catalytic efficiency by ~ 400-fold.

Introducing flexibility into the design procedure is required for enhancing prediction accuracy. However, the methods used to obtain backbone ensembles will result to different backbone conformations, thereby influencing prediction accuracy.⁵¹ Consequently, it is imperative to generate ensembles that effectively model the conformational flexibility of the enzyme in solution.

1.6. Thesis objectives

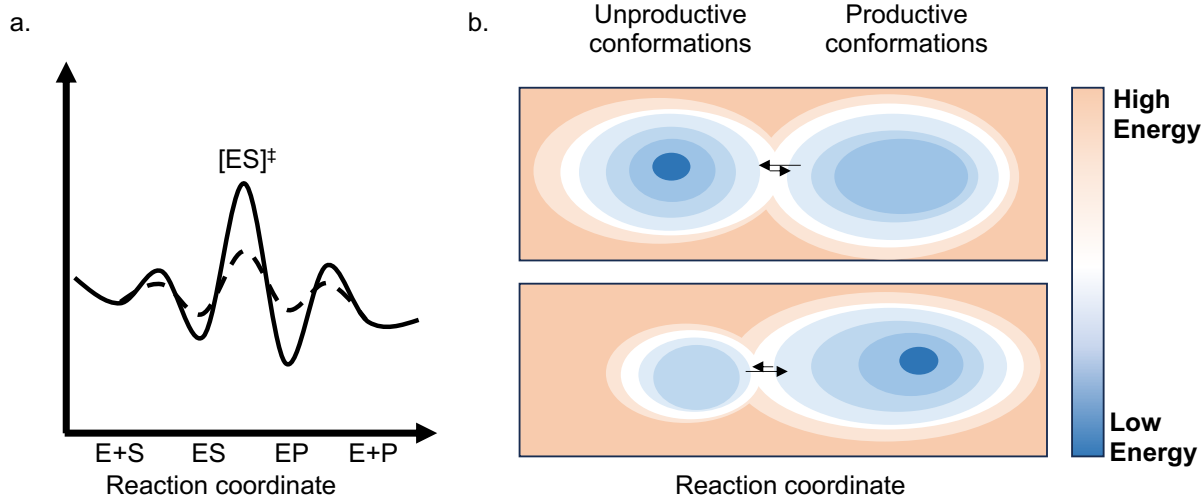


Figure 1.12. Ensemble-based enzyme design goal. a) the ultimate goal of the ensemble-based enzyme design approach is to flatten the energy landscape of the enzyme reaction by improving (1) substrate entrance, (2) active site preorganization through destabilizing the substrate ground state, (3) stabilizing the transition state, and (4) destabilizing the enzyme-product complex to (5) facilitate product release. b) Using the ensemble-based method, we aim to stabilize existing productive conformations to improve the efficiency of artificial enzymes

In this thesis, the main objective is to devise a robust computational methodology for improving the catalytic efficiency of low-activity enzymes by improving prediction accuracy. Using this ensemble-based enzyme design methodology, we aim to use an ensemble of backbones capable of sampling the enzyme's conformational states in solution (Figure 1.12.a). Moreover, enzymes with low activity often exhibit a dynamic interplay between productive and unproductive states. Leveraging an ensemble of diverse enzyme backbones makes it possible to explore multiple catalytically active states. These states can be further stabilized by introducing mutations during the design process, ultimately facilitating a shift towards active conformations (Figure 1.12.b).²² By allowing precise catalytic side chain orientation, we aim to address the limitations enumerated in earlier sections, allowing for stronger TS stabilization, better TS active site shape complementarity, better active site preorganization and enrichment of catalytically active conformational states.

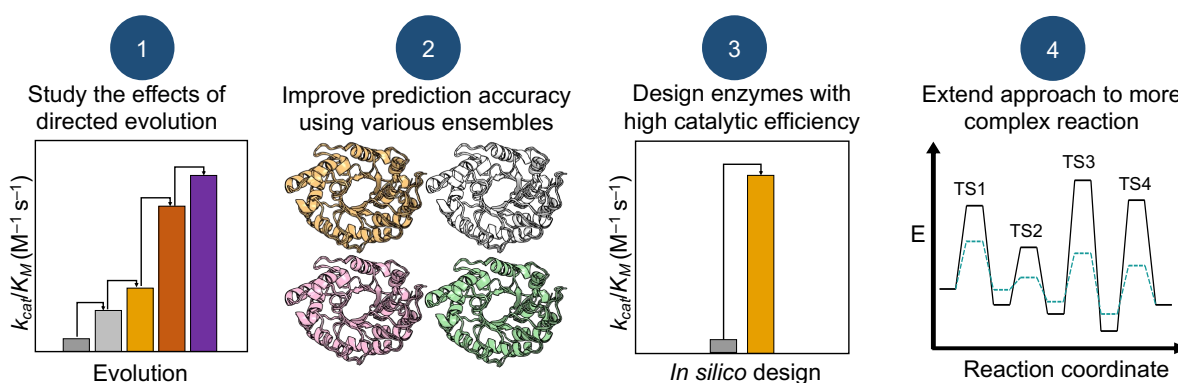


Figure 1.13. Thesis objectives. To achieve the goal of ensemble-based design, the first objective of my thesis is to (1) study the evolutionary trajectory of a low-activity enzyme using structural information from directed evolution, (2) creating a benchmark of ensembles to find the best ensembles for improving the prediction (3) using the

information from directed evolution and from the benchmark to design efficient artificial enzymes (4) using a benchmark for ensemble-based design to stabilize multiple transition states from a more complex reaction.

To achieve the goal of my thesis, my first objective is to study the crystal structure of the variants from the evolutionary trajectory of HG3, ^{36,60} a weakly active enzyme, to study the structural effects of directed evolution that lead to the increase in catalytic efficiency. Following this study, in a second objective, a benchmark of ensembles from four distinct methods will be generated to find the ensemble that allows the most accurate prediction of the active site of the best variants from four designed Kemp eliminases. In this chapter, ensembles from AlphaFold2, ^{70,71}, MD simulations, ensemble refinement, ^{72,73} and PertMin, ^{51,52} will be generated from the crystal structure used for the design, the crystal structure of the unbound designed enzyme and the crystal structure of the bound designed enzyme when available. The ensembles generated will be compared and discussed. In the third objective, I will apply the ensemble-based method to improve the catalytic efficiency of the computationally designed enzyme HG3. Lastly, I aim to extend the methodology developed in the previous chapters to establish a benchmark on a more complex reaction, the Morita-Baylis-Hillman reaction. Finally, I will discuss the success of my methodology and its limitations and perspectives.

The successful development of a robust methodology to design efficient artificial enzymes will have two significant impacts. First, it addresses the increasing demands of various industrial applications. Second, it contributes to a deeper understanding of enzymatic processes, shedding light on the fundamental workings of enzymes.

References

1. Gurung, N., *et al.* A Broader View: Microbial Enzymes and Their Relevance in Industries, Medicine, and Beyond. *BioMed Res. Int.* **2013**, 329121 (2013).
2. Robinson, P. K. Enzymes: principles and biotechnological applications. *Essays Biochem.* **59**, 1–41 (2015).
3. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. **2001**, (2001).
4. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
5. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
6. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **329**, 309–313 (2010).
7. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
8. Bjelic, S. *et al.* Computational design of enone-binding proteins with catalytic activity for the morita-baylis-hillman reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).
9. Richter, F. *et al.* Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).
10. Edwards, D. R., *et al.* Catalytic proficiency: The extreme case of S-O cleaving sulfatases. *J. Am. Chem. Soc.* **134**, 525–531 (2012).
11. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch. Chem. Ges.* **27**, 2985–2993 (1894).
12. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
13. Leonor Michaelis, M. L. M. Die Kinetik der Invertinwirkung. *Biochem. Z.* **49**, 333–369 (1913).
14. Keith Laidler & King, C. The development of Transition State Theory. **87**, 2657–2664 (1983).
15. Pauling, L. CHEMICAL ACHIEVEMENT AND HOPE FOR THE FUTURE. *Am. Sci.* **36**, 51–58.
16. Frauenfelder, H., *et al.* Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **280**, 558–563 (1979).
17. Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117–121 (2005).
18. Eisenmesser, E. Z., *et al.* Enzyme Dynamics During Catalysis. *Science* **295**, 1520–1523 (2002).
19. Volkman, B. F., *et al.* Two-State Allosteric Behavior in a Single-Domain Signaling Protein. *Science* **291**, 2429–2433 (2001).
20. Smith, A. J. T. *et al.* Structural Reorganization and Preorganization in Enzyme Active Sites: Comparisons of Experimental and Theoretically Ideal Active Site Geometries in the Multistep Serine Esterase Reaction Cycle. *J. Am. Chem. Soc.* **130**, 15361–15373 (2008).
21. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).
22. Romero-Rivera, A., *et al.* Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **7**, 8524–8532 (2017).
23. Agarwal, P. K., *et al.* Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci.* **99**, 2794–2799 (2002).
24. Süel, G. M., *et al.* Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59–69 (2003).
25. Böde, C. *et al.* Network analysis of protein dynamics. *FEBS Lett.* **581**, 2776–2782 (2007).
26. O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).
27. Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
28. Joerger, A. C., *et al.* Mimicking natural evolution in vitro: An N-acetylneuraminate lyase mutant with an increased dihydrodipicolinate synthase activity. *Proc. Natl. Acad. Sci.* **100**, 5694–5699 (2003).
29. Poelarends, G. J. *et al.* Evolution of Enzymatic Activity in the Tautomerase Superfamily. *Biochemistry* **45**, 7700–7708 (2006).
30. Wise, E. L., *et al.* Evolution of enzymatic activities in the orotidine 5'-monophosphate decarboxylase suprafamily: structural basis for catalytic promiscuity in wild-type and designed mutants of 3-keto-L-gulonate 6-phosphate decarboxylase. *Biochemistry* **44**, 1816–1823 (2005).
31. Arnold, F. H. Directed evolution: Creating biocatalysts for the future. *Chem. Eng. Sci.* **51**, 5091–5102 (1996).

32. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed Engl.* **57**, 4143–4148 (2018).
33. Coelho, P. S., *et al.* Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome P450 enzymes. *Science* **339**, 307–310 (2013).
34. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. <https://www.science.org/doi/10.1126/science.aah6219> doi:10.1126/science.aah6219.
35. Kan, S. B. J., *et al.* Genetically programmed chiral organoborane synthesis. *Nature* **552**, 132–136 (2017).
36. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
37. Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475 (1998).
38. Altman, M. D., *et al.* Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. *Proteins* **70**, 678–694 (2008).
39. Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656–659 (2006).
40. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223–230 (1973).
41. Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
42. Schrauber, H., *et al.* Rotamers: To be or not to be?: An Analysis of Amino Acid Side-chain Conformations in Globular Proteins. *J. Mol. Biol.* **230**, 592–612 (1993).
43. Lovell, S. C., *et al.* The penultimate rotamer library. *Proteins Struct. Funct. Bioinforma.* **40**, 389–408 (2000).
44. Dunbrack, R. L. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681 (1997).
45. Gordon, D. B., *et al.* Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* 509–513 (1999) doi:10.1016/j.sbi.2007.03.006.
46. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10383–10388 (2000).
47. Voigt, C. A., *et al.* Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design | Edited by J. Thornton. *J. Mol. Biol.* **299**, 789–803 (2000).
48. Allen, B. D. & Mayo, S. L. An Efficient Algorithm for Multistate Protein Design Based on FASTER. (2009) doi:10.1002/jcc.
49. Metropolis, N., *et al.* Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (2004).
50. Desmet, J., *et al.* The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542 (1992).
51. Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins Struct. Funct. Bioinforma.* **82**, 771–784 (2014).
52. Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Sci.* **21**, 1241–1252 (2012).
53. Lauck, F., *et al.* RosettaBackrub—a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res.* **38**, W569–W575 (2010).
54. Khersonsky, O. *et al.* Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.* **407**, 391–412 (2011).
55. Khersonsky, O. *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10358–10363 (2012).
56. Tantillo, D. J., *et al.* Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).
57. St-jacques, A. D. Computational Enzyme Design : Successes , Challenges , and Future Directions. 88–116.
58. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
59. Lassila, J. K., *et al.* Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16710–16715 (2006).
60. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790–3795 (2012).
61. Hong, N. S. *et al.* The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.* **9**, (2018).

62. Bunzel, H. A. *et al.* Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **141**, 11745–11748 (2019).
63. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
64. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
65. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
66. Vaissier, V., *et al.* Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **8**, 219–227 (2018).
67. Kiss, G., Röthlisberger, D., Baker, D. & Houk, K. N. Evaluation and ranking of enzyme designs. *Protein Sci.* **19**, 1760–1773 (2010).
68. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
69. Mandell, D. J. & Kortemme, T. Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420–428 (2009).
70. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
71. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
72. Burnley, B. T., *et al.* Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).
73. Burnley, B. T. & Gros, P. phenix.ensemble_refinement: a test study of apo and holo BACE1. *Comput. Crystallogr. Newsl.* **4**, 51–58 (2013).

Chapter 2. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico

Copyright: Aron Broom*, Rojo V. Rakotoharisoa*, Michael C. Thompson, Niayesh Zarifi, Erin Nguyen, Nurzhan Mukhametzhanov, Lin Liu, James S. Fraser & Roberto A. Chica, *Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico*, Nature Communications, 2020, Springer Nature

*Authors contributed equally

2.1. Statement of contribution

A.B. and R.A.C. conceived the project. A.B., N.Z., E.N., N.M., and L.L. purified proteins. A.B. and N.Z. performed enzyme kinetics experiments. R.A.C. and M.C.T. crystallized proteins and performed X-ray diffraction experiments. A.B., R.V.R., M.C.T., and R.A.C. performed refinements. M.C.T. and J.S.F. designed X-ray crystallography experiments. R.V.R. and A.B. performed computational design experiments. R.A.C. wrote the manuscript. A.B., R.V.R., and M.C.T. edited the manuscript.

2.2. Abstract

The creation of artificial enzymes is a key objective of computational protein design. Although *de novo* enzymes have been successfully designed, these exhibit low catalytic efficiencies, requiring directed evolution to improve activity. Here, we use room-temperature X-ray crystallography to study changes in the conformational ensemble during evolution of the designed Kemp eliminase HG3 (k_{cat}/K_M 146 M⁻¹s⁻¹). We observe that catalytic residues are increasingly rigidified, the active site becomes better pre-organized, and its entrance is widened. Based on these observations, we

engineer HG4, an efficient biocatalyst ($k_{\text{cat}}/K_{\text{M}}$ 103,000 M⁻¹s⁻¹) containing key first and second-shell mutations found during evolution. HG4 structures reveal that its active site is pre-organized and rigidified for efficient catalysis. Our results show how directed evolution circumvents challenges inherent to enzyme design by shifting conformational ensembles to favor catalytically-productive sub-states, and suggest improvements to the design methodology that incorporate ensemble modeling of crystallographic data.

2.3. Introduction

Enzymes are the most efficient catalysts known, accelerating chemical reactions by up to 26 orders of magnitude,¹ while displaying unmatched selectivity. The ability to create, from scratch, an efficient artificial enzyme for any desired chemical reaction (i.e., a *de novo* enzyme) is a key objective of computational protein design. Progress towards this goal has been made over the past few decades following the development of computational enzyme design algorithms.^{1,2} These methods have been used to create *de novo* enzymes for a variety of model organic transformations including the Kemp elimination,^{1,2} retro-aldol,^{1,2} Diels-Alder,¹ ester hydrolysis,¹ and Morita-Baylis-Hilman¹ reactions. Although successful, catalytic activities of *de novo* enzymes have been modest, with $k_{\text{cat}}/K_{\text{M}}$ values being several orders of magnitude lower than those of natural enzymes.^{1,2} In addition, structural analyses of designed enzymes have revealed important deficiencies in the computational methodologies, resulting in inaccurate predictions of catalytic and ligand-binding interactions,¹ and thereby low success rates,¹⁻³ emphasizing the need for continued development of robust enzyme design algorithms.

To improve the catalytic activity of designed enzymes, researchers have used directed evolution. This process has yielded artificial enzymes displaying catalytic efficiencies approaching those of their natural counterparts, and provided valuable information about the structural determinants of efficient catalysis.¹⁻³ During evolution, active-site residues, including designed catalytic amino acids, were often mutated, leading to enhanced catalysis via introduction of new catalytic groups, optimization of catalytic contacts and ligand binding modes, and enhanced transition-state complementarity of the binding pocket.¹⁻³ Directed evolution has also yielded beneficial mutations at positions remote from the active site. Distal mutations have been shown to enhance catalysis by shifting the populations of conformational sub-states that enzymes sample on

their energy landscape towards those that are more catalytically active.¹⁻³ Therefore, a better understanding of enzyme conformational ensembles, including the effect of mutations on the population of sub-states, could provide valuable insights to aid in the development of robust computational enzyme design methodologies.

Here, we study changes in the conformational ensemble along the evolutionary trajectory of the *de novo* Kemp eliminase HG3 ($k_{\text{cat}}/K_{\text{M}}$ 146 M⁻¹s⁻¹) using room-temperature X-ray crystallography. We observe that during evolution, catalytic residues were increasingly rigidified through improved packing, the active site became better pre-organized to favor productive binding of the substrate, and the active-site entrance was widened to facilitate substrate entry and product release. Based on these observations, we generate a variant that contains all mutations necessary to establish these structural features, which are found at positions within or close to the active site. This variant, HG4, is >700-fold more active than HG3, with a catalytic efficiency on par with that of the average natural enzyme ($k_{\text{cat}}/K_{\text{M}}$ 103,000 M⁻¹s⁻¹). Crystallographic analysis of HG4 reveals that mutations proximal to the active site are sufficient to alter the conformational ensemble for enrichment of catalytically-competent sub-states. Lastly, we demonstrate that HG4 can be successfully designed using a crystallographically-derived ensemble of backbone templates approximating conformational flexibility, but not with the single template used to design HG3, offering insights for improving enzyme design methodologies.

2.4. Results

2.4.1. HG series of Kemp eliminases

Perhaps the most successful example of the improvement of a *de novo* enzyme by directed evolution has been the engineering of HG3.17, the most active Kemp eliminase reported to date.¹

This artificial enzyme catalyzes the concerted deprotonation and ring-opening of 5-nitrobenzisoxazole into the corresponding *o*-cyanophenolate (Figure 2.1a) with a reported catalytic efficiency of $2.3 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$.¹ HG3.17 was evolved from HG3, a higher-activity mutant (S265T) of the *in silico* design HG2 (Supplementary Figure 2.1) that was engineered post-design to reduce the active-site conformational heterogeneity observed by molecular dynamics analysis of HG2.¹ Over an evolutionary trajectory that yielded the HG3.3b, HG3.7, and HG3.14 intermediates (Figure 2.1b, Supplementary Figure 2.1), a total of 17 mutations were introduced into HG3 to produce HG3.17, resulting in a catalytic efficiency increase of approximately three orders of magnitude (Table 2.1, Supplementary Figure 2.2). Of these mutations, 11 occurred at positions within or close to the active site, including 8 at positions that were optimized during computational design of HG2 (Table 2.1). One of the key active-site mutations occurred at position 50, which was mutated twice during evolution, first from lysine to histidine (HG3 to HG3.3b) and then from histidine to glutamine (HG3.3b to HG3.7), resulting in a novel catalytic residue ideally positioned for stabilizing negative charge buildup on the phenolic oxygen at the transition state (Figure 2.1a).

Table 2.1. Kinetic parameters of Kemp eliminases

Enzyme	Mutations from HG3 ^a	$k_{\text{cat}} / K_{\text{M}}$ ($\text{M}^{-1} \text{s}^{-1}$) ^b
HG3	-	146 ± 6 (1300)
HG3.3b	V6I <i>K50H M84C</i> S89R <i>Q90D A125N</i>	2200 ± 100 (5400)
HG3.7	V6I Q37K <i>K50Q M84C</i> S89R <i>Q90H A125N</i>	27000 ± 2000 (37000)
HG3.14	V6I Q37K <i>K50Q</i> G82A <i>M84C Q90H</i> T105I <i>A125T</i> T142N T208M T279S D300N	52000 ± 1000 (70000)
HG3.17	V6I Q37K <i>N47E K50Q</i> G82A <i>M84C</i> S89N <i>Q90F</i> T105I <i>A125T</i> T142N T208M <i>F267M W275A R276F</i> T279S D300N	126000 ± 9000 (230000)
HG4	<i>K50Q</i> G82A <i>M84C Q90F A125T F267M W275A R276F</i>	103000 ± 4000

^a Mutations in italics occurred at sites optimized during computational design of HG2.⁵

^b Individual parameters K_{M} and k_{cat} could not be determined accurately because saturation was not possible at the maximum substrate concentration tested (2 mM), which is the substrate's solubility limit (Supplementary Fig. 2). Catalytic efficiencies ($k_{\text{cat}}/K_{\text{M}}$) were calculated from the slope of the linear portion ($[\text{S}] \ll K_{\text{M}}$) of the Michaelis-Menten model ($v_0 = (k_{\text{cat}}/K_{\text{M}})[\text{E}_0][\text{S}]$). $n = 2$ independent experiments for HG3, HG3.3b, HG3.7 and HG3.17. $n = 3$ independent experiments for HG3.14 and HG4. Errors of linear regression fitting, which represent the absolute measure of the typical distance that each data point falls from the regression line, are provided. Values in parentheses are from Blomberg *et al.*¹

Comparison of the crystal structure of the *in silico* design HG2 (PDB ID: 3NYD)¹ with that of a double mutant of HG3.17, in which surface mutations N47E and D300N were reverted to the corresponding amino acids found in HG2 to facilitate crystallization (HG3.17-E47N/N300D, PDB ID: 4BS0),¹ revealed that catalytic activity was also enhanced via optimized alignment of the transition-state analogue 6-nitrobenzotriazole (6NT) with the catalytic base Asp127 (Figure 2.1c), and improved active-site complementarity to this ligand (Figure 2.1d). Given that subtle changes to the conformational ensemble of an enzyme can lead to significant rate enhancements,¹⁻³ it is possible that mutations in HG3.17 also contributed to enhanced catalytic efficiency by altering the conformational landscape to enrich catalytically-competent sub-states. However, the structures of HG2 and HG3.17-E47N/N300D were solved in the presence of bound 6NT and at cryogenic temperatures, which could have shifted the conformational ensemble towards a single predominant sub-state, thereby limiting our ability to evaluate changes to the conformational landscape during directed evolution.

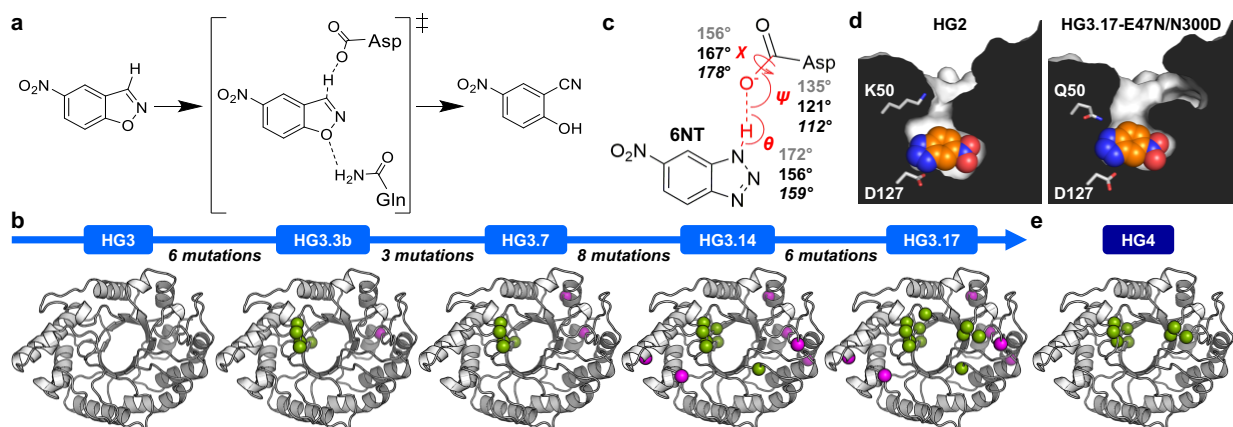


Figure 2.1. HG series of Kemp eliminases. (a) HG enzymes catalyze the Kemp elimination reaction using a catalytic dyad consisting of a base (Asp127) that deprotonates 5-nitrobenzisoxazole, and an H-bond donor (Gln50) that stabilizes negative charge buildup on the phenolic oxygen at the transition state ($\ddagger$). This reaction yields 4-nitro-2-cyanophenol. (b) Directed evolution of HG3, a higher-activity mutant (S265T) of the *in silico* design HG2. A total of 17 mutations (shown as spheres) were introduced during evolution, including 11 at positions within or close to the active site (green) and 6 at distal sites (magenta). (c) Angles describing the hydrogen bonding interaction between the transition state analogue 6-nitrobenzotriazole (6NT) and Asp127 in the HG2 (PDB ID: 3NYD)¹ and HG3.17-E47N/N300D (PDB ID: 4BS0)¹ crystal structures are indicated in grey and black, respectively. Values in italics are

optimal angles calculated for hydrogen bonding interactions between acetamide dimers.¹ (d) Cut-away view of the active-site pocket shows that its structural complementarity with 6NT (spheres) is improved in the higher activity variant HG3.17-E47N/N300D. Key active-site residues are shown as sticks. (e) The HG4 variant engineered in this study contains 8 first- and second-shell mutations found during evolution.

2.4.2. Room-temperature crystal structures

To evaluate changes to the HG3 conformational ensemble along its evolutionary trajectory, we solved room-temperature (277 K) X-ray crystal structures of all HG-series Kemp eliminases, both in the presence and absence of bound 6NT. Room-temperature X-ray crystallography can reveal conformational heterogeneity in protein structures that would not be visible at cryogenic temperatures and thereby provide insights into the conformational ensemble that is sampled by a protein in solution.¹ All five enzymes yielded crystals under similar conditions (Supplementary Table 2.2), and these diffracted at resolutions of 1.35–1.99 Å (Supplementary Table 2.3). All unit cells corresponded to space group $P2_12_12_1$ with two protein molecules in the asymmetric unit, except that of HG3.17, whose asymmetric unit was half the volume of the others and contained only one polypeptide chain although the space group was also $P2_12_12_1$. This result is in contrast with the deposited structure of HG3.17-E47N/N300D, which contains two molecules in the asymmetric unit, with identical space group and similar unit cell dimensions to those of all other HG variants reported here.¹ This discrepancy between our structure of HG3.17 and the previously published structure of HG3.17-E47N/N300D is likely caused by the presence of the Asn47 surface residue in all variants except for HG3.17, since this amino acid is involved in crystal packing interactions.

All HG-series enzymes bound 6NT in the same catalytically productive pose (Figure 2.2a) as that observed in HG2 and HG3.17-E47N/N300D (Figure 2.1d). In this pose, the acidic N–H bond of 6NT that mimics the cleavable C–H bond of the substrate is located within hydrogen-

bonding distance to the carboxylate oxygen of Asp127 (2.5–2.6 Å distance between heavy atoms), while

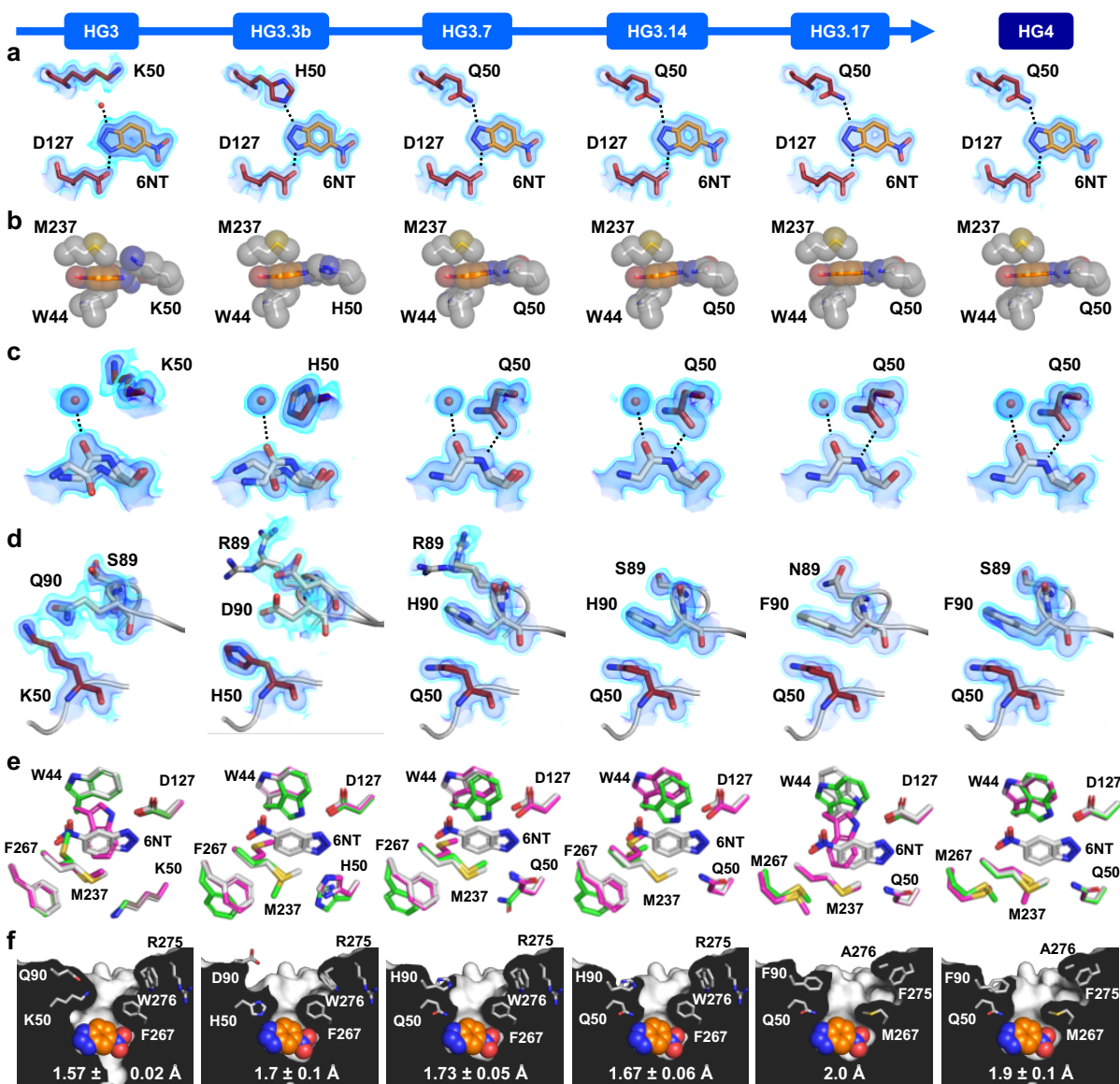


Figure 2.2. Crystal structures of HG-series Kemp eliminases. In all cases, only atoms from chain A are shown. (a) Binding pose of the 6NT transition-state analogue (orange). Hydrogen bonds are shown as dashed lines. The red sphere represents a water molecule. The 2Fo-Fc map is shown in volume representation at two contour levels: 0.5 eÅ⁻³ and 1.5 eÅ⁻³ in light and dark blue, respectively. (b) 6NT (orange) is sandwiched between the hydrophobic side chains of Trp44 and Met237. (c) The peptide bond between residues 83 and 84 can adopt *cis* or *trans* conformations. Hydrogen bonds are shown as dashed lines. The 2Fo-Fc map is shown in volume representation at two contour levels: 0.5 eÅ⁻³ and 1.5 eÅ⁻³ in light and dark blue, respectively. (d) Conformational changes to loop formed by residues 87–90 over the course of the evolutionary trajectory. The 2Fo-Fc map is shown in volume representation at two contour levels: 0.50 eÅ⁻³ and 1.5 eÅ⁻³ in light and dark blue, respectively. (e) Superposition of the 6NT-bound structure (white) with the highest (magenta) and lowest (green) occupancy conformers of the unbound structure for each Kemp eliminase. From HG3 to HG3.14, the unbound state is never pre-organized for catalysis as both Trp44 and Met237 adopt conformations that would prevent productive binding of the transition state. In HG3.17 and HG4 however, only Trp44 adopts a non-productive conformation in the unbound state, with an occupancy of 62% or 26%, respectively.

(f) Cut-away view of the active site shows that its entrance (top) becomes widened during evolution, as indicated by an increasing bottleneck radius (reported as the average radius $\pm$ s.d. calculated using the highest occupancy conformers from both chain A and B, except for HG3.17, which contains a single chain). 6NT is shown as orange spheres. Bottleneck radii were calculated using the PyMOL plugin Caver 3.0.1

the basic nitrogen atom corresponding to the phenolic oxygen of the transition state forms an H-bond with either a water molecule (HG3), the N_ε atom of His50 (HG3.3b), or the side-chain amide nitrogen of Gln50 (HG3.7, HG3.14, HG3.17). In addition to being held in place by these polar interactions, 6NT is sandwiched between the hydrophobic side chains of Trp44 and Met237 (Figure 2.2b), which are part of a mostly hydrophobic binding pocket that also includes the side chains of Ala21, Met/Cys84, Met172, Leu236, Thr265, and Phe/Met267, as well as the backbone of Gly83 and Pro45 (Supplementary Table 2.3). Interestingly, the *cis* peptide bond formed between residues 83 and 84 that is present in the *Thermoascus aurantiacus* xylanase 10A structure used as template for computational design (PDB ID: 1GOR¹) is maintained in all HG structures (Figure 2.2c) even though both residues were mutated to obtain HG3 (H83G and T84M). In addition to adopting a *cis* conformation, which is stabilized by hydrogen bonding to an ordered water molecule, this peptide bond also adopts the *trans* conformation in the structures of 6NT-bound HG3 and HG3.3b (Figure 2.2c, Supplementary Figure 2.4). However, starting at HG3.7, the peptide bond is found exclusively in the *cis* conformation in the 6NT-bound structures because it is stabilized by an additional hydrogen bond with the Gln50 side-chain carbonyl oxygen. This hydrogen bonding interaction helps to lock Gln50 in a conformation that is properly oriented to stabilize negative charge buildup on the phenolic oxygen at the transition state. Introduction of this new catalytic group in a catalytically-productive conformation likely accounts for the majority of the 12-fold enhancement in $k_{\text{cat}}/K_{\text{M}}$ observed in HG3.7 relative to HG3.3b, a hypothesis that is supported by the 16-fold decrease in $k_{\text{cat}}/K_{\text{M}}$ observed when the Q50H mutation is introduced into HG3.17.¹

From HG3.7 to HG3.17, no further changes in catalytic residues occurred during evolution. Yet, the catalytic efficiency increased by approximately 5-fold (Table 2.1). To evaluate whether this increase in activity was caused by changes to the conformational ensemble, we analyzed the B-factors of catalytic residues, which can be interpreted as a measure of the average displacement of an atom, or group of atoms, in the crystal. Since both conformational heterogeneity and crystalline disorder can contribute to atomic B-factors, with the latter effect potentially varying between different crystals, we calculated the Z-scores of the atomic B-factors and compared those across our crystal structures of different HG variants. This Z-score analysis allowed us to evaluate the variation of B-factors relative to the mean value within an individual crystal, and showed that rigidity of the Asp127 side chain did not vary significantly during evolution (Figure 2.3a). By contrast, the side chain of residue 50 became increasingly rigidified over the course of the evolutionary trajectory. Increasing rigidity at position 50 is expected when this residue is mutated from a lysine to a histidine (HG3 to HG3.3b), given the lower number of degrees of freedom in the latter amino acid. This trend is also expected when histidine at position 50 is mutated to a glutamine (HG3.3b to HG3.7) given the ability of glutamine but not histidine to hydrogen-bond with the *cis* peptide formed by residues Gly83 and Cys84 (Figure 2.2c). However, rigidity continues to increase at this position between HG3.7 and HG3.17, even though the side-chain rotamer of Gln50 in the presence of bound 6NT remains the same (Figure 2.2a). This result suggests that other structural features contribute to the increased rigidity observed at this position.

To verify the underlying cause of the increased rigidity at position 50, we calculated the average Z-score of atomic B-factors for each residue. We observed a trend whereby the loop formed by residues 87–90, which is located directly on top of residue 50, becomes increasingly rigidified during evolution (Figure 2.3b). Interestingly, two residues forming this loop (89 and 90)

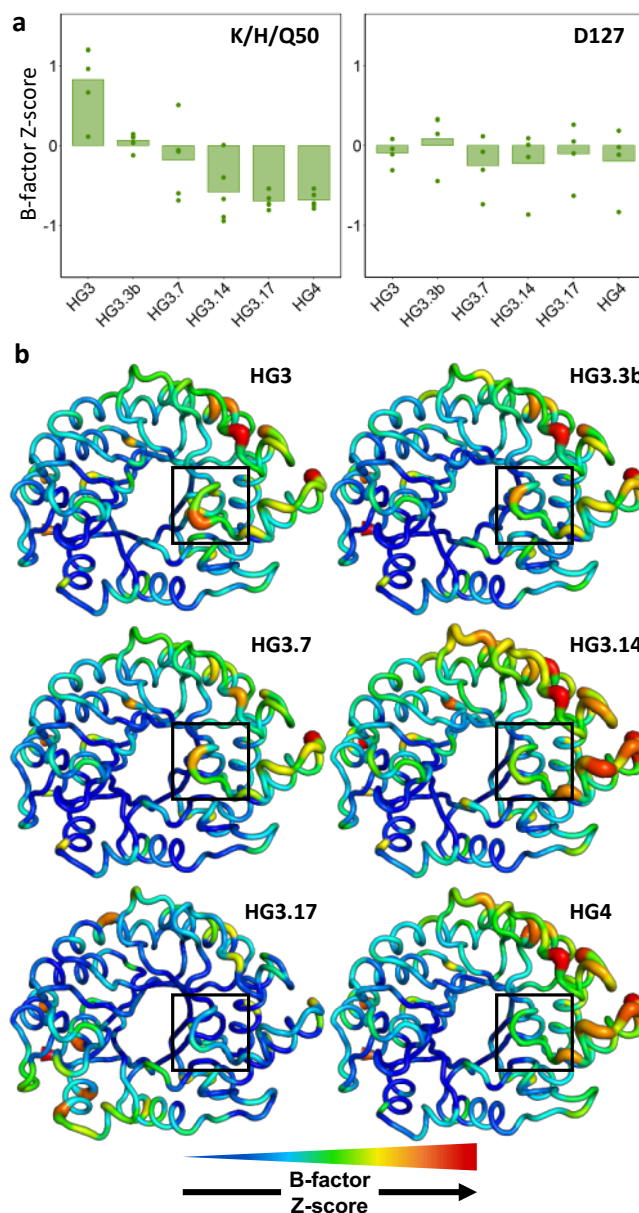


Figure 2.3. Conformational heterogeneity. (a) B-factor Z-scores for the residue at position 50 in the absence of bound 6NT decrease over the course of the evolutionary trajectory, while those for Asp127 do not change significantly. Z-scores of individual side-chain atoms are shown as dots, while the average is indicated by the bar. Positive and negative Z-scores indicate increased flexibility or rigidity relative to the average residue in the protein, respectively. (b) B-factor Z-scores for all protein residues plotted on a model backbone for each Kemp eliminase. Thickness of the sausage plot increases with the B-factor Z-score, indicating increased flexibility. The loop formed by residues 87–90 (boxed) becomes more rigid during evolution.

were mutated multiple times over the course of the evolutionary trajectory (Table 2.1). These mutations induce a conformational change in the loop that moves it closer to the active site, which results in a pi-stacking interaction between the phenyl and carboxamide groups of Phe90 and Gln50 that increases rigidity of the catalytic residue (Figure 2.2d, Supplementary Figure 2.5).

Although increasing rigidity of the Gln50 catalytic residue from HG3.7 to HG3.17 likely contributes to enhanced catalysis, other structural effects were investigated. A key determinant of efficient enzyme catalysis is active site pre-organization, which enables enzymes to bind substrates in a geometry close to that of the transition state. To evaluate changes in active site pre-organization during evolution, we compared the structures of HG-series Kemp eliminases in the presence and absence of bound 6NT. In all enzymes except for HG3.17, the unbound state is never pre-organized for catalysis as both Trp44 and Met237 adopt conformations that would prevent productive binding of 6NT (Figure 2.2e). In addition, the His50 and Gln50 catalytic residues in HG3.3b and HG3.7, respectively, adopt a low-occupancy, catalytically non-productive conformation in the unbound state that cannot interact favorably with 6NT. Interestingly, the non-productive conformation of Gln50 in the HG3.7 unbound state (26% occupancy) cannot stabilize the *cis* peptide bond formed by residues 83 and 84 via a hydrogen bonding interaction, and accordingly, the *trans* peptide conformation is also observed in this structure (25% occupancy) (Supplementary Figure 2.4 and Supplementary Figure 2.5b).

In contrast with all other HG variants, the unbound state of HG3.17 is correctly pre-organized for catalysis in a large portion of the molecules in the crystal, with only Trp44 adopting a non-productive conformation at 62% occupancy (Figure 2.2e). In this variant, Met237 adopts exclusively the productive conformer in the unbound state, which is stabilized by packing interactions with the neighboring Met267 side chain, a mutation that was introduced late in the

evolutionary trajectory (HG3.14 to HG3.17). Overall, three of the four residues that are key for binding and stabilizing 6NT (Gln50, Asp127, Met237) adopt a catalytically productive conformation in the HG3.17 unbound state, resulting in approximately 40% of the molecules in the crystal being correctly pre-organized for efficient catalysis.

Enhanced complementarity to the transition state is another important feature of efficient catalysis. Therefore, computational enzyme design algorithms aim to optimize packing of the transition state. However, transition-state overpacking may reduce catalytic efficiency by creating a high-energy barrier preventing substrate entry and product release. To evaluate whether active-site accessibility changed during evolution, we calculated the active-site entrance bottleneck radius on 6NT-bound structures.¹ We observed that during evolution, the active-site bottleneck formed by the side chains of residues 50 and 267, became widened (Figure 2.2f), as did the mouth of the substrate entry channel formed by residues Arg275 and Trp276, which were mutated to smaller amino acids. This widening of the active site entrance could help to eliminate high-energy barriers to substrate entry and product release that could have been caused by tighter packing of 6NT in higher activity HG variants.

2.4.3. HG4, an efficient artificial enzyme

All of the structural features that enhance activity described above are caused primarily by residues within or close to the active site, which suggests that mutagenesis far from the active site may not be essential to create an efficient artificial enzyme. To test this hypothesis, we generated a variant of HG3 that contains all HG3.17 mutations found within 7.5 Å of 6NT, with the exception of N47E, which we omitted to favor the formation of a unit cell similar to that of HG3. We also included the second-shell W275A and R276F mutations found to widen the active site entrance.

This yielded HG4, a variant of HG3 containing 8 mutations (Figure 2.1e). Kinetic analysis of HG4 revealed that its catalytic efficiency is >700-fold higher than that of HG3 (Table 2.1, Supplementary Figure 2.2), and equivalent to that of the average natural enzyme ($\sim 10^5 \text{ M}^{-1}\text{s}^{-1}$).¹ Crystallographic analysis of HG4 (Supplementary Table 2.2 and Supplementary Table 2.3) showed that its structure is highly similar to that of HG3.17 but with an active site that is better pre-organized (Figure 2.2 and Figure 2.3, Supplementary Figure 2.3 and Supplementary Figure 2.5). However, HG4 is approximately 20% less active than HG3.17, demonstrating that the additional 9 mutations found in the latter enzyme, most of which are distal to the active site, play a role in enhancing catalytic efficiency.

