



***De novo* enzyme design by customizing minimal TIM barrels**

Benjamin Smith

Thesis submitted to the University of Ottawa in partial fulfillment of
the requirements for the degree of Master of Science in Chemistry

Department of Chemistry and Biomolecular Sciences

Faculty of Science

University of Ottawa

Abstract

Enzymes are the most efficient and selective catalysts known, and the ability to design an enzyme for a target reaction is a primary goal of protein engineering. Among natural enzymes, the TIM barrel represents the most prevalent and versatile protein fold, using extended loops and other structural elements to facilitate substrate binding, catalysis, and