2.4.4. Computational design of HG4

Given that all but one mutation (G82A) in HG4 are found at sites that were optimized during design of HG2,¹ we investigated whether the HG4 structure could be accurately predicted using a computational protocol similar to the one that produced HG2 (Methods, Supplementary Table 2.4, Supplementary Table 2.5 and Supplementary Table 2.6). To do so, we first performed a positive control calculation in which rotamers for the HG4 sequence were optimized on the crystal structure backbone of 6NT-bound HG4. This calculation yielded an *in silico* model of HG4 with an energy score and a predicted rotameric configuration in excellent agreement with the crystal structure (Figure 2.4a). This control demonstrates that the combination of energy function, rotamer library, and search algorithm used in this protocol is sufficiently accurate for recapitulating the structure of HG4, provided that the correct template, binding pose, and catalytic dyad are allowed. By contrast, when we replaced the HG4 backbone template with the *Thermoascus aurantiacus* xylanase 10A backbone used to design HG2 (PDB ID: 1GOR),¹ we obtained a

structural model that differs significantly from the HG4 crystal structure and that is destabilized by approximately 45 kcal/mol (Figure 2.4b). This result demonstrates that the 1GOR backbone template is not well-suited to accommodate the HG4 sequence, as evidenced by differences between the 1GOR-derived model and the HG4 crystal structure. Specifically, the backbone at position 83 is shifted by 1.1 Å in the HG4 crystal structure relative to its position in the 1GOR template, causing the transition state to adopt an alternate binding pose that minimizes steric clashes with Gly83, which is accompanied by repacking of several residues around the transition state, including Gln50. Use of our HG3 crystal structures with or without 6NT as the design template causes similar, but less severe, structural and energetic effects (Figure 2.4c,d). These results highlight the impact of small backbone geometry variations on predictions made by enzyme design.

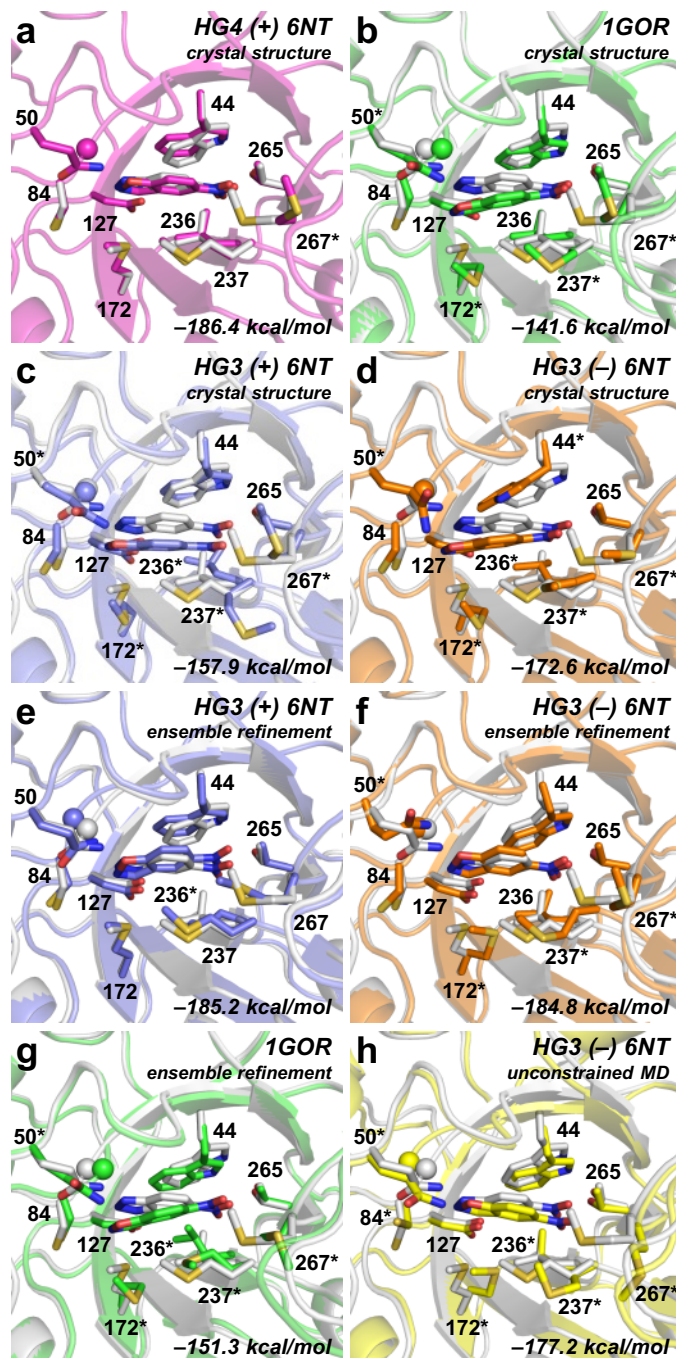


Figure 2.4. Computational design of HG4 on various backbone templates. The HG4 crystal structure with bound 6NT (white) is overlaid on the HG4 design models (colored) obtained using the crystal structure of (a) HG4 with bound 6NT, (b) *Thermoascus aurantiacus* xylanase 10A (PDB ID: 1GOR), (c) HG3 with bound 6NT, or (d) HG3 without 6NT. Panels (e–h) show the HG4 design models obtained using the template prepared by ensemble refinement or unconstrained molecular dynamics (MD) that gave the best energy following repacking. PHOENIX energies of design models after repacking are indicated at bottom right. For reference, the energy of the HG4 crystal structure with bound transition state is -186.7 kcal/mol. In all cases, the transition state and transition state analogue are shown at the center of barrel. Side chains of all residues forming the binding pocket are shown with the exception of Ala21 and Pro45, which were omitted for clarity. The sphere shows the alpha carbon of Gly83. Asterisks indicate residues

that adopt side-chain rotamers varying by >20 degrees around one or more side-chain dihedrals between the design model and crystal structure.

To address issues arising from the use of a single fixed backbone template, we generated backbone ensembles using molecular dynamics restrained by the HG3 or 1GOR diffraction data, also known as ensemble refinement (Methods), and used the resulting templates to optimize rotamers for the HG4 sequence. We were able to recapitulate the correct transition-state binding mode on several individual ensemble members derived from the HG3 crystallographic data, with energies comparable to that of the HG4 crystal structure (Figure 2.4e–f, Supplementary Figure 2.6). However, use of an ensemble derived from the 1GOR diffraction data did not allow recapitulation of the crystallographic transition-state binding mode (Figure 2.4g) although it did yield several computational models displaying improved energy (Supplementary Figure 2.6). The inferior performance of the 1GOR-derived ensemble compared to the HG3 ensembles likely results from differences in conformational heterogeneity within the ensemble, specifically at position 83 (Supplementary Figure 2.7). To evaluate the effect of restraints imposed by the diffraction data, we generated an ensemble using unconstrained molecular dynamics starting from the unbound HG3 crystal structure (Methods), and used it to optimize rotamers for the HG4 sequence. Use of this ensemble resulted in an improved structural model (Figure 2.4h) compared to the one obtained from the corresponding crystal structure (Figure 2.4d) that is however less structurally accurate and stable than the one obtained from ensemble refinement (Figure 2.4f). The better predictive ability of the HG3-derived ensembles prepared using crystallographic restraints likely results from their lower deviation from the HG4 crystal structure (≈ 0.4 Å, Supplementary Figure 2.7), which we previously showed to be necessary for an ensemble to represent a physically valid model of the target protein fold.¹ Overall, these results suggest that computational enzyme design with a crystallographically-derived backbone ensemble derived from a low activity enzyme

could obviate the need for directed evolution by allowing catalytically-competent sub-states to be sampled during the design procedure.

2.5. Discussion

In this work, we followed changes to the conformational ensemble that occur during evolution of an enzyme with *de novo* biocatalytic function. Unlike previous examples where the active sites of *de novo* enzymes were completely remodeled during evolution,^{25,26} or where the binding pose of the substrate or transition state analogue was significantly altered,^{13,17} we observed only subtle changes to the active site geometry or 6NT binding pose in the HG-series of Kemp eliminases. By contrast, many of the structural changes that contribute to enhanced catalysis in the HG series are dynamic in nature: the Gln50 catalytic residue became more rigid even though its average structure did not vary substantially, and the active site became better pre-organized via enrichment of catalytically-productive conformations of 6NT-binding residues that were already present in the unbound state. These observations illustrate how small changes to the active site conformational ensemble can drive large changes in catalytic efficiency. Since these changes can be subtle and difficult to predict computationally, directed evolution can help increase activity by selecting for mutations that enrich catalytically-competent sub-states.^{17,18}

Despite the challenges inherent to enzyme design, which are highlighted by our observations of the effects of mutations in the HG series of Kemp eliminases, our results suggest that *de novo* enzymes with native-like catalytic efficiencies can be computationally-designed, without the need to rely on subsequent improvement by laboratory directed evolution. Indeed, all mutations found in HG4 relative to the wild-type *Thermoascus aurantiacus* xylanase 10A template from which it is derived (PDB ID: 1GOR) are found at either first or second-shell residues, and

these sites were all optimized during the original design of HG2.⁵ Yet, Privett *et al.* designed the lower activity enzyme HG2 instead of HG4. While Gln50 was not sampled as part of the catalytic dyad during design of HG2, the combination of the Asp127/Gln50 dyad with the productive transition-state binding pose would have scored poorly on the 1GOR template regardless. However, our approach to computational enzyme design that utilized an experimentally-derived ensemble of backbone templates yielded HG4 models with energies and binding modes comparable to that of the HG4 crystal structure. These results suggest an iterative approach to computational enzyme design that could circumvent the need for directed evolution by introducing an additional round of design that utilizes a backbone ensemble generated from experimental structural data obtained for an initial, low-activity enzyme. In the case of evolution, mutations are not selected for in the context of a single backbone conformation but instead across an entire conformational ensemble.¹⁸ Our ensemble design approach should therefore be more accurate than traditional approaches relying on a single backbone template because it allows the accessible conformational ensemble to be represented in the scoring of sequences. The incorporation of experimental restraints in the generation of the ensemble ensures that the computational procedure is applied to the true conformational ensemble that is sampled by the enzyme.

The results reported here provide additional support for the well-known fact that enzymes are plastic molecules whose backbone conformation can change upon introduction of mutations (as seen when comparing the 1GOR and HG-series crystal structures), and suggest improvements to the enzyme design protocol that can account for this property. This could be achieved by incorporating flexible backbone design algorithms during the repacking step,^{27,28} or by using pre-generated ensembles of energetically-accessible backbone templates,^{24,29} as was done here. While these methodological changes may improve the design of the enzyme transition state, it is likely

that the creation of *de novo* enzymes with native-like catalytic efficiencies for more complex reactions will require a holistic approach where every possible state that the enzyme samples along its reaction coordinate is included in the design calculation. This could be achieved by the implementation of multistate approaches to computational protein design that allow the design of protein energy landscapes,³⁰ rather than single structures. We expect that the structures reported here, especially those of HG4 and HG3, will be helpful to benchmark these future enzyme design protocols.

2.6. Methods

2.6.1. Protein expression and purification

Codon-optimized and his-tagged (C-terminus) genes for HG-series Kemp eliminases (Supplementary Table 2.6) cloned into the pET-11a vector (Novagen) via *Nde*I and *Bam*HI were obtained from Genscript. Enzymes were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth (LB) supplemented with 100 $\mu\text{g mL}^{-1}$ ampicillin. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of 0.3, at which point the incubation temperature was reduced to 18 °C. At an OD₆₀₀ of 0.6, protein expression was initiated with 1 mM isopropyl β -D-1-thiogalactopyranoside. Following incubation for 16 hours at 18 °C with shaking (250 rpm), cells were harvested by centrifugation, resuspended in 10 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 were further subjected to 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/>).

2.6.2. 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-nitrobenzisoazole (AstaTech) dissolved in methanol (10% final concentration, pH of reaction mixture adjusted to 7.0 after addition of methanol-solubilized substrate) were initiated by addition of approximately 2 μ M HG3, 50 nM HG3.3b, 10 nM HG3.7/HG3.14, or 5 nM HG3.17/HG4. Product formation was monitored spectrophotometrically at 380 nm ($\epsilon = 15,800 \text{ M}^{-1} \text{ cm}^{-1}$)⁵ in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax 384Plus plate reader (Molecular Devices). 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).³¹ Linear phases of the kinetic traces were used to measure initial reaction rates. 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.

2.6.3. Crystallization

Enzyme variants were prepared in 50 mM sodium citrate buffer (pH 5.5) at the concentrations listed in Supplementary Table 2. For samples that were co-crystallized with the transition state analogue, a 100 mM stock solution of 6NT (AstaTech) was prepared in dimethyl sulfoxide (DMSO) and diluted 20-fold in the enzyme solutions for a final concentration of 5 mM 6NT (5% DMSO). For each enzyme variant, we carried out initial crystallization trials in 15-well hanging drop format using EasyXtal crystallization plates (Qiagen) and a crystallization screen that was designed to explore the chemical space around the crystallization conditions reported by Blomberg *et al.*¹⁵ 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. The mother liquor solutions contained ammonium sulfate as a precipitant in sodium acetate buffer (100 mM), and the specific growth conditions that yielded the crystals used for X-ray data collection are provided in Supplementary Table 2. In some cases, a microseeding protocol was required to obtain high-quality crystals. Microseeds were prepared by vortexing crystals in their mother liquor in the presence of glass beads (0.5 mm), and were subsequently diluted into the mother liquor solutions used to form the crystallization drops.

2.6.4. X-ray data collection and processing

Prior to X-ray data collection, crystals were mounted in polyimide loops and sealed using a MicroRT tubing kit (MiTeGen). Single-crystal X-ray diffraction data was 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. Crystals were maintained at 277 K throughout

the course of data collection. Each data set was collected using a total X-ray dose of 200 kGy or less, and covered a 180° wedge of reciprocal space. Multiple data sets were collected for each enzyme variant either from different crystals, or if their size permitted, from unique regions of larger crystals.

X-ray data was processed with the Xia2 0.5.492 program (<https://doi.org/10.1107/S0021889809045701>), which performed indexing, integration, and scaling with the 20180126 version of XDS and XSCALE,³² followed by merging with Pointless as distributed in CCP4 7.0.053.³³ For each variant, multiple individual data sets were merged to obtain the final set of reduced intensities, and the resolution cutoff was taken where the $CC_{1/2}$ and $\langle I/\sigma I \rangle$ values for the merged intensities fell to approximately 0.5 and 1.0 respectively. We determined which individual data sets should be combined by evaluating the overall effects of adding or removing individual data sets on the $CC_{1/2}$ and I/σ in the high-resolution bins of the merged data set. Information regarding data collection and processing is presented in Supplementary Table 3. The reduced diffraction data were analyzed with phenix.xtriage (http://www.ccp4.ac.uk/newsletters/newsletter43/articles/PHZ_RWGK_PDA.pdf) to check for crystal pathologies, and no complications were identified.

2.6.5. Structure determination

We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser,³⁴ as implemented in v1.13.2998 of the PHENIX suite.³⁵ Several different HG-series enzymes were used as molecular replacement search models. All members of the HG-series of enzymes crystallized in the same crystal form, containing two copies of the molecule in the crystallographic asymmetric unit, except for HG3.17, which crystallized

with only one molecule in the asymmetric unit. To avoid model bias that could originate from using other members of the HG-series as molecular replacement search models, we applied random coordinate displacements ($\sigma = 0.5 \text{ \AA}$) to the atoms, and performed coordinate refinement against the structure factor data before proceeding to manual model building.

Next, we performed iterative steps of manual model rebuilding followed by refinement of atomic positions, atomic displacement parameters, and occupancies using a translation-libration-screw (TLS) model, a riding hydrogen model, and automatic weight optimization. All model building was performed using Coot 0.8.9.2³⁶ and refinement steps were performed with phenix.refine within the PHENIX suite (v1.13-2998).^{35,37} Restraints for 6NT were generated using phenix.elbow,³⁸ starting from coordinates available in the Protein Data Bank (PDB ligand ID: 6NT).³⁹ Further information regarding model building and refinement, as well as PDB accession codes for the final models, are presented in Supplementary Table 2.2. Time-averaged ensembles were generated for 1GOR, and HG3 with or without ligand, with phenix.ensemble_refinement implemented in PHENIX v.1.15.2-3472. To prepare the structures for ensemble refinement, low-occupancy conformers were removed, and occupancies adjusted to 100% using phenix.pdbtools. Hydrogen atoms were then added using phenix.ready_set. This procedure yielded ensembles containing 80, 84, or 50 templates from the 1GOR, HG3 (+) 6NT, or HG3 (–) 6NT crystal structures, respectively.

2.6.6. Unconstrained molecular dynamics

All simulations were performed using GROMACS 2019.3 (<http://www.gromacs.org>) with the AMBER99SB forcefield.⁴⁰ Long-range electrostatics ($> 12 \text{ \AA}$) were modeled using the particle mesh Ewald method,⁴¹ and the LINCS algorithm⁴² was used to treat all bonds as constraints,

allowing a time step of 2 fs. Heavy atom coordinates of the major conformer from chain A were extracted from the crystal structure of HG3 in the unbound state (PDB ID: 5RG4). Following coordinate extraction, hydrogen atoms were added using Reduce,⁴³ and the resulting protein molecule was placed in an dodecahedral box with periodic boundary conditions where the distance between the protein surface and the box edges was set to 14 Å. After addition of explicit TIP3P⁴⁴ water molecules, charges on protein atoms were neutralized with Na⁺ and Cl⁻ counter-ions at a concentration of 0.15 M. The structure was then energy minimized with the steepest descent method to a target maximum force of 1000 kJ mol⁻¹ nm⁻¹. The system was equilibrated under an NVT ensemble for 125 ps at a temperature of 300 K using a Nose-Hoover thermostat,⁴⁵ while applying position restraints for protein heavy atoms. A second equilibration step under an NPT ensemble was performed for 1 ns with a constant pressure and temperature of 1 bar and 300 K, respectively, using the Berendsen barostat.⁴⁶ Following removal of the position restraints, a 500-ns production run under Parrinello-Rahman pressure coupling⁴⁷ was initiated from the final snapshot of the NPT equilibration. At the end of the simulation, 50 snapshots separated by 10 ns along the production trajectory were extracted. This 50-member ensemble was energy-minimized with a gradient-based tolerance of 0.1 kcal mol⁻¹ with the Cartesian minimization application included in the Triad protein design software (Protabit, Pasadena, CA, USA) using the PHOENIX energy function⁵ with added covalent terms from the DREIDING forcefield.⁴⁸

2.6.7. Computational enzyme design

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. The crystal structure of *Thermoascus aurantiacus* xylanase 10A was obtained from the Protein

Data Bank (PDB code: 1GOR²⁰) and further refined as described above to fix modeling issues with Thr84. Structures of HG3 with and without 6NT, HG4 with 6NT, and ensembles of 1GOR or HG3-derived templates were obtained from refinement of crystallographic data as described above. Following extraction of protein heavy-atom coordinates for the highest occupancy conformer from chain A, hydrogen atoms were added using the *addH.py* application in Triad. The Kemp elimination transition state (TS) structure⁴⁹ was built using the parameters described by Privett and coworkers.⁵ Residue positions surrounding Asp127 were mutated to Gly (Supplementary Table 2.3), with the exception of position 50, which was mutated to Gln. The 2002 Dunbrack backbone-independent rotamer library⁵⁰ 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 placement² using the contact geometries listed in Supplementary Table 5. TS pose energies were calculated using the PHOENIX energy function,⁵ which consists of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field⁴⁸ 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,⁵¹ an electrostatic energy term modelled using Coulomb's law with a distance-dependent dielectric of 10, an occlusion-based solvation potential with scale factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial,⁵² and a secondary structural propensity term.⁵³ During the energy calculation step, TS–side-chain interaction energies were biased to favor interactions that satisfy contact geometries (Supplementary Table 2.5) as described by Lassila *et al.*²

Following ligand placement, the 10 lowest energy TS poses found on each template (HG4 with 6NT, 1GOR, HG3 with 6NT, and HG3 without 6NT) were selected as starting points for repacking of the HG4 sequence. For individual members of the 1GOR and HG3-derived

ensembles, only the single lowest energy TS pose was used for repacking. In the repacking calculation, the TS structure was translated ± 0.4 Å in each Cartesian coordinate in 0.2-Å steps, and rotated 10° about all three axes (origin at TS geometric center) in 5° steps for a total combinatorial rotation/translation search size of 5⁶ or 15,625 poses. Residues that were converted to Gly in the ligand placement step were allowed to sample all conformations of the amino acid found at that position in the HG4 sequence (Supplementary Table 2.3). The identities of the catalytic residues were fixed and allowed to sample all conformations of that amino-acid type. Side-chain–TS interaction energies were biased to favor those contacts that satisfy the geometries as done during the ligand placement step (Supplementary Table 2.5). Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. The single lowest energy repacked structure on each backbone template was used for analysis. To compare energies of the HG4 models obtained on the various templates, we calculated the energy difference between each repacked structure and the corresponding all-Gly structure obtained after ligand placement, and these energies are reported throughout the figures and text.

2.7. Supplementary information

Supplementary Figure 2.1. Amino-acid sequences of HG-series Kemp eliminases

Enzyme	# mutations from HG3	Sequence ^a
HG3	–	MAEAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIQADFGMVWPENS MKWDATEPSQGNFNFAGADYLVNWAQQNGKLIGGGMLVWHSQLPSWVSSI TDKNTLTNVMKNHITTLMTTRYKGIKIRAWDVVGEAFNEDGSLRQTVFLNVI GEDYIPIAFQTARAADPNKLYIMDYNLDSASYPKTQAIVNRVKQWRAAG VPIDGIGSQTLSAGQGAGVLQALPLLASAGTPEVSIIMLDVAGASPTDY VNVVNACLVQSCVGITVFGVADPDSWRASTTPLLFDGNFNPKPAYNAIV QDLQQGSIEGRGHHHHHH9

HG3.3b	6	MAEAAQ S I DQLIKARGKVYFGVATDQNRLTTGKNAA I IQADFGMVWPENS M H WDATEPSQGNFN F AGADYLVNWAQQNGK L IGGG C LVW H R D L PSWVSSI TDKNTLTNV M KNHITTL M TRYKGKIR N WDVVGEAFNEDGSLRQTVFLNVI GEDYIPIAFQTARAADPN A KLYIMDYNLDSASYPKTQ A IVNRVKQWRAAG VPIDGIGSQ T HL S AGQGAGVLQAL P LLASAGTPEV S ILMLDVAGASPTDY VNVVNACLN V QSCVGITVFGVADPDSWRAS T TP L LF D GNF N PKPAYNA I V QDLQQGSIEGRGHHHHHH
HG3.7	7	MAEAAQ S I DQLIKARGKVYFGVATDQNRLTTGKNAA I I KADFGMVWPENS M Q WDATEPSQGNFN F AGADYLVNWAQQNGK L IGGG C LVW H R H L PSWVSSI TDKNTLTNV M KNHITTL M TRYKGKIR N WDVVGEAFNEDGSLRQTVFLNVI GEDYIPIAFQTARAADPN A KLYIMDYNLDSASYPKTQ A IVNRVKQWRAAG VPIDGIGSQ T HL S AGQGAGVLQAL P LLASAGTPEV S ILMLDVAGASPTDY VNVVNACLN V QSCVGITVFGVADPDSWRAS T TP L LF D GNF N PKPAYNA I V QDLQQGSIEGRGHHHHHH
HG3.14	12	MAEAAQ S I DQLIKARGKVYFGVATDQNRLTTGKNAA I I KADFGMVWPENS M Q WDATEPSQGNFN F AGADYLVNWAQQNGK L IG A G C LVW H S H L PSWVSSI TDKNTLT I N V MKNHITTL M TRYKGKIR T WDVVGEAFNEDGSLR Q N V FLNVI GEDYIPIAFQTARAADPN A KLYIMDYNLDSASYPKTQ A IVNRVKQWRAAG VPIDGIGSQ M H L SAGQGAGVLQAL P LLASAGTPEV S ILMLDVAGASPTDY VNVVNACLN V QSCVGITVFGVADPDSWRAS S T P LL F DGNF N PKPAYNA I V Q N LQQGSIEGRGHHHHHH
HG3.17	17	MAEAAQ S I DQLIKARGKVYFGVATDQNRLTTGKNAA I I KADFGMVW P E S M Q WDATEPSQGNFN F AGADYLVNWAQQNGK L IG A G C LVW H N F LPSWVSSI TDKNTLT I N V MKNHITTL M TRYKGKIR T WDVVGEAFNEDGSLR Q N V FLNVI GEDYIPIAFQTARAADPN A KLYIMDYNLDSASYPKTQ A IVNRVKQWRAAG VPIDGIGSQ M H L SAGQGAGVLQAL P LLASAGTPEV S ILMLDVAGASPTDY VNVVNACLN V QSCVGIT V M GVADPDS A F A S T TP L LF D GNF N PKPAYNA I V Q N LQQGSIEGRGHHHHHH
HG4	8	MAEAAQ S V DQLIKARGKVYFGVATDQNRLTTGKNAA I IQADFGMVWPENS M Q WDATEPSQGNFN F AGADYLVNWAQQNGK L IG A G C LVW H S F LPSWVSSI TDKNTLTNV M KNHITTL M TRYKGKIR T WDVVGEAFNEDGSLRQTVFLNVI GEDYIPIAFQTARAADPN A KLYIMDYNLDSASYPKTQ A IVNRVKQWRAAG VPIDGIGSQ T HL S AGQGAGVLQAL P LLASAGTPEV S ILMLDVAGASPTDY VNVVNACLN V QSCVGIT V M GVADPDS A F A S T TP L LF D GNF N PKPAYNA I V QDLQQGSIEGRGHHHHHH

^a Mutations from HG3 are highlighted in bold. All sequences contain a His-tag at the C-terminus.

Supplementary Table 2.1. Crystallization conditions

Enzyme ^a	6NT ^b	pH	Protein (mg mL ⁻¹)	(NH ₄) ₂ SO ₄ (M)
HG3	(-)	4.6	4	1.8
	(+)	4.6	4	1.8
HG3.3b	(-)	5.4	5	2.0
	(+)	5.4	5	2.0
HG3.7	(-)	4.0	6	1.6
	(+)	5.0	12	2.0
HG3.14	(-)	4.5	3	1.6
	(+)	5.0	3	2.4
HG3.17	(-)	5.0	5	1.6
	(+)	5.0	10	1.6
HG4	(-)	4.6	10	1.6
	(+)	4.6	10	1.6

^a All enzymes were crystallized in 100 mM sodium acetate buffer at the indicated pH.

^b 6-nitrobenzotriazole (6NT) was dissolved in pure dimethyl sulfoxide (DMSO) and added to crystallization solution at a final concentration of 5 mM (5% DMSO).

Supplementary Table 2.2. Crystallographic data and refinement statistics for room-temperature structures (277 K)

	HG3		HG3.3b		HG3.7		HG3.14		HG3.17		HG4	
6NT	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)
PDB ID	5RG4	5RGA	5RG5	5RGB	5RG6	5RGC	5RG7	5RGD	5RG8	5RGE	5RG9	5RGF
Data collection^a												
Resolution (Å)	41.52–1.99	55.15–1.89	37.09–1.62	48.16–1.44	60.34–1.35	79.81–1.39	79.83–1.47	48.94–1.40	46.17–1.73	38.46–1.77	79.99–1.47	41.50–1.40
Space group	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁
Cell params.												
a b c (Å)	76.14 79.97 99.06	76.24 79.88 99.05	76.23 80.03 98.91	76.24 79.92 98.71	76.20 79.85 98.81	76.26 79.81 98.72	76.31 79.83 99.11	76.32 79.81 98.94	51.35 58.17 92.35	50.90 57.98 95.69	76.29 79.99 98.98	75.97 77.99 98.03
α β γ (°)	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90
Molecules per asymm. unit	2	2	2	2	2	2	2	2	1	1	2	2
R_{pin}	0.125 (0.375)	0.067 (0.332)	0.047 (0.407)	0.043 (0.540)	0.029 (0.246)	0.029 (0.359)	0.029 (0.372)	0.046 (0.391)	0.088 (0.734)	0.101 (0.606)	0.048 (0.550)	0.028 (0.375)
CC_{1/2}	0.974 (0.679)	0.994 (0.745)	0.997 (0.690)	0.998 (0.524)	0.998 (0.826)	0.999 (0.681)	0.999 (0.586)	0.998 (0.697)	0.996 (0.587)	0.990 (0.499)	0.997 (0.609)	0.999 (0.743)
I/σI	3.7 (1.0)	6.8 (1.0)	8.9 (1.0)	8.9 (1.1)	13.0 (2.3)	12.6 (1.3)	13.6 (1.4)	9.4 (1.4)	8.8 (1.1)	5.5 (1.0)	8.2 (1.1)	29.4 (3.5)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	99.8 (99.8)	99.0 (91.8)	100.0 (99.6)	99.8 (99.2)	100.0 (100.0)	100.0 (99.5)	99.5 (99.6)	99.7 (99.6)	99.8 (99.0)
Multiplicity	6.4 (6.5)	12.7 (9.3)	12.9 (10.9)	6.4 (6.4)	6.3 (4.7)	12.7 (10.0)	18.5 (10.4)	13.0 (13.5)	16.5 (6.3)	6.5 (6.1)	5.8 (5.9)	17.3 (13.0)
Wilson B-factor (Å²)	17.21	17.58	14.72	14.85	11.43	13.07	13.31	14.34	14.08	13.30	14.07	15.40
# reflections (total unique)	42219	51523	118990	113916	131151	121461	103204	80367	29625	27504	103164	114707
Refinement												
R work/free	0.1767/ 0.2164	0.1482/ 0.1863	0.1421/ 0.1665	0.1360/ 0.1665	0.1271/ 0.1412	0.1255/ 0.1409	0.1319/ 0.1480	0.1329/ 0.1518	0.1397/ 0.1615	0.1501/ 0.1794	0.1434/ 0.1633	0.1317/ 0.1497
No. atoms												
Protein	4754	5017	5261	5376	5661	5703	5407	5066	2398	2387	5164	5180
Ligand	–	24	–	24	–	24	–	24	–	12	–	24
Water	424	483	463	498	543	527	511	496	267	228	474	488
Averaged B-factors (Å²)												
Protein	25.85	25.65	24.74	22.83	18.78	21.11	23.73	24.92	20.75	20.94	25.38	25.40
Ligand	–	29.10	–	20.17	–	15.46	–	18.52	–	15.41	–	17.54
Water	33.56	35.97	35.69	36.07	33.62	35.81	37.51	37.24	33.28	31.83	36.97	37.27
RMSD												
bond lengths (Å)	0.005	0.008	0.004	0.005	0.008	0.008	0.008	0.005	0.004	0.003	0.004	0.004
bond angles (°)	0.831	0.910	0.784	0.880	1.029	0.986	0.973	0.835	0.731	0.631	0.773	0.822
MolProbity clashscore	2.97	3.09	3.35	4.67	3.89	2.02	2.32	2.26	1.26	1.26	2.91	2.70

^a Highest resolution shell is shown in parentheses.

Supplementary Table 2.3. Amino-acid positions optimized during computational design of HG4

Ligand placement ^a	Repacking ^b
16, 17, 21, 42, 44, 46, 47, 79, 81, 83, 84, 87, 90, 125, 130, 170, 172, 207, 209, 234, 236, 237, 239, 265, 267, 275, 276	V16, Y17, A21, M42, W44, E46, N47, Q50, L79, G81, A82, G83, C84, W87, F90, T125, D127, G130, Y170, M172, Q207, H209, S234, L236, M237, D239, T265, M267, A275, F276

^a Positions that were mutated to Gly during ligand placement. Catalytic residues D127 and Q50 were allowed to sample alternate rotamers. All other residues were kept fixed.

^b Positions and amino-acid types that were allowed to sample alternate rotamers during repacking. All other residues were kept fixed.

Supplementary Table 2.4 Geometric definitions for generation of transition-state poses off the side chains of catalytic residues

Contact	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
Asp127	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
Gln50	Distance	1HE2 or 2HE2	O1			1.2, 1.5, 1.7, 1.9, 2.1, 2.3
	Angle	NE2	1HE2 or 2HE2	O1		145, 148, 151, 154, 157
	Angle	2HE2	O1	N2		120, 125, 135, 145, 155
	Torsion	CD	NE2	1HE2 or 2HE2	O1	115, 120, 135, 140, 145
	Torsion	NE2	1HE2 or 2HE2	O1	N2	180, 190, 200, 210, 220, 230
	Torsion	1HE2 or 2HE2	O1	N2	C3	150, 160, 170, 180, 190, 200

^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 2.5. Geometric constraints used to define catalytic contacts during HG4 computational design

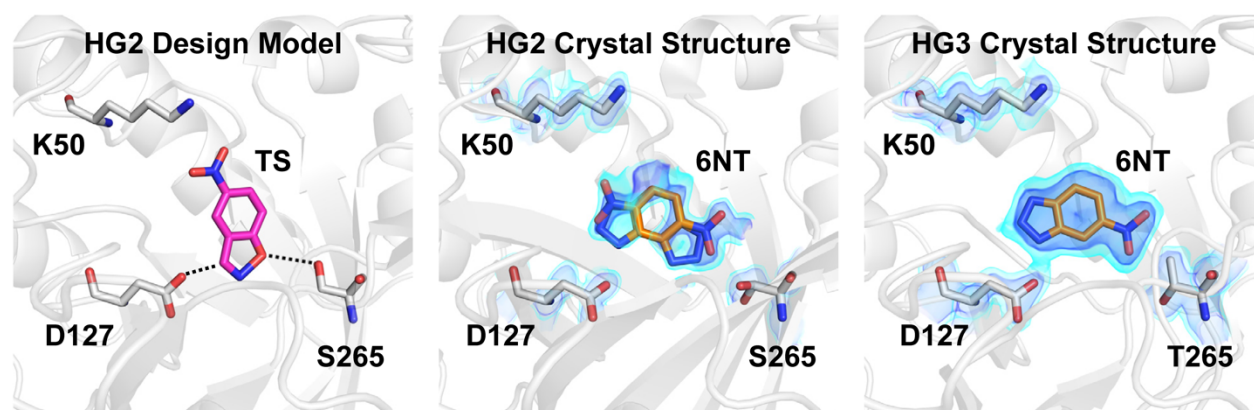
Contact	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max _b
Asp127	Distance	OD1 or OD2	H3			1.0 (1.0)	1.6 (1.6)
	Angle	CG	OD1 or OD2	H3		109 (109)	131 (131)
	Angle	OD1 or OD2	H3	C3		159 (159)	180 (180)
	Torsion	CB	CG	OD1 or OD2	H3	-21 (-21)	21 (21)
Gln50	Distance	1HE2 or 2HE2	O1			1.2 (1.2)	2.3 (3.2)
	Angle	NE2	1HE2 or 2HE2	O1		147 (130)	157 (180)
	Angle	1HE2 or 2HE2	O1	N2		120 (111)	140 (151)
	Torsion	1HE2 or 2HE2	O1	N2	C3	160 (130)	200 (200)

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

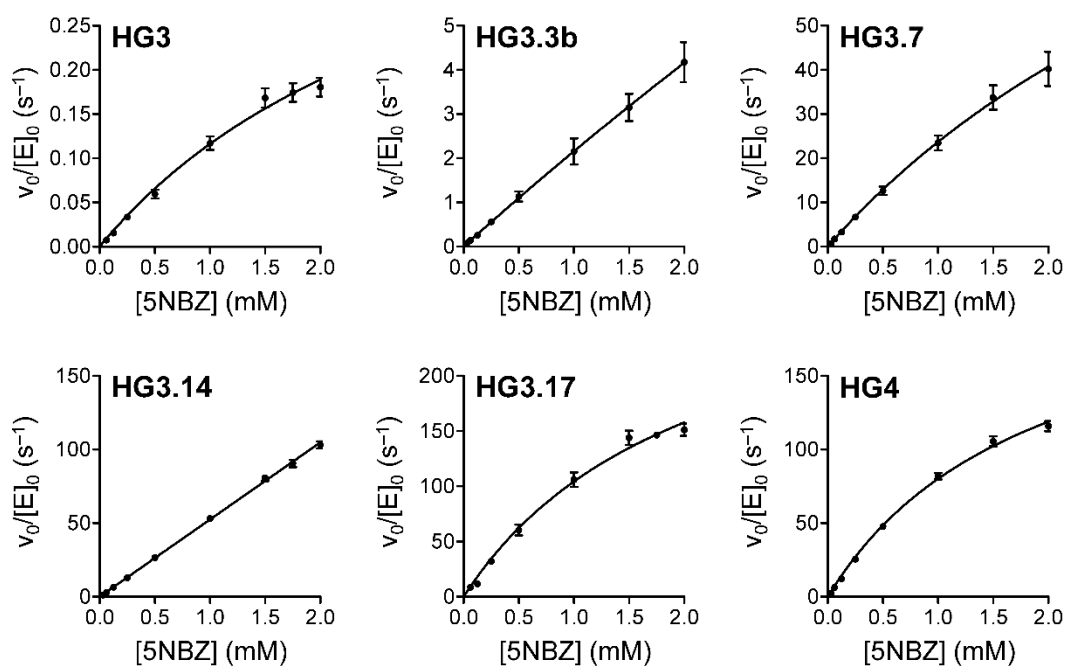
^b Distance measurements given in Å, all others in degrees. Values in parentheses are for the repacking step, while the others are for ligand placement.

Supplementary Table 2.6. Codon-optimized genes of HG-series Kemp eliminases

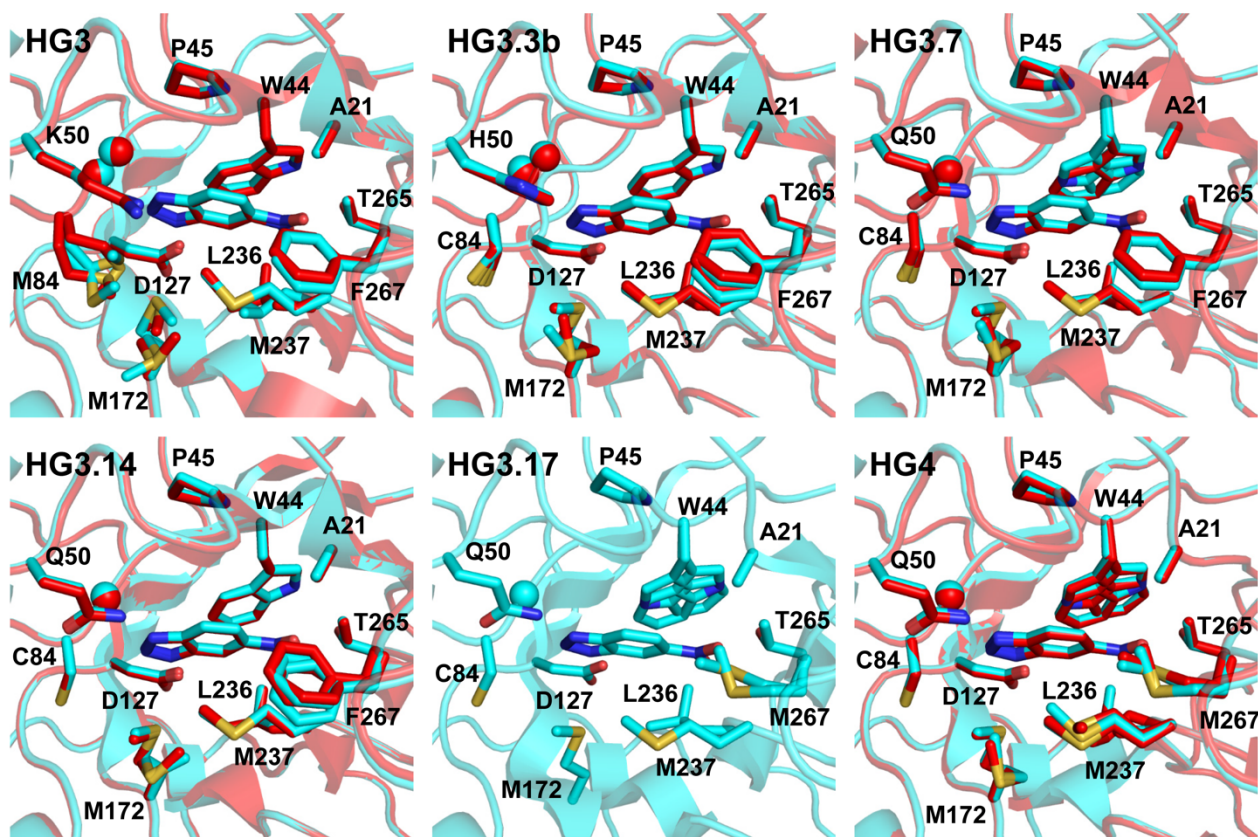
Enzyme	DNA sequence
HG3	ATGGCGGAGGCGGCGCAGAGCGTGGACCAACTGATCAAGGCGCGTGGCAAGGTTTACTTTGGCGTGGCGACCGACCAGAATCGTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGCATGGTGTGGCCGGAGAACAGCATGAAATGGGATGCGACCGAACCGAGCCAGGGTAAC TTCAACTTTGCGGGCGCGGACTACCTGGTTAACTGGGCGCAGCAAAAACGGCAAGCTGATCGGTGCGGGTATGCTGGTGTGGCAGACCCAGCTG CCGAGCTGGGTAGCAGCATTACCGATAAAGAACACCTGACCAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGTAAA ATTCGTGCGTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGCAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATCATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATTGTGAACCGTGTAAACAATGGCGTGGCGCGGGTGTGCCGATTGATGGTATTGGTAGCCAGACCCATCTGAGCGCGGGTCAGGGT GCGGGCGTTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGTATTACCGTGTGGTGTGCGGACCCCGGATAGCTGGCGTGGG AGCACCACCCGCTGCTGTTTCGATGGCAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAGATCTGCAACAGGGTAGCATCGAGGGT CGTGGTTCATCATCATCATCACTAA
HG3.3b	ATGGCGGAGGCGGCGCAGAGCATCGACCAACTGATTAAGGCGCGTGGTAAAGTGTAAGTTCGGCGTTGCGACCGATCAGAACCCTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGCATGGTGTGGCCGGAGAACAGCATGCAGTGGGATGCGACCGAACCGAGCCAGGGTAAC TTCAACTTTGCGGGCGCGGACTATCTGGTTAACTGGGCGCAGCAAAAACGGTAACTGATCGGTGCGGGTATGCTGGTGTGGCAGCCGTGACCTG CCGAGCTGGGTAGCAGCATTACCGATAAAGAACACCTGACCAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGTAAA ATTCGTAAGTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGCAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATCATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATTGTGAACCGTGTAAACAATGGCGTGGCGCGGGTGTGCCGATTGATGGTATTGGTAGCCAGACCCATCTGAGCGCGGGTCAGGGT GCGGGCGTTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGTATTACCGTGTGGTGTGCGGACCCCGGATAGCTGGCGTGGG AGCACCACCCGCTGCTGTTTCGATGGCAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAGATCTGCAACAGGGTAGCATCGAGGGT CGTGGCCACCACCACCACCACCTAA
HG3.7	ATGGCAGAAGCGGCGCAGAGCATCGACCAACTGATTAAGGCGCGTGGCAAGGTTTACTTTGGCGTGGCGACCGACCAGAACCCTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGCATGGTGTGGCCGGAGAACAGCATGCAGTGGGATGCGACCGAACCGAGCCAGGGTAAC TTCAACTTTGCGGGCGCGGACTACCTGGTTAACTGGGCGCAGCAAAAACGGCAAGCTGATCGGTGCGGGTATGCTGGTGTGGCAGCCGTGACCTG CCGAGCTGGGTAGCAGCATTACCGATAAAGAACACCTGACCAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGTAAA ATTCGTAAGTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGCAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATCATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATTGTGAACCGTGTAAACAATGGCGTGGCGCGGGTGTGCCGATTGATGGTATTGGTAGCCAGACCCATCTGAGCGCGGGTCAGGGT GCGGGCGTTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGTATTACCGTGTGGTGTGCGGACCCCGGATAGCTGGCGTGGG AGCACCACCCGCTGCTGTTTCGATGGCAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAGATCTGCAACAGGGTAGCATCGAGGGT CGTGGTCAACCATCATCACCACCTAA
HG3.14	ATGGCGGAAGCGGCGCAGAGCATCGACCAACTGATCAAGGCGCGTGGCAAGGTGTACTTTGGCGTGGCGACCGACCAGAACCCTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGTATGGTGTGGCCGGAGAACAGCATGCAGTGGGATGCGACCGAACCGAGCCAGGCAAC TTCAACTTTGCGGGTGGGACTACCTGGTTAACTGGGCGCAGCAAAAACGGCAAGCTGATGGTGGCGGGTATGCTGGTGTGGCAGCCACCTG CCGAGCTGGGTAGCAGCATCACCAGATAAAGAACACCTGATTAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGCAAA ATTCGTAACTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGTAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATTATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATCGTGAACCGTGTAAACAGTGGCGTGGCGCGGGTGTGCCGATTGATGGTATTGGCAGCCAAATGCACCTGAGCGCGGGTCAAGGT GCGGGTGTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGCATCACCCTGTTGGTGTGCGGACCCCGGATAGCTGGCGTGGG AGCAGCACCCGCTGCTGTTTCGATGGTAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAAATCTGCAACAGGTAGCATCGAGGGT CGTGGTTCATCATCATCATCACTAA
HG3.17	ATGGCGGAGGCGGCGCAGAGCATCGACCAACTGATTAAGGCGCGTGGCAAGGTGTACTTCGGTGTGCGACCGATCAGAACCCTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGTATGGTGTGGCCGGAGAAAGCATGCAGTGGGATGCGACCGAACCGAGCCAGGGTAAC TTCAACTTTGCGGGTGGGACTACCTGGTTAACTGGGCGCAGCAAAAACGGCAAGCTGATCGGTGCGGGTATGCTGGTGTGGCAGCAACTTCCTG CCGAGCTGGGTAGCAGCATCACCAGATAAAGAACACCTGATTAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGCAAA ATTCGTAACTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGTAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATTATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATCGTGAACCGTGTAAACAGTGGCGTGGCGCGGGTGTGCCGATTGACGGTATTGGCAGCCAGATGCACCTGAGCGCGGGTCAAGGT GCGGGTGTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGCATCACCCTGTTGGTGTGCGGACCCCGGATAGCTGGCGTGGG AGCAGCACCCGCTGCTGTTTCGATGGCAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAAACCTGCAACAGGTAGCATCGAAGGC CGTGGTTCACCAACCACCACCACCTAA
HG4	ATGGCGGAGGCGGCGCAGAGCGTGGACCAACTGATCAAGGCGCGTGGCAAGGTTTACTTTGGCGTGGCGACCGACCAGAATCGTCTGACCACC GGCAAGAACGCGGCGATCATTTAAGCGGACTTCGGCATGGTGTGGCCGGAGAACAGCATGCAGTGGGATGCGACCGAACCGAGCCAGGGTAAC TTCAACTTTGCGGGCGCGGACTACCTGGTTAACTGGGCGCAGCAAAAACGGCAAGCTGATCGGTGCGGGTATGCTGGTGTGGCAGAGCTTCTG CCGAGCTGGGTAGCAGCATTACCGATAAAGAACACCTGACCAACGTGATGAAAAACCACATCACCACCCTGATGACCCGTTATAAGGGTAAA ATTCGTACGTGGGACGTGGTTGGCGAGGCGTTCAACGAAGATGGTAGCCTGCGTCAAGCCGTGTTCTGAACGTTATCGGCGAGGACTACATC CCGATTGCGTTTCAGACCGCGCGTGGCGCGGACCCGAACGCGAACTGTACATCATGGACTATAACCTGGATAGCGCGAGCTATCCGAAGACC CAGGCGATTGTGAACCGTGTAAACAATGGCGTGGCGCGGGTGTGCCGATTGATGGTATTGGTAGCCAGACCCATCTGAGCGCGGGTCAGGGT GCGGGCGTTCTGCAAGCGCTGCCGCTGCTGGCGAGCGCGGGTACCCCGGAAGTGAGCATTCTGATGCTGGATGTTGCGGGTGCAGCCCCGACC GATTACGTTAAGCTGGTTAACGCGTGCCTGAACGTGCAAAGCTGCGTTGGCATCACCCTGATGGGTGTTGCGGACCCCGGATAGCTGGCGTGGG AGCACCACCCGCTGCTGTTTCGATGGCAACTTTAACCAGGCGCGGTATAACGCGATTGTTCAAAACCTGCAACAGGTAGCATCGAAGGC CGTGGTTCATCATCATCATCACTAA



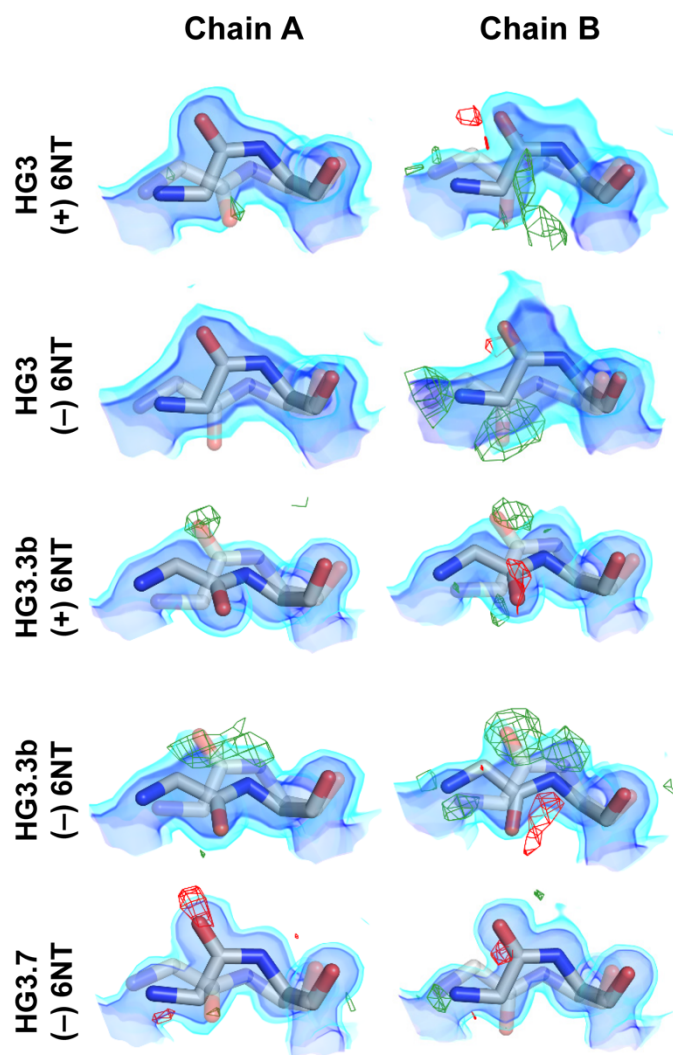
Supplementary Figure 2.2. HG2 is the direct precursor to HG3. For the crystal structures of HG2 and HG3, the 2Fo-Fc map is shown in volume representation at two contour levels: $0.5 \text{ e}\text{\AA}^{-3}$ and $1.5 \text{ e}\text{\AA}^{-3}$ in light and dark blue, respectively. HG2 was designed *in silico* to stabilize the transition state (TS) via catalytic contacts (dashed lines) with the D127 base and the S265 hydrogen bond donor (Left panel). However, its crystal structure (PDB ID: 3NYD) showed that the transition state analogue 6-nitrobenzotriazole (6NT) was bound in two alternate orientations (Middle panel). In the catalytically productive pose, the acidic N–H bond of 6NT that mimics the cleavable C–H bond of the substrate is located within H-bonding distance to the carboxylate oxygen of D127, and the nitro group is close to S265 (found to adopt two conformations in the crystal structure). In the catalytically non-productive pose, 6NT is flipped, which positions its nitro group closer to K50, and its acidic N–H bond far from the side chain of D127. To increase activity, Privett *et al.* introduced the S265T mutation into HG2 post-design to reduce the active-site conformational heterogeneity that was observed by molecular dynamics analysis, leading to the higher-activity variant HG3 (Right panel). Of note, no density for the non-productive binding pose of 6NT was observed in the HG3 crystal structure reported here.



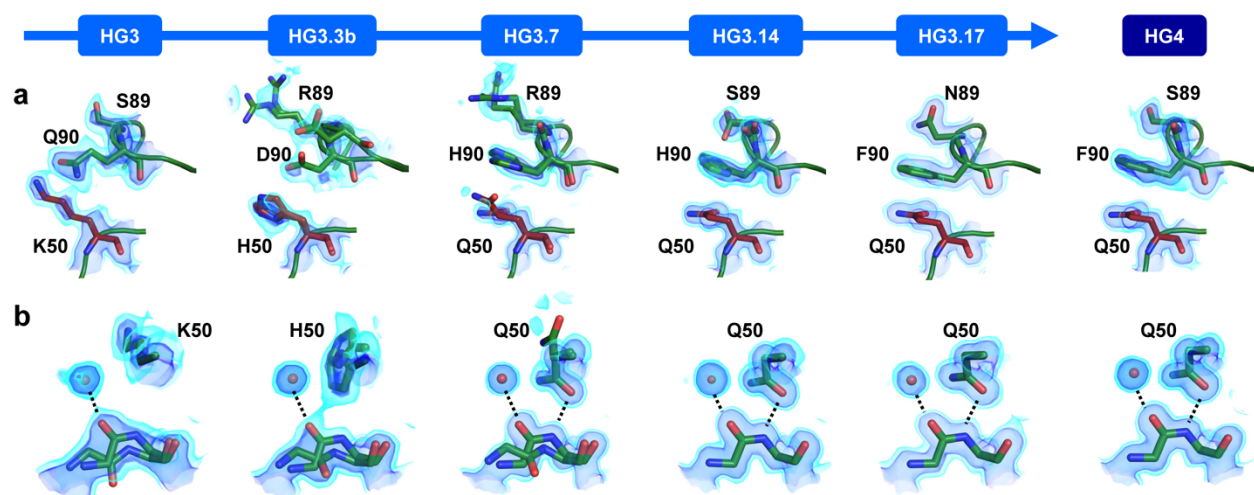
Supplementary Figure 2.3. Steady-state kinetics. Michaelis–Menten plots of normalized initial rates as a function of 5-nitrobenzisoxazole (5NBZ) concentrations are shown. Data represent the average of six or nine individual replicate measurements from two or three independent protein batches, with error bars indicating the SEM ($n = 2$ independent experiments for HG3, HG3.3b, HG3.7 and HG3.17, $n = 3$ independent experiments for HG3.14 and HG4, mean $\pm$ SEM in all cases). Saturation was not achieved for any enzyme at the substrate’s solubility limit (2 mM). Therefore, only k_{cat}/K_M values are reported on Table 1, and these values were calculated using linear regression of rates measured at the three or four lowest substrate concentrations. Under these conditions, $[S] \ll K_M$, and the Michaelis-Menten equation becomes $v_0 = (k_{cat}/K_M)[E_0][S]$. Source data are provided as a source data file.



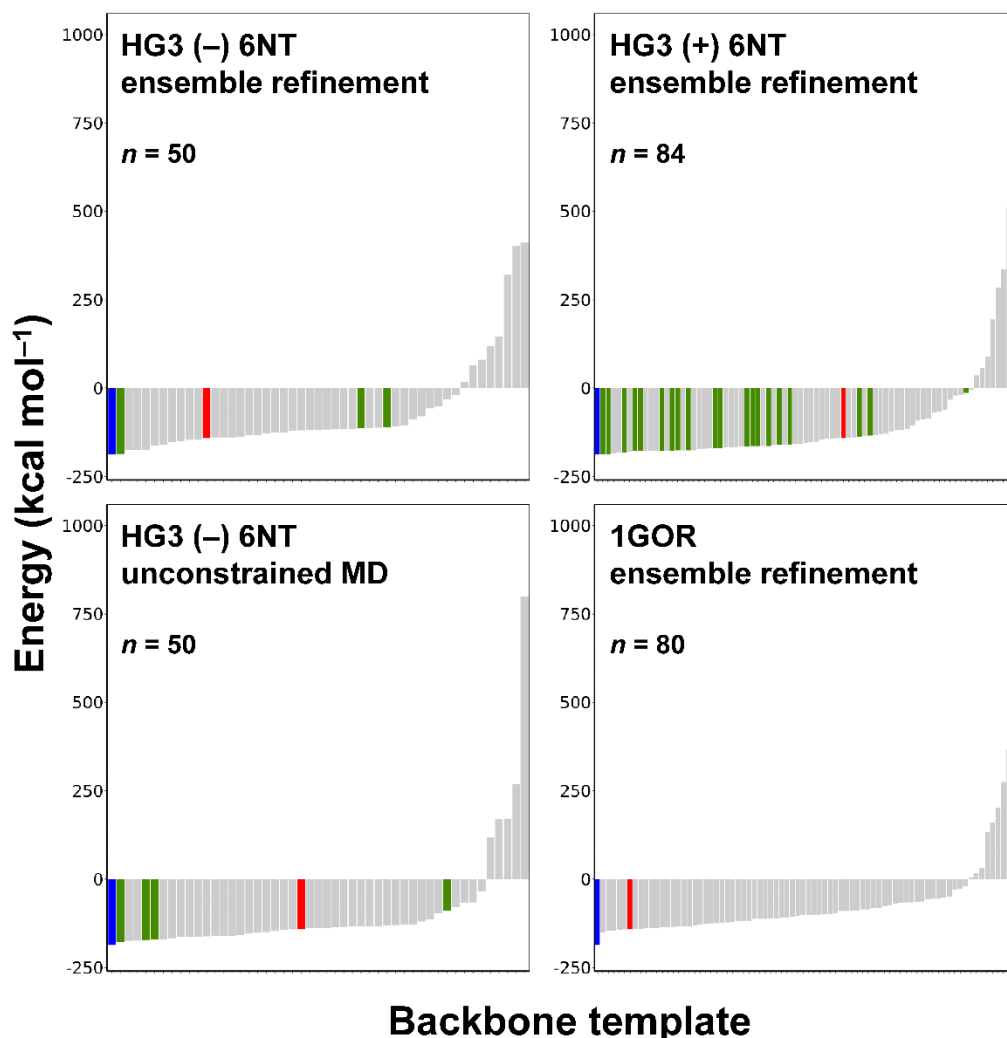
Supplementary Figure 2.4. 6NT-binding pocket. Overlay of chains A (cyan) and B (red) showing residues forming the 6NT-binding pocket in all HG-series Kemp eliminases. For HG3.17, only chain A is shown as its asymmetric unit contained a single polypeptide chain. In all cases, 6NT is bound in the middle of the barrel. Spheres indicate alpha carbons of Gly83 (two conformers, cis/trans, are observed for peptide bond between Gly83 and Met/Cys84 in HG3 and HG3.3b).



Supplementary Figure 2.5. Difference density increases around the 83–84 peptide bond after refinement in the absence of minor conformer. The peptide bond between residues 83 and 84 adopts both *cis* and *trans* conformations in HG3, HG3.3b, and HG3.7. To confirm the presence of the alternate peptide, the structure was re-refined in the absence of the minor conformer (transparent sticks), resulting in difference density for one or both chains in the asymmetric unit of HG3 and HG3.3b. A similar effect was also seen for HG3.7 in the unbound state, although the difference features corresponding to the minor conformation were weaker in comparison to HG3 and HG3.3b.

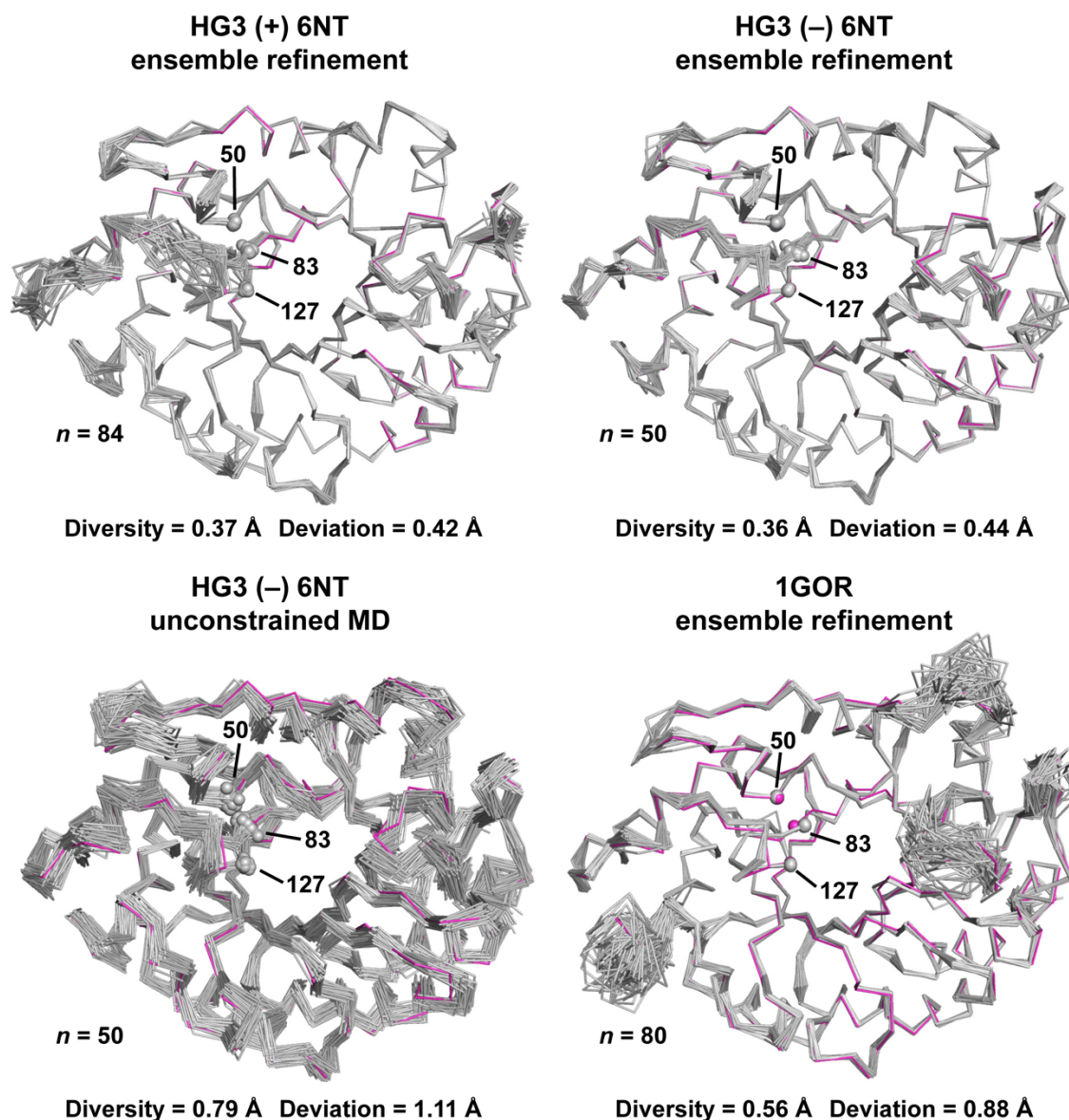


Supplementary Figure 2.6. Crystal structures of HG-series Kemp eliminases in the unbound state. In all cases, only atoms from chain A are shown. The 2Fo-Fc maps are shown in volume representation at two contour levels: 0.5 eÅ⁻³ and 1.5 eÅ⁻³ in light and dark blue, respectively. (a) Conformational changes to loop formed by residues 87–90 over the course of the evolutionary trajectory. (b) The peptide bond between residues 83 and 84 adopts both cis and trans conformations in HG3, HG3.3b, and HG3.7, but only the cis conformation in the higher activity variants. Ordered water molecules are shown as red spheres, and hydrogen bonds as dashed lines.



Supplementary Figure 2.7. Energy of HG4 design models generated on various backbone templates.

Rotamers for the HG4 sequence and its associated transition state binding pose were optimized (Methods) on individual backbone templates (bars). These ensembles of backbone templates were generated using molecular dynamics (MD) constrained or not by the diffraction data. In all cases, n indicates the total number of templates in the ensemble. Green bars indicate templates that yielded design models with transition-state binding poses within 0.7 Å root-mean-square deviation from the crystallographic binding pose. Blue and red bars indicate design models obtained from the HG4 with bound 6NT (−186.4 kcal/mol) or 1GOR (−141.6 kcal/mol) crystal structures, respectively. All HG3-derived ensembles yielded at least one HG4 design model with an accurate transition-state binding pose (i.e., root-mean-square deviation < 0.7 Å) and a more favorable energy than that obtained on the 1GOR template. However, the ensemble prepared from the HG3 (+) 6NT data outperforms all others: it yielded the lowest energy model and the highest number of models with accurate transition-state binding poses (21 out of 84).



Supplementary Figure 2.8. Backbone ensembles. The crystal structure of HG4 (+) 6NT (magenta) is superimposed with members of ensembles (grey) generated by ensemble refinement or unconstrained molecular dynamics (MD) starting from the HG3 (+) 6NT, HG3 (-) 6NT, or 1GOR crystal structures. Spheres indicate alpha carbons of key active-site residues. Ensemble properties, such as the average root-mean-square backbone coordinate deviation between pairs of ensemble members (diversity) or average root-mean-square backbone coordinate deviation from the HG4 (+) 6NT crystal structure (deviation) are indicated. In all cases, n indicates the number of templates per ensemble. Several templates from HG3-derived ensembles enable recapitulation of the crystallographic transition-state binding pose (Figure 4e,f,h and Supplementary Figure 6) due to the ability of the residue 83 backbone to sample conformations that eliminate steric clashes with this ligand, and that are similar to the one observed in the HG4 (+) 6NT crystal structure (i.e., there is overlap of magenta and grey spheres at position 83). This is not the case for the 1GOR-derived ensemble, in which the position of the residue 83 backbone does not shift substantially, preventing it from adopting a conformation necessary to allow the crystallographic transition-state binding pose found in the HG4 (+) 6NT crystal structure (i.e., there is no overlap of magenta and grey spheres at position 83). Of note, the ensemble generated by unconstrained MD (bottom left) samples a range of conformations with higher diversity and deviation than the corresponding one generated from ensemble refinement (top right), and with evenly-distributed conformational

heterogeneity throughout the structure. By contrast, conformational heterogeneity in ensembles generated using crystallographic restraints is unevenly distributed, being higher in specific structural elements.

References

- (1) Edwards, D. R.; Lohman, D. C.; Wolfenden, R. Catalytic proficiency: the extreme case of S-O cleaving sulfatases *J Am Chem Soc* **134**,525-531 (2012).
- (2) 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 U S A* **103**,16710-16715 (2006).
- (3) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Rothlisberger, D.; Baker, D. New algorithms and an in silico benchmark for computational enzyme design *Protein Sci* **15**,2785-2794 (2006).
- (4) Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp elimination catalysts by computational enzyme design *Nature* **453**,190-195 (2008).
- (5) 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 U S A* **109**,3790-3795 (2012).
- (6) Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; R  thlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F., 3rd; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De novo computational design of retro-aldol enzymes *Science* **319**,1387-1391 (2008).
- (7) 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).
- (8) 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).
- (9) Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J.; 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).
- (10) Bjelic, S.; Niv  n, L. G.;   lebi-  l  m, 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).
- (11) St-Jacques, A. D.; Gagnon, O.; Chica, R. A. In *Modern Biocatalysis: Advances Towards Synthetic Biological Systems*; The Royal Society of Chemistry: 2018, p 88-116.
- (12) Mak, W. S.; Siegel, J. B. Computational enzyme design: transitioning from catalytic proteins to enzymes *Curr Opin Struct Biol* **27**,87-94 (2014).
- (13) Khersonsky, O.; Kiss, G.; Rothlisberger, 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 U S A* **109**,10358-10363 (2012).
- (14) Khersonsky, O.; Rothlisberger, 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).
- (15) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R.; Grutter, 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).
- (16) Romero-Rivera, A.; Garcia-Borras, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity *Acs Catalysis* **7**,8524-8532 (2017).
- (17) Hong, N. S.; Petrovic, 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).
- (18) 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).
- (19) Biel, J. T.; Thompson, M. C.; Cunningham, C. N.; Corn, J. E.; Fraser, J. S. Flexibility and Design: Conformational Heterogeneity along the Evolutionary Trajectory of a Redesigned Ubiquitin *Structure* **25**,739-749 e733 (2017).

- (20) Lo Leggio, L.; Kalogiannis, S.; Eckert, K.; Teixeira, S. C.; 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).
- (21) Kries, H.; Bloch, J. S.; Bunzel, H. A.; Pinkas, D. M.; Hilvert, D. Contribution of Oxyanion Stabilization to Kemp Eliminase Efficiency *Acs Catalysis* **10**,4460-4464 (2020).
- (22) 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).
- (23) 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).
- (24) Davey, J. A.; Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles *Proteins* **82**,771-784 (2014).
- (25) 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).
- (26) Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R.; 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).
- (27) 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).
- (28) Babor, M.; Mandell, D. J.; Kortemme, T. Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody Herceptin-HER2 interface *Protein Sci* **20**,1082-1089 (2011).
- (29) Davey, J. A.; Chica, R. A. Multistate Computational Protein Design with Backbone Ensembles *Methods Mol Biol* **1529**,161-179 (2017).
- (30) Davey, J. A.; Damry, A. M.; Goto, N. K.; Chica, R. A. Rational design of proteins that exchange on functional timescales *Nat Chem Biol* **13**,1280-1285 (2017).
- (31) Lampinen, J.; Raitio, M.; Perälä, A.; Oranen, H.; Harinen, R.-R. In *Technical Reports Thermo Fisher Scientific* Available at: <http://www.thermoscientific.com>, 2012.
- (32) Kabsch, W. Xds *Acta Crystallogr D Biol Crystallogr* **66**,125-132 (2010).
- (33) Evans, P. Scaling and assessment of data quality *Acta Crystallogr D Biol Crystallogr* **62**,72-82 (2006).
- (34) 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).
- (35) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution *Acta Crystallogr D Biol Crystallogr* **66**,213-221 (2010).
- (36) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot *Acta Crystallogr D Biol Crystallogr* **66**,486-501 (2010).
- (37) Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards automated crystallographic structure refinement with phenix.refine *Acta Crystallogr D Biol Crystallogr* **68**,352-367 (2012).
- (38) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation *Acta Crystallogr D Biol Crystallogr* **65**,1074-1080 (2009).
- (39) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures *J Mol Biol* **112**,535-542 (1977).
- (40) 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* **78**,1950-1958 (2010).
- (41) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems *The Journal of Chemical Physics* **98**,10089-10092 (1993).
- (42) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations *Journal of Computational Chemistry* **18**,1463-1472 (1997).
- (43) 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 orientation *J Mol Biol* **285**,1735-1747 (1999).

- (44) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water *The Journal of Chemical Physics* **79**,926-935 (1983).
- (45) Evans, D. J.; Holian, B. L. The Nose–Hoover thermostat *The Journal of Chemical Physics* **83**,4069-4074 (1985).
- (46) 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 *The Journal of Chemical Physics* **81**,3684-3690 (1984).
- (47) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method *Journal of Applied Physics* **52**,7182-7190 (1981).
- (48) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. Dreiding - a Generic Force-Field for Molecular Simulations *Journal of Physical Chemistry* **94**,8897-8909 (1990).
- (49) Na, J.; Houk, K. N.; Hilvert, D. Transition state of the base-promoted ring-opening of isoxazoles. Theoretical prediction of catalytic functionalities and design of haptens for antibody production *Journal of the American Chemical Society* **118**,6462-6471 (1996).
- (50) Dunbrack, R. L., Jr. Rotamer libraries in the 21st century *Curr Opin Struct Biol* **12**,431-440 (2002).
- (51) 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).
- (52) 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 U S A* **107**,20257-20262 (2010).
- (53) Shortle, D. Propensities, Probabilities, and the Boltzmann Hypothesis *Protein Sci* **12**,1298-1302 (2003).
- (54) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations *Proc Natl Acad Sci U S A* **101**,6946-6951 (2004).

Chapter 3. Improving the Accuracy of De Novo Enzyme Design Using a Variety of Backbone Ensembles

Rakotoharisoa R.V., Pham N.T.H., Dementyev I.S., Chica R.A.

3.1. Statement of contribution

R.V.R wrote the chapter, R.V.R designed the molecular dynamics (MD) simulations experiment as well as the computational design experiments. R.V.R, P.N.T.H and D.I.S ran the MD simulations. P.N.T.H and D.I.S and R.V.R generated PertMin ensembles. P.N.T.H generated the ensembles from AlphaFold2 (AF). R.V.R generated ensembles from ensemble refinement (ER) and R.V.R ran the computational design experiments.

3.2. Introduction

In chapter 2, we have seen that subtle backbone rearrangements occurring during directed evolution of the low activity *de novo* enzyme HG3 contribute to the increase in catalytic efficiency observed with the most active variant HG3.17. Generating backbone ensembles of the crystal structure of HG3 bound with the transition state analogue at room temperature using the ensemble refinement (ER) of crystallographic data, allowed the sampling of a backbone enabling accurate prediction accuracy of the highly active variant HG4 sequence. In this chapter, we are interested to benchmark different methods to generate backbone ensembles to investigate which type of backbone movements are important to allow high prediction accuracy.

Designing artificial enzymes for any desired reaction is of great interest for many industrial applications, allowing for natural enzymes high proficiency to catalyze reactions sustainably. Over the past decades, organic model reactions were introduced successfully in natural protein scaffolds

devoid of the desired function¹⁻⁷. Although this procedure enables the design of artificial enzymes, their efficiency falls short of their natural counterparts and prevents their use for a larger scale application. To reach activity similar to natural enzymes, they need to be optimized by several rounds of directed evolution^{1,8-13}. Structural analyses of low-activity *de novo* enzymes and variants from directed evolution revealed discrepancies between the computationally predicted models and their experimental data^{10,11,14,15}.

Traditional methods use a single crystal structure from the Protein Data Bank (PDB), in which an active site is carved to accommodate and stabilize the calculated transition state of the reaction from QM/MM¹⁶. Discrepancies between the crystal structures of the *de novo* enzymes and their computational models showed that the predicted catalytic contacts with the transition state and the binding poses were different^{5,14}. In addition, different molecular dynamic simulations timescales showed high flexibility of the designed enzyme's active site⁵, and important loop dynamics in its vicinity, preventing accurate prediction^{8,9}. Further structural analysis showed the presence of unproductive alternative conformations and low electronic density in positions close to the active site^{1,14,15}. Taken together, these shortcomings contribute to the weak binding of the transition state and prevent higher catalytic efficiency. Enzymes are dynamic molecules that adopt different conformational states during their catalytic cycles^{17,18}. Therefore, using a single conformational state in the design procedure can introduce artifacts and prevent accurate positioning of the theozyme (Figure 3.1), thus increasing the proportion of false positive and false negative designs^{8,19}. To optimize the catalytic efficiency of KE70, Khersonsky et al. 2011⁸, introduced flexibility in their procedure by generating an ensemble of backbones from Backrub motions²⁰. It resulted in variants with catalytic efficiency improved by only 2-4 fold compared to the designed KE70, showing that introducing flexibility is insufficient to achieve significant

improvement. In work conducted by Davey et al. 2014, to improve the prediction accuracy of a designed Gβ1 protein, various ensembles of the input structure were generated from popular methods such as molecular dynamics simulations at different temperatures, perturbation and minimization (PertMin) developed previously in our group^{19,21}, Backrub ensembles²⁰, and solution NMR ensembles. The best prediction was achieved when ensembles were extracted from PertMin, as this method generates ensembles the most structurally similar to the input structure¹⁹. In Broom et al. 2020, an ensemble of backbones generated from ensemble refinement of a low activity variant enables the accurate prediction of HG4, a highly active variant from the HG-series¹⁴. In previous work, we observed that ensembles generated with unconstrained 500-ns molecular dynamics simulations at 300K led to a poorer prediction than ensemble refinement. We hypothesized that ensemble refinement better approximates the enzyme's conformational state in solution, as local atomic motions are modelled by molecular dynamics and are restrained by electron density²². Thus, generating an ensemble of backbones from a method that better approximates the enzyme dynamics in solution is ideal to improve prediction accuracy and holds the potential to increase the enzyme's activity.

In this study, we aim to create an *in-silico* benchmark of our ensemble-based computational enzyme design method. This benchmark involves generating ensembles of structures to recapitulate the binding pose and the active site conformation of active variants obtained from the optimization of HG3, catalyzing the Kemp eliminase reaction (Figure 3.1.a). We selected the Kemp elimination reaction for the benchmark due to its simplicity and extensive work on prior studies. We chose this enzyme as its catalytic activity has undergone optimization through directed evolution^{1,9,10}. Furthermore, crystallographic structures are available for its optimized variant in the bound state¹⁴ (see Table 3.1). Based on chapter 2, we assumed that, using the same scoring

function, search algorithm, and rotamer library, we should be capable of reproducing the active site conformation of the optimized variants when employing a template that shares structural similarities with it. Given this assumption, we defined as our positive control the model from the recapitulation of the optimized variant's active site on its own crystal structure and the negative control as the recapitulation of the crystal structure of the wild type (WT) originally used for the design. Using the available crystal structures of the initial designs in the unbound and bound states, as well as the crystal structure of the WT, we generated ensembles using ensemble refinement (ER) of crystallographic data from PHENIX^{22,23}, a method previously developed in our group called PertMin^{19,21}, and 20 replicates of 1-ns molecular dynamics (MD) simulations at 300 K. Additionally, we aimed to generate ensembles based on their sequences using AlphaFold2^{24,25} (AF). With the ensembles we generated, we hypothesized that the best method to generate ensembles would be the one with the highest proportion of templates capable of recapitulating the active site of the optimized variant. For HG3, ER ensembles generate the most accurate prediction, thanks to the structural similarity of the ensemble's active site with the crystal structure of its optimized variant, HG4. Likewise, PertMin ensembles give accurate predictions, with a high proportion of ensembles similar to HG4. Conversely, MD and AF ensembles yield less accurate predictions, which is attributed to their lower structural similarity with the HG4 active site. The benchmark established in this study will serve as a guide for computational enzyme design methodologies to enhance prediction accuracy and enable the development of enzymes with natural-like activity.

3.3. Results

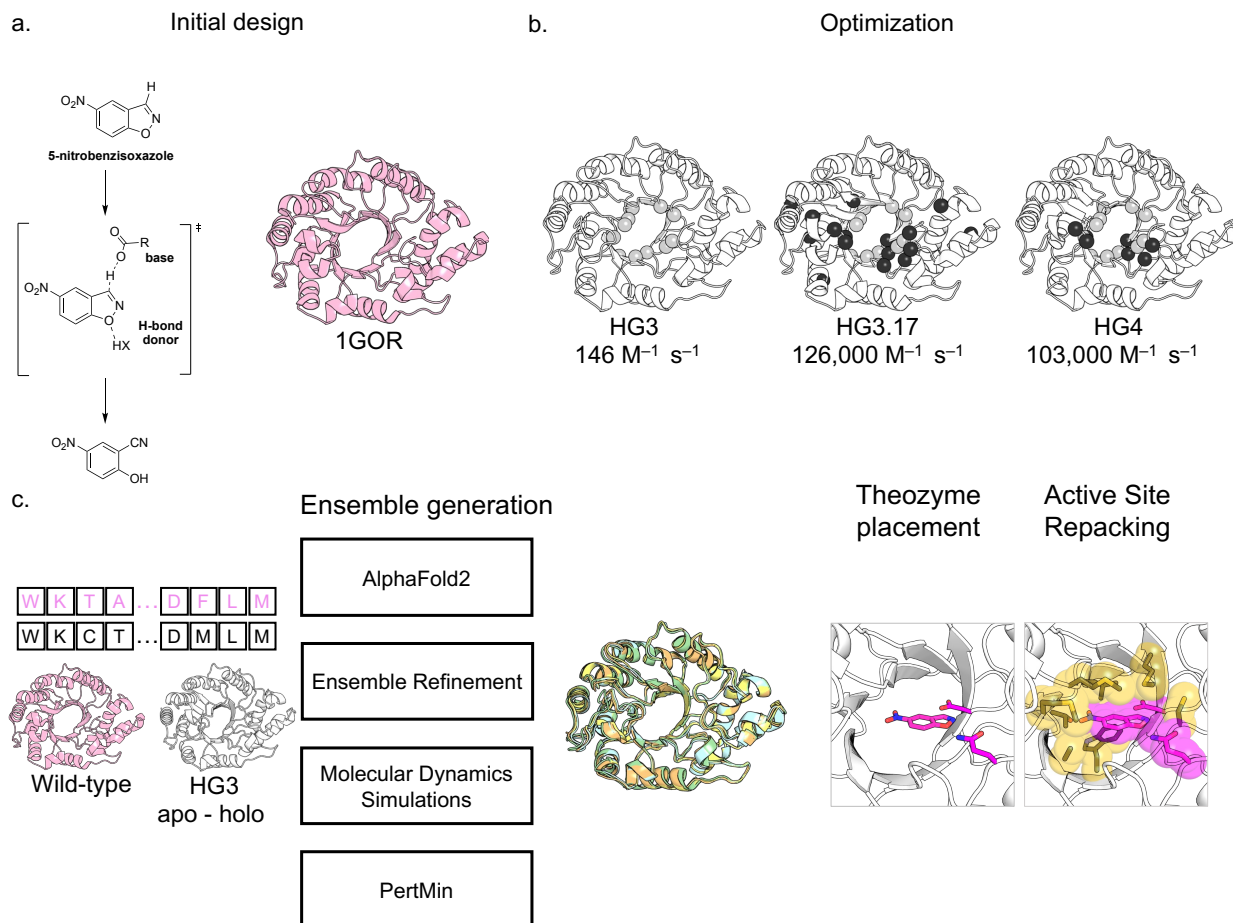


Figure 3.1. Method for Kemp Eliminate benchmark. a) Initial design, showing the Kemp elimination reaction introduced in 1GOR. b) crystal structure of HG3, its variant from directed evolution HG3.17 and a highly active variant HG4 used as a positive control. Grey spheres correspond to designed mutations from the input crystal structure, and black spheres correspond to mutations introduced by directed evolution. For each enzyme, k_{cat}/K_M is given at the bottom. c) Procedure for the benchmark. Ensembles are generated using AlphaFold2 using the input sequence of the wild-type template used to design HG3. Ensembles from ensemble refinement, molecular dynamics simulations and PertMin methods were generated from the crystal structures of the wild-type template of HG3, the crystal structures of the designed enzymes in the unbound and bound states when available. The design procedure involves two steps. The first step, called theozyme placement (see methods), is conducted to place the theozyme of the highly active variant (represented in sticks and coloured in magenta) on each individual template from the ensembles. When a stable theozyme configuration from the first step is found, a second step of active site repacking is conducted, where positions in the first shell and second shell of the active site are mutated to the sequence of the highly active variant.

3.3.1. Hypothesis and general strategy

In this study, using the benchmark of our ensemble-based computational enzyme design method, we aimed to identify a method for generating backbone ensembles that allow an accurate

prediction of the binding pose and the active site conformation of optimized variants of a *de novo* Kemp eliminase called HG3. We hypothesized that a method generating an ensemble that would perform the best would be an ensemble with high structural similarity with the crystal structure of the optimized variant. To test our hypothesis, we generated backbone ensembles from various methods to assess which ensemble will enable accurate prediction. We used the crystal structure of HG3 in the unbound and bound state as input structures for the ensemble generation from ER, PertMin, and MD. At the same time, the amino acid sequence was used as input for generating AF ensembles. Additionally, we generated control ensembles using the crystal structure of the WT as input structure for generating backbone ensembles from ER, PertMin and MD, or the WT amino acid sequence for input in generating AF ensembles. Previous studies conducted to assess the prediction accuracy of existing sequences' stability^{19,26} demonstrated that ensembles generated from unconstrained MD simulations lead to sequences adopting non-native fold and are then called off-target ensembles, while ensembles generated from constrained MD simulations lead to on-target ensembles as their motions are restricted to physical models. According to these assumptions, we hypothesized that ensembles generated from ER and PertMin will provide the most accurate results as their local motions are restricted by electron density for ER ensembles and by small perturbations and minimization steps for PertMin ensembles. On the contrary, ensembles from the 20 short 1-ns MD simulations at 300 K will lead to predictions with low accuracy. Due to the ability of AF to predict protein structures with high accuracy^{24,27,28}, we found it interesting to investigate the accuracy of AF to generate ensembles predicting correct binding poses.

Table 3.1 List of enzymes used for the benchmark and a description of their crystal structures.

Family	Alias	PDB ID	Resolution (Å)	Data collection Temperature (K)	Crystal structures
HG3	HG3_WT	1GOR	1.70	100	Thermostable Xylanase I From <i>Thermoascus Aurantiacus</i> bound with Xylobiose
	HG3_DA	5RG4	1.99	277	HG3 in unbound state
	HG3_DH	5RGA	1.86	277	HG3 bound with 6-nitrobenzotriazole

3.3.2. Ensemble generation and characterization

To enable direct comparison, the number of backbone templates generated for each method and each enzyme depended on the number of backbone templates generated by ER (Table 3.2 and Supplementary Table 3.3), as we cannot control the final number of templates generated by this method. To ensure that the ensembles obtained by ER represent the experimental data, we implemented multiple ensemble generation runs using an array of parameters (see methods). Ensembles were selected based on their R_{work} and R_{free} values, ensuring their values were lower than those of the input crystal structure (Supplementary Table 3.3). Finally, we extracted individual backbones for the calculation. The PertMin ensembles are generated from the random coordinate perturbations of an input structure, allowing the generation of structures adopting different trajectories. Following perturbation, the structures are energy-minimized to yield ensembles in different local energy minima^{19,21}. The diversity of the ensembles generated by the PertMin protocol is controlled by the number of iterations used during energy minimization and the number of atoms in the system¹⁹. For MD simulations, we decided to use 20 short 1-ns production runs instead of longer simulations, as previous studies have shown that multiple short MD simulations allow a better sampling of locally accessible conformational states, in contrast to longer simulations, which can trap the system in local minima, preventing it from accessing other conformational states^{23,24}. Finally, AF ensembles were generated using the webserver ColabFold³¹

(<https://github.com/sokrypton/ColabFold>). To ensure the ensembles generated backbones with the highest prediction quality, extracted templates have pLDDT³² score in the range of 88 to 97. Using these methods to generate ensembles, we obtained 856 backbones for HG3 with all ensembles combined (Table 3.2).

Table 3.2. Diversity and deviation

Ensemble	Ensemble size (n)	Ensemble Refinement		PertMin		AlphaFold		Molecular Dynamics	
		Diversity (Å)	Deviation (Å)	Diversity (Å)	Deviation (Å)	Diversity (Å)	Deviation (Å)	Diversity (Å)	Deviation (Å)
HG3_WT	80	0.9 ± 0.2	0.8 ± 0.2	0.2 ± 0.03	0.3 ± 0.02	0.5 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.04
HG3_DA	50	0.4 ± 0.1	0.3 ± 0.1	0.2 ± 0.03	0.2 ± 0.02	0.4 ± 0.1	0.7 ± 0.1	0.9 ± 0.1	0.9 ± 0.1
HG3_DH	84	0.4 ± 0.1	0.3 ± 0.1	0.2 ± 0.03	0.2 ± 0.02	0.5 ± 0.1	0.7 ± 0.1	0.9 ± 0.2	0.9 ± 0.1

Diversity corresponds to the average pairwise RMSD between each ensemble member.

Deviation corresponds to the average pairwise RMSD between each ensemble member and the input crystal structure.

To characterize the structural similarity of each ensemble member from each other, and the structural similarity of the ensembles with the input crystal structure, we calculated their diversity and deviation. The diversity of an ensemble is the average root mean square deviation (RMSD) of backbone atoms (N, C, and CA) of each pair of templates. It informs us of the structural similarity of individual members within the same ensemble. Deviation is the average RMSD of backbone atoms (N, C, and CA) of each template with the input crystal structure¹⁹. To calculate the diversity and deviation of each ensemble, we truncated the N-terminus and C-terminus of individual backbones to prevent their values from being biased towards high fluctuations of terminal residues.

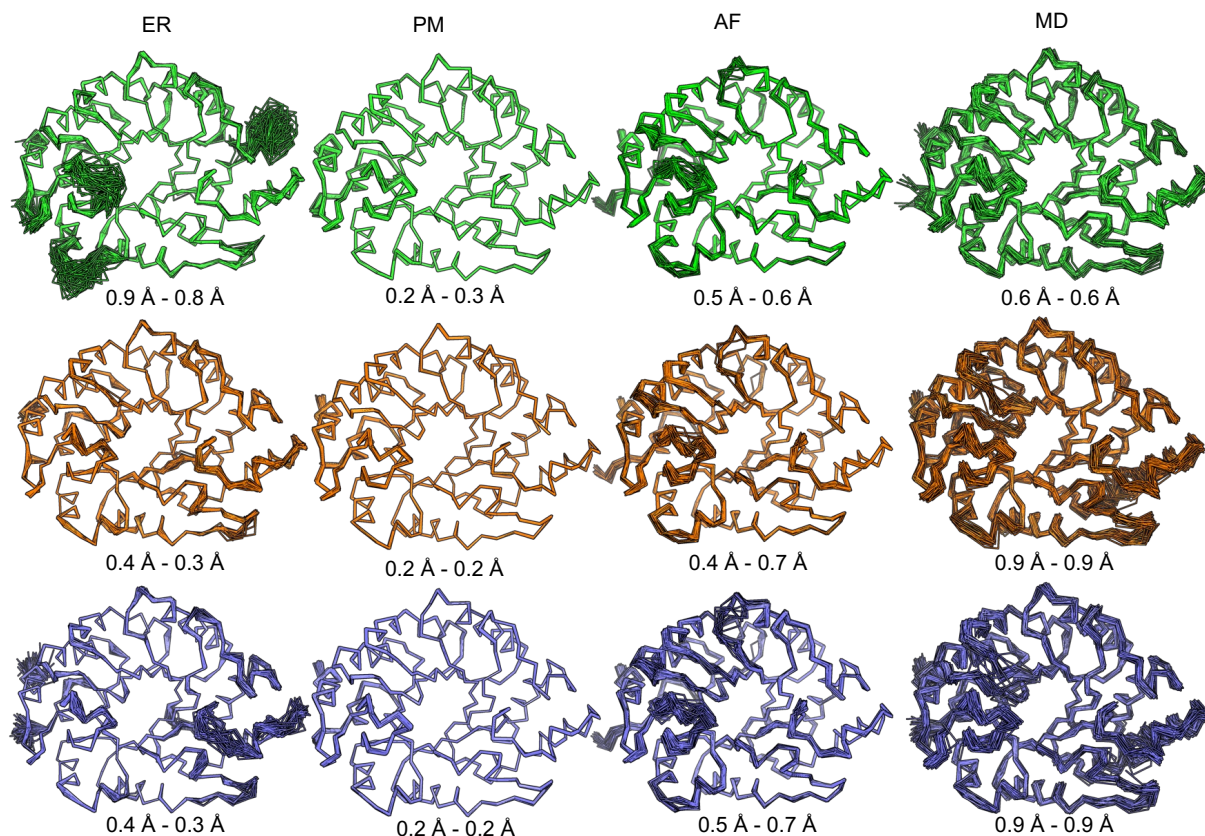


Figure 3.2. Overlay of all the ensembles from HG3. Diversity and deviation are shown at the bottom of each superposed ensemble, their standard deviation is shown in Table 3.2. The first row corresponds to ensembles of the wild-type template (green). The second row corresponds to the ensembles from HG3 in the unbound state (orange), and the last row corresponds to HG3 in the bound state (blue). From left to right are ensembles generated from ensemble refinement (ER), PertMin (PM), AlphaFold2 (AF) and Molecular dynamics simulations (MD).

As expected, ER and PertMin methods yielded ensembles with the lowest diversity and deviation for each enzyme, with PertMin generating ensembles with the lowest diversity (ranging from 0.2 Å to 0.3 Å) and the lowest deviation from the input crystal structure for all the enzymes in the three families (ranging from 0.2 Å to 0.3 Å). Except for ER ensembles generated from HG3_WT, in general, ER generates ensembles with low diversity (ranging from 0.3 Å to 0.7 Å) and low deviation (ranging from 0.3 Å to 0.7 Å) (Figure 3.2). For ER ensembles from HG3_WT, the diversity and deviation values are 0.9 Å and 0.8 Å, respectively. These high diversity and deviation are attributed to high fluctuations in loops (Figure 3.2, Table 3.2). In the case of AF, the

diversity and deviation of all the ensembles are higher compared to PertMin and ER but still lower compared to the ensembles generated by MD simulations. Interestingly, for HG3_WT, diversity and deviation are smaller than the other ensembles in its specific family. This observation could be attributed to the inherent stability of the protein, which potentially renders it more rigid.

From these generated ensembles, we expected that ensembles with low deviation from the HG3_EV, such as PertMin or ER ensembles, would provide better prediction overall. Since enzyme design procedure focuses on positions within the active site, we decided to calculate the diversity of the active site of each ensemble, and we presented their distribution (Figure 3.3 and Supplementary Figure 3.1). Furthermore, since we aim to determine if ensembles with low deviation from the optimized variants result in more accurate predictions, we plotted the distribution showing the deviation of the ensembles from the respective optimized variants for each enzyme as depicted in Figure 3.3.

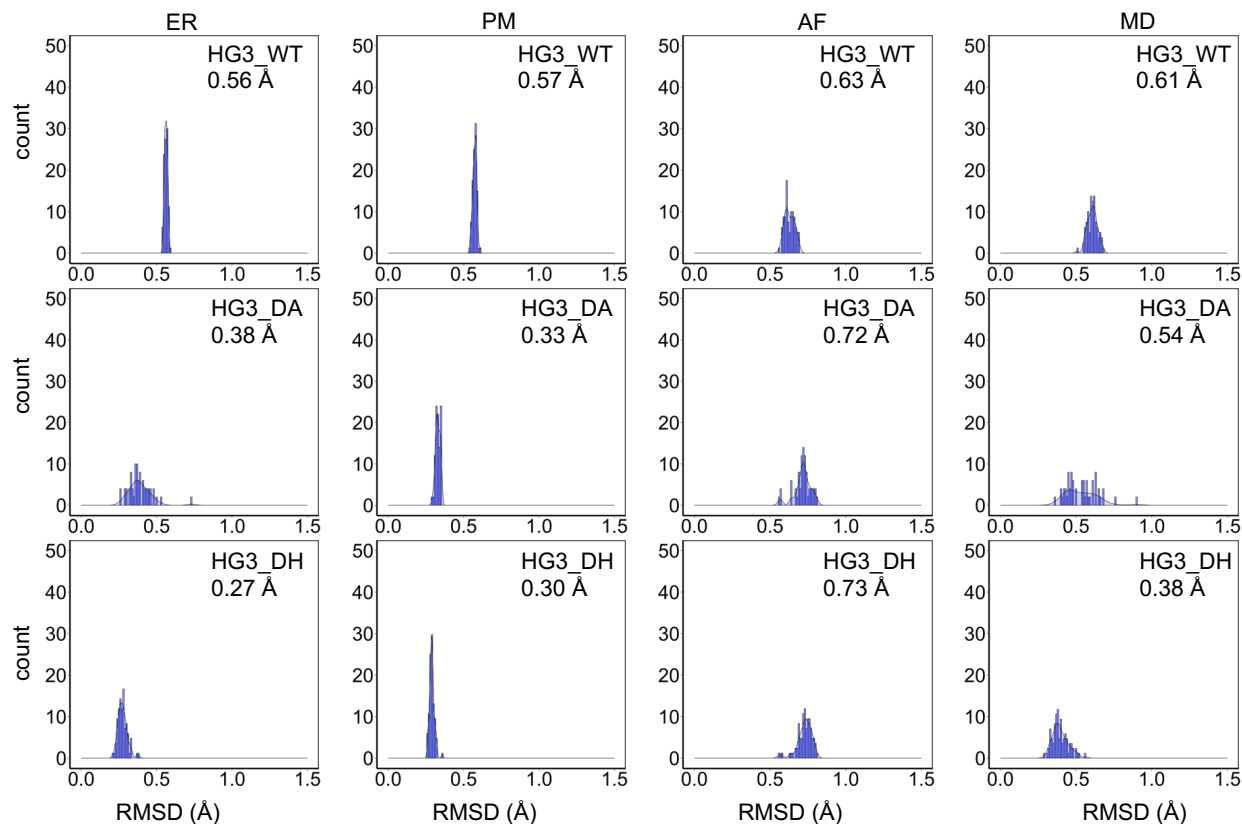


Figure 3.3. Distribution of the deviation of ensembles from HG3_EV crystal structure. Backbone atoms (N, C and CA) of active site positions (listed in Supplementary Table 3.1) are used for the calculation. Respective medians are shown at the top of the graph. From left to right, ensembles generated from ensemble refinement (ER), PertMin (PM), AlphaFold2 (AF) and molecular dynamics simulations (MD) are presented. From top to bottom, the ensembles generated from the crystal structures of the wild type used for the design (HG3_WT), the designed enzyme in the unbound state (HG3_DA), and the bound state HG3_DH. Medians are shown at the top of each graph.

We investigated structural similarities of the active site of each ensemble with the optimized variants by calculating the deviation of the backbone atoms (N, C and CA) of the first shell residues of the ensemble with the optimized variant (Figure 3.3). In the HG3 family, ER, PertMin and MD ensembles, the distribution of deviation is consistently centred around a median value that decreases as the ensemble is structurally similar to HG4 (Figure 3.3). This is the case of ensembles generated from the crystal structure of the designed enzyme in the bound state. However, for AF2 ensembles, the medians of deviation in each ensemble across the enzymes tested does not follow the same trend, in which 1GOR median (0.69 Å) is lower than HG3_DA and HG3_DH (0.73 Å) (Figure 3.3).

From these observations, we postulated that ensembles produced by ER and PertMin for the HG3 family will result in a better prediction accuracy compared to those generated by MD and AF ensembles.

3.3.3. Computational Ensemble-based Design

Controls. Designs are conducted as described in the method section. For the HG3 family, the positive control chosen is HG4¹⁴, a highly active variant of the HG-series, crystallized with the transition state analogue. We used this optimized variant for our positive control, as this enzyme harbours active site mutations of the highly active variant HG3.17 from directed evolution³³, and its activity is only 20% lower than that of HG3.17¹⁴. For simplification, throughout this study, HG4 will be called HG3_EV.

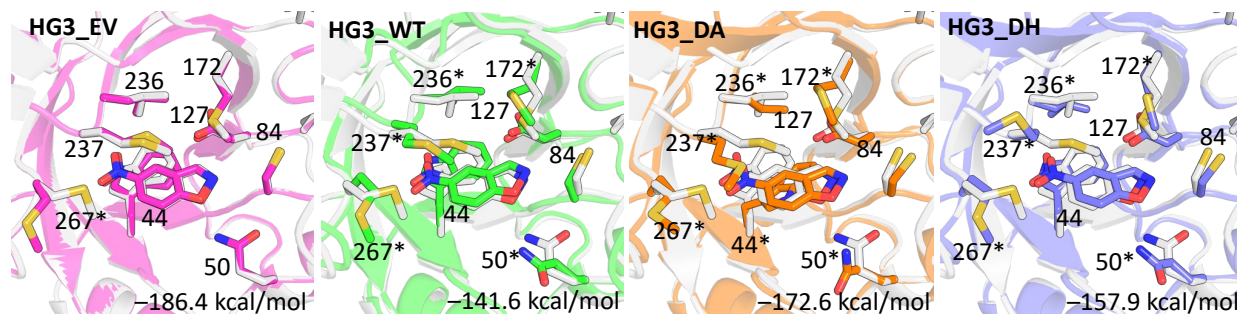


Figure 3.4. Computational design on crystal structure. Overlay of the crystal structure of the optimized variant HG3_EV (in grey) and the computation model obtained after the repacking step on its own template in magenta and designated as positive control. Computational models obtained using the crystal structure of the WT used for the initial design is shown in green, the model obtained using the crystal structure of HG3 in the unbound state in orange and the bound state in slate. Positions optimized during the repacking step and within 4 Å of the TS are represented in sticks.

To test the validity of our prediction, we conducted the prediction of HG3_EV active site and binding pose on its crystal structures bound with the TSA. To measure the ability of the method to predict the correct binding pose, we considered the binding energy (ΔE_H), the Root Mean Square Deviation of the TS position (RMSD_{TS}), and the percentage of correct rotamers (see methods). As

shown in Figure 3.4, the binding pose of the model obtained with the design on the positive control HG3_EV superpose well with its bound crystal structures shown in grey, and its RMSD_{TS} is low (0.42 Å), while the percentage of correct rotamers is above 60%. In addition, its ΔE_{H} of -186.4 kcal/mol demonstrating favourable interactions between active site residues and the TS.

We generated additional controls to test whether we could have accurate prediction given the crystal structure of the WT as our negative control, and the crystal structure of the designed enzymes both in unbound and bound states. As shown in (Figure 3.4), the negative control yielded to an unfavourable ΔE_{H} , high RMSD_{TS} and low percentage of correct rotamers. We observed similar results when we use the crystal structures of the designed enzymes. Taken together, these results confirm that given the correct template, the correct search algorithm, the correct rotamer library, and the correct forcefield, we should be able to get accurate prediction.

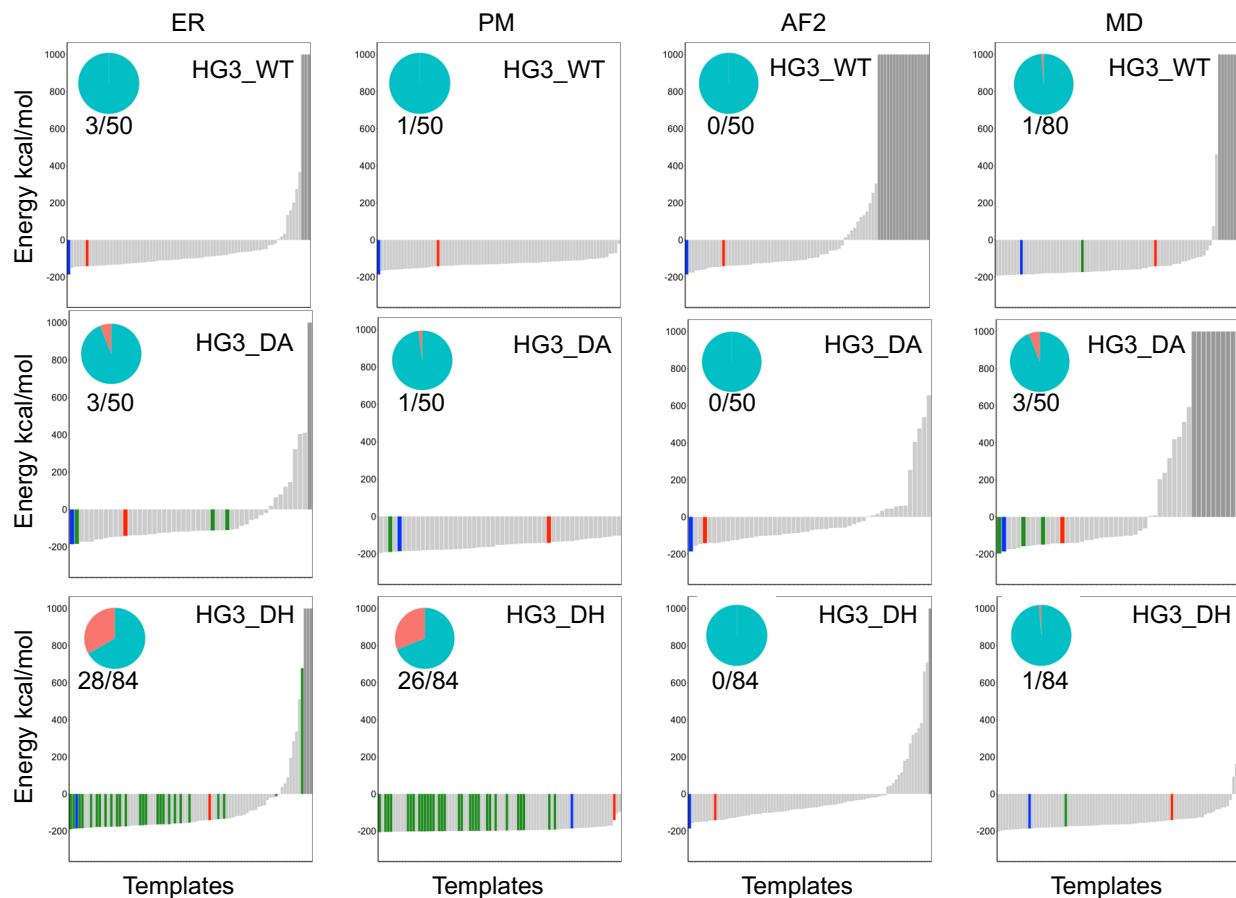


Figure 3.5. Bar plot showing repacking results. a) bar plots showing the templates generated from each method and their energies. The red bar corresponds to the design on the HG3_WT template, and the blue bar corresponds to the design on the template of the positive control. Bars coloured in green correspond to templates with RMSD_{TS} lower than 0.7 Å. A pie chart is superposed to show the number of templates with high similarity to the positive control (and orange pie) or low similarity (light blue). A similar template will be based on energy lower than 10 kcal/mol of the positive control (ΔE_{H} of the positive control is -186.4 kcal/mol), RMSD lower than 0.7 Å, and % correct rotamers higher than 50%)

3.3.4. Recapitulation of the transition-state binding pose.

Table 3.3. HG3 designs best results

Template	Energy (kcal/mol)	RMSD _{TS} (Å) ^a	% correct rotamers ^b
HG3_EV	−186.4	0.5	89
HG3_WT	−141.6	1.5	56
HG3_DA	−172.6	1.9	44
HG3_DH	−157.9	1.9	56
HG3_WT_ER_40	−143.5	1.2	44
HG3_WT_PM_18	−160.8	1.5	44
HG3_WT_AF_005	−176.2	1.4	56
HG3_WT_MD_17	−186.1	1.0	67
HG3_DA_ER_38	−184.8	0.4	56
HG3_DA_PM_038	−190.8	0.7	56
HG3_DA_AF_037	−154.6	3.4	44
HG3_DA_MD_40	−197.6	0.6	56
HG3_DH_ER_05	−185.2	0.5	78
HG3_DH_PM_013	−198.8	0.3	56
HG3_DH_AF_006	−157.8	1.4	44
HG3_DH_MD_81	−192.7	1.2	56

^a RMSD_{TS} is the RMSD between the TSA atoms in the crystal structure of HG3_EV and the TS atoms of the designed model.

^b % correct rotamers is the percentage of rotamers adopting the same rotamer as the one found in the crystal structure of HG3_EV. Residues selected to calculate correct rotamers are first shell positions. Ala and Gly were not included in the calculation as they do not have χ angles (see Supplementary Table 3.1)

Best results correspond to model in which their energy is at most 10 kcal/mol higher than the energy of the positive control (−186.4 kcal/mol), lowest RMSD_{TS} and the highest percentage of correct rotamers.

To assess the ability of the method to predict the transition-state position, we measured the Root Mean Square Deviation of all atoms of TS in the designed model and all the corresponding atoms of the TSA from the crystal structure of HG3_EV (RMSD_{TS}). In Figure 1.3, each data point corresponds to an individual template whose position is determined by its RMSD_{TS} as a function of its binding energy (ΔE_H). Ensembles generated from HG3_WT with the four methods resulted in templates giving higher ΔE_H and RMSD_{TS} than the positive control. In the case of HG3_DA, we can observe a similar trend with the AF ensembles, while results are improved with the MD ensembles, in which we observe a few templates with ΔE_H and RMSD_{TS} approaching those of the positive control. With the PertMin ensembles, we observe favourable ΔE_H for all the templates despite having high RMSD_{TS}. ER ensembles, on the other hand, have more templates with

favourable energy and low RMSD_{TS}. For HG3_DH, AF ensembles provide similar results to those from HG3_WT and HG3_DA. For MD ensembles, a higher proportion of templates has favourable energy compared to HG3_WT MD ensembles, and a small proportion of templates has RMSD_{TS} approaching the positive control values. In the case of HG3_DH, PertMin and ER ensembles generate templates with more favourable energies and templates with low RMSD_{TS}. Based on RMSD_{TS} and ΔE_H , HG3_DH ensembles from PertMin and ER are the two methods providing both favourable ΔE_H and low RMSD_{TS}. In addition to RMSD_{TS}, we calculated the percentage of correct rotamers of the active site residues on the individual templates. For the WT ensembles, we can see that the method giving the highest percentage of correct rotamers is the MD method. And the worst one is PertMin. In the case of the ensembles generated from HG3_DA, we observe that ensembles from ER yielded results with the highest percentage of correct rotamers across all the methods. In contrast, MD simulations ensembles still allow a high percentage of correct rotamers. In the case of HG3_DH, we can observe similar trends in which ER gives the highest proportion of ensembles with the highest percentage of correct rotamers. We can conclude here that for HG3, ER ensembles allow the generation of ensembles with a higher proportion of templates with a high % of correct rotamers. In Figure 3.5, the best models from ER and PertMin ensemble yield a design with a good agreement with the crystal structure of the optimized variant. In contrast, models with AF and MD ensembles provide a design with a lower similarity to the crystal structure of HG3_EV. These results agree with the hypothesis that ensembles from ER and PertMin will give the best variants.

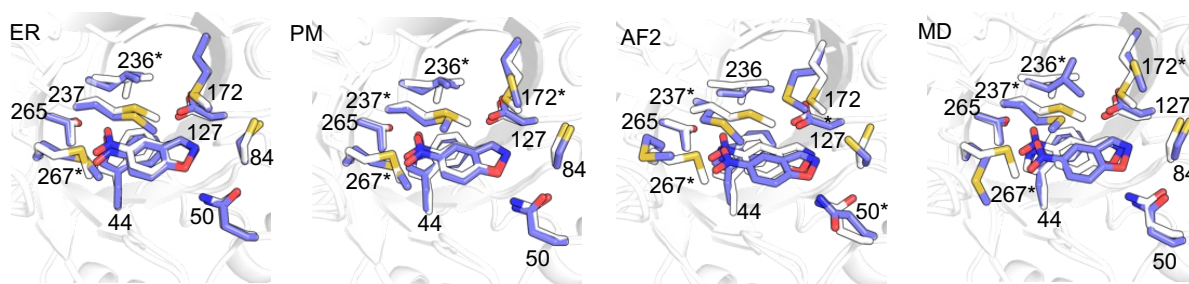


Figure 3.6. Best results obtained from HG3_DH ensembles. Models obtained from the calculations (in blue) are overlaid with the crystal structure of the positive control HG4 (in grey). Residues optimized during the computational design and in first shell positions are shown. Residue with * are residues with incorrect rotamers (see methods).

We observed that ER enables the design of a bound transition state similar to the coordinates of the TSA in the crystal structure. However, PertMin of the ensemble from HG3_DH gives the highest proportion of snapshots with low RMSD.

Regarding energy, PertMin ensembles give a higher proportion of templates with favourable energy. All ensembles obtained from the WT, HG3_DA and HG3_DH do not show any templates with positive energy. Interestingly, HG3_DH PertMin generates designs with a great proportion of templates with energies more favourable than the positive control (Figure 3.4 and Table 3.3). In the case of ER, HG3_DH also gives a high proportion of templates with favourable energies. In contrast to PertMin, few templates have energies that are more favourable than the positive control. Overall, we observed a trend where HG3_DH provides the highest proportion of templates with favourable energy, low RMSD_{TS} of the TS and high percentage of correct rotamers.

3.4. Discussion

3.4.1. Properties of the best ensembles

Our results demonstrated that ensembles obtained from the crystal structure of the designed enzyme bound with the ligand yield the most accurate prediction. Design on ER and PertMin ensembles consistently resulted in prediction with favourable binding energy and similar binding

poses in this study. MD ensembles yielded predictions with favourable binding energy, but inconsistent binding poses. AF ensembles generated low prediction accuracy. According to our results, the ensemble yielding the most accurate prediction corresponds to ensembles with the lowest deviation from the crystal structure of their optimized variant. This is supported by median values $<0.3 \text{ \AA}$ for ensembles yielding the most accurate predictions.

Results are consistent with our hypothesis, in which designs on the crystal structures of the WT and the design with the unbound structure provide less accurate predictions. By comparing their structural similarity with the optimized variant, we observed that their median values are lower compared to ensembles generated from the bound structure (Figure 3.3), reflecting high similarity of HG3_EV with ensembles obtained from ER and PertMin.

3.4.2. Parameters used for the prediction

In this study, we used a combination of binding energy, RMSD_{TS} , and percentage of correct rotamers to evaluate the designs.

Binding energy. In general, overall energies on each backbone member of the ensembles from the designed enzymes bound with the ligand yield favourable energy compared to the unbound structures and the WT. Interestingly, PertMin ensembles provide most of their templates with binding energy lower than 0 kcal/mol . The PertMin ensembles are generated by small random perturbations and energy minimization, in which the number of iterations guides the diversity of the ensemble. From this procedure, the ensemble generated will be structurally similar to its input structure and between each other, thus generating templates with energies similar across all individual members.

RMSD_{TS}. The RMSD_{TS} parameter used in this study allows us to evaluate the binding poses. In general, a low RMSD_{TS} is desirable. But this value has to be complemented with the percentage of correct rotamers, which tells us how similar the binding poses are compared to the crystal structure of the optimized variant. Although PertMin ensembles provides templates with a high proportion of low RMSD_{TS}, the percentage of correct rotamers are lower compared to the ensembles generated by ER. This comparison informs us about the precision of the prediction, in which overall ER ensembles provide the best results.

3.4.3. Limitations

As we have observed throughout this study, the prediction accuracy highly depends on the input structure we choose. For this reason, our ensemble-based method is limited with the obtention of a high-resolution crystal structure bound with the TSA. This limitation can be important if obtaining crystal is challenging. This is what we observe with the KE70 design, in which obtaining the crystal bound with the TSA was challenging⁸. Otherwise, one can still use the ensembles obtained from the enzymes in the unbound state. However, the results will inevitably be less precise, but they can still yield a more accurate prediction than the WT template ensembles. One alternative for using the crystal structure is using ensembles from machine learning prediction, such as AF. However, as we observed in this study, using the AF ensembles we generated resulted in a lower prediction accuracy. AF has good prediction accuracy when generating a single model^{26,27,30}. However, the ensembles we generated in this study may not accurately represent the conformational ensembles of the enzyme in solution. It could be possible to use an ensemble from machine learning in the condition that the ensembles generated have low diversity, and the structure of the active site should be the most similar to the actual structure of the designed enzyme

with the bound state. The accuracy of this method is limited with high confidence in domain prediction but not for loops and linkers²⁷, which, in the case of enzymes, are very important. In addition, this method will be more accurate if it can integrate experimental enzyme dynamic information.

Prediction of Gβ1 variants sequence showed that ensembles with high similarity to the initial input template give the best prediction. In contrast, ensembles with low structural similarity with the input structure give low prediction accuracy¹⁹. Allen et al. showed that MD ensembles would lead to proteins adopting non-native fold. However, MD simulations were used in the design procedure to evaluate and rank the designs^{5,11,35}. To our knowledge, no studies involved the use of ensembles from MD simulations directly in the design procedure. Earlier work of the Baker group used Backrub ensembles to optimize their design of KE70. However, the catalytic improvement only spanned 2 to 4-fold. The optimization procedure was coupled with directed evolution without evaluating the precision of the design⁸. Designing artificial enzymes increasingly involves the use of machine learning. The most recent example is the design of a highly efficient artificial luciferase ($k_{\text{cat}}/K_{\text{M}} \sim 10^6$). The use of a deep-learning-based family-wide hallucination was used to generate scaffolds with catalytic sites that stabilize the high-energy states of reactions involving two distinct substrates, diphenylterazine or 2-deoxycoelenterazine³⁶. However, high catalytic efficiency is only achieved after substitutions from site-saturated mutagenesis are introduced. The advantage of using machine learning for designing artificial enzymes lies in its ability to create scaffolds with high stability, which can be explained by the high solubility of the artificial luciferases. However, enzyme stability is not correlated with higher activity³⁷⁻³⁹, and stable proteins tend to exhibit more rigidity. Furthermore, the *de novo* protein scaffolds created in the study have shortened loops,

enabling higher stability. However, critical regions like loops can be important for catalysis^{4,40}, in addition to the intricate relationship between dynamic and enzyme functions^{17,18}.

Although machine learning is a promising route for designing artificial enzymes, introducing flexibility in the design procedure for higher geometric precision is very important.

3.4.4. Recommendations and perspectives

Enzymes acquire high catalytic activity through the precise alignment of the TS with the catalytic residues. The most active variant HG3.17, from the trajectory of HG3 evolution, showed that the base used for the design adopts a rotamer that is precisely positioned for catalysis, where there is only a small difference between the ideal geometries from the calculated angles of the acetamide dimer³³. Based on previous work and the results we obtained here, using a single crystal structure will lead to inaccurate prediction as it does not account for the dynamic properties of the enzyme.

MD ensembles we generated with the WT templates enabled a better prediction accuracy of the binding pose than its other ensembles. We can postulate that using the parameters we used for MD simulation would enable the sampling of a larger conformational space to access the energy of the activities' energy landscape. A good ensemble enriches the active site conformation and binding pose of the optimized variant. We can establish that to allow precise prediction, we should use ER and PertMin ensembles of the crystal structure bound with the TSA, as they yield templates with low active site diversity and low active site deviation from the optimized variant, which in turn enriches templates with low RMSD_{TS}.

3.5. Methods

Ensemble generation. For each family of enzymes, we generated ensembles from the crystal structures of the wild template used for the design (PDB ID: 1GOR), the designed enzymes in the unbound and bound states with the transition state analog or the product (PDB ID: 5RG4 and 5RGA). Alternative conformations were removed, while only the major conformations were kept. For the enzymes bound with the TSA or the product, The number of templates generated for each ensemble is guided by the number of templates generated by ensemble refinement (Supplementary Table 3.3). This method applies an ensemble reduction step, in which the smallest number of structures that reproduce R_{free} within 0.1%⁴¹ are kept to avoid redundant structures.

Ensemble refinement. Ensembles of backbone templates were generated using the ensemble refinement protocol⁴¹ implemented in the PHENIX macromolecular structure determination software (v1.18.2-3874)⁴². In ensemble refinement, local atomic fluctuations are sampled using molecular dynamics simulations accelerated and restrained by electron density to produce ensemble models fitted to diffraction data. Input crystal structures were edited to remove low-occupancy conformers and assign an occupancy of 1.0 to the remaining conformer. Following the 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 the translation-libration-screw fitting, τ_x is the relaxation time and $w_{\text{x-ray}}$ is the coupled b_{ath} offset, which controls the X-ray weight such that the simulation runs at the target temperature of 300 K. The ensemble generated from each crystal structure that displayed the lowest R_{free} value while maintaining its $R_{\text{free}} - R_{\text{work}}$ difference of less than five percent was selected for design and analysis.

AlphaFold2. AlphaFold2 is a deep learning-based artificial intelligence system that predicts protein conformations from its amino acid sequence, through its neural network architecture and trained data from known protein structures incorporating evolutionary, physical and geometric constraints. It takes a large genomic sequence database to generate a multiple-sequence alignment (MSA) from the input amino acid sequence and a large database of existing protein structures to incorporate pairwise features²⁶. The spatial and evolutionary relationship is computed from the input information as a distance matrix that integrates distance predictions into a heuristic system that produces the final 3D coordinates prediction. Ensembles are generated using ColabFold³¹ (<https://github.com/sokrypton/ColabFold>), an open-source platform combining Mmseqs2^{43,44} for fast homology search and AlphaFold2 architecture for structure prediction. The diversity of the ensemble was controlled by reducing max_msa to 16:32 and number of recycles to 1, while enabling the dropout option and withholding templates during prediction^{26,45,46}. To minimize the computational cost, we used fifty seeds and five models per seed for a final two hundred and fifty models and no final relaxation. Here, we use pLDDT (predicted Local Distance Difference Test) score and TM-score (Template Modelling score) to infer the quality of the prediction³². pLDDT evaluates the local accuracy by evaluating the local distance differences of all atoms in the model, regardless of domain motions. TM-score assesses the accuracy of the entire protein structure and its similarity to the reference by considering the distances and orientation of the alignment of C α atoms. Models were ranked based on the predicted pLDDT, and ensembles were selected from the top results, where the cut-off number used for design is based on the number of templates obtained by ensemble refinement. The ensemble members' pLDDT scores range from 88 to 97, and TM scores from 0.88 to 0.93, highlighting the accuracy of the generated ensembles.

Molecular Dynamics Simulation. 20 short MD simulations of one nanosecond each were performed using Amber 2020 (<http://ambermd.org/>) with the AMBER19SB forcefield⁴⁷. Long-range electrostatics (>10 Å) were modelled using the particle mesh Ewald method⁴⁸, and a time step of 2 fs was used for the production phase. Parameters for the ligands were generated using the Antechamber^{47,49} package.

Crystal structures in the bound and unbound states were protonated using the H++⁵⁰ webserver (<http://newbiophysics.cs.vt.edu/H++/index.php>) at pH 7.0 and a salt concentration of 0.15 M. Each unbound or ligand-bound protein molecule was placed in a dodecahedral box with periodic boundary conditions where the distance between the protein surface and the box edges were set to 10 Å. Explicit water molecules from the crystal structure were kept along with additional TIP3P water molecules⁵¹, charges on protein atoms were neutralized with Na⁺ and Cl⁻ counter-ions at a concentration of 0.15 M. The structures were then energy minimized with the steepest descent method to a target maximum force of 1000 kJ mol⁻¹ nm⁻¹. Before equilibration, the system was heated with iterative steps of 50 ps starting from 0 K to 300 K. The system was then equilibrated under an NVT ensemble for one ns at a temperature of 300 K using a Nose-Hoover thermostat⁵², while applying position restraints for heavy protein atoms. A second equilibration step under an NPT ensemble was performed for one ns with constant pressure and temperature of 1 bar and 300 K, respectively, using the Berendsen barostat⁵³. Following the removal of position restraints, 1-ns production runs under Parrinello-Rahman pressure coupling⁵⁴ were initiated from the final snapshot of the NPT equilibration. The ensemble was generated by extracting models from the concatenation of snapshots in 20 simulations. The number of extracted models was determined based on the number of templates obtained by ensemble refinement.

PertMin ensembles. Input templates for PertMin ensembles were prepared using the *addH.py* application (described below). A PertMin ensemble is generated by randomly perturbing the coordinates of all heavy atoms of the protonated input crystal structure by ± 0.001 Å along each Cartesian coordinate axis¹⁹. The perturbed structures were then energy minimized using 3000 iteration steps of a truncated Newton minimization algorithm.

Ensemble preparation. For ensemble refinement, hydrogens from all the templates were stripped off and replaced by hydrogens from the *addH.py* application from the Triad protein design software (Protabit, Pasadena, CA, USA). This method places and optimizes the geometry of the hydrogen using an energy-based criterion. It optimizes polar hydrogens, allows sidechain flips of amide groups in Asn/Gln/His, optimizes His protonation states, and rotates hydroxyl, sulfhydryl, methyl, and amino groups in Ser/Thr, Cys, Met, and Lys, respectively. No minimization was included in the template preparation for ensemble refinement. For MD simulations and AlphaFold2 templates, the *proteinProcess.py* application was used to prepare the ensembles. This process optimizes hydrogen placement as described previously in the *addH.py* application. Additionally, the *proteinProcess.py* protocol adds a step to fix clashes and bad angles by using *proteinHealth.py*. For PertMin, no additional preparations were included.

Computational enzyme design. The Kemp elimination transition state (TS) was built using parameters described by Privett and coworkers⁵. For HG3 designs, Asp127 was used as the catalytic base, and Gln50 was used as hydrogen-bond donors at position 50 (Supplementary Table 3.4). Except for catalytic residues, all the positions in the active site were mutated to Gly during

theozyme placement to avoid steric clashes with designed catalytic residues (Supplementary Table 3.1). The 2002 Dunbrack backbone-independent rotamer library⁵⁵ 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 placement¹⁶ using the contact geometries listed in (Supplementary Table 3.1). TS pose energies were calculated using a modified version of the Phoenix energy function⁵⁶ consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field⁵⁷ 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 Å⁵⁸, and an electrostatic energy term modelled using Coulomb's law with a distance-dependent dielectric of 10. During the energy calculation step, TS–side-chain interaction energies were biased to favour interactions that satisfy contact geometries (Supplementary Table 3.4) by applying an energy benefit of –100 kcal mol⁻¹ as described by Lassila *et al.*¹⁶.

Next, active-site repacking calculations were performed on the structure, stabilizing the structure the TS. In these calculations, the TS was translated ± 0.4 Å in each Cartesian coordinate in 0.2-Å increments and rotated 10° about all three axes (origin at TS geometric center) 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. Residues mutated to Gly in the theozyme placement step were allowed to sample rotamers from HG3_EV (Supplementary Table 3.1). Sidechain–TS interaction energies were biased to favour contacts satisfying geometries as done during the theozyme placement step (Supplementary Table 3.4). Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. After active-site repacking, we calculated energy differences between each repacked structure and the corresponding structure in the absence of the TS where all

designed residues (including catalytic residues) were converted to Gly. This energy subtraction was done to allow comparison of sequence energies across backbone templates.

Evaluation of the best ensemble. To assess the validity of our method, we used RMSD_{TS} , corresponding to the root mean square deviation (RMSD) of the transition state (TS) from the model and the TS that we superposed with the. Transition state analogue (TSA) in the crystal structure for HG3 of the optimized variant. ΔE_{H} (binding energy) corresponds to the energy of interaction between the TS and the active site residues. To calculate the % correct rotamers, we measured the difference between the χ angles of the individual residues in the active site of the optimized variant and the model. For one residue, if the difference is greater than a cut-off (40°), the rotamer of that specific residue is considered different. Finally, we counted the number of residues in the active site with correct rotamers to have the percentage of correct rotamers.

3.6. Supplementary information

Supplementary Table 3.1 List of amino-acid positions optimized during the computational design procedure

	Designed positions
Catalytic base	D127
H-bond donor	Q50
Pi-stacking	W44
First shell residues ^a	A21, W44, G83, C84, M172, L236, M237, T265, M267
Second shell residues ^b	V16, Y17, M42, E46, N47, L79, G81, W87, Q90, T125, G130, Q207, H209, S234, D239, A275, R276

^afirst shell residues correspond to any residues with at least one atom within 4 Å from the ligand.

^bsecond shell residues correspond to any residues with at least one atom within 8 Å from the ligand and residues 4 Å away from the catalytic residues

Supplementary Table 3.2. AlphaFold2 ensembles scores

Family	Alias	Predicted sequence	Ensemble size (n)	pLDDT _{max} – pLDDT _{min}	pTM _{max} – pTM _{min}
HG3	HG3_WT	QAAQSV DQLIKARGKVYFGVATDQNRLTTGKNAAI IQAD FGQVTPENSMKWDATEPSQGNFNFAGADYLVNWAQQNGK LIRGHTLVWHSQLPSWVSSITDKNTLTNVMKNHITLMT RYKGKIRAWDVVNEAFNEDGSLRQTVFLNVIGEDYIPIA FQTARAADPNAKLYINDYNLDSASYPKTQAIVNRVKQWR AAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVA ITELDVAGASPTDYVNVVNACLVQSCVGITVWGVADPD SWRASTPPLLFDGNFNPKPAYNAIVQDLQQ	80	96.5 – 94.5	0.93 – 0.93
	HG3_DA	EAAQSV DQLIKARGKVYFGVATDQNRLTTGKNAAI IQAD FGMVWVPEENSMKWDATEPSQGNFNFAGADYLVNWAQQNGK LIGGGMLVWHSQLPSWVSSITDKNTLTNVMKNHITLMT RYKGKIRAWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIA FQTARAADPNAKLYIMDYNLDSASYPKTQAIVNRVKQWR AAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVS ILMLDVAGASPTDYVNVVNACLVQSCVGITVFGVADPD SWRASTPPLLFDGNFNPKPAYNAIVQDLQQ	50	95.8 – 94.4	0.93 – 0.92
	HG3_DH	EAAQSV DQLIKARGKVYFGVATDQNRLTTGKNAAI IQAD FGMVWVPEENSMKWDATEPSQGNFNFAGADYLVNWAQQNGK LIGGGMLVWHSQLPSWVSSITDKNTLTNVMKNHITLMT RYKGKIRAWDVVGEAFNEDGSLRQTVFLNVIGEDYIPIA FQTARAADPNAKLYIMDYNLDSASYPKTQAIVNRVKQWR AAGVPIDGIGSQTHLSAGQGAGVLQALPLLASAGTPEVS ILMLDVAGASPTDYVNVVNACLVQSCVGITVFGVADPD SWRASTPPLLFDGNFNPKPAYNAIVQDLQQ	84	95.8 – 93.6	0.93 – 0.92

Supplementary Table 3.3. Ensemble refinement results

Structure	PDB ID	Resolution (Å)	# chains	R _{work} /R _{free} before ensemble refinement	R _{work} /R _{free} after ensemble refinement	# templates in ensemble
HG3_WT	1GOR	1.70	1	0.1740/0.2000	0.1168/0.1674	80
HG3_DA	5RG4	1.99	2	0.1770/0.2160	0.1600/0.2113	50
HG3_DH	5RGA	1.86	2	0.1482/0.1863	0.1332/0.1814	84

Supplementary Table 3.4. HG3 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 (1.0)	1.6 (1.6)
		Angle	CG	OD1 or OD2	H3		109 (109)	131 (131)
		Angle	OD1 or OD2	H3	C3		159 (159)	180 (180)
		Torsion	CB	CG	OD1 or OD2	H3	-21 (-21)	21 (21)
H-bond donor	Gln	Distance	1HE2 or 2HE2	O1			1.2 (1.2)	2.3 (2.3)
		Angle	NE2	1HE2 or 2HE2	O1		145 (145)	157 (157)
		Angle	1HE2 or 2HE2	O1	N2		120 (120)	140 (140)
		Torsion	1HE2 or 2HE2	O1	N2	C3	160 (160)	200 (200)

^aAtoms from the TS are in bold, the other atoms are from the interacting sidechain of His23

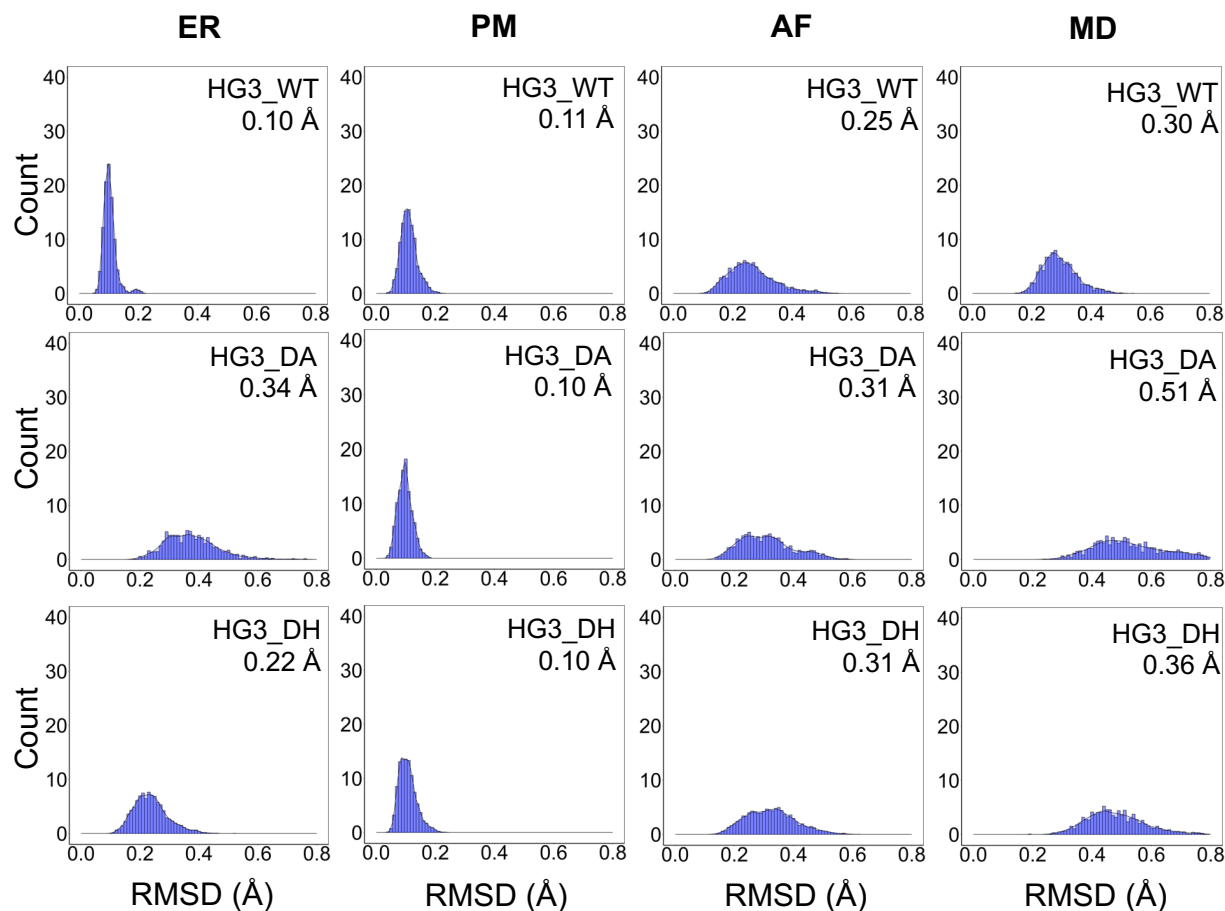
^bDistances are measured in Å, angles and torsions are in ° (degree). Values in parenthesis are the cut-off used for repacking

Supplementary Table 3.5. Geometric definitions for generation of transition-state poses off the side chain of catalytic base

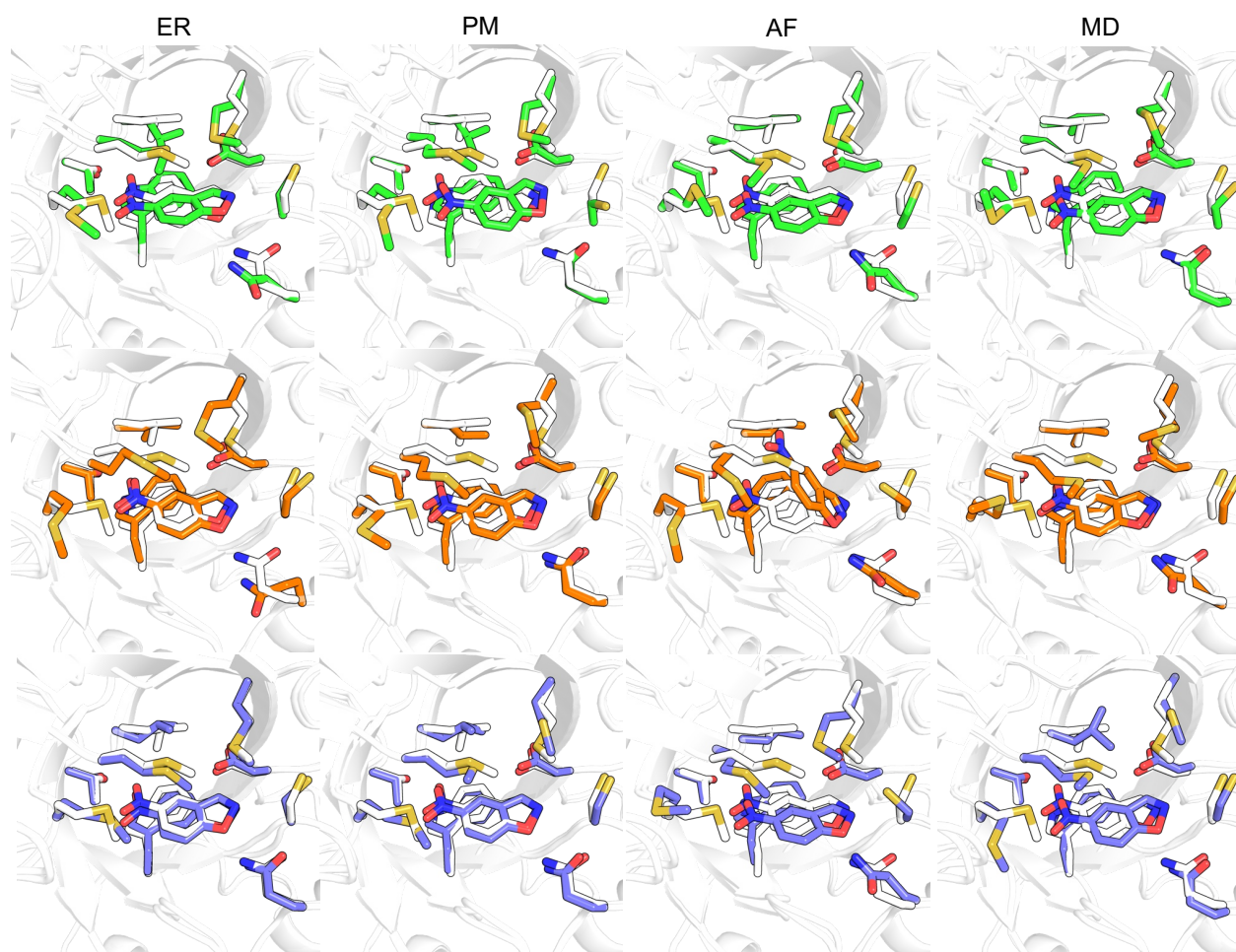
Enzymes	Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values ^b
HG3	Asp	Distance	OD1 or OD2	H3			1.0, 1.2, 1.5
		Angle	CG	OD1 or OD2	H3		112, 117, 122
		Angle	OD1 or OD2	H3	C3		159, 164, 169, 174, 179
		Torsion	CB	CG	OD1 or OD2	H3	0, 5, 10, 170, 175, 180
		Torsion	CG	OD1 or OD2	H3	C3	170, 175, 180, 185, 190
		Torsion	OD1 or OD2	H3	C3	N2	0, 5, 170, 175, 180

^aAtoms from the TS are in bold, the other atoms are from the interacting sidechain of His23

^bDistances are measured in Å, angles and torsions are in ° (degree)



Supplementary Figure 3.1. HG3 diversity distribution of active site positions. Backbone atoms (N, C and CA) of active site positions (Supplementary Table 3.1) are used for the calculation. Respective medians are shown at the top of the graph. From left to right, ensembles generated from ensemble refinement (ER), PertMin (PM), AlphaFold2 (AF) and molecular dynamics simulations (MD) are presented. From top to bottom, the ensembles generated from the crystal structures of the wild type used for the design (HG3_WT), the designed enzyme in the unbound state (HG3_DA), and the bound state (HG3_DH)



Supplementary Figure 3.2. Best designs obtained from the design of HG3 family. From left to right ER, PM, AF, and MD ensembles. From top to bottom, designs on HG3_WT, HG3_DA and HG3_DH. Residues optimized during computational design are shown in sticks.

References

1. Röthlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. *Nature* 453, 190–195 (2008).
2. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* 319, 1387–1391 (2008).
3. Siegel, J. B. et al. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* 329, 309–313 (2010).
4. Althoff, E. A. et al. Robust design and optimization of retroaldol enzymes. *Protein Science* 21, 717–726 (2012).
5. Privett, H. K. et al. Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* 109, 3790–3795 (2012).
6. Richter, F. et al. Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *Journal of the American Chemical Society* 134, 16197–16206 (2012).
7. Bjelic, S. et al. Computational design of enone-binding proteins with catalytic activity for the morita-baylis-hillman reaction. *ACS Chemical Biology* 8, 749–757 (2013).
8. Khersonsky, O. et al. Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *Journal of Molecular Biology* 407, 391–412 (2011).
9. Khersonsky, O. et al. 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).
10. Blomberg, R. et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* 503, 418–421 (2013).
11. Hong, N. S. et al. The evolution of multiple active site configurations in a designed enzyme. *Nature Communications* 9, (2018).
12. Burke, A. J. et al. Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* 570, 219–223 (2019).
13. Crawshaw, R. et al. Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* 14, 313–320 (2022).
14. Broom, A. et al. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nature Communications* 11, (2020).
15. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis | *Science*. <https://www.science.org/doi/10.1126/science.abd3623>.
16. Lassila, J. K., et al. Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* 103, 16710–16715 (2006).
17. Eisenmesser, E. Z., et al. Enzyme Dynamics During Catalysis. *Science* 295, 1520–1523 (2002).
18. Eisenmesser, E. Z. et al. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438, 117–121 (2005).
19. Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins: Structure, Function and Bioinformatics* 82, 771–784 (2014).
20. Lauck, F., Smith, et al. RosettaBackrub—a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res* 38, W569–W575 (2010).
21. Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Science* 21, 1241–1252 (2012).
22. Burnley, B. T., et al. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* 1, e00311 (2012).
23. Burnley, B. T. & Gros, P. phenix.ensemble_refinement: a test study of apo and holo BACE1. *Computational crystallography newsletter* 4, 51–58 (2013).
24. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
25. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596 (2021).
26. Allen, B. D., et al. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* 107, 19838–19843 (2010).
27. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710 (2020).
28. Ourmazd, A., et al. Structural biology is solved — now what? *Nat Methods* 19, 24–26 (2022).

29. Caves, L. S. D., et al. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Science* 7, 649–666 (1998).
30. Wang, P., et al. Using High-Throughput Molecular Dynamics Simulation to Enhance the Computational Design of Kemp Elimination Enzymes. *J. Chem. Inf. Model.* 63, 1323–1337 (2023).
31. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022).
32. Mariani, V., et al. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 29, 2722–2728 (2013).
33. Blomberg, R. et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* 503, 418–421 (2013).
34. Kiss, G., et al. Evaluation and ranking of enzyme designs. *Protein Science* 19, 1760–1773 (2010).
35. Yeh, A. H.-W. et al. De novo design of luciferases using deep learning. *Nature* 614, 774–780 (2023).
36. Hou, Q., et al. Enzyme Stability-Activity Trade-Off: New Insights from Protein Stability Weaknesses and Evolutionary Conservation. *J. Chem. Theory Comput.* 19, 3664–3671 (2023).
37. Stark, C., et al. Systematic investigation of the link between enzyme catalysis and cold adaptation. *Elife* 11, e72884 (2022).
38. Struvay, C. & Feller, G. Optimization to low temperature activity in psychrophilic enzymes. *Int J Mol Sci* 13, 11643–11665 (2012).
39. Romero-Rivera, A., et al. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catalysis* 7, 8524–8532 (2017).
40. Tom Burnley, B., et al. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* 2012, 1–29 (2012).
41. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Cryst D* 68, 352–367 (2012).
42. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35, 1026–1028 (2017).
43. Mirdita, M., et al. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35, 2856–2858 (2019).
44. Casadevall, G., et al. Estimating conformational heterogeneity of tryptophan synthase with a template-based AlphaFold2 approach. *Protein Sci* 31, e4426 (2022).
45. Del Alamo, D., et al. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* 11, e75751 (2022).
46. Wang, J., et al. Development and testing of a general amber force field. *J Comput Chem* 25, 1157–1174 (2004).
47. Darden, T., et al. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics* 98, 10089–10092 (1993).
48. Wang, J., et al. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25, 247–260 (2006).
49. Gordon, J. C. et al. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 33, W368–W371 (2005).
50. Jorgensen, W. L., et al. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79, 926–935 (1983).
51. You, A., et al. The Nose–Hoover thermostat. 4069, (2007).
52. Berendsen, H. J. C., et al. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* 81, 3684–3690 (1984).
53. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* 52, 7182–7190 (1981).
54. Dunbrack, R. L. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology* 12, 431–440 (2002).
55. 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).
56. Mayo, S. L., et al. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* 94, 8897–8909 (1990).
57. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* 94, 10172–10177 (1997).

Chapter 4. Harnessing the conformational ensemble to design efficient artificial enzymes

Rajo V. Rakotoharisoa, Behnouth Seifinoferest, Niayesh Zarifi, Jack Miller, Joshua M. Rodriguez, Michael C. Thompson, Roberto A. Chica

4.1. Statement of contribution

R.V.R. performed the computational design experiments of HG3 and N.Z. performed computational design experiments of KE70. R.V.R. purified HG3 proteins and J.M. purified KE70 proteins. R.V.R. performed enzyme kinetics experiments on HG3 and J.M. performed enzyme kinetics experiments on KE70. B.S. crystallized proteins. J.M.R. and M.C.T. performed X-ray diffraction experiments. B.S., R.V.R., and R.A.C. performed refinements. M.C.T. designed X-ray crystallography experiments. R.A.C. and R.V.R. wrote the manuscript. N.Z., B.S., and M.C.T. edited the manuscript.

4.2. Abstract

The ability to create efficient artificial enzymes for any desired chemical reaction is of great interest. Here, we describe an ensemble-based computational design method for increasing catalytic efficiency of *de novo* enzymes to a level comparable to their natural counterparts without relying on directed evolution. Using structural ensembles generated from dynamics-based refinement against of the crystal structures of Kemp eliminases HG3 ($k_{\text{cat}}/K_{\text{M}}$ 125 M⁻¹ s⁻¹) and KE70 ($k_{\text{cat}}/K_{\text{M}}$ 57 M⁻¹ s⁻¹), we design from each enzyme ≤ 10 sequences predicted to more efficiently catalyze this reaction. The most active designs display $k_{\text{cat}}/K_{\text{M}}$ values improved by 100–250-fold, comparable to mutants obtained after screening thousands of variants in multiple rounds

of directed evolution, and crystal structures show excellent agreement with computational models. Our work shows how ensemble-based design can generate efficient artificial enzymes by exploiting the true conformational ensemble to accurately model the transition state.

4.3. Introduction

In chapter 3, a benchmark of different backbone ensembles generated by different methods showed that for HG3, Ensemble Refinement (ER) and PertMin allowed us to get higher proportion of templates with high prediction accuracy in the case of HG3. In this chapter, we aim to use the ensembles from ER we generated for HG3 to design enzyme variants with high catalytic efficiency.

The ability to create efficient artificial enzymes for any desired chemical reaction is of great interest as it would open the door to a broad range of applications in industry and medicine. Algorithms for *de novo* enzyme design^{1,2} can be used to create artificial enzymes for reactions for which no naturally occurring enzyme is known. While this computational design process has been used to produce *de novo* enzymes for various model organic transformations³⁻⁷, downstream directed evolution has been necessary to convert these enzyme prototypes into catalysts as efficient as their natural counterparts⁸⁻¹³. However, directed evolution is costly, time-consuming and impractical for many enzymatic reactions of interest for which high-throughput screening assays are unavailable. Therefore, a computational design procedure to enhance catalytic efficiency of *de novo* enzymes would be beneficial as it could bypass directed evolution and accelerate the creation of efficient artificial enzymes.

Structural analyses of high activity variants evolved from *de novo* enzymes revealed several mechanisms by which catalytic function can be augmented. These mechanisms include

optimization of hydrogen-bond geometries between catalytic residues and the transition state⁸, enhancement of active-site complementarity to the transition state^{8,9,14}, improvement of active-site preorganization¹⁴, and enrichment of catalytically competent substates in the conformational ensemble¹⁵⁻¹⁷. These structural effects are challenging to design using current enzyme design protocols, which rely on the use of a single protein structure as scaffold^{3,7,9}, because they are not sufficiently precise⁸ and do not consider the dynamic nature of enzyme structures¹⁵. Recently, new methodologies have been developed that take advantage of the spatiotemporal averaging inherent to X-ray crystallography – crystallographic data represent billions of crystallized molecules that are all sampling different conformations during data collection – to model the conformational ensembles of proteins with high fidelity^{18,19}. New design strategies that could exploit this experimental information should facilitate the design of efficient artificial enzymes by enabling accurate modeling of catalytically productive conformational substates.

Here, we report a computational procedure for enhancing the catalytic efficiency of *de novo* enzymes that exploits crystallographically-derived ensemble models for increased accuracy. Starting from the non-cryogenic crystal structure of *de novo* Kemp eliminase HG3⁷, we generate a structural ensemble fitted to the diffraction data^{18,20} and use it to design mutant sequences predicted to efficiently catalyze the Kemp elimination. We filter these sequences using a variety of structural criteria aiming to reproduce molecular mechanisms known to augment enzyme catalysis, and experimentally characterize 10 sequences. The most active design displays a catalytic efficiency improved by approximately 250-fold, a result comparable to those obtained by seven rounds of directed evolution and screening of >7000 variants⁸. We solve crystal structures of several high-activity variants, which show excellent agreement with the computational models and confirm that catalytic efficiency is enhanced via the intended molecular mechanisms. We also

demonstrate the general applicability of our approach by applying it to an unrelated *de novo* Kemp eliminase, KE70³, which yields comparable results. Our work shows how ensemble-based design can generate efficient artificial enzymes by exploiting information derived from X-ray crystallography to identify productive substates for efficient catalysis.

4.4. Results

4.4.1. Ensemble-based enzyme design

To create efficient *de novo* enzymes without relying on iterative rounds of mutagenesis and high-throughput screening, we implemented an ensemble-based computational enzyme design pipeline that uses protein backbones generated by ensemble refinement of diffraction data¹⁸ as templates for design (Figure 4.1). Because the diffraction data and calculated electron density maps are derived from crystals that contain billions of molecules, which are all sampling different conformational states throughout data collection, the resulting ensemble models are expected to represent the entire equilibrium conformational ensemble of the target protein fold²¹. Thus, we postulated that performing crystallographic ensemble refinement on a low activity *de novo* enzyme would allow sampling of conformational substates that can more accurately accommodate the transition state and its interactions with catalytic residues than the single conformation modeled using traditional refinement methods, thereby facilitating the design of improved active sites. To test our hypothesis, we applied ensemble-based design to increase catalytic efficiency of the *de novo* enzyme HG3⁷ that catalyzes the Kemp elimination (Supplementary Figure 4.1a), a well-established model organic transformation for benchmarking computational enzyme design methods. We selected HG3 as a case study for three reasons. First, X-ray diffraction data collected at non-cryogenic temperatures (277 K) were available¹⁴, which is ideal for ensemble refinement

because cryocooling is known to narrow the conformational ensemble of crystallized proteins²², reducing the number of conformations that are represented in the ensemble-averaged data. Second, these data were collected from crystals with a bound transition-state analogue that is resolved in the electron density¹⁴, allowing inclusion of the transition state in the ensemble models. Finally, HG3 was subjected to an extensive directed evolution campaign that increased catalytic efficiency by almost three orders of magnitude⁸, allowing us to compare results with those from purely experimental approaches.

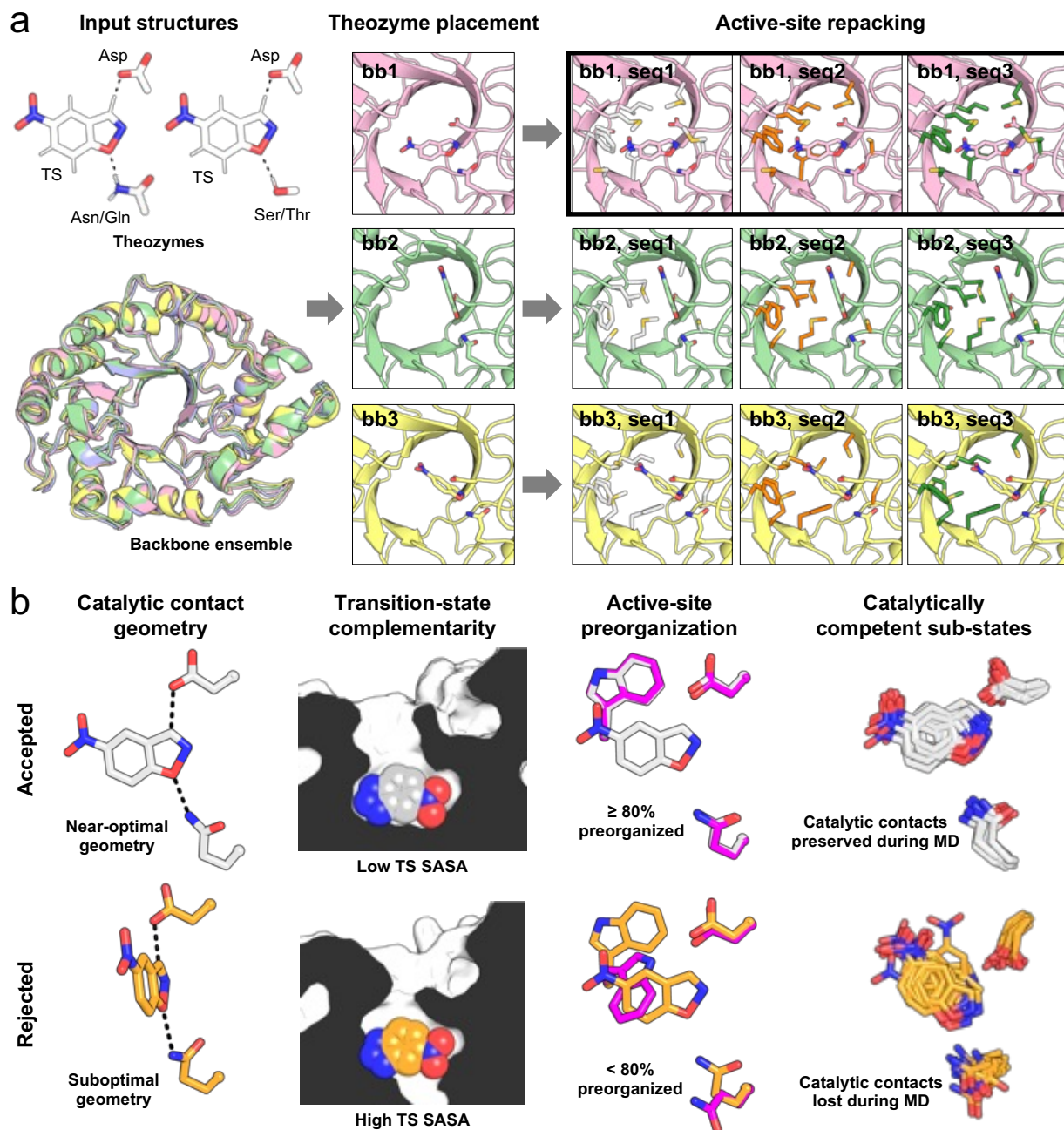


Figure 4.1. Ensemble-based computational enzyme design. (a) This procedure requires as inputs one or more theozymes, which are combinations of catalytic groups arranged in a unique geometry to stabilize a transition state (TS), and a backbone ensemble that approximates the intrinsic flexibility of the target protein fold. The algorithm proceeds in two stages: placement of each theozyme onto every member of the backbone ensemble (bb1–3), followed by repacking of active-site residues to identify sequences (seq1–3) that can further stabilize the TS. During repacking, both the identity and conformation of designed active-site residues are allowed to vary, and the TS is rotated and translated to reduce steric clashes with repacked side chains. In this toy example, only the theozyme containing a Gln hydrogen-bond donor is shown for clarity. After active-site repacking, designs from one or more theozyme/template combinations are selected based on energy (bold outline) and filtered using structural criteria. (b) Filtering of designs is performed to reproduce activity-enhancing mechanisms of directed evolution. Accepted designs are those that have

(1) near-ideal contact geometries between the TS and catalytic side chains (dashed lines), (2) active sites with high structural complementarity to the TS, as calculated by low TS solvent-accessible surface area (SASA), (3) highly preorganized active sites with >80% of designed residues predicted to adopt the same rotameric configuration in the presence (white/orange) and absence (magenta) of TS, and (4) enriched catalytically competent sub-states, which are snapshots that preserve catalytic contacts over the course of short molecular dynamics (MD) simulations.

We generated an ensemble of 84 backbone templates (Supplementary Table 4.1) from the crystal structure of HG3 bound with the 6-nitrobenzotriazole (6NT, Supplementary Table 4.1.b) transition-state analogue (PDB ID: 5RGA¹⁴) and used these structures for placement of four theozymes containing an Asp residue as catalytic base, and either a Gln, Asn, Ser or Thr to act as hydrogen-bond donor for stabilizing negative charge buildup on the phenolic oxygen at the transition state (Figure 4.1a). We kept the base position fixed to that of HG3 (D127) but allowed the algorithm to place the hydrogen-bond donor at multiple positions in the active site given that no amino acid residue in HG3 forms this catalytic contact with 6NT¹⁴. We then repacked the active site around the most stable theozyme placed on each backbone template by optimizing the identity and side-chain conformation of neighbouring residues to maximize beneficial packing interactions with the transition state, which has been shown to contribute to increased catalytic activity⁸.

After repacking, we filtered designs using a series of structural metrics aiming to reproduce mechanisms known to augment enzyme catalysis (Figure 4.1b, Supplementary Table 4.2). First, we measured angles defining the hydrogen bond between D127 and the transition state for the top-scoring design obtained on each backbone template, and selected sequences from the single backbone template that gave the most stable sequence with angles that were closest to the optimal geometries calculated for acetamide dimers²³ (Supplementary Figure 4.2, Methods). The selected backbone, which we name bb05 (Supplementary Table 4.3), yielded sequences that contain the Q50 hydrogen-bond donor that emerged during evolution of HG3 and that is present in the highest activity variants of this enzyme family⁸. Second, we evaluated active-site complementarity to the

transition state by calculating the solvent-accessible surface area of this ligand in the computational models generated for 1000 sequences designed on backbone bb05. Only designs where $\geq 95\%$ of the transition state surface is buried were kept. Third, we evaluated active-site preorganization by calculating the percentage of designed active-site residues that were predicted to adopt similar rotamers in the absence and presence of the transition state, which we consider to be preorganized for catalysis. Sequences whose active-sites were predicted to be $>70\%$ preorganized were kept. We also eliminated sequences with calculated energies over 10 kcal mol^{-1} higher than that of an HG3 single point mutant containing the D127/Q50 catalytic dyad, as well as those that contained more than four methionines or more than one cysteine in the active site to avoid excessive conformational heterogeneity and prevent formation of disulfide bonds, respectively. This filtering process narrowed the original list of 1000 designs obtained from backbone bb05 down to 55 designs. A final filtering step consisting of a 20-ns molecular dynamics simulation was used to identify designs that were enriched in catalytically-competent conformational substates, which we define as snapshots that make the designed catalytic contacts (Supplementary Figure 4.3). This final filtering resulted in nine sequences containing four or five mutations each (Table 4.1), which we experimentally characterized.

Table 4.1. Kinetic parameters of Kemp eliminases

Enzyme	Mutations ^a	# Mutations ^a	k_{cat} (s ⁻¹) ^b	K_M (mM) ^b	k_{cat}/K_M (M ⁻¹ s ⁻¹) ^{b,c}	Relative k_{cat}/K_M
HG series – Controls						
HG3	–	–	N.D.	N.D.	125 ± 8 (146 ± 6)	1
HG3.3b	V6I K50H M84C S89R Q90D A125N	6	N.D.	N.D.	3400 ± 700 (2200 ± 100)	27
HG3-K50Q/M84C	K50Q M84C	2	N.D.	N.D.	4300 ± 300	34
HG3.7	V6I Q37K K50Q M84C S89R Q90H A125N	7	N.D.	N.D.	24000 ± 3000 (27000 ± 2000)	192
HG series – Designs						
HG378	A21C K50Q Q90L A125V W275V	5	N.D.	N.D.	70 ± 5	0.6
HG57	A21C K50Q Q90L A125V	4	N.D.	N.D.	120 ± 8	1
HG44	A21C K50Q Q90F A125V	4	N.D.	N.D.	250 ± 10	2
HG603	K50Q Q90F A125V W275V	4	N.D.	N.D.	800 ± 10	6.4
HG493	K50Q Q90F A125T W275T	4	N.D.	N.D.	2100 ± 100	17
HG185	A21C K50Q Q90F A125T W275V	5	N.D.	N.D.	2300 ± 100	18
HG198	K50Q M84C A125T F267M	4	N.D.	N.D.	8000 ± 300	64
HG259	K50Q M84C Q90L A125T F267M	5	N.D.	N.D.	13000 ± 400	104
HG630	K50Q M84C A125T F267M W275V	5	N.D.	N.D.	15000 ± 500	120
HG694	K50Q M84C Q90L A125T F267M W275V	6	N.D.	N.D.	20000 ± 700	160
HG649	K50Q M84C Q90F A125T F267M W275V	6	N.D.	N.D.	32000 ± 3000	256
KE70 series – Controls						
KE70	–	–	N.D.	N.D.	57 ± 3 (78 ± 14)	1
KE70-S20a/A239a ^d	S20a A239a	2	N.D.	N.D.	150 ± 7	2.6
R2 7/12F	D23G, Y48F, D212E, H251Y	4	0.50 ± 0.01	1.00 ± 0.06	500 ± 30 (1330 ± 13)	9
R4 8/5A ^d	S20a, L29N, Y48F, W72C, H166Y, L197N, A204V, G239a	8	3.0 ± 0.1	1.0 ± 0.1	3000 ± 300 (4380 ± 350)	53
KE70 series – Designs ^e						
KE707 (26)	^d H16D S20a D45N Y48F W72I G101A A103S S138A H166N V168A A204I A239a	12	0.020 ± 0.001	0.40 ± 0.04	50 ± 5	1
KE700 (28)	^d H16D S20a D45Q Y48F R70K W72Y G101S A103S S138V H166N V168A A204I A239a	13	N.D.	N.D.	41.3 ± 0.9	0.7
KE708 (22)	^d H16D S20a D45Q Y48F W72C G101S S138Q H166N V168A A204V A239a	11	0.040 ± 0.001	0.10 ± 0.02	400 ± 80	7
KE703 (21)	^d H16D S20a D45N Y48F W72V G101M S138Q H166N V168A A204V A239a	11	0.50 ± 0.01	0.20 ± 0.02	2500 ± 300	44
KE703b (21OG)	^d S20a Y48F W72V G101M S138Q H166N V168A A204V A239a	9	ND	ND	5900 ± 100	96

^a Mutations are relative to the starting template (HG3 or KE70).

^b Data represent the average of six to fifteen individual replicate measurements from two to five independent protein batches. 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 (2 mM, Supplementary Figure 4.4 and Supplementary Figure 4.5), which is the substrate's solubility limit. In these cases, catalytic efficiencies (k_{cat}/K_M) were calculated from the slope of the linear portion ($[S] \ll K_M$) of the Michaelis-Menten model ($v_0 = (k_{cat}/K_M)[E_0][S]$). 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 Values in parentheses are from Broom *et al.* (1) (HG series), or Röthlisberger *et al.* (2) and Khersonsky *et al.* (3) (KE70 series).

^d These enzymes contain two insertions, S20a and A239a/G239a.

^e Six designs were either inactive (KE706 and KE709) or precipitated upon addition of substrate (KE701, KE702, KE704 and KE705), which did not allow kinetic parameters to be determined.

4.4.2. Kinetic characterization of designs

All nine designs catalyzed the Kemp elimination (Supplementary Figure 4.4). The most active of these variants, HG630, had a catalytic efficiency of $15,000 \pm 500 \text{ M}^{-1} \text{ s}^{-1}$, a value that is 120-fold higher than that of the starting template HG3 (Table 4.1). Furthermore, the three most active designs (HG198, HG259 and HG630) were more catalytically efficient than the HG3.3b variant obtained after three rounds of directed evolution and screening of >3000 mutants⁸. These designs all contain the M84C mutation that was previously shown to increase catalytic efficiency by one order of magnitude when combined with K50Q to yield the HG3-K50Q/M84C double mutant¹⁵ (Table 4.1). Analysis of mutant sequences suggested that introduction of the Q90F mutation into HG630 would further enhance its catalytic efficiency, as it doubles $k_{\text{cat}}/K_{\text{M}}$ when introduced into HG57 to yield HG44. Introduction of this mutation into HG630 resulted in a hexamutant (HG649) that is 256-fold more catalytically efficient than HG3 ($k_{\text{cat}}/K_{\text{M}} = 32,000 \pm 3000 \text{ M}^{-1} \text{ s}^{-1}$) and more active than HG3.7, a variant found after seven rounds of directed evolution involving the screening of >7000 mutants⁸ (Table 4.1, Figure 4.2a). HG649 shares a majority of its mutation sites with HG3.7 but the combination of specific amino-acid substitutions is distinct, with only two mutations in common (Figure 4.2b). These results demonstrate how ensemble-based enzyme design produces similar activity increases as directed evolution but at a fraction of the experimental screening cost. Interestingly, HG649 was one of the 55 designs that passed all filtering criteria except for the final molecular dynamics analysis, which led to its elimination due to its low predicted catalytic competency (Supplementary Figure 4.3). This false negative, along with the two false positives (Table 4.1, HG378 and HG57), indicate that improvements are required to make the molecular dynamics filtering more robust and suggests that other eliminated sequences may also display high catalytic efficiency.

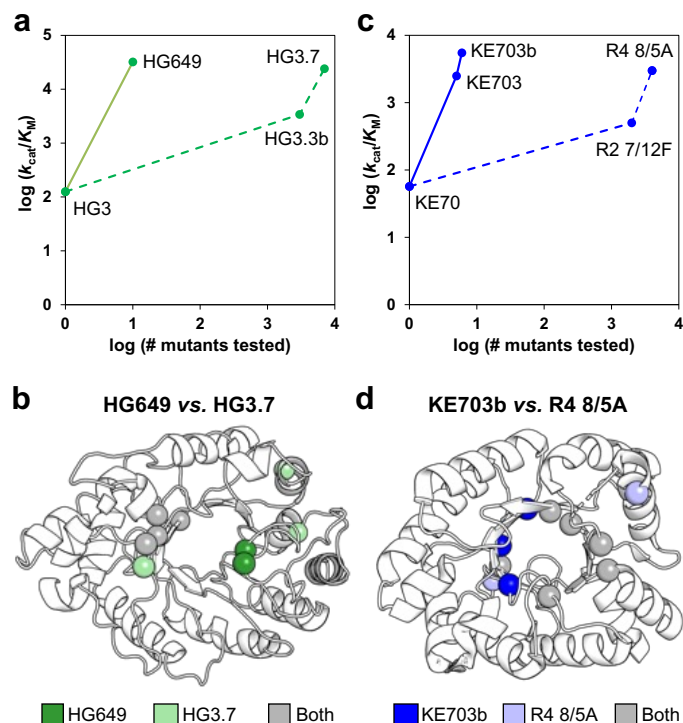


Figure 4.2. Ensemble-based design vs. directed evolution. (a) Ensemble-based design yields catalytic efficiency improvements comparable to those achieved by directed evolution with a fraction of the experimental screening effort. Full and dashed lines show ensemble-based design and directed evolution trajectories, respectively. (b) Ensemble-based design identifies a majority of the mutation sites (spheres) found by directed evolution (4/6) but mutation combinations of designed (HG649) and evolved (HG3.7) enzymes are distinct, as these share only two mutations (K50Q and M84C). (c) Ensemble-based design of KE70 yields catalytic efficiency improvements comparable to those achieved by directed evolution with a fraction of the experimental screening effort. KE703b differs from KE703 by reversion of its D16/N45 residues to the H16/D45 catalytic dyad found in KE70 and its evolved variants. (d) Designed variant KE703b shares 6 of the 9 mutation sites found by directed evolution but has only three mutations in common with evolved variant R4 8/5A (S20a, Y48F, A204V).

4.4.3. Crystal structures

To evaluate whether the catalytic efficiency of our designs was increased via the predicted molecular mechanisms (Figure 4.1b), we solved room-temperature (280 K) X-ray crystal structures of four of the most active designs (HG185, HG198, HG630 and HG649) in the presence and absence of the 6NT transition-state analogue. We used room-temperature X-ray crystallography for direct comparison with the HG3 non-cryogenic structure used to generate the ensemble and because this method can reveal conformational heterogeneity in protein structures

that would not be visible at cryogenic temperatures, providing insights into the conformational ensemble that is sampled by a protein^{21,24}. Crystals of HG185, HG630, and HG649 were obtained in the absence of 6NT, and soaked into solutions containing this molecule prior to X-ray data collection to obtain the analogue-bound structures (Supplementary Table 4.4). These three enzyme variants all crystallize in space group $P 2_1$ with two molecules in the asymmetric unit, roughly matching the crystal form reported for HG3¹⁴. On the other hand, HG198 could only be crystallized in the presence of 6NT, and its unbound structure was derived from serially soaking these crystals in solutions lacking the transition-state analogue prior to data collection. In the presence of 6NT, HG198 crystallizes in the low symmetry space group $P 1$ with three molecules in the asymmetric unit, but removing the transition-state analogue via serial soaking transforms the crystals to space group $P 2_1$ with a single molecule in the asymmetric unit. This unusual transformation of the crystal system was consistent over multiple crystal specimens. All crystals diffracted at resolutions of 1.32–1.90 Å, which enabled us to create high-quality structural models (Supplementary Table 4.5).

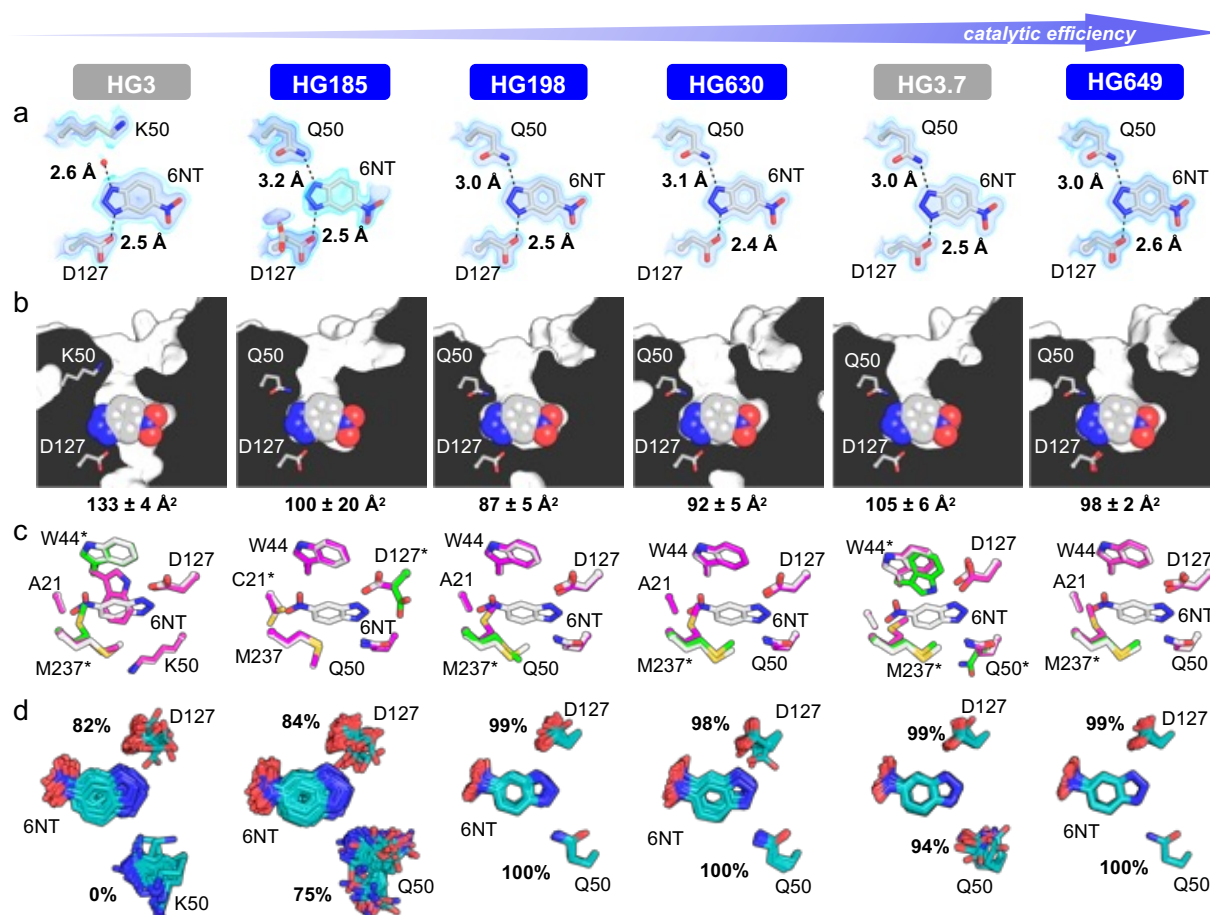


Figure 4.3. Crystal structures of HG-series Kemp eliminases. In all cases, only atoms from chain A are shown. Structures of HG3 (PDB ID: 5RG4 and 5RGA) and its evolved variant HG3.7 (PDB ID: 5RG6 and 5RGC) were previously published¹⁴. (a) Hydrogen bonds to the 6-benzotriazole (6NT) transition-state analogue are shown as dashed lines and distances are indicated. The red sphere represents a water molecule. The 2Fo-Fc map is shown in volume representation at two contour levels: 0.5 eÅ⁻³ and 1.5 eÅ⁻³ in light and dark blue, respectively. In the case of HG185, occupancy of 6NT in the crystal is approximately 60%, which is reflected by its lower electron density. (b) Cut-away view shows that active sites of designs are more complementary to the transition-state analogue than that of the HG3 template, as indicated by lower solvent-accessible surface areas for 6NT (mean ± s.d. across all chains). 6NT is shown as spheres. (c) Superposition of the 6NT-bound structure (white) with the highest (magenta) and lowest (green) occupancy conformers of the unbound structure for each Kemp eliminase. Active-site residues that are not preorganized are indicated by an asterisk. In the starting template HG3 and evolved variant HG3.7, the unbound state is never preorganized for catalysis as both W44 and M237 in the major and minor conformer, respectively, adopt side-chain rotamers that would prevent productive binding of the transition state. In the HG185 design, C21 is never preorganized in the unbound state, and the catalytic base D127 is only preorganized in the major conformer (80% occupancy). In the HG198, HG630 and HG649 designs, only M237 adopts a non-productive conformation in the unbound state, with an occupancy ranging from 52 to 59%. (d) Ensemble refinement of 6NT-bound structures show that designed variants HG198, HG630 and HG649, as well as evolved variant HG3.7, display lower structural heterogeneity than HG3 and a higher percentage of ensemble members whose catalytic side chains (D127 and Q50) make hydrogen bonds with 6NT. In the case of HG185, the lower occupancy of 6NT in the crystal (approximately 60%) likely contributes to its comparable heterogeneity to HG3 and lower percentage of ensemble members making catalytic contacts with 6NT.

All designs bound 6NT in the same catalytically productive pose (Figure 4.3) as that observed in HG3, in excellent agreement with the computational models generated by ensemble-based design (Figure 4.4) but in stark contrast with those generated using the average crystal structure (Supplementary Figure 4.5). In this pose, the acidic N–H bond of 6NT that mimics the cleavable C–H bond of the substrate is located within hydrogen-bonding distance to the carboxylate oxygen of D127, while the basic nitrogen atom corresponding to the phenolic oxygen of the transition state forms a hydrogen bond with the N_ε atom of Q50, similar to what was observed for evolved variant HG3.7¹⁴. By contrast, HG3 does not form this second hydrogen bond because its K50 residue does not adopt a suitable side-chain conformation and instead it is a water molecule that hydrogen-bonds to the basic nitrogen atom of 6NT. In all cases, angles defining the hydrogen bond between D127 and the transition-state analogue were improved or within error compared to HG3 (Supplementary Table 4.3). In contrast to all other enzymes, the 6NT transition-state analogue of HG185 refined to an occupancy of approximately 60%, indicating that its crystal comprises a mixture of enzyme molecules in their bound and unbound forms. Given that the soaking procedure we used to generate crystals of HG185 bound to 6NT was identical to the one used to soak isomorphous crystals of HG630 and HG649 (Supplementary Table 4.4), which yielded 6NT bound at 100% occupancy, it is likely that the lower occupancy of this ligand in HG185 reflects its weaker binding in the active site, in agreement with the enzyme's 6.5–14-fold lower catalytic efficiency compared to HG630 and HG649 (Table 4.1). The lower occupancy of 6NT in the HG185 structure is reflected by weaker electron density in the active site and the presence of an alternate D127 rotamer that does not hydrogen-bond with the transition-state analogue (Figure 4.3a). Regardless, the active site of all designed variants is more complementary

to the transition state than that of HG3, as indicated by the lower solvent-accessible surface area for 6NT when bound in the active site of these variants (Figure 4.3b).

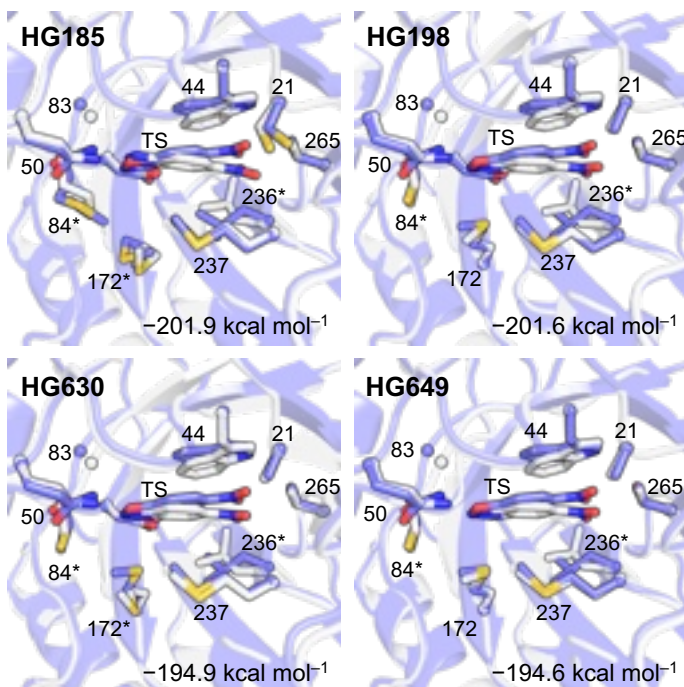


Figure 4.4. Ensemble-based design accurately predicts crystal structures of designed enzymes. Crystal structures of HG-series Kemp eliminases with bound 6NT (white) are overlaid on the corresponding design models obtained using ensemble-based design (blue). Computed energies of design models after repacking are indicated at bottom right. For reference, calculated energies of the HG185, HG198, HG630 and HG649 crystal structures with bound transition-state analogue 6NT are $-230.9 \text{ kcal mol}^{-1}$, $-200.8 \text{ kcal mol}^{-1}$, $-210.8 \text{ kcal mol}^{-1}$, and $-234.0 \text{ kcal mol}^{-1}$, respectively. In all cases, the transition state and transition-state analogue are shown at the center of the image. Side chains of all residues forming the binding pocket are shown with the exception of P45, which was omitted for clarity. The sphere shows the alpha carbon of G83. Asterisks indicate residues that adopt side-chain rotamers varying by >20 degrees around one or more side-chain dihedrals between the design model and crystal structure.

Next, we evaluated active-site preorganization by comparing structures in the presence and absence of bound 6NT. In HG3, the unbound form is never preorganized for catalysis as both W44 and M237 adopt conformations that would prevent productive binding of 6NT in the major or minor conformations, respectively (Figure 4.3c). This is also the case for evolved variant HG3.7¹⁴. HG185 is also never preorganized in the unbound form because C21 exclusively adopts a rotamer that would clash with 6NT. Furthermore, the D127 catalytic residue of HG185 adopts a low-

occupancy (7% and 14% for chains A and B, respectively), catalytically non-productive conformation in the unbound form that cannot interact favorably with 6NT, likely contributing to its inferior activity compared with the other variants. In contrast, active sites of higher activity variants HG198, HG630 and HG649 are correctly preorganized for catalysis in a substantial portion of the molecules in the crystal, with only M237 adopting a non-productive conformation at 47–64% occupancy (Figure 4.3c). Overall, crystal structures confirmed that activity of designed HG variants was increased by several of the mechanisms known to enhance enzyme catalysis and that we aimed to reproduce using our filtering criteria.

4.4.4. Enrichment of catalytically competent sub-states

The ability of an enzyme to frequently sample productive substates in its conformational ensemble is another important feature of efficient catalysis^{15,17,25,26}. To evaluate this property for our designs, we performed ensemble refinement of their 6NT-bound structures and evaluated the proportion of ensemble members that make the designed catalytic contacts with the transition-state analogue. We observed that in HG3, the hydrogen bond between the catalytic base and 6NT is formed in 84% of ensemble members, and no interaction is observed between the ligand and K50 (Figure 4.3d). In HG185, interactions between 6NT and catalytic residues D127 and Q50 are present in 84% and 75% of ensemble members, respectively, likely due to the lower occupancy of 6NT. In the case of the most active variants HG198, HG630 and HG649, catalytic contacts to both D127 and Q50 are present in $\geq 98\%$ of ensemble members, and their ensembles are less structurally diverse than those of HG3 or HG185 (Figure 4.3d, Figure 4.5a, Supplementary Table 4.6). The higher percentage of hydrogen bonds between 6NT and catalytic residues and narrower

conformational ensembles of high activity HG variants confirm a higher population of catalytically-competent substates than in the ensemble of HG3.

Having established that the conformational ensembles of efficient HG variants were enriched in catalytically competent substates, we investigated whether the bb05 template that we extracted from the HG3 ensemble and used as design template is a better approximation of the transition-state ensemble than the HG3 crystal structure obtained by traditional refinement of a single, major conformation. To do so, we calculated the pairwise backbone root-mean-square deviations of designed active-site residues (Supplementary Table 4.8) between every member of the HG variants' ensembles and bb05. Distributions (Figure 4.5b) and median values (Supplementary Table 4.6) of root-mean-square deviations show that the bb05 active-site backbone is as or more structurally similar to the ensembles of designed HG variants than to its own ensemble. By contrast, the HG3 crystal structure with bound 6NT is more similar to its own ensemble than to those of higher activity HG variants (Figure 4.5c and Supplementary Table 4.6). These results strongly support our hypothesis that ensemble refinement performed on the crystal structure of a low activity *de novo* enzyme allows sampling of catalytically-productive conformational substates for more accurate accommodation of the transition state than the single crystal structure.

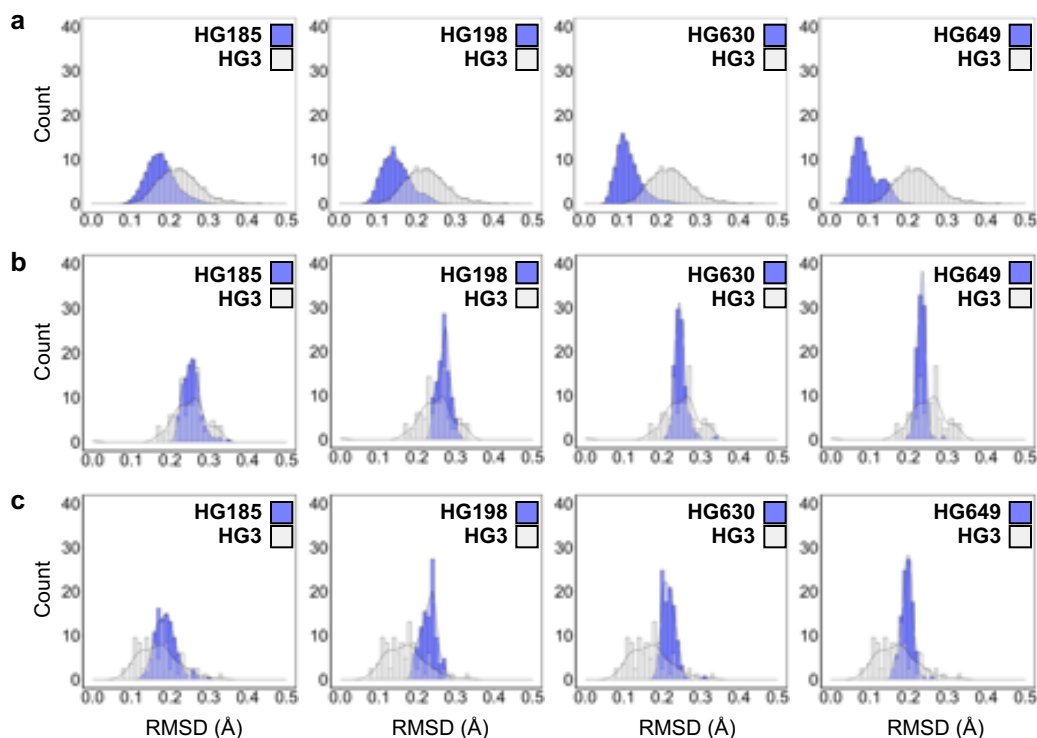


Figure 4.5. Enrichment of catalytically-competent sub-states in the conformational ensemble. Distributions of pairwise backbone root-mean-square deviations (RMSD) for designed active site residues (listed in Supplementary Table 4.8) calculated between each backbone template obtained by ensemble refinement of various HG-series Kemp eliminases with bound 6NT and (a) the HG3_05 template used for computational design, (b) the HG3 crystal structure with bound 6NT (PDB ID: 5RGA) or (c) each template within the same ensemble. Distributions show that (a) the HG3_05 template is equally or more similar to the ensembles of high-activity HG variants (blue) than to its own ensemble (white), whereas (b) the HG3 crystal structure is more similar to its own ensemble (white) than to those from higher activity HG enzymes (blue). (c) Diversity of ensembles (defined as the backbone RMSD between every pair of ensemble members) show that ensemble members are more similar in the ensembles of higher activity variants than of HG3.

4.4.5. KE70 designs

To test the general applicability of our ensemble-based enzyme design method, we applied it to another Kemp eliminase that was designed on a distinct protein scaffold. This enzyme, KE70, contains a His/Asp dyad that acts as catalytic base and a Ser hydrogen-bond donor to stabilize the phenolic oxygen of the transition state³. To allow direct comparison to our HG variants, we designed variants using a theozyme comprising an Asp residue as catalytic base and a Gln as hydrogen-bond donor. Starting from the available unbound cryogenic crystal structure of KE70 (PDB ID: 3NPV¹⁰), we generated an 84-member ensemble (Supplementary Table 4.1) and used it

to design variants predicted to efficiently catalyze the Kemp elimination using a similar protocol as that used for HG variants but without the final filtering step involving molecular dynamics (Methods). Three variants were selected for kinetic analysis, and these contain between 11 and 13 mutations (Supplementary Table 4.7), including the D16 and Q138 catalytic motif as well as two insertions in active site loops (following residues 19 and 238) that we introduced after design to allow for comparison with evolved KE70 variants that contain these insertions¹⁰. Two of the three tested designs were active (Table 4.1, Supplementary Figure 4.6). The most active variant, KE703, displays a k_{cat}/K_M of $2500 \pm 300 \text{ M}^{-1} \text{ s}^{-1}$, which is 44-fold higher than KE70 and approximately 17-fold higher than that of a KE70 variant containing the two loop insertions (Table 4.1). To compare with evolved KE70 variants that contain the His/Asp dyad as catalytic base, we introduced the D16H and N45D mutations into KE703, generating the KE703b variant. This variant has a k_{cat}/K_M of $5500 \pm 100 \text{ M}^{-1} \text{ s}^{-1}$, which is approximately 100-fold higher than KE70 and two-fold higher than that of the evolved variant R4 8/5A that was isolated after 4 rounds of directed evolution and screening of approximately 4000 clones¹⁰ (Figure 4.2c). KE703b shares a majority of its mutation sites with R4 8/5A but its specific amino-acid substitutions are distinct, with only three mutations in common (Figure 4.2d). These results confirm the general applicability of ensemble-based design and its ability to generate artificial enzymes that are orders of magnitude more active than the original *de novo* enzymes at a fraction of the experimental effort required by directed evolution. Furthermore, our results show that ensemble-based design does not require an ensemble generated from a crystal structure with a bound transition-state analogue to be successful, which is advantageous given that such structures are less common than unbound ones.

4.5. Discussion

In this work, we used ensemble-based design to increase the catalytic efficiency of the HG3 and KE70 *de novo* Kemp eliminases by two orders of magnitude, yielding artificial enzymes with $k_{\text{cat}}/K_{\text{M}}$ values that are within the 10^3 – 10^6 $\text{M}^{-1} \text{s}^{-1}$ range encompassing a majority of natural enzymes²⁷, and only 4- or 10-fold lower than those of the most active variants evolved from these enzyme prototypes (e.g., $k_{\text{cat}}/K_{\text{M}}$ of 126,000 $\text{M}^{-1} \text{s}^{-1}$ and 53,100 $\text{M}^{-1} \text{s}^{-1}$ for HG3.17¹⁴ and KE70 R8 12/12B¹⁰, respectively). To achieve these results, we tested a number of variants that is several orders of magnitude lower than those previously used to achieve comparable improvements by directed evolution^{8,10}. Further increases to catalytic efficiency could be achieved by incorporating beneficial mutations outside the active site, which have been shown to augment catalysis via alterations to allosteric networks²⁸. Although more challenging to predict a priori, mutational hotspots distal from the active site can be identified using nuclear magnetic resonance spectroscopy²⁹ or by applying correlation-based tools to molecular dynamics trajectories³⁰. In theory, it should be possible to create artificial enzymes for any desired reaction with catalytic efficiencies comparable to that of the average natural enzyme²⁷ using purely rational approaches, and the ensemble-based design procedure described here can expedite this process.

The success of our ensemble-based design relied on exploiting the spatiotemporally-averaged information contained in X-ray crystallographic data to overcome the backbone sampling challenge that is responsible for much inaccuracy in enzyme design⁸. While others have previously attempted to address the backbone sampling issue by using ensembles generated via Monte-Carlo-based backrub sampling and incorporating loop remodelling during the enzyme design process, activity improvements were modest (<5-fold)¹⁰. These results, in combination with our previous observation that use of an ensemble generated by unconstrained molecular dynamics does not

allow accurate recapitulation of the active-site configuration and catalytic contacts for an efficient artificial enzyme¹⁴, suggest that to obtain large activity increases, it is not sufficient to introduce backbone flexibility in the design process. Instead, the correct backbone conformation must be sampled. Crystallographic ensemble refinement of the initial *de novo* enzyme satisfies this stringent requirement in that the resulting model represents the true conformational ensemble sampled by the enzyme, which should include productive substates that are better suited for stabilization of the transition state than the average crystal structure. Generating the ensemble from a structure with a bound transition-state analogue likely helps to enrich such productive substates, and could partially explain why the KE70 ensemble, which was derived from an unbound structure, led to a lower overall activity improvement. Although the use of an ensemble derived from enzymes in their bound and unbound forms led to activity improvements, one important disadvantage of our approach is that it requires a crystal structure of the original *de novo* enzyme to generate the ensemble. Since *de novo* enzymes are typically generated from small, monomeric protein scaffolds whose crystal structures have been determined to high resolution, we do not anticipate that structure determination will create a substantial bottleneck in most cases.

Our results support the idea that to design efficient artificial enzymes, it is essential to consider residual conformational heterogeneity present in the designs, and whether it leads to the population of catalytically unproductive conformations. In other words, dynamic aspects of enzyme structure, such as active-site preorganization and enrichment of catalytically-competent substates, must also be considered. This is supported by the fact that the most active designs described here, HG630 and HG649, were ranked lower than less active ones such as HG44 and HG57 based on their predicted ability to stabilize the transition state (as reflected by computed energies, Supplementary Table 4.2), which has hitherto been the most important metric used to

pick designs for experimental testing. Recently, researchers have proposed that enzyme design should be viewed as a population shift problem, where relative stabilities of conformational substates are redistributed by the introduction of mutations to stabilize productive substates^{15,28}. Previous work on the HG series of Kemp eliminases by our team¹⁴ and others¹⁵ demonstrated that directed evolution improves activity through essentially the same mechanism, with higher activity variants having reduced flexibility and greater active-site preorganization, both of which favor the population of catalytically active substates. The similarity is further highlighted by the fact that our ensemble-based design approach and directed evolution identify activity enhancing mutations at many of the same positions.

Moving forward, it should be possible to further enhance the population of catalytically-competent conformations for a designed enzyme ensemble using the information obtained from ensemble refinement against X-ray diffraction data. Specifically, we anticipate using these ensembles to identify sequences that can stabilize productive substates while simultaneously destabilizing non-productive ones in a multistate design approach³¹. The ability to remodel an enzyme conformational landscape³² in this way, using negative design strategies to widen the energy gap between productive and non-productive conformations, may open the door to the design of more efficient artificial enzymes than previously possible. Beyond the scope of enzyme design, we expect that a similar strategy, which takes advantage of crystallographically-derived ensemble models, could find utility in a broad range of protein design applications. For example, directed evolution following computational design of protein-protein interactions has been shown to enhance affinity by eliminating conformations from the ensemble that are non-productive for binding²⁴. Additionally, conformational heterogeneity has been shown to be detrimental to the

functions of computationally designed fluorescent proteins³³ and ligand-binding proteins³⁴. Each of these observations suggests a potential application of our proposed approach.

4.6. Methods

Ensemble generation. Ensembles of backbone templates were generated using the ensemble refinement protocol¹⁸ implemented in the PHENIX macromolecular structure determination software (v1.18.2-3874)³⁵. In ensemble refinement, 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_{ath} 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).

Computational enzyme design. 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. Individual backbone templates from each structural ensemble 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 sidechains if necessary, and rotated hydroxyl, sulfhydryl, methyl, and amino groups in Ser/Thr, Cys, Met, and Lys, respectively.

The Kemp elimination transition-state (TS) was built using parameters described by Privett and coworkers⁷. For HG3 designs, D127 was used as the catalytic base, and Gln/Asn/Thr/ Ser were tested as hydrogen-bond donors at various positions (Supplementary Table 4.8). For KE70 designs, D16 served as the catalytic base, Q137 as hydrogen-bond donor, and F47 was utilized for introduction of a pi-stacking interaction with the TS (Supplementary Table 4.9). For both HG3 and KE70 designs, all other positions in the active site were mutated to Gly during theozyme placement to avoid steric clashes with designed catalytic residues (Supplementary Table 4.8 and Supplementary Table 4.9). The 2002 Dunbrack backbone-independent rotamer library³⁶ 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 placement¹ using the contact geometries listed in Supplementary Table 10. TS pose energies were calculated using a modified version of the Phoenix protein design energy function³⁷ consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field³⁸ 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 Å³⁹, and an electrostatic energy term modeled using Coulomb's law with a distance-dependent dielectric of 10. During the energy calculation step, TS–side-chain interaction energies were biased to favor interactions that satisfy contact geometries (Supplementary Table 4.11) by applying an energy benefit of –100 kcal mol⁻¹ as described by Lassila *et al.*¹. A Triad patch file for theozyme placement is provided as a Supplementary File.

Next, active-site repacking calculations were performed on the theozyme-containing templates. In these calculations, the TS was translated ± 0.4 Å in each Cartesian coordinate in 0.2-Å increments and rotated 10° about all three axes (origin at TS geometric center) 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. Residues that were converted to Gly in the theozyme placement step were allowed to sample rotamers of selected amino acids (Supplementary Table 4.8 and Supplementary Table 4.9) to form a mostly hydrophobic active site. Side-chain–TS interaction energies were biased to favor contacts satisfying geometries as done during the theozyme placement step (Supplementary Table 4.11). Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. A Triad patch file for active-site repacking is provided as a Supplementary File. After active-site repacking, we calculated energy differences between each repacked structure and the corresponding structure in the absence of the TS where all designed residues (including catalytic residues) were converted to Gly. This energy subtraction was done to allow comparison of transition-state binding energies for sequences across multiple backbone templates. We also measured angles defining the hydrogen-bond contact between the catalytic base and the TS of the top-scoring sequence obtained from each backbone template. Sequences from the HG3 and KE70 backbone template that gave the lowest energy top-ranked sequence with angles between its catalytic base and the transition state within 3% or 3.5% error of the optimal geometries calculated for acetamide dimers²³ (Supplementary Figure 4.2 and Supplementary Table 4.3), respectively, were selected for library design and cleaning, as described below.

Computational library design. Library design was performed with the CLEARSS algorithm⁴⁰ using the energies obtained after active-site repacking on the single template selected as described above. For a specific library size configuration, which is the specific number of amino acids at each position in the protein, the highest probability set of amino acids at each position (based on the sum of their Boltzmann weights) were included in the library. A combinatorial library of approximately 1000 sequences derived from the selected HG3 or KE70 ensemble member was generated. To ensure that the lowest energy rotameric configuration was obtained for each library sequence, we used the *cleanSequences.py* app in Triad to “clean” their structures by optimizing rotamers on the input backbone template using the active-site repacking protocol described above but without modification of catalytic residue rotameric configuration. To compare transition-state binding energies of “cleaned” sequences, we calculated the energy difference between each optimized structure and the corresponding structure in the absence of the TS where all designed residues (including catalytic residues) were converted to Gly. These energies are reported throughout the figures and text.

Design filtering. To decrease the number of sequences to test experimentally, “cleaned” structures were filtered using geometrical criteria. We kept sequences with (i) a TS solvent-accessible surface area lower than 100 Å² or 102 Å² for HG3 and KE70 designs, respectively, and (ii) ≥70 % of all designed active-site residues predicted to be preorganized, which was determined as those that adopted the same rotamer in the absence and presence of TS. Furthermore, sequences that contained ≥5 Met residues or ≥2 Cys residues were discarded to reduce conformational heterogeneity of the active site and prevent formation of disulfide bonds, respectively. A final

filtering step consisting of short molecular dynamics simulations was performed for HG3 designs, as described below.

Unconstrained molecular dynamics simulation for filtering of HG3 designs. All simulations were performed using the Amber 2020 software (<http://ambermd.org/>) with the AMBER19SB forcefield⁴¹. Long-range electrostatics (>10 Å) were modeled using the particle mesh Ewald method⁴², and a time step of 2 fs was used for the production phase. Parameters for the 5-nitrobenzisoxazole substrate were generated using the Antechamber⁴³ package. “Cleaned” computational models generated by Triad for the 55 sequences that passed the filtering procedure described above were prepared for molecular dynamics. The TS structure was converted to that of the substrate, hydrogen atoms were added using Reduce⁴⁴, and each substrate-bound protein molecule was placed in a dodecahedral box with periodic boundary conditions where the distance between the protein surface and the box edges were set to 10 Å. After explicit TIP3P water molecules⁴⁵ were added, charges on protein atoms were neutralized with Na⁺ and Cl⁻ counter-ions at a concentration of 0.15 M. The structures were then energy minimized with the steepest descent method to a target maximum force of $1000 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. Before equilibration, the system was heated to a target temperature of 300 K for 100 ps. The system was then equilibrated under an NVT ensemble for 100 ps at a temperature of 300 K using a Nose-Hoover thermostat⁴⁶, while applying position restraints for heavy protein atoms. A second equilibration step under an NPT ensemble was performed for 100 ps with constant pressure and temperature of 1 bar and 300 K, respectively, using the Berendsen barostat⁴⁷. Following removal of position restraints, 20-ns production runs under Parrinello-Rahman pressure coupling⁴⁸ were initiated from the final snapshot of the NPT equilibration. After molecular dynamics, designs were evaluated and ranked

according to their catalytic competency, which is the percentage of snapshots in which catalytic contacts between substrate and catalytic side chains were maintained, as calculated using distance and angle metrics listed on Supplementary Figure 3.

Protein expression and purification. Codon-optimized (*E. coli*) and his-tagged (C-terminus) genes for Kemp eliminases (Supplementary Tables 12–13) 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, 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 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 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-3K centrifugal filter units (EMD Millipore) and quantified by measuring the absorbance of the denatured protein in 6 M guanidinium chloride at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPASy ProtParam tool (<https://web.expasy.org/protparam>).

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-μL reactions with varying concentrations of freshly prepared 5-nitrobenzisoazole (Aablocks) 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 0.02–5.0 μM or 0.03–16.0 μM for HG or KE70 variants, respectively. Product formation was monitored spectrophotometrically at 380 nm ($\epsilon = 15,800 \text{ M}^{-1} \text{ cm}^{-1}$) in individual 96-well plates (Greiner Bio-One) wells using a Biotek Synergy H1 multimode plate reader (Agilent). 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 (2 mM), 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.

Crystallization. Enzyme variants were prepared in 50 mM sodium citrate buffer (pH 5.5) at the concentrations listed in Supplementary Table 4. For samples that were co-crystallized with the

transition-state analogue, a 100 mM stock solution of 6-nitrobenzotriazole (6NT, AstaTech) was prepared in dimethyl sulfoxide (DMSO) and diluted 20-fold in the enzyme solutions for a final concentration of 5 mM (5% DMSO). For each enzyme variant, we carried out initial crystallization trials in 15-well hanging drop format using EasyXtal crystallization plates (Qiagen) and a crystallization screen that was designed to explore the chemical space around the crystallization conditions reported by Broom *et al.*¹⁴ 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. This procedure was successful for HG185, HG198, and HG630; however, crystallization of HG649 required identification of a new crystallization condition using sparse matrix screening. We used commercial screening kits (NeXtal) to prepare crystallization drops in 96-well sitting drop plates 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 100 μ L of the mother liquor solution. After identifying a suitable mother liquor solution from screening experiments, large crystals suitable for X-ray diffraction experiments were obtained in 15-well hanging-drop format as described for the other HG variants. The specific growth conditions that yielded the crystals of each HG variant used for X-ray data collection are provided in Supplementary Table 4. HG185, HG630, and HG649 crystallized in the absence of the 6NT transition-state analogue, and the resulting crystals were isomorphous with the HG3 crystal form reported in the PDB (PDB ID: 5RGA and 5RG4). For these variants, crystals with bound 6NT were obtained by soaking the unbound crystals in mother liquor containing 10 mM 6NT for at least 30 minutes prior to data collection. On the other hand, HG198 could not be crystallized without 6NT present, and the crystals that were obtained belonged to space group P1 with three polypeptides in the asymmetric unit. To prepare crystals of HG198 without 6NT bound,

we serially soaked 6NT-bound crystals in three drops of crystallization mother liquor without 6NT. Each sequential soaking step was carried out for at least 30 minutes. After removal of 6NT by soaking, the P1 crystals undergo a transition to space group $P2_1$, with a smaller unit cell and only one molecule per asymmetric unit. Although we could not determine the cause of this unusual change in crystal symmetry upon soaking, we observed that it occurred with high reproducibility. We measured 10 individual crystals of 6NT-bound HG198, which all belonged to space group P1 and were roughly isomorphous. Additionally, we measured four individual crystals of HG198 without 6NT, and found that all of them demonstrated the transition to monoclinic symmetry.

X-ray data collection and processing. Prior to X-ray data collection, crystals were mounted in polyimide loops and sealed using a MicroRT tubing kit (MiTeGen). 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. Crystals were maintained at 280 K throughout the course of data collection. Each data set was collected using a total X-ray dose of 50 kGy or less, and covered a 180° wedge of reciprocal space. Multiple data sets were collected for each enzyme variant either from different crystals, or if their size permitted, from non-overlapping volumes of larger crystals, in order to identify optimally diffracting samples. X-ray data were processed with the Xia2 0.5.492 program (<https://doi.org/10.1107/S0021889809045701>)⁴⁹, which performed indexing, integration, and scaling with DIALS⁵⁰, followed by merging with Pointless⁵¹. For each data set, a resolution cutoff was taken where the CC1/2 and $\langle I/\sigma I \rangle$ values for the merged intensities fell to approximately 0.5 and 1.0, respectively. Information regarding data collection and processing is presented in

Supplementary Table 5. The reduced diffraction data were analyzed with phenix.xtriage to check for crystal pathologies, and no complications were identified.

Structure determination. We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser⁵², as implemented in v1.13.2998 of the PHENIX suite⁵³. Several different models of HG variants, including crystal structures and design models, were used for molecular replacement. For all data sets except that of HG630 (+) 6NT, we performed iterative steps of manual model rebuilding followed by refinement of atomic positions, atomic displacement parameters, and occupancies using a translation-libration-screw (TLS) model, a riding hydrogen model, and automatic weight optimization. For HG630 (+) 6NT, refinement was done as described above but without usage of a translation-libration-screw (TLS) model. All model building was performed using Coot 0.8.9.236⁵⁴ and refinement steps were performed with phenix.refine within the PHENIX suite (v1.13-2998)^{53,55}. Restraints for 6NT were generated using phenix.elbow⁵⁶, starting from coordinates available in the Protein Data Bank (PDB ligand ID: 6NT). Further information regarding model building and refinement, as well as PDB accession codes for the final models, are presented in Supplementary Table 5.

4.7. Supplementary Information

Supplementary Table 4.1. Ensemble refinement results

Structure	PDB ID	Resolution (Å)	# chains	$R_{\text{work}}/R_{\text{free}}$ before ensemble refinement	$R_{\text{work}}/R_{\text{free}}$ after ensemble refinement	Ensemble size
HG3 (+) 6NT	5RGA	1.86	2	0.1482/0.1863	0.1332/0.1814	84
HG185 (+) 6NT	TBD	1.55	2	0.1610/0.1831	0.1387/0.1740	100
HG198 (+) 6NT	TBD	1.60	3	0.1540/0.1900	0.1402/0.1890	50
HG630 (+) 6NT	TBD	1.32	2	0.1190/0.1476	0.1316/0.1589	127
HG649 (+) 6NT	TBD	1.45	2	0.1373/0.1567	0.1312/0.1573	127
KE70 (–) 6NT	3NPV	1.48	2	0.1430/0.1700	0.1290/0.1549	84

^a TBD = to be determined upon structure deposition.

Supplementary Table 4.2. Design filtering – HG series

Design	Mutations from HG3	Energy (kcal mol ⁻¹)	SASA (Å ²) ^a	# Cys	# Met	# Preorganized residues (on 29) ^b	Residues not preorganized
HG44	A21C K50Q Q90F A125V	-206.8	86.2	1	4	23	Cys21, Trp44, Gln50, Val125, Leu236, Asn239
HG57	A21C K50Q Q90L A125V	-205.9	86.2	1	4	21	Cys21, Trp44, Gln50, Val125, Leu236, Met237, Asn239, Trp275
HG185	A21C K50Q Q90F A125T W275V	-201.9	87.5	1	4	26	Cys21, Gln50, Met237
HG198	K50Q M84C A125T F267M	-201.6	99.8	1	4	23	Trp44, Gln50, Thr125, Gln207, Met237, Met267
HG259	K50Q M84C Q90L A125T F267M	-200.5	99.8	1	4	23	Trp44, Gln50, Thr125, Gln207, Met237, Met267
HG378	A21C K50Q Q90L A125V W275V	-198.7	87.5	1	4	23	Cys21, Trp44, Gln50, Val125, Leu236, Met237
HG493	K50Q Q90F A125T W275T	-197.1	87.1	0	4	24	Trp44, Gln50, Thr125, Leu236, Met237
HG603	K50Q Q90F A125V W275V	-195.5	87.1	0	4	24	Trp44, Gln50, Thr125, Leu236, Met237
HG630	K50Q M84C A125T F267M W275V	-194.9	99.8	1	4	22	Trp44, Asn47, Gln50, Thr125, Gln207, Met237, Met267
HG649	K50Q M84C Q90F A125T F267M W275V	-194.6	99.8	1	4	23	Trp44, Gln50, Thr125, Gln207, Met237, Met267

^a Solvent-accessible surface area of the transition state in the design model.^b Preorganized residues are those predicted to adopt the same rotamer in the presence and absence of transition state.

Supplementary Table 4.3. Angles defining the hydrogen bond between D127 and the transition state or transition-state analogue for various Kemp eliminase structures

Template	PDB ID	Angle (°)			Deviation from ideal (°)		
		Ψ^a	θ^a	χ^a	Ψ	θ	χ
Crystal structures ^b							
HG3	5RGA	134 ± 1	173 ± 2	182 ± 3	22	14	4.5
HG3.7	5RGC	125 ± 1	167 ± 1	174 ± 2	13	8	3.5
HG185	TBD	125 ± 3	163 ± 3	175 ± 9	13	4	2.5
HG198	TBD	124 ± 2	168 ± 1	186 ± 5	12	9	8.5
HG630	TBD	125 ± 1	170 ± 1	173 ± 5	13	11	4.5
HG649	TBD	124 ± 2	169 ± 2	181 ± 9	12	10	3.5
Ensemble members ^c							
HG3 bb05	–	113.6	164.0	175.2	1.3	4.6	2.3
KE70 bb06	–	116.2	159.3	173.0	3.9	0.1	4.5

^{a a} Average angles across all chains are reported (mean ± s.d.). Ideal values for ψ , θ and χ are 112.3, 159.4 and 177.5 degrees, respectively.

Atoms defining ψ , θ and χ are shown on Supplementary Figure 2.

^b Crystal structures contain the 6NT transition-state analogue.

^c Ensemble members contain the transition state..

Supplementary Table 4.4. Crystallization conditions

Enzyme	6NT	Method ^a	Buffer	pH	Protein (mg mL ⁻¹)	Precipitant #1	Precipitant #2
HG185	(–)	Crystallization	100 mM sodium acetate	4.6	12.8	1.6 M ammonium sulfate	–
	(+)	Soaking	100 mM sodium acetate	4.6	12.8	1.6 M ammonium sulfate	–
HG198	(–)	Soaking	100 mM sodium acetate	4.6	6.0	1.8 M ammonium sulfate	–
	(+)	Co-crystallization	100 mM sodium acetate	4.6	6.0	1.6 M ammonium sulfate	–
HG630	(–)	Crystallization	100 mM sodium acetate	4.8	10.6	1.4 M ammonium sulfate	–
	(+)	Soaking	100 mM sodium acetate	4.8	10.6	1.4 M ammonium sulfate	–
HG649	(–)	Crystallization	–	n.d.	12.5	0.3 M ammonium sulfate	30% PEG-4000
	(+)	Soaking	–	n.d.	12.5	0.3 M ammonium sulfate	30% PEG-4000

^a Crystals of ligand-bound HG185, HG630 and HG649 were obtained by soaking ligand-free crystals in 1 μ L drops of mother liquor containing 5 mM 6NT and 5% DMSO in a single step. Crystals of ligand-free HG198 were obtained by sequentially soaking crystals obtained by co-crystallization with 6NT in three new drops of 1 μ L mother liquor containing no ligand (30 min incubation per drop).

Supplementary Table 4.5. Crystallographic data and refinement statistics

	HG185	HG185	HG198	HG198	HG630	HG630	HG649	HG649
6NT	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)
PDB ID	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD
Data collection ^a								
Temperature (K)	280	280	280	280	280	280	280	280
Resolution (Å)	1.60	1.55	1.90	1.60	1.37	1.32	1.55	1.45
Space group	P2 ₁	P2 ₁	P2 ₁	P1	P2 ₁	P2 ₁	P2 ₁	P2 ₁
Cell params.								
a b c (Å)	51.12 77.44 93.75 90.00	51.11 77.49 93.85 90.00	45.68 58.72 51.30 90.00	46.65 51.03 86.30 83.58	51.12 77.51 93.88 90.00	51.02 77.51 93.90 90.00	51.13 77.47 93.78 90.00	51.07 77.47 93.81 90.00
$\alpha \beta \gamma$ (°)	105.55 90.00	105.51 90.00	95.36 90.00	81.12 87.01	105.42 90.00	105.27 90.00	105.39 90.00	105.28 90.00
Chains per asymm. unit	2	2	1	3	2	2	2	2
R_{pin}	0.071 (0.542)	0.065 (0.470)	0.169 (0.517)	0.123 (0.547)	0.062 (0.443)	0.052 (0.585)	0.070 (0.513)	0.059 (0.552)
CC_{1/2}	0.992 (0.474)	0.994 (0.531)	0.955 (0.493)	0.981 (0.466)	0.990 (0.551)	0.995 (0.477)	0.991 (0.505)	0.996 (0.452)
I/σI	10.3 (1.3)	11.1 (1.5)	6.0 (1.7)	7.7 (1.7)	8.2 (1.0)	9.4 (1.1)	7.9 (1.4)	8.7 (1.0)
Completeness (%)	96.4 (94.1)	99.7 (99.4)	99.9 (99.6)	94.3 (91.8)	99.4 (94.4)	99.7 (98.9)	99.4 (98.6)	99.7 (99.7)
Multiplicity	3.4 (3.5)	3.3 (3.3)	3.4 (3.1)	1.8 (1.8)	3.2 (2.7)	3.3 (3.2)	3.3 (3.1)	3.2 (3.1)
Wilson B-factor (Å²)	14.880	14.190	9.580	10.130	14.910	13.060	14.580	12.510
# unique reflections	89399 (4304)	101929 (5042)	21360 (1030)	97130 (4660)	147144 (6954)	164672 (8177)	101581 (4971)	124369 (6172)
Refinement								
R work/free	0.1605/ 0.1833	0.1610/ 0.1831	0.1585/ 0.1991	0.1540/ 0.1900	0.1274/ 0.1408	0.1190/ 0.1476	0.1471/ 0.1727	0.1373/ 0.1567
No. atoms								
Protein	5068	5116	2408	7489	5156	5166	5188	5260
Ligand	0	24	0	36	10	24	0	24
Water	543	362	294	820	557	557	743	748
Averaged B-factors (Å²)								
Protein	19.74	19.04	14.02	14.54	19.37	17.87	19.77	17.50
Ligand	—	32.92	—	13.70	81.47	17.95	—	17.02
Water	35.18	29.85	27.68	30.53	36.59	39.27	37.66	35.87
RMSD								
bond lengths (Å)	0.005	0.007	0.003	0.004	0.007	0.007	0.005	0.007
bond angles (°)	0.834	0.848	0.607	0.724	1.003	0.886	0.811	0.885
Molprobrity statistics								
Ramachand. outliers (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ramachand. allowed (%)	1.67	1.51	1.66	1.56	1.51	1.34	1.84	1.17
Ramachan. favored (%)	98.33	98.49	98.34	98.44	98.49	98.66	98.16	98.83
Rotamer outliers (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MolProbrity clashscore	4.15	1.67	1.67	1.73	2.52	2.32	3.57	3.03

^a Highest resolution shell is shown in parentheses.

Supplementary Table 4.6. Median pairwise backbone RMSDs for designed active-site residues

Ensemble ^a	Reference structure		
	Self ^b	bb05 ^c	HG3 _{x_{tal}} ^d
HG3	0.22	0.25	0.17
HG185	0.17	0.25	0.19
HG198	0.14	0.27	0.23
HG630	0.11	0.24	0.23
HG649	0.09	0.23	0.20

Median pairwise backbone RMSDs for designed active-site residues (listed on Supplementary Table 8) are reported in Å.

^a Ensembles were generated by ensemble refinement of 6NT-bound crystal structures (Supplementary Table 1).

^b Median pairwise backbone RMSD between every pair of backbones from the same ensemble.

^c Template from the HG3 ensemble that was used to design HG185, HG198, HG630 and HG649.

^d HG3 crystal structure bound with 6NT (PDB ID: 5RGA, Chain A).

Supplementary Table 4.7. Design filtering – KE70 series

Design	Mutations from KE70	Energy (kcal mol ⁻¹)	SASA (Å ²) ^a	# Cys	# Met	# Preorganized residues ^b	Residues not preorganized ^c
KE701	H16D S20a D45Q Y48F W72C G101S S138Q H166N V168A A204V A239a	-29.4	101.0	1	0	12	Phe47, Trp71, Gln137
KE702	H16D S20a D45Q Y48F R70K W72Y G101M S138Q H166N V168A I202V A204M A239a	-17.9	41.4	0	2	12	Asp16, Gln137, Met203
KE703	H16D S20a D45N Y48F W72V G101M S138Q H166N V168A A204V A239a	-17.8	101.6	0	1	12	Phe47, Trp71, Gln137

^a Solvent-accessible surface area of the transition state in the design model.

^b Preorganized residues are those predicted to adopt the same rotamer in the presence and absence of transition state.

^c Numbering based on crystal structure of KE70 (PDB ID: 3NPV).

Supplementary Table 4.8. Amino-acid positions optimized during computational design of HG3 variants

Design objective	Designed residues ^a	Allowed amino acids
Catalytic base	D127	D
H-bond donor	A21 W44 K50 M84 W87 Q90 A125 Y170 M172 L236 M237 T265 F267	N, Q, S, T
Transition-state packing	A21 W44 M84 W87 Q90 A125 Y170 M172 L236 M237 T265 F267	A, C, F, I, L, M, V, T, WT ^b
Active-site bottleneck ^c	W275	A, S, V, T, WT ^b
Floated positions ^d	V16 Y17 M42 E46 N47 L79 G130 Q207 H209 S234 D239 R276	WT ^b

^a Numbering based on crystal structure of HG3 (PDB ID: 5RGA).

^b WT indicates the amino acid found at that position in the HG3 sequence (e.g., N, Q, and Y at positions 47, 90, and 170, respectively)

^c The bulky W275 residue was mutated to small amino acids to widen the active-site entrance bottleneck.

^d Amino acids at these positions were allowed to change conformation but not identity.

Supplementary Table 4.9. Amino-acid positions optimized during computational design of KE70 variants

Design objective	Designed residues ^a	Allowed amino acids
Catalytic base	H16	D
H-bond donor	S137	Q
Pi-stacking	Y47	F
Removal of original catalytic dyad ^b	D44	N, Q
Transition-state packing ^c	A18 W71 G100 A102 I139 H165 V167 S169 I201 A203	A, F, I, L, M, N, V, Y
Surface residue ^d	R69	K, R

^a Numbering based on crystal structure of KE70 (PDB ID: 3NPV).

^b KE70 contains the H16/D44 catalytic dyad. The D44N and D44Q mutations were introduced to make this residue compatible with the newly designed catalytic base (D16).

^c C and S were also allowed at positions W71 and G100 instead of A.

^d This surface residue is located next to D44.

Supplementary Table 4.10. 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 4.11. 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 (1.0)	1.6 (1.6)
		Angle	CG	OD1 or OD2	H3		109 (109)	131 (131)
		Angle	OD1 or OD2	H3	C3		159 (159)	180 (180)
		Torsion	CB	CG	OD1 or OD2	H3	-21 (-21)	21 (21)
H-bond donor	Gln	Distance	1HE2 or 2HE2	O1			1.2 (1.2)	2.3 (2.3)
		Angle	NE2	1HE2 or 2HE2	O1		145 (145)	157 (157)
		Angle	1HE2 or 2HE2	O1	N2		120 (120)	140 (140)
		Torsion	1HE2 or 2HE2	O1	N2	C3	160 (160)	200 (200)
	Asn	Distance	1HD2 or 2HD2	O1			1.2 (1.2)	2.3 (2.3)
		Angle	ND2	1HD2 or 2HD2	O1		145 (145)	157 (157)
		Angle	1HD2 or 2HD2	O1	N2		120 (120)	140 (140)
		Torsion	1HD2 or 2HD2	O1	N2	C3	160 (160)	200 (200)
	Thr	Distance	HG1	O1			1.2 (1.2)	2.3 (2.3)
		Angle	OG1	HG1	O1		145 (145)	157 (157)
		Angle	HG1	O1	N2		120 (120)	140 (140)
		Torsion	HG1	O1	N2	C3	160 (160)	200 (200)
	Ser	Distance	HG	O1			1.2 (1.2)	2.3 (2.3)
		Angle	OG	HG	O1		145 (145)	157 (157)
		Angle	HG	O1	N2		120 (120)	140 (140)
		Torsion	HG	O1	N2	C3	160 (160)	200 (200)
Pi-stacking ^c	Phe	Distance between aromatic planes	CE1 CD2	C4 C9			3.5 (3.5)	4 (4)
		Angle between aromatic planes	CG CE1 CR2	C5 C8 N2			0 (120)	40 (180)

^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. Values are for theozyme placement, while those in parentheses are for active-site repacking.

^c This catalytic contact was used for KE70 designs only.

Supplementary Table 4.12. Amino-acid sequences of Kemp eliminases from HG3 series

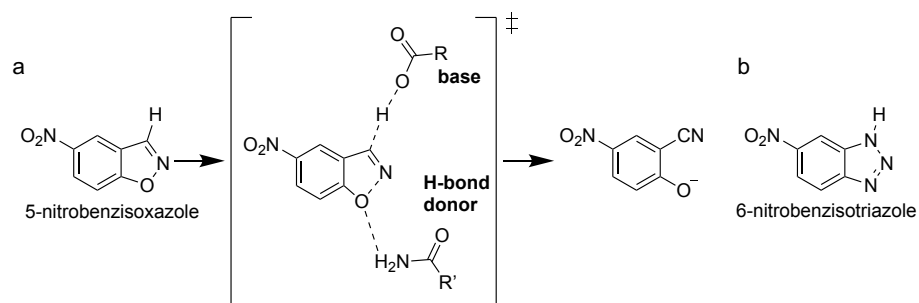
Enzyme	# Mutations from HG3	Sequences
HG44	4	MAEAAQSV D QLIKARGKVYFGVCTDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIRVWDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSWRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG57	4	MAEAAQSV D QLIKARGKVYFGVCTDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIRVWDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSWRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG185	5	MAEAAQSV D QLIKARGKVYFGVCTDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSVRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG198	4	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGCLVWHSQ L PSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVMGVADPDSWRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG259	5	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGCLVWHSQ L PSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVMGVADPDSWRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG378	5	MAEAAQSV D QLIKARGKVYFGVCTDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSQ L PSWVSSITDKNTLT N VMKNHITTLTRYKGKIRVWDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSVRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG493	4	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSTRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG603	4	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGMLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIRVWDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVFGVADPDSVRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG630	5	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGCLVWHSQ L PSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVMGVADPDSVRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH
HG649	6	MAEAAQSV D QLIKARGKVYFGVATDQ N R L TTGKNA A I I QADFGMVWPEN S MQWDATEPSQGNF N FAGADYLVNWAQ Q N GKLIGGGCLVWHSFLPSWVSSITDKNTLT N VMKNHITTLTRYKGKIR T WDVVGEAFNEDGSLRQ T VFLNVIGEDYIP IAFQTARAADPN A KLYIMDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQ T HL S AGQGAGVLQALPL L ASAGT P E VSILMLDVAGASPTDYVNVNACLN V QSCVGITVMGVADPDSVRAS T TP L LF D GNF N PKPAYNAIVQDLQ Q GSIEGRG HHHHHH

^a Mutations from HG3 are highlighted in bold and underlined. All sequences contain a His-tag at the C-terminus.

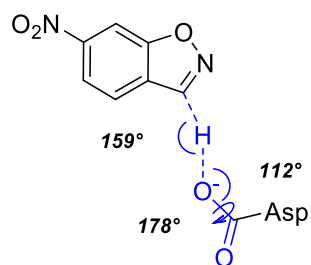
Supplementary Table 4.13. Amino-acid sequences of Kemp eliminases from KE70 series

Enzyme	# Mutations from KE70	Sequence
KE70_S20a A239a	2	MTDLKASSLRALKLMHLAT <u>S</u> ANDDDTDEKVIALCHQAKTPVGTTDAIYIYPRFIPIARKTLKEQGTPEIRI WTSTNFPHGNDIDIALAETRAAIAYGADGVAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLLSVIIE TGELKDEALIRKASEISIKAGADHIVTSTGKVAVGATPESARIMMEVIRDMGVEKTVGFIPAGGVRTAEDA QKYLAIADDELFGADWADARHYAFGAS <u>A</u> SLLASLLKALGHGDGKSASSYGSLEHHHHHH
KE701	11	MTDLKASSLRALKLM <u>D</u> LAT <u>S</u> ANDDDTDEKVIALCHQAKTPVGTT <u>Q</u> A <u>I</u> <u>F</u> IYPRFIPIARKTLKEQGTPEIRI <u>C</u> TSTNFPHGNDIDIALAETRAAIAYGAD <u>S</u> VAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLL <u>Q</u> VIIE TGELKDEALIRKASEISIKAGAD <u>N</u> IATSTGKVAVGATPESARIMMEVIRDMGVEKTVGFIP <u>V</u> GGVRTAEDA QKYLAIADDELFGADWADARHYAFGAS <u>A</u> SLLASLLKALGHGDGKSASSYGSLEHHHHHH
KE702	13	MTDLKASSLRALKLM <u>D</u> LAT <u>S</u> ANDDDTDEKVIALCHQAKTPVGTT <u>N</u> A <u>I</u> <u>F</u> IYPRFIPIARKTLKEQGTPEI <u>K</u> I <u>Y</u> TSTNFPHGNDIDIALAETRAAIAYGAD <u>M</u> VAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLL <u>Q</u> VIIE TGELKDEALIRKASEISIKAGAD <u>N</u> IATSTGKVAVGATPESARIMMEVIRDMGVEKTVGFIP <u>M</u> GGVRTAEDA QKYLAIADDELFGADWADARHYAFGAS <u>A</u> SLLASLLKALGHGDGKSASSYGSLEHHHHHH
KE703	11	MTDLKASSLRALKLM <u>D</u> LAT <u>S</u> ANDDDTDEKVIALCHQAKTPVGTT <u>N</u> A <u>I</u> <u>F</u> IYPRFIPIARKTLKEQGTPEIRI <u>V</u> TSTNFPHGNDIDIALAETRAAIAYGAD <u>M</u> VAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLL <u>Q</u> VIIE TGELKDEALIRKASEISIKAGAD <u>N</u> IATSTGKVAVGATPESARIMMEVIRDMGVEKTVGFIP <u>V</u> GGVRTAEDA QKYLAIADDELFGADWADARHYAFGAS <u>A</u> SLLASLLKALGHGDGKSASSYGSLEHHHHHH
KE703b	9	MTDLKASSLRALKLMHLAT <u>S</u> ANDDDTDEKVIALCHQAKTPVGTTDAI <u>F</u> IYPRFIPIARKTLKEQGTPEIRI <u>V</u> TSTNFPHGNDIDIALAETRAAIAYGAD <u>M</u> VAVVFPYRALMAGNEQVGFDLVKACKEACAAANVLL <u>Q</u> VIIE TGELKDEALIRKASEISIKAGAD <u>N</u> IATSTGKVAVGATPESARIMMEVIRDMGVEKTVGFIP <u>V</u> GGVRTAEDA QKYLAIADDELFGADWADARHYAFGAS <u>A</u> SLLASLLKALGHGDGKSASSYGSLEHHHHHH

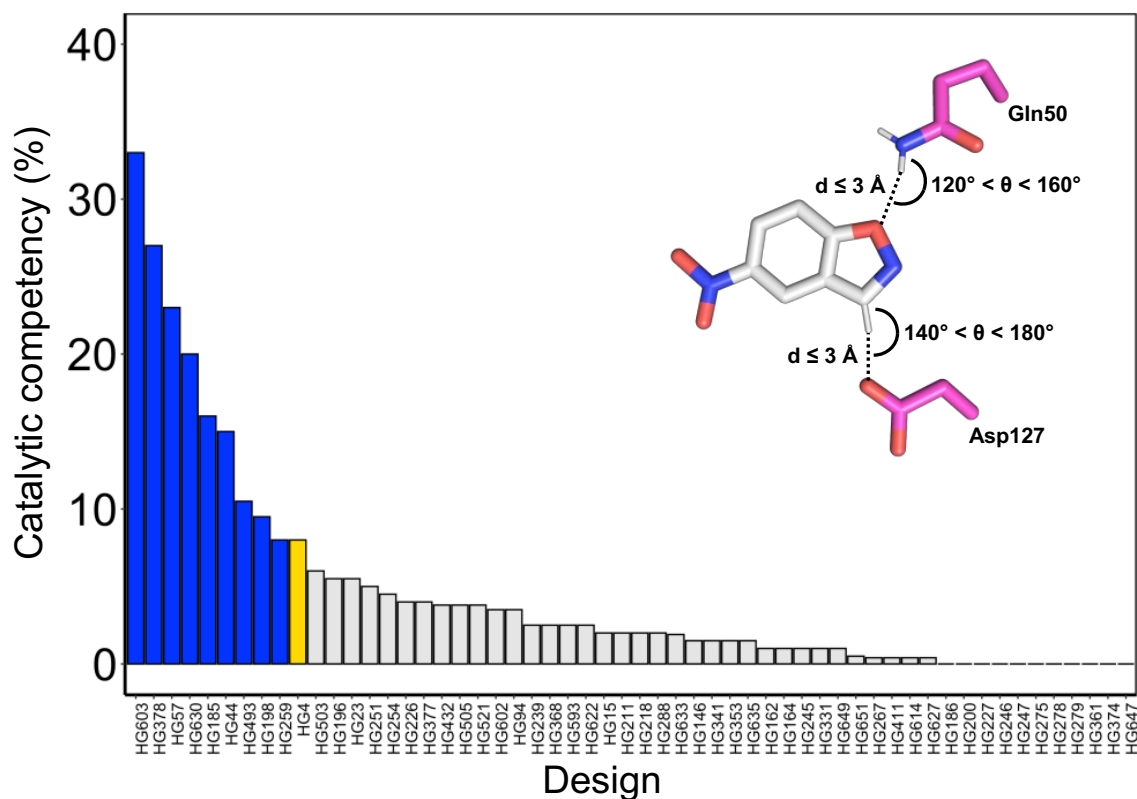
^a Mutations from KE70 are highlighted in bold and underlined. All sequences contain a His-tag at the C-terminus.



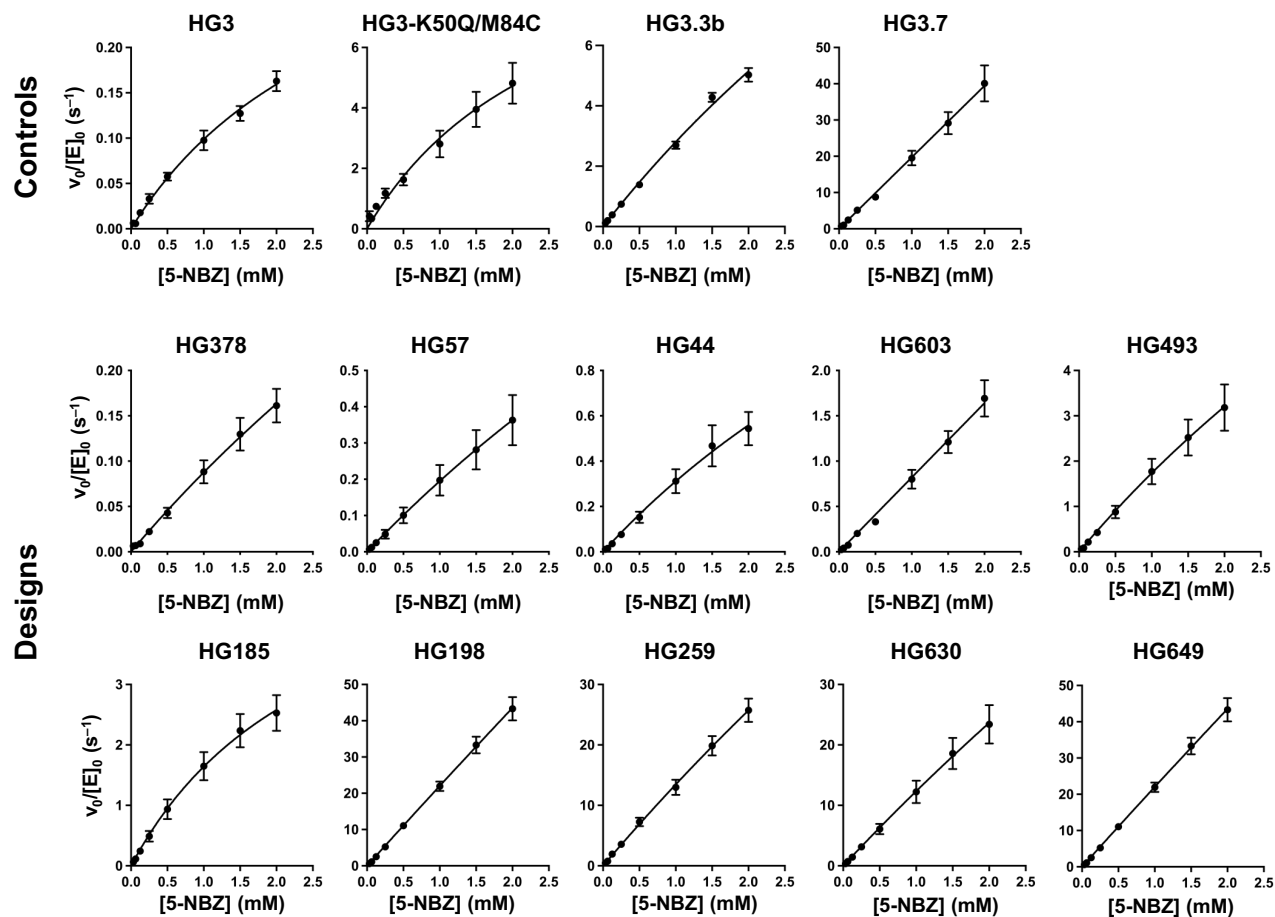
Supplementary Figure 4.1. The Kemp elimination reaction. HG-series Kemp eliminases catalyze the Kemp elimination using a catalytic motif consisting of a base (Asp) that deprotonates 5-nitrobenzisoxazole, and an H-bond donor (Gln) that stabilizes negative charge buildup on the phenolic oxygen at the transition state ($\ddagger$). This reaction yields 4-nitro-2-cyanophenol. (b) Structure of the 6NT transition-state analogue.



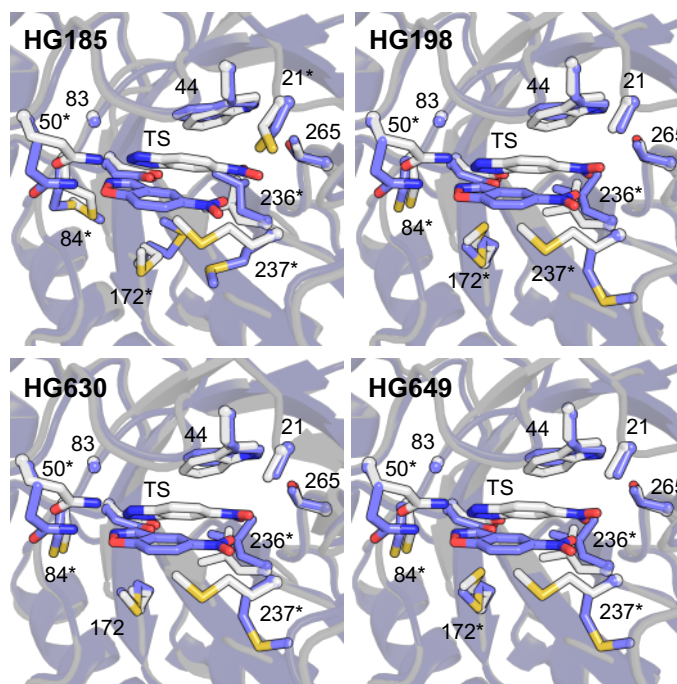
Supplementary Figure 4.2. Ideal geometry for catalytic base contact. Angles describing the hydrogen-bond interaction between the transition state and designed Asp catalytic base. Values in italics are optimal angles calculated for hydrogen bonding interactions between acetamide dimers.



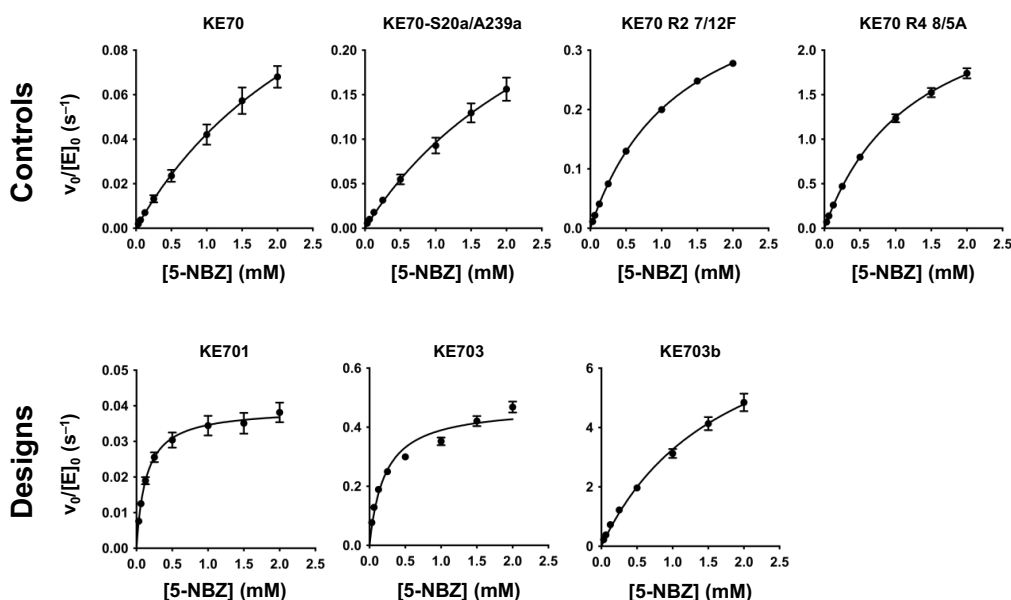
Supplementary Figure 4.3. Design filtering using molecular dynamics. Following molecular dynamics, angles (θ) and distances (d) defining catalytic contacts between the 5-nitrobenzisoxazole substrate and the designed catalytic residues (Asp127 and Gln50) were calculated on each snapshot. Snapshots whose θ and d values fall within the allowed ranges were considered to be catalytically competent. Catalytic competency is the percentage of snapshots in the trajectory that are catalytically competent. Only 9 variants (blue) have a catalytic competency that is greater than that of our positive control HG4 (yellow, PDB ID: 5RGF), an HG3 variant that is approximately 700-fold more catalytically efficient than HG3¹⁴.



Supplementary Figure 4.4. Steady-state kinetics of HG-series Kemp eliminases. Michaelis–Menten plots of normalized initial rates as a function of 5-nitrobenzisoxazole (5-NBZ) concentrations 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. Saturation was not achieved for any enzyme at the substrate’s solubility limit (2 mM). Therefore, only $k_{\text{cat}}/K_{\text{M}}$ values are reported on Table 1, and these values were calculated using linear regression of rates measured at the three or four lowest substrate concentrations.



Supplementary Figure 4.5. Design on HG3 crystal structure does not accurately predict crystal structures of designed enzymes. Crystal structures of HG-series Kemp eliminases with bound 6NT (white) are overlaid on the corresponding design models (blue) obtained using the average crystal structure (PDB ID: 5RGA) as template. In all cases, the transition state and transition-state analogue are shown at the center of the image. Side chains of all residues forming the binding pocket are shown with the exception of P45, which was omitted for clarity. The sphere shows the alpha carbon of G83. Asterisks indicate residues that adopt side-chain rotamers varying by >20 degrees around one or more side-chain dihedrals between the design model and crystal structure. In all design models generated from the average crystal structure, the catalytic hydrogen-bond donor Q50 is predicted to adopt an alternate rotamer, which causes an erroneous prediction of the transition-state binding pose in the active site.



Supplementary Figure 4.6. Steady-state kinetics of KE70-series Kemp eliminases. Michaelis–Menten plots of normalized initial rates as a function of 5-nitrobenzisoxazole (5-NBZ) concentrations 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. Saturation was not achieved for some enzymes at the substrate’s solubility limit (2 mM). In these cases, only k_{cat}/K_M values are reported on Table 1, and these values were calculated using linear regression of rates measured at the three or four lowest substrate concentrations.

References

- (1) 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 U S A* **103**,16710-16715 (2006).
- (2) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Rothlisberger, D.; Baker, D. New algorithms and an in silico benchmark for computational enzyme design *Protein Sci* **15**,2785-2794 (2006).
- (3) Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp elimination catalysts by computational enzyme design *Nature* **453**,190-195 (2008).
- (4) 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., 3rd; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De novo computational design of retro-aldol enzymes *Science* **319**,1387-1391 (2008).
- (5) 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 U S A* **109**,3790-3795 (2012).
- (6) 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).
- (7) 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).
- (8) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R.; Grutter, 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).
- (9) Yeh, A. H.; 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).
- (10) Khersonsky, O.; Rothlisberger, 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).
- (11) Khersonsky, O.; Kiss, G.; Rothlisberger, 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 U S A* **109**,10358-10363 (2012).
- (12) 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).
- (13) Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R.; 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) 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).
- (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) Otten, R.; Padua, 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).
- (17) 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).
- (18) 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 *Nature Chemistry* **13**,1017-1022 (2021).
- (19) Burnley, B. T.; Afonine, P. V.; Adams, P. D.; Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement *Elife* **1**,e00311 (2012).

- (20) Keedy, D. A.; Fraser, J. S.; van den Bedem, H. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit *PLoS Comput Biol* **11**,e1004507 (2015).
- (21) Thompson, M. C. In *Methods in Enzymology*; Academic Press: 2023.
- (22) Levin, E. J.; Kondrashov, D. A.; Wesenberg, G. E.; Phillips, G. N., Jr. Ensemble Refinement of Protein Crystal Structures: Validation and Application *Structure* **15**,1040-1052 (2007).
- (23) Keedy, D. A.; van den Bedem, H.; Sivak, D. A.; Petsko, G. A.; Ringe, D.; Wilson, M. A.; Fraser, J. S. Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR *Structure* **22**,899-910 (2014).
- (24) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations *Proc Natl Acad Sci U S A* **101**,6946-6951 (2004).
- (25) Biel, J. T.; Thompson, M. C.; Cunningham, C. N.; Corn, J. E.; Fraser, J. S. Flexibility and Design: Conformational Heterogeneity along the Evolutionary Trajectory of a Redesigned Ubiquitin *Structure* **25**,739-749 e733 (2017).
- (26) Maria-Solano, M. A.; Serrano-Hervas, E.; Romero-Rivera, A.; Iglesias-Fernandez, J.; Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function *Chem Commun (Camb)* **54**,6622-6634 (2018).
- (27) Yabukarski, F.; Doukov, T.; Pinney, M. M.; Biel, J. T.; Fraser, J. S.; Herschlag, D. Ensemble-function relationships to dissect mechanisms of enzyme catalysis *Science Advances* **8**,eabn7738 (2022).
- (28) Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design *WIREs Computational Molecular Science* **11**,e1502 (2021).
- (29) 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).
- (30) Maria-Solano, M. A.; Kinateter, T.; Iglesias-Fernandez, J.; Sterner, R.; Osuna, S. In Silico Identification and Experimental Validation of Distal Activity-Enhancing Mutations in Tryptophan Synthase *Acs Catalysis* **11**,13733-13743 (2021).
- (31) Romero-Rivera, A.; Garcia-Borras, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity *Acs Catalysis* **7**,8524-8532 (2017).
- (32) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold *Nature* **596**,583-589 (2021).
- (33) Peccati, F.; Alunno-Rufini, S.; Jiménez-Osés, G. Accurate Prediction of Enzyme Thermostabilization with Rosetta Using AlphaFold Ensembles *Journal of Chemical Information and Modeling* **63**,898-909 (2023).
- (34) Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. Estimating conformational heterogeneity of tryptophan synthase with a template-based Alphafold2 approach *Protein Science* **31**,e4426 (2022).
- (35) Davey, J. A.; Chica, R. A. Multistate approaches in computational protein design *Protein Sci* **21**,1241-1252 (2012).
- (36) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, 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 Crystallographica Section D* **75**,861-877 (2019).
- (37) Dunbrack, R. L., Jr. Rotamer libraries in the 21st century *Curr Opin Struct Biol* **12**,431-440 (2002).
- (38) 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 U S A* **107**,19838-19843 (2010).
- (39) 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 U S A* **107**,20257-20262 (2010).
- (40) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. Dreiding - a Generic Force-Field for Molecular Simulations *Journal of Physical Chemistry* **94**,8897-8909 (1990).
- (41) 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).

- (42) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field *Proteins* **78**,1950-1958 (2010).
- (43) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems *The Journal of Chemical Physics* **98**,10089-10092 (1993).
- (44) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations *Journal of Molecular Graphics and Modelling* **25**,247-260 (2006).
- (45) 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 orientation *J Mol Biol* **285**,1735-1747 (1999).
- (46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water *The Journal of Chemical Physics* **79**,926-935 (1983).
- (47) Evans, D. J.; Holian, B. L. The Nose–Hoover thermostat *The Journal of Chemical Physics* **83**,4069-4074 (1985).
- (48) 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 *The Journal of Chemical Physics* **81**,3684-3690 (1984).
- (49) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method *Journal of Applied Physics* **52**,7182-7190 (1981).
- (50) Kabsch, W. Xds *Acta Crystallogr D Biol Crystallogr* **66**,125-132 (2010).
- (51) Evans, P. Scaling and assessment of data quality *Acta Crystallogr D Biol Crystallogr* **62**,72-82 (2006).
- (52) 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).
- (53) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution *Acta Crystallogr D Biol Crystallogr* **66**,213-221 (2010).
- (54) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot *Acta Crystallogr D Biol Crystallogr* **66**,486-501 (2010).
- (55) Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards automated crystallographic structure refinement with phenix.refine *Acta Crystallogr D Biol Crystallogr* **68**,352-367 (2012).
- (56) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation *Acta Crystallogr D Biol Crystallogr* **65**,1074-1080 (2009).

Chapter 5. Ensemble-based computational enzyme design for stabilizing the transition states of a more complex reaction

Rakotoharisoa R.V., Pham N.T.H., Chica R.A.

5.1. Statement of contribution

R.V.R. wrote the chapter, R.V.R. generated ensembles from ensemble refinement (ER) for BH32, BH32.6, and BH32.12, P.N.T.H. generated ensembles from ER of BH32.7, ER ensembles of BH32.8 were provided by B.S., from our collaboration with the Thompson group at UC Merced. R.V.R. performed the design on BH32.12 ensembles and P.N.T.H. performed the design on the other ensembles.

5.2. Introduction

In chapter 1, we observed that small backbone rearrangements occurring during directed evolution are responsible for the activity increase of the most active variant from HG3 evolutionary trajectory. In chapter 2, we then observed that ensemble refinement (ER) and PertMin methods generate backbone ensembles that more accurately predict the conformation of the optimized variant in the case of HG3. In chapter 3, we used ER ensembles from the crystal structure of HG3 bound with the transition state analogue to design sequences displaying natural-like activity. In this chapter, we aim to expand our ensemble-based enzyme design methodology to establish a benchmark of backbone ensembles to predict the four transition states of the Morita-Baylis-Hillman reaction catalyzed by the evolved variant BH32.12 from the *de novo* enzyme BH32 evolutionary trajectory.

The primary objective of computational enzyme design is to create enzymes for non-natural reactions. These enzymes exhibit remarkable efficiency, regioselectivity, and stereoselectivity, making them highly desirable catalysts for industrial applications. Given the limited range of reactions catalyzed by natural enzymes, creating artificial enzymes capable of catalyzing any desired reaction is immensely important. Introducing catalytic functions into protein scaffolds paves the way for more sustainable and efficient industrial processes. Thanks to the continuous advancement in computational enzyme design methodologies, various model reactions have been successfully incorporated into inert protein scaffolds.¹⁻⁶

The Morita-Baylis-Hillman (MBH) reaction represents the carbon-carbon bond-forming transformation between an activated alkene and a carbon electrophile by a catalytic nucleophile. The significance of this reaction lies in its diverse industrial applications owing to the extensive array of substrates it can accommodate. Furthermore, the MBH reaction is atom economical. It minimizes waste, which is highly desirable from a green chemistry perspective. Additionally, the product obtained from this reaction is highly functionalized, making it well-suited for further modifications and transformations in synthetic processes.

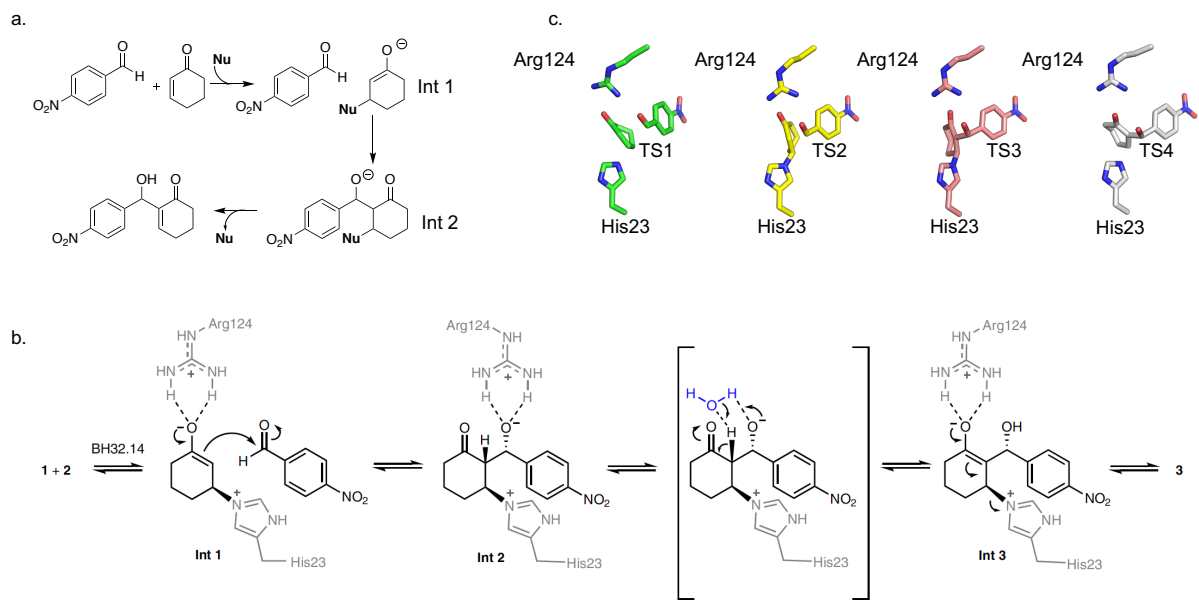


Figure 5.1. MBH reaction between 4-nitrobenzaldehyde and cyclohexenone. An activated nucleophile attacks the β -carbon of the α,β -unsaturated carbonyl compound, which leads to the formation of a carbon-carbon bond between the α -carbon of the Michael acceptor and the nucleophile. This attack triggers the addition of the carbon-carbon bond between the carbon at the β position of 2-cyclohexenone and the electrophilic carbon from 4-nitrobenzaldehyde. b) BH32.14 theoretical catalytic mechanism. Here, Arg124 acts as a H-bond donor to stabilize each intermediate covalently bound with His23 (Figure adapted from Crawshaw et al., ⁷). A water molecule has also been shown to stabilize proton transfer on the alkoxyde of intermediate 2. c) Sticks representation of the four theozymes we stabilized in this study.

Bjelic et al., ⁶ introduced the MBH reaction into the natural protein scaffold of a haloacid dehalogenase from *Pyrococcus horikoshii*, in which a nucleophilic histidine was built to catalyze the C-C bond formation on the α position of 2-cyclohexenone and the carbonyl carbon of 4-nitrobenzaldehyde (Figure 5.1). This artificial enzyme called BH32 exhibited a modest activity, where its rate enhancement over the uncatalyzed reaction in buffer ($k_{\text{cat}}/k_{\text{uncat}}$) is only 54. ⁶ This enzyme has been optimized through directed evolution by the group of Anthony Green, in which the activity increased by 1000-fold for the best variant BH32.14. The backbone scaffold has been extensively remodelled throughout the evolution, and 23 positions were mutated. Analyzing the mutation introduced during evolution revealed that the initially designed hydrogen bond donor Gln128 was abandoned as it was mutated to a proline. Structural analysis and density functional

theory (DFT) of one of the best variants of the evolutionary trajectory of BH32, BH32.12 showed that Arg124 is well placed for a bidentate hydrogen bonding with the developing charges on the oxyanion on the first intermediate and the second intermediate of the reaction.⁷ The reaction mechanism obtained from DFT starts with a Michael addition triggered by a nucleophilic attack from His23 at the β -carbon position of 2-cyclohexenone. An enolate intermediate is formed in which Arg124 is positioned to stabilize the developing negative charge on the alkoxide by bidentate hydrogen bonding. It is followed by addition of the enolate to the 4-nitrobenzaldehyde carbonyl, during which diastereoselective carbon-carbon bond formation generates a second oxyanion intermediate with R-configuration. This second alkoxide is also stabilized by Arg124. In the next step, a third intermediate is formed by water-mediated proton transfer, and its oxyanion is stabilized by hydrogen-bonding to Arg124. Finally, product release proceeds by elimination of the catalytic nucleophile.

These studies highlighted the need for directed evolution to improve the catalytic efficiency of *de novo* enzymes to allow their use for industrial applications. To build the active site of BH32, the Baker lab used a composite transition state, which is the superimposition of four transition states and a single crystal structure as input template. These oversimplifications do not consider the dynamic nature of enzymes, and a more precise prediction could be achieved by considering each transition state individually in the design process. More sophisticated approaches should be devised to improve the design methodology and account for the ensemble of conformations an enzyme adopts during its catalytic reaction cycle. In previous chapters, we highlighted the need to use an ensemble of backbones generated by ensemble refinement to account for the dynamic behaviour of the enzyme in solution. This method uses molecular dynamics simulations restrained by electron density to model local fluctuation. It gives the advantage of generating models that

more accurately predict the designed enzyme conformations.

In this chapter, our primary objective was to extend the ensemble-based design methodology by establishing a benchmark of backbone ensembles obtained from the ensemble refinement protocol of crystallographic data of five variants from the BH32 evolutionary trajectory (BH32, BH32.6, BH32.7, BH32.8, and BH32.12). We aimed to predict the theozyme obtained from the DFT calculation of the reaction of 4-nitrobenzaldehyde and 2-cyclohexenone catalyzed by the most active variant from our benchmark, BH32.12. As no crystal structure in the bound state was obtained, its crystal structure in the unbound state docked with the substrates was provided for the DFT calculation of the reaction. Designs resulted in the best prediction obtained from ensembles of BH32.12, while ensembles from earlier variant, including BH32, BH32.6 and BH32.7 did not enable the placement of the transition state, leading to a design with binding energy above 500 kcal/mol. Designs on the ensemble of BH32.8 resulted in high binding energy for TS3 and TS4 and only TS1 and TS2 allowed favourable energy, albeit higher than those of BH32.12 templates. These results highlight the backbone rearrangements that occurred during directed evolution, and confirm that BH32.12 will enable the best prediction on its own template. The result of this benchmark would serve as a reference for flattening the enzyme's catalytic reaction cycle and enhancing the rate of MBH reaction.

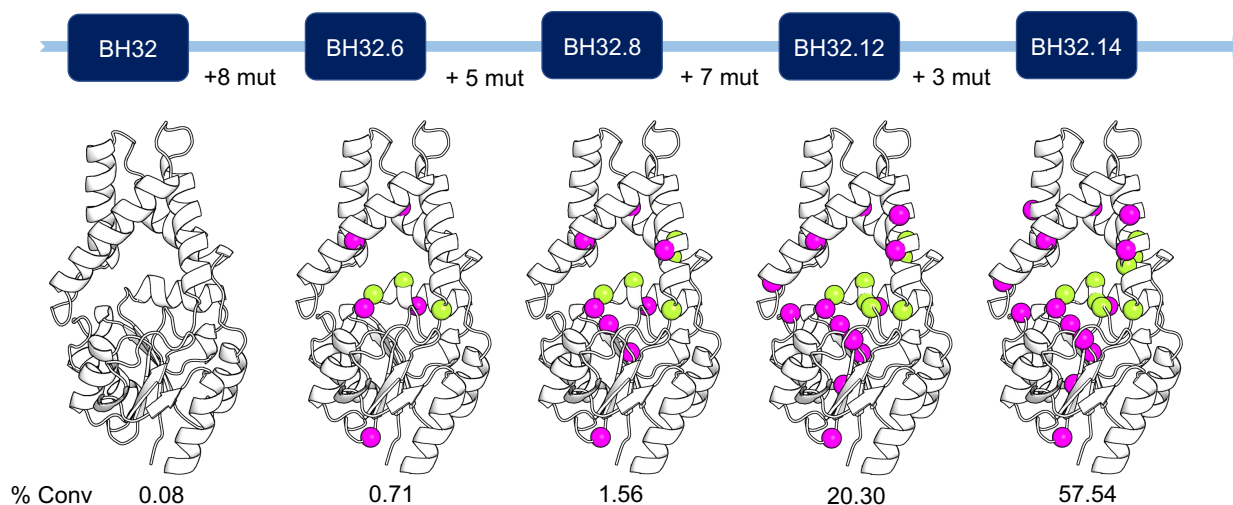


Figure 5.2. Mutations introduced in variants from the evolutionary trajectory of BH32. The names of enzymes are shown along with the number of mutations occurring during evolution and their corresponding percentage of conversion (% Conv). The % Conv is defined as the percentage of substrate molecules that are converted to product at the end of an enzyme reaction. Green spheres are mutations of positions proximal to the active site, and magenta spheres correspond to distal positions. Crystal structure used for visualization is from BH32.12 (PDB ID: 6Z1L).¹

5.3. Results

5.3.1. Ensemble generation

BH32 design has been optimized through fourteen rounds of directed evolution by the Greene Lab¹. Intermediate variants from the evolutionary trajectory of BH32 have been crystallized to uncover structural information leading to the improvement of activity. The backbone scaffold was extensively remodelled, and 10 % of the enzyme sequence was mutated (Figure 5.2). For the benchmark, we selected all available crystal structures of intermediates from the trajectory of BH32 (BH32 bound with an inhibitor, BH32.6 unbound, BH32.7 unbound, BH32.8 bound with an inhibitor, BH32.12 unbound) (Table 5.1). The most active variant, BH32.14, did not crystallize and was not included in the benchmark. BH32.12 only has three mutations that differ from BH32.14 (Ala19Thr, Glu70Arg, Pro128Leu), where only Ala19Thr is in direct contact with the ligand. Using the ensemble refinement protocol (see material and methods), we generated ensembles from each crystal structure. To ensure we obtained ensembles that most approximate

the enzyme conformations in solution, we selected ensembles with R_{work} and R_{free} lower than the single crystal structure (Table 5.1). We hypothesized that the BH32.12 ensemble would sample templates capable of stabilizing the calculated transition state with favourable energy and low root mean square deviation (RMSD) of the theozymes. In contrast, we assumed that ensembles derived from less active variants (BH32, BH32.6, BH32.7 and potentially BH32.8), in which active site positions are substituted with BH32.12 sequence would result in less favourable energy and higher RMSD.

Table 5.1. Crystal structures information and $R_{\text{work}}/R_{\text{free}}$ of crystal structures before and after ensemble refinement

Enzymes	PDB ID	Resolution (Å)	Crystal structure properties	$R_{\text{work}}/R_{\text{free}}$ before ER ^a	$R_{\text{work}}/R_{\text{free}}$ after ER ^a
BH32	6Q7N	2.02	Bound with 2-phenylethanone	0.197/0.236	0.164/0.217
BH32.6	6Z1K	1.48	Unbound state	0.157/0.177	0.151/0.173
BH32.7	7O1D	1.80	Bound state	0.204/0.255	0.170/0.227
BH32.8	Not published	1.89	Unbound state	0.179/0.206	0.164/0.199
BH32.12	6Z1L	2.29	Unbound state	0.233/0.262	0.192/0.235

^aER: Ensemble refinement

5.3.2. Design on BH32.12

We introduced each transition state in the individual backbone template as described in the materials and methods section. Using our geometries for building the poses, we have not successfully placed the transition states in the crystal structure of BH32.12, as the binding energy exceeded the penalty we included in the design and no solutions were found (see materials and methods). When we used the ensembles from ensemble refinement, we were able to place each transition state in many templates. To assess the effectiveness of our method, we superimposed each calculated theozymes on top of the models we obtained from the prediction. Figure 5.3 represents the best result of the repacking of each transition state on the individual template of BH32.12 ensembles. The calculated binding poses superimpose well with the calculated theozyme,

with a theozyme RMSD between 0.3 Å to 0.9 Å for all four theozymes. Repacking resulted in favourable energy, in which TS3 and TS4 yielded the highest energy. Interestingly, different templates are capable of stabilizing the four theozymes (Figure 5.3), highlighting the importance of using an ensemble for the stabilization of the MBH transition states.

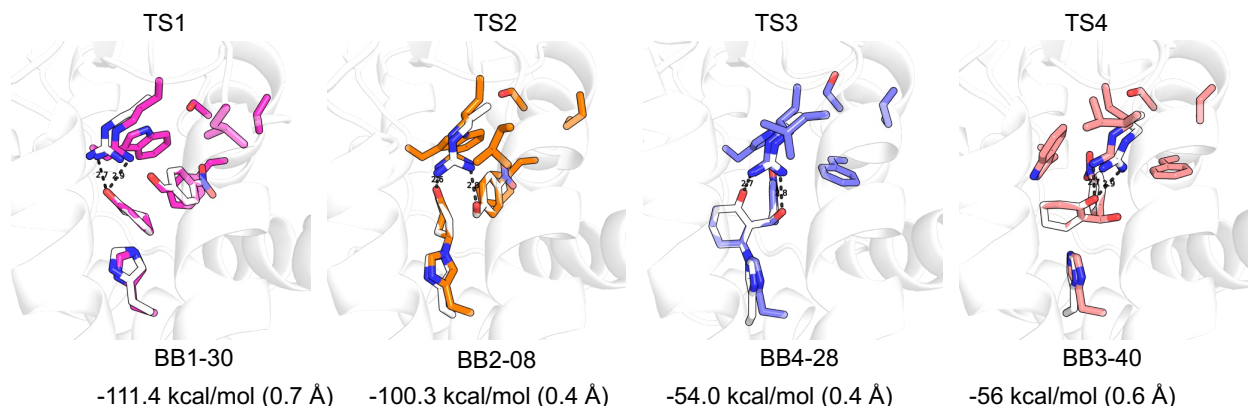


Figure 5.3. Repacking result of the individual transition state placement. Active sites of each template with the best energy and lowest root mean square deviation (RMSD) of catalytic residues from the model and the calculated transition state (TS) are presented. The name of each template is provided at the bottom, along with their binding energy and RMSD. Each template (coloured in magenta for the template stabilizing TS1, orange for TS2, purple for TS3, and salmon for TS4) is superposed with the corresponding calculated transition state coloured in grey. The distance between the hydrogen bond donor and the transition state functional groups is shown in dashed lines. Active site residues in direct contact with the transition states are also represented.

5.3.3. Benchmark on different templates

Using the information we got from the BH32.12 prediction, we aimed to use the same protocol to place each theozyme into the different ensembles we generated (Figure 5.1, and Supplementary Figure 5.1). We used the ensembles we generated from ER for BH32 bound with the inhibitor, BH32.6 unbound, BH32.7 unbound, and BH32.8 unbound, to stabilize the transition states and with BH32.12 sequence. We hypothesized that the prediction would not be successful for ensembles from less active variants because of backbone remodelling occurring during evolution. As expected, results obtained from the prediction on the less active variants resulted in

high energy binding (**Figure 5.4**). Templates from BH32 and BH32.6 did not result in favourable binding energy (>500 kcal/mol) for all states (**Figure 5.4** and Supplementary Table 5.1). BH32.7 ensembles allowed the stabilization of the TS4 with high energy (329 kcal/mol) but not the other TSs. In the case of BH32.8, no solutions were found for TS3 stabilization, TS4 stabilization was with high energy (354 kcal/mol). TS1 and TS2 were stabilized more favourably (**Figure 5.4** and Supplementary Table 5.1).

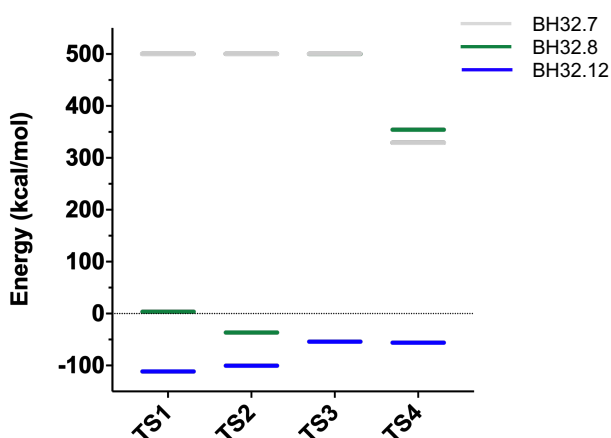


Figure 5.4. Energy diagram of the results obtained with the ensemble-based method. The blue energy level corresponds to the prediction on the BH32.7 best template, the dark green level corresponds to the prediction on the BH32.8 ensemble, and dark blue represents the prediction on the ensemble of BH32.12. Ensembles that did not generate results were considered to have energies above 500 kcal/mol.

These results showed that BH32.12 ensembles is the best ensemble to predict the calculated theozymes whereas ensembles from less active variants do not sample templates that allow favourable stabilization of the theozymes.

5.4. Discussions

Building upon the work in previous chapters, we have successfully placed the four theozymes using templates obtained from the ensemble refinement of BH32.12, one of the most active

variants obtained from the directed evolution of BH32. The results also showed that using ensembles from less active variants has not yielded templates predicting binding poses with favourable energy (Figure 5.4), highlighting the sensitivity of the method on conformational backbone changes occurring during evolution. Interestingly, the templates we obtained giving us the prediction are different across the four theozymes, despite their small structural differences. On the other hand, it was not possible to introduce the theozymes using the single crystal structure of BH32.12. First, the crystal structure obtained was in the unbound state and might not have the correct conformation to bind any of the transition states. Second, earlier work has highlighted the possibility of the haloacid dehalogenase backbone undergoing domain motion in other variants.⁸ These results highlight the need to use ensembles to stabilize theozymes for a multi-step reaction. While our end goal is to improve the catalytic efficiency of BH32.12, our approach is limited by the use of a crystal structure from the variant in the unbound state, which might lead to ensembles yielding less accurate predictions, as we pointed out in Chapter 2 and Chapter 3.

Our approach is significantly different from earlier attempts to stabilize multiple transition states,^{2,4,6} as a composite transition state was used to account for the multiple transition states of the reaction on a single crystal structure. Using this approach inevitably leads to weak binding and inaccurate prediction,^{1,4,9,10} as it did not account for the dynamic behaviour of the enzyme. Furthermore, as pointed out in this study, the different theozymes were stabilized by different templates instead of one single template. Highlighting the inaccuracy that the composite transition state approximation introduces.

MBH is an interesting model, not only due to its relevance in organic synthesis,^{11–13} but also because it challenges our capacity to design more efficient enzymes. Evolution has extensively remodelled the backbone conformation of BH32,⁷ and the Gln128 initially designed for H-bond

to stabilize the transition states, which was abandoned for an Arg124. This is on par with other enzymes optimized by directed evolution.^{14,15} Crossley et al.¹⁶ improved the reaction rate of MBH by mutating the catalytic His23 to a more reactive methylated histidine or MeHis in BH32.8 variant using *in vitro* translation with an engineered pyrrolysyl-tRNA synthetase_tRNA pair. They performed directed evolution on this new mutant, and the highest variant led to BH_{MeHis}1.8, a more active enzyme compared to BH32.14. Leading to another evolutionary pathway, a new hydrogen bond Glu26 appeared, and the initial hydrogen bond donor Arg124 was abandoned. This result reflects the influence of the enzyme active sites' microenvironment and shows the complexity of the enzyme design procedure. To consider the electronic environment of the active site, energy functions must be improved, and the choice of allowed mutations, as well as the optimized position has to be meticulously thought.

With the results we obtained from this benchmark, we aim to generate new sequences from the ensemble of BH32.12 capable of stabilizing each individual transition state to increase the MBH reaction rate and test the ability of our method to design more efficient catalysts (Figure 5.5).

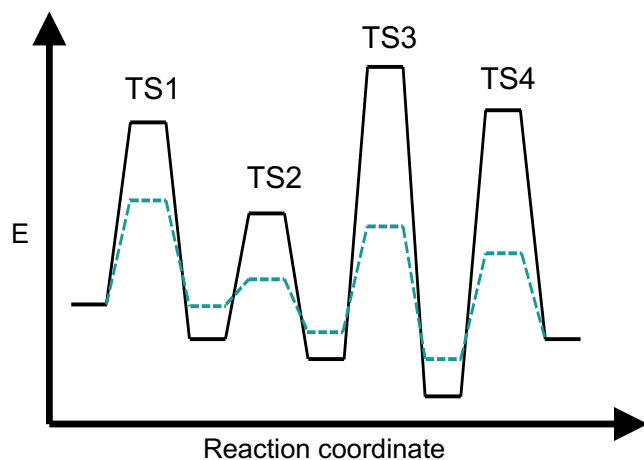


Figure 5.5 Future work for designing efficient MBHases. In a future work, this benchmark will enable the design of more efficient catalysts with multiple transition states, by flattening the energy landscape (dashed green lines is the objectif)

5.5. Methods

Ensemble Refinement. Ensembles of backbone templates were generated using the ensemble refinement protocol¹⁷ implemented in the PHENIX macromolecular structure determination software (v1.18.2-3874)¹⁸. In ensemble refinement, 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 the 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 the translation-libration-screw fitting, τ_x is the relaxation time and $w_{\text{x-ray}}$ is the coupled b -bath offset, which controls the X-ray weight such that the simulation runs at the 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.

Computational enzyme design. The four Morita Baylis Hillman transition-states (TS) were obtained from the work conducted by Crawshaw et al and coworkers⁷. His23 was used as a nucleophile and Arg124 was used to stabilize the oxyanion hole of intermediate one and intermediate two (Supplementary Table 5.2). Except for catalytic residues, all the positions in the active site were mutated to Gly during theozyme placement to avoid steric clashes with designed catalytic residues (Supplementary Table 5.2). The 2002 Dunbrack backbone-independent rotamer

library¹⁹ 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 placement²⁰ using the contact geometries listed in Supplementary Table 5.3. TS pose energies were calculated using a modified version of the Phoenix energy function¹³ consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field²² 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 Å²³, 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 5.4) by applying an energy benefit of –100 kcal mol⁻¹ as described by Lassila *et al.*²⁰. A penalty of 500 kcal/mol was applied to poses that did not satisfy the biases.

Next, active-site repacking calculations were performed on each template, stabilizing mostly the structure bound with the TS. In these calculations, the TS was translated ± 0.4 Å in each Cartesian coordinate in 0.2-Å increments and rotated 10° about all three axes (origin at TS geometric center) in 5° increments. The identities of catalytic residues were fixed and allowed to sample all rotamers of that amino-acid type. Residues mutated to Gly in the theozyme placement step were allowed to sample rotamers from the sequence of BH32.12. Sidechain–TS interaction energies were biased to favour contacts satisfying geometries as done during the theozyme placement step. Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. After active-site repacking, we calculated energy differences between each repacked structure and the corresponding structure in the absence of the TS where all designed residues (including catalytic residues) were converted to Gly. This energy subtraction was done to allow comparison of sequence energies across backbone templates.

5.6. Supplementary information

Supplementary Table 5.1. Repacking results. Energy on the best template obtained from each ensemble. Root mean square deviation (RMSD) of the catalytic residues and the theozymes are shown in parenthesis.

Templates	TS1 energy (kcal/mol)	TS2 energy (kcal/mol)	TS3 energy (kcal/mol)	TS4 energy (kcal/mol)
BH32	>500	>500	>500	>500
BH32.6	>500	>500	>500	>500
BH32.7	>500	>500	>500	329 (0.50 Å)
BH32.8	3.4 (0.40 Å)	-36.7 (0.28 Å)	>500	354 (0.54 Å)
BH32.12	-111.4 (0.70 Å)	-100.3 (0.40 Å)	-54.0 (0.4 Å)	-56.0 (0.60 Å)

Supplementary Table 5.2. Amino acid and positions included in the design procedure

Design objective	Designed residues
Nucleophile	H23
H-bond donor	R124
BH32	L10, N14, A19, A20, S22, L24, I26, M27, L42, Y45, E46, T49, F53, L64, L68, W88, S91, L92, S95, T122, D123, D125, Q128, A129, F132
BH32.6	L10, I14, A19, A20, S22, L24, I26, M27, L42, Y45, E46, A49, F53, L64, L68, W88, S91, L92, S95, T122, N123, D125, P128, A129, F132
BH32.7	L10, I14, A19, A20, V22, L24, I26, M27, L42, Y45, E46, A49, F53, L64, L68, W88, S91, L92, A95, L122, N123, D125, P128, A129, F132
BH32.8 sequence prediction	L10, I14, A19, Y20, V22, L24, I26, M27, L42, Y45, E46, A49, F53, L64, L68, W88, S91, L92, A95, L122, N123, D125, P128, A129, F132
BH32.12 sequence prediction	W10, I14, A19, Y20, A22, F24, I26, M27, L42, Y45, E46, A49, F53, L64, L68, W88, S91, L92, A95, L122, N123, D125, P128, S129, F132
BH32.14 sequence prediction	W10, I14, T19, Y20, V22, F24, I26, M27, L42, Y45, E46, A49, F53, L64, L68, W88, S91, L92, A95, L122, N123, D125, L128, S129, F132

Supplementary Table 5.3. Geometric definitions for generation of transition-state poses off the side chain of the nucleophile for each transition state

Transition state	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Values
TS1	Distance	NE2	C5			1.5, 1.6, 1.8, 1.9, 2.0, 2.2, 2.3, 2.4, 2.5
	Angle	CE1	NE2	C5		102, 117, 122, 127, 132
	Angle	NE2	C5	C6		97, 102, 107, 112, 117
	Torsion	ND1	CE1	NE2	C5	−178, −173, −168, −158, 158, 168, 173, 178
TS2	Torsion	CE1	NE2	C5	C6	−13.4, −8.4, −3.4, 1.6, 6.6
	Torsion	NE2	C5	C6	C1	−98, −102, −107, −112, −117
	Distance	NE2	C5			1.3, 1.4, 1.5
	Angle	CE1	NE2	C5		125, 127, 129
	Angle	NE2	C5	C6		108, 110, 112
	Torsion	ND1	CE1	NE2	C5	−188, −183, −178, −173, 168, 168, 173, 178, 183, 188
TS3	Torsion	CE1	NE2	C5	C6	29, 34, 39, 44, 49
	Torsion	NE2	C5	C6	C1	−145, −140, −135, −130, −125
	Distance	NE2	C5			1.3, 1.4, 1.5
	Angle	CE1	NE2	C5		125, 127, 129
	Angle	NE2	C5	C6		110, 112, 114
	Torsion	ND1	CE1	NE2	C5	−178, −172, −167, −157, 157, 162, 167, 172, 178
TS4	Torsion	CE1	NE2	C5	C6	25, 30, 35, 40, 45
	Torsion	NE2	C5	C6	C1	−155, −150, −145, −140, −135
	Distance	NE2	C5			1.8, 2.0, 2.2
	Angle	CE1	NE2	C5		113, 118, 123, 128, 133
	Angle	NE2	C5	C6		93, 98, 103, 108, 113
	Torsion	ND1	CE1	NE2	C5	−188, −183, −178, −173, −168, 168, 173, 178, 183, 188
	Torsion	CE1	NE2	C5	C6	40, 45, 50, 55, 60
	Torsion	NE2	C5	C6	C1	−106, −101, −96, −91, −86

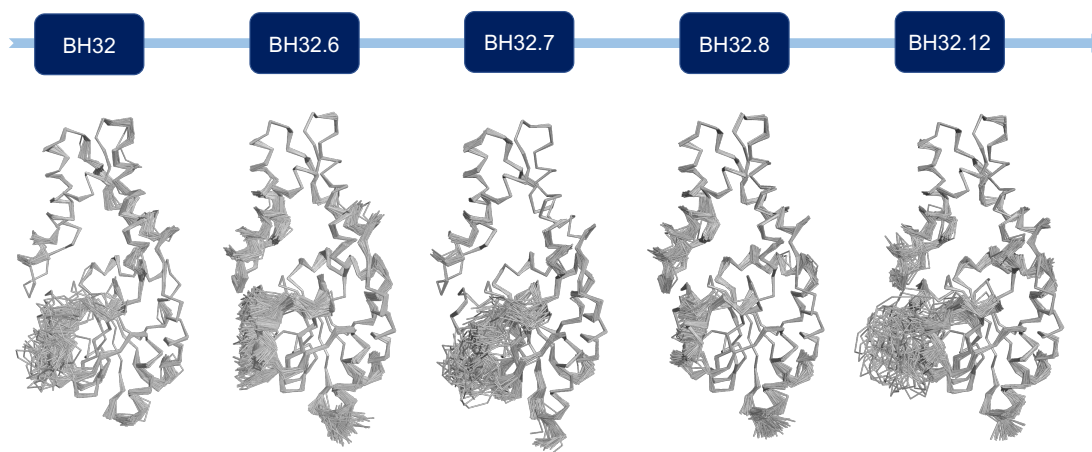
^aAtoms from the TS are in bold, the other atoms are from the interacting sidechain of His23

Side

Supplementary Table 5.4. Geometric constraints used to define catalytic contacts during the design procedure for each transition state

Transition state	Contact	Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max ^b
TS1	Nucleophile	H23	Distance	NE2	C5			1.5	2.5
			Angle	CE1	NE2	C5		102	132
			Angle	NE2	C5	C6		97	117
			Torsion	ND1	CE1	NE2	C5	158	178
			Torsion	CE1	NE2	C5	C6	-13	7
			Torsion	NE2	C5	C6	C1	-117	-97
	H-bond	R124	Distance	NH1	O1			2.6	3.0
			Distance	NH2	O1			2.6	3.0
			Distance	NH2	O7			2.8	3.2
			Angle	CZ	NH1	O1		86	106
			Angle	CZ	NH2	O7		87	107
			Torsion	NH1	O1	C1	C6	58	76
			Distance	NE2	C5			1.3	1.5
			Angle	CE1	NE2	C5		125	128
TS2	Nucleophile	H23	Angle	NE2	C5	C6		108	112
			Torsion	ND1	CE1	NE2	C5	168	188
			Torsion	CE1	NE2	C5	C6	29	49
			Torsion	NE2	C5	C6	C1	-145	-125
	H-bond	R124	Distance	NH1	O1			2.5	3.9
			Distance	NH2	O7			2.5	2.9
			Angle	CZ	NH1	O1		95	115
			Angle	CZ	NH2	O7		114	134
			Angle	NH1	O1	C1		142	162
			Torsion	NH2	CZ	NH1	O1	-58	-37.7
	Nucleophile	H23	Distance	NE2	C5			1.3	1.5
			Angle	CE1	NE2	C5		125	129
			Angle	NE2	C5	C6		110	114
			Torsion	ND1	CE1	NE2	C5	157	177
			Torsion	CE1	NE2	C5	C6	25	45
			Torsion	NE2	C5	C6	C1	155	135
TS3	H-bond	R124	Distance	NH1	O1			2.5	2.9
			Distance	NH2	O7			2.6	3.0
			Angle	CZ	NH1	O1		92	112
			Angle	CZ	NH2	O7		116	136
			Angle	NH1	O1	C1	C6	-80	-60
			Distance	NE2	C5			1.8	2.2
	Nucleophile	H23	Angle	CE1	NE2	C5		113	133
			Angle	NE2	C5	C6		93	113
			Torsion	ND1	CE1	NE2	C5	168	188
			Torsion	CE1	NE2	C5	C6	40	60
			Torsion	NE2	C5	C6	C1	-106	-86
	H-bond	R124	Distance	NH1	O1			2.6	3.0
			Distance	NH2	O1			2.6	3.0
			Distance	NH2	O7			3.5	3.9
			Angle	CZ	NH1	O1		86	106
			Angle	CZ	NH2	O7		129	149
			Angle	NH2	CZ	NH1	O1	-159	-170

^aAtoms from the TS are in bold, the other atoms are from the interacting sidechain of His23^bDistances are measured in Å, angles and torsions are in ° (degree)



Supplementary Figure 5.1. Ensembles from ensemble refinement (ER) of BH32 variants.

Ensembles are obtained from the following crystal structures: BH32 PDB ID: 6Q7N, BH32.6 PDB ID: 6Z1K, BH32.7 PDB ID: 7O1D, BH32.8 unpublished data, BH32.12 PDB ID: 6Z1L. Variants are ordered by their order of evolution and their activity

References

1. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
2. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
3. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **329**, 309–313 (2010).
4. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Science* **21**, 717–726 (2012).
5. Richter, F. *et al.* Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *Journal of the American Chemical Society* **134**, 16197–16206 (2012).
6. Bjelic, S. *et al.* Computational design of enone-binding proteins with catalytic activity for the morita-baylis-hillman reaction. *ACS Chemical Biology* **8**, 749–757 (2013).
7. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
8. Burke, A. J. *et al.* Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* **570**, 219–223 (2019).
9. Khersonsky, O. *et al.* Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *Journal of Molecular Biology* **407**, 391–412 (2011).
10. Khersonsky, O. *et al.* 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).
11. Basavaiah, D., *et al.* Recent Contributions from the Baylis–Hillman Reaction to Organic Chemistry. *Chem. Rev.* **110**, 5447–5674 (2010).
12. Basavaiah, D., *et al.* Recent Advances in the Baylis–Hillman Reaction and Applications. *Chem. Rev.* **103**, 811–892 (2003).
13. Wei, Y. & Shi, M. Recent Advances in Organocatalytic Asymmetric Morita–Baylis–Hillman/aza-Morita–Baylis–Hillman Reactions. *Chem. Rev.* **113**, 6659–6690 (2013).
14. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790–3795 (2012).
15. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nature Chemistry* **9**, 50–56 (2017).
16. Crossley, A. *et al.* A Non-Canonical Nucleophile Unlocks a New Mechanistic Pathway in a Designed Enzym. Preprint at <https://doi.org/10.21203/rs.3.rs-2922796/v1> (2023).
17. Tom Burnley, B., *et al.* Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **2012**, 1–29 (2012).
18. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Cryst D* **68**, 352–367 (2012).
19. Dunbrack, R. L. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology* **12**, 431–440 (2002).
20. Lassila, J. K., *et al.* Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 16710–16715 (2006).
21. 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).
22. Mayo, S. L., *et al.* DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
23. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10172–10177 (1997).

Chapter 6. Discussion and conclusion

6.1. Summary

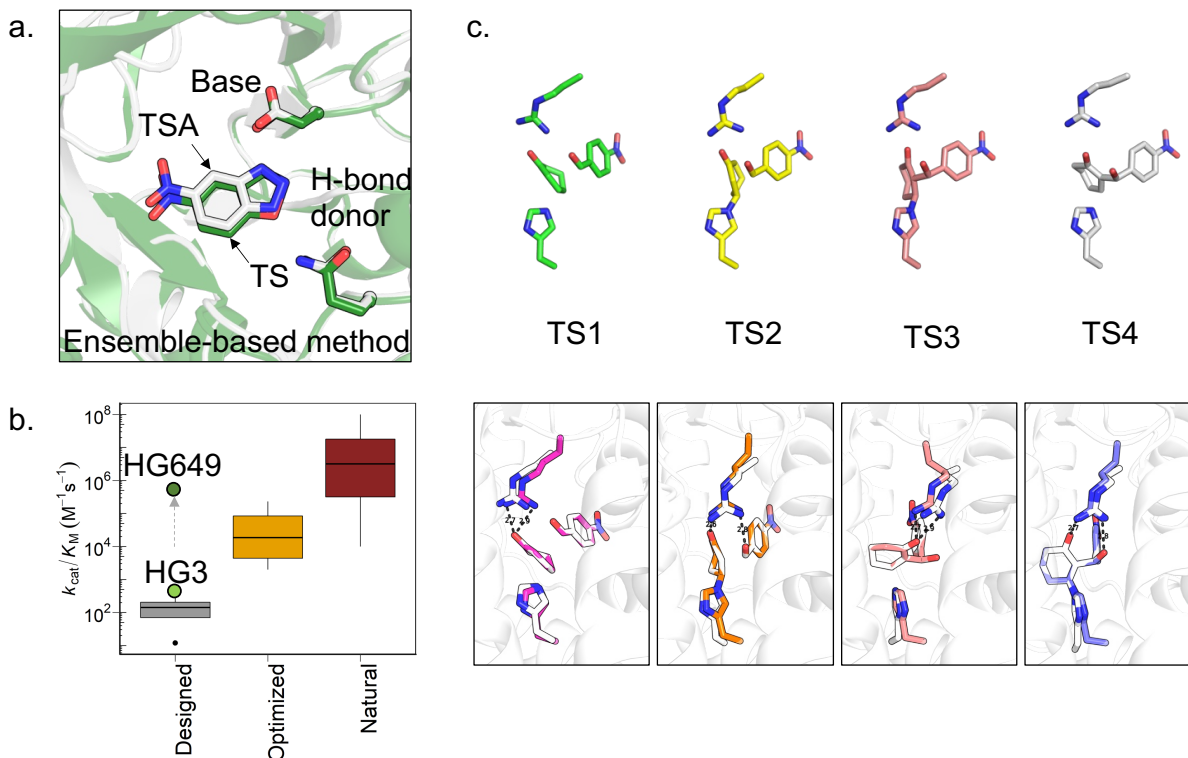


Figure 6.1. Summary of findings. a) Overlay of crystal structure in grey, with the designed model (pale green or dark green). The predicted base and the predicted H-bond donor are represented. The transition state (TS) from the model and the transition state analogue (TSA) used for crystallization are also represented. This panel shows the design prediction using our ensemble-based method. b) k_{cat}/K_M of designed enzymes, optimized enzymes and natural enzymes. The green sphere represents the catalytic efficiency of HG3 and the dark green sphere corresponds to the most active variant of HG3 designed in this project. c) the transition states of the MBH reaction between 4-nitrobenzaldehyde and cyclohexenone² are shown. With our ensemble-based enzyme design approach, we successfully placed the four transition states using different backbone templates.

The design of *de novo* enzymes for tailored reactions offers a wide range of applications and deepens our comprehension of enzyme catalysis. While computational enzyme design methods have facilitated the creation of artificial enzymes,^{1,3–7} a significant challenge has been their inability to consider protein dynamics, resulting in less accurate predictions and weaker substrate binding.^{3,8–11} An innovative approach to address this limitation involves introducing flexibility into the design framework by utilizing an ensemble of backbones, as opposed to relying on a single

crystal structure, with the potential to enhance prediction accuracy.^{12–14} Enzyme evolution utilizes the remodelling of backbone conformations to optimize catalytic functions.¹⁵ However, incorporating these backbone rearrangements into the design framework has remained a challenge.

¹¹ With this thesis project, (1) we developed an ensemble-based computational enzyme design method to enhance the prediction accuracy of a highly active enzyme (Figure 6.1.a), (2) we established a benchmark for our ensemble-based design method to determine the most suitable ensemble for accurate predictions, (3) we applied our ensemble-based enzyme design method to enhance the activity of a low-catalytically active enzyme (Figure 6.1.b), and (4) we extended our methodology to stabilize multiple transition states of a more complex reaction (Figure 6.1.c).

Chapter 2. A comprehensive study into the changes during the evolution of a low-activity enzyme guided the development of our ensemble-based design method. Variants from the evolutionary trajectory of a low activity *de novo* enzyme HG3,^{1,16} were crystallized to investigate the resulting backbone rearrangements. These subtle changes led to the stabilization of heterogeneous loops, improved active site preorganization, enhanced transition state active site shape complementarity, and widened the mouth of the active site. To develop our ensemble-based enzyme design method, we generated a backbone ensemble from the ensemble refinement of crystallographic data of HG3 bound with the transition state analogue. With this ensemble, we improved the prediction accuracy of a highly active variant of HG3, called HG4, and reproduced the effects of directed evolution. This project highlighted that evolution-induced mutations enriched pre-existing productive conformational states within the conformational landscape of HG3. Ensemble refinement can sample these lower energy states, which can be harnessed with our computational methodology to predict highly active enzymes.

Chapter 3. We established a benchmark for our ensemble-based computational enzyme design method, aiming to identify the most effective ensembles for predicting the binding pose of highly active enzymes. This benchmark incorporated one designed enzyme, HG3 in both the unbound and bound states. We generated ensembles for each enzyme state using molecular dynamics, ensemble refinement of crystallographic data, and PertMin ensembles, and ensembles from AlphaFold2 based on the sequences of these enzymes. Our benchmark analysis revealed that the most accurate results were achieved through the ensemble refinement of crystallographic data derived from the crystal structure of the enzyme in its bound state. By accommodating local motions constrained by electron density within the molecular dynamic simulations, this method facilitated the sampling of conformational states that closely approximated the true enzyme conformation in a solution. These insights have provided valuable guidance for our efforts in designing efficient enzymes.

Chapter 4. Building upon the insights from earlier chapters, we refined our computational enzyme design methodology by introducing backbone flexibility into the design procedure to enhance the catalytic efficiency of two low-activity *de novo* enzymes, HG3 and KE70. For this purpose, we utilized ensembles from ensemble refinement of crystallographic data to stabilize the transition state of the Kemp elimination reaction. Sequences selected for experimental characterization met specific criteria, including theozyme geometry close to the ideal configuration, favourable energy, 95% of transition state burial, and a percentage of preorganized active site residues over 80%. Interactions of catalytic residues with the substrate were ranked through either a 20-ns molecular dynamics simulation or molecular docking. The catalytic efficiency of the most active variants improved by 100 to 250-fold, which is comparable to those obtained by screening thousands of mutants from multiple rounds of directed evolution. Crystallographic data of HG3 confirmed that

the structural features we predicted from our applied filter were maintained. Our ensemble-based method successfully enhanced the catalytic efficiency of enzymes tested, achieving natural-like efficiency by enriching existing productive conformational states.

Chapter 5. In this chapter, we aimed to extend our computational methodology by establishing a benchmark for a more complex reaction, Morita Baylis Hillman (MBH) reaction. The model reaction involves the carbon-carbon bond formation of cyclohexenone and 4-benzaldehyde, comprising three intermediates and four transition states. Our approach involved placing the four transition states into the backbone templates of ensembles from the evolutionary trajectory of BH32. As anticipated, the four transition states were effectively incorporated into the backbone template of an active variant, BH32.12. Using ensembles from less active variants resulted in less accurate prediction and higher energy. To complement our benchmark, we optimized sequences that emerged from the evolutionary trajectory of BH32.14 on individual backbones of the ensemble of BH32.12. BH32.14 sequence turned out to have the most favourable energy during optimization. With a favourable prediction of BH32.14 sequence on BH32.12 templates we are convinced that our algorithm is capable of predicting a high catalytic activity sequence, if we have the correct backbone ensemble.

Throughout the projects in this thesis, we developed a robust methodology for designing highly efficient enzymes, hoping that it will not only unlock the potential for creating more effective artificial catalysts but also pave the way for a better understanding of enzyme catalysis.

6.2. Perspective and future directions

6.2.1. Advantages of the ensemble-based design approach

With our ensemble-based computational enzyme design approach, we achieved substantial improvements in the catalytic efficiency of two low-activity enzymes, HG3 and KE70, with minimal experimental effort. The best variant of HG3 obtained from our approach displayed a 250-fold increase in catalytic efficiency, nearing natural-like levels ($k_{\text{cat}}/K_{\text{M}} \sim 30,000 \text{ M}^{-1} \text{ s}^{-1}$). This fold improvement is comparable to the results of extensive screening and directed evolution. Our use of backbone ensembles derived from ensemble refinement of crystallographic data of the designed enzyme offers the advantage of sampling pre-existing productive states within the enzyme's conformational landscape. Expanding on established computational methodologies,^{17,18} we performed theozyme placement and repacking in the individual backbone of the ensemble. This approach enabled us to identify a productive backbone template among the ensemble members from ER, in which TS is interacting with catalytic residues in a geometry close to the ideal configuration calculated by DFT. Using this same backbone, optimizing active site positions to further stabilize the TS enhanced catalytic activity by up to 100 or 250-fold compared to the initial design of KE70 or HG3, respectively. This study aligns with the work done by Otten et al.¹⁹ By using a combination of NMR spectroscopy, cryo- and high temperature crystallography of HG3, a variant of the evolutionary trajectory, HG3.7 and the most active variant HG3.17, revealed the existence of an equilibrium between productive and unproductive conformational sub-states, where HG3.17 has a lower population of unproductive substates compared to HG3 and HG3.7. This study revealed that productive conformational ensemble was stabilized upon introducing mutation through directed evolution. Our ensemble-based enzyme design method uses this hypothesis to sample a productive state from ER ensembles of a low activity enzyme, such as HG3,

in which the algorithm introduces mutations to stabilize the highly active conformational ensemble.

Based on the findings of this study, we propose that our ensemble-based enzyme design method can generate enzymes with catalytic efficiencies approaching those of natural enzymes and should be broadly applicable for different reactions provided that a theozyme model is available.

6.2.2. Limitations

Our approach, while straightforward in its implementation, is constrained by the requirement for a high-resolution crystal structure of the designed enzyme bound with the transition state analogue. As our project has demonstrated, prediction accuracy is influenced by the backbone conformation employed for the design. Subtle changes in backbone conformations can trigger the lever-arm effect,²⁰ where subtle backbone conformational changes can lead to large changes in the position of sidechain atoms far from the backbone, and significantly reducing prediction accuracy. Additionally, in the case of HG3, we had complete crystallographic data available to allow the generation of ensembles from ensemble refinement. However, for KE70, no crystal structure bound with the TSA was available at the time of the study, which likely accounts for the modest improvement achieved by our methods for this enzyme. Another limitation of our method is that it requires minimal activity for the reaction of interest. When the chosen input structure comes from a scaffold lacking the activity of interest, it may be advisable to employ ensembles from methods that explore a wider conformational space within the protein energy landscape, such as backbone ensembles generated from the MD simulation parameters outlined in Chapter 3.

Despite our success in enhancing the catalytic efficiency of HG3, we have not reached the efficiency of the highly active variant HG4 using our method (k_{cat}/K_M). Although most substitutions sampled in our computational library design were similar to those found in HG4, two positions were never sampled to find the HG4 sequence (Arg276Phe and Gly82Ala). First, Arg276 was not part of the optimized position, as it was included in our cut-off for defining first-shell or second-shell positions. Second, the algorithm and forcefield we used never allowed the position Gly82 to sample alanine during the repacking step. It will be interesting to include these substitutions in another study to investigate the effect of these two positions individually.

To date, directed evolution is still important for improving low-activity enzymes due to its efficiency in optimizing distal mutations, which current computational methods cannot yet predict efficiently.^{3,8,10,11,21,22} However, it is still challenging to define which positions to optimize as mutations at certain positions may have an adverse impact on the enzyme activity. One solution is to use molecular dynamics techniques to identify positions important for catalysis.²¹

Finally, the computational enzyme design approach centers around stabilizing the TS to enhance the reaction rate.²³ However, in solution, enzymes adopt a transition state ensemble, which computational approaches often simplify by approximating them as a single conformation,^{24,25}. This approximation may result in less effective TS stabilization. Therefore, refining the design procedure to account for the TS as an ensemble of states can enhance the development of more efficient catalysts.

6.2.3. The ensemble-based enzyme design method and other methods

Several methods were developed in conjunction with the current study to improve the catalytic efficiency of *de novo* enzymes. Molecular dynamics simulation was used in several studies to rank and evaluate the designs.^{1,26,27} Especially, a study conducted by Wang et al.²⁷, using multiple short MD simulations, improved the catalytic efficiency of HG3.17,¹⁶ by 65%. This high-throughput MD simulation approach to evaluate the designs can be useful for eliminating false positives and false negatives from the sequences generated by the design procedures.

Bhattacharya et al.²⁸ used an NMR-guided directed evolution methodology to increase the catalytic efficiency of an artificial Kemp eliminase designed in a haem protein (Mb). The improved catalytic efficiency reached $15,721,000 \text{ M}^{-1} \text{ s}^{-1}$ for the best variant. While this achievement is impressive, it relies on the presence of a haem in the reduced form in the active site. With the same procedure, they improved another artificial Kemp eliminase called AlleyCat,²⁹ which also catalyzes the Kemp elimination reaction. Still, the catalytic improvement was less impressive and reached $4,378 \text{ M}^{-1} \text{ s}^{-1}$. While NMR-guided directed evolution is a promising method, it needs a high protein concentration and requires a high-resolution NMR signal to interpret the data.

It is also possible to increase catalytic efficiency by introducing a more reactive non-canonical amino acid. Burke et al.³⁰ used a Me-His as a catalytic nucleophile for *de novo* BH32 enzyme catalyzing the Morita Baylis Hilman reaction. Introducing Me-His reduces the energy barrier of an acyl-enzyme intermediate and leads to an increase in catalytic efficiency.³⁰ Integrating non-canonical amino acids (ncAAs) as catalytic residues can also enable the design of more efficient catalysts. However, introducing ncAAs in the protein scaffold requires post-translational modification and more cost-effective enzyme production containing ncAAs.³¹

It can be possible to include machine learning in the design of artificial enzyme methodologies. Yeh et al.³² successfully designed a highly efficient artificial luciferase ($k_{\text{cat}}/K_{\text{M}} \sim 10^6$). To avoid being restricted to the backbone conformation of natural proteins, they developed an approach called deep-learning-based family-wide hallucination to generate *de novo* scaffolds with catalytic sites stabilizing the high-energy states of reactions involving two distinct substrates, diphenylterazine or 2-deoxycoelenterazine.³² This approach led to highly soluble proteins, in which catalytic activity was enhanced by introducing substitutions with site-saturated mutagenesis. Machine learning can create scaffolds with high stability. However, enzyme stability is not correlated with higher activity,^{33–35} and very stable proteins tend to exhibit more rigidity, overlooking enzyme dynamics and catalysis.^{10,36–38} With the continuous development of machine learning algorithms to generate ensembles that integrate protein dynamics from learning protein energy landscape,^{39–42} the accuracy of ensembles generated will be improved in the near future. It can be more reliably used for our ensemble-based method.

6.3. Future directions

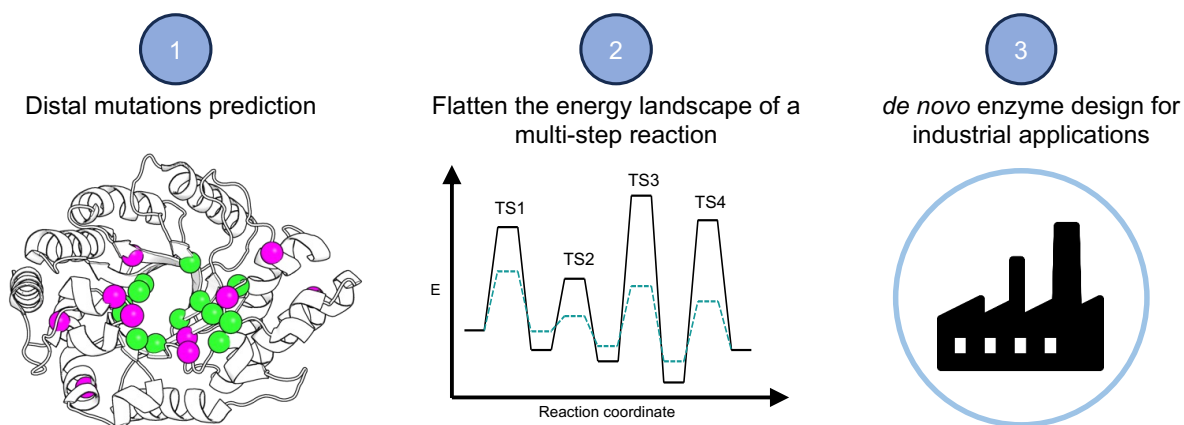


Figure 6.2. Future directions. With the ensemble-based method, we aim to (1) enhance the methodology by implementing distal mutation (in magenta) prediction in the procedure. (2) flatten the energy landscape of the MBH reaction to improve its catalytic efficiency, (3) design *de novo* enzymes for reactions useful for industrial applications.

In future work, we aim to devise a methodology to predict distal mutations for exploring a wider conformational state and improve the catalytic efficiency of designed enzymes. Our ensemble-based method can be used in conjunction with the Shortest Path Map (SPM),³⁶ method to predict positions important for catalysis. Using molecular dynamics simulations, this method generates a correlation map that identifies pairs of residues that are correlated. This information will serve as a guide for optimizing these positions during the enzyme design process.

The design of the *de novo* MBH was achieved by placing a composite transition state inside a single backbone template.⁶ However, as pointed out by this study, using this approach is insufficient to account for the multiple transition states of the MBH reaction. Thus, by using the ensemble-based enzyme design method, we should be able to generate variants of BH32.12 with enhanced catalytic efficiency, by stabilizing the individual transition states and flattening its energy landscape.

With our method, we aim to design artificial enzymes capable of catalyzing reactions important for industrial applications.

References

1. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790–3795 (2012).
2. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
3. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
4. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
5. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **329**, 309–313 (2010).
6. Bjelic, S. *et al.* Computational design of enone-binding proteins with catalytic activity for the morita-baylis-hillman reaction. *ACS Chemical Biology* **8**, 749–757 (2013).
7. Richter, F. *et al.* Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *Journal of the American Chemical Society* **134**, 16197–16206 (2012).
8. Khersonsky, O. *et al.* 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).
9. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Science* **21**, 717–726 (2012).
10. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nature Chemistry* **9**, 50–56 (2017).
11. Khersonsky, O. *et al.* Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *Journal of Molecular Biology* **407**, 391–412 (2011).
12. Allen, B. D., *et al.* Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 19838–19843 (2010).
13. Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Science* **21**, 1241–1252 (2012).
14. Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins: Structure, Function and Bioinformatics* **82**, 771–784 (2014).
15. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the ($\beta\alpha$)8-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).
16. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
17. Lassila, J. K., *et al.* Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 16710–16715 (2006).
18. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* **15**, 2785–2794 (2006).
19. Otten, R. *et al.* How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).
20. Polizzi, N. F. & DeGrado, W. F. A defined structural unit enables de novo design of small-molecule-binding proteins. *Science* **369**, 1227–1233 (2020).
21. Romero-Rivera, A., *et al.* Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catalysis* **7**, 8524–8532 (2017).
22. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1–20 (2020) doi:10.1002/wcms.1502.
23. Pauling, L. CHEMICAL ACHIEVEMENT AND HOPE FOR THE FUTURE. *American Scientist* **36**, 51–58.
24. Keith Laidler & King, C. The development of Transition State Theory. *87*, 2657–2664 (1983).
25. Ma, B., *et al.* Transition-state Ensemble in Enzyme Catalysis: Possibility, Reality, or Necessity? *Journal of Theoretical Biology* **203**, 383–397 (2000).
26. Kiss, G., *et al.* Evaluation and ranking of enzyme designs. *Protein Science* **19**, 1760–1773 (2010).
27. Wang, P., *et al.* Using High-Throughput Molecular Dynamics Simulation to Enhance the Computational Design of Kemp Elimination Enzymes. *J. Chem. Inf. Model.* **63**, 1323–1337 (2023).

28. Bhattacharya, S. *et al.* NMR-guided directed evolution. *Nature* 610, 389–393 (2022).
29. Korendovych, I. V. *et al.* Design of a switchable eliminase. *Proceedings of the National Academy of Sciences of the United States of America* 108, 6823–6827 (2011).
30. Burke, A. J. *et al.* Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* 570, 219–223 (2019).
31. Birch-Price, Z., *et al.* Engineering enzyme activity using an expanded amino acid alphabet. *Protein Eng Des Sel* 36, gzac013 (2022).
32. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* 614, 774–780 (2023).
33. Hou, Q., *et al.* Enzyme Stability-Activity Trade-Off: New Insights from Protein Stability Weaknesses and Evolutionary Conservation. *J. Chem. Theory Comput.* 19, 3664–3671 (2023).
34. Stark, C., *et al.* Systematic investigation of the link between enzyme catalysis and cold adaptation. *Elife* 11, e72884 (2022).
35. Struvay, C. & Feller, G. Optimization to low temperature activity in psychrophilic enzymes. *Int J Mol Sci* 13, 11643–11665 (2012).
36. Romero-Rivera, A., *et al.* Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catalysis* 7, 8524–8532 (2017).
37. Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438, 117–121 (2005).
38. Eisenmesser, E. Z., *et al.* Enzyme Dynamics During Catalysis. *Science* 295, 1520–1523 (2002).
39. Janson, G., *et al.* Direct generation of protein conformational ensembles via machine learning. *Nat Commun* 14, 774 (2023).
40. Gkeka, P. *et al.* Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* 16, 4757–4775 (2020).
41. Lindorff-Larsen, K. & Kragelund, B. B. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *Journal of Molecular Biology* 433, 167196 (2021).
42. Noé, F., *et al.* Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology* 60, 77–84 (2020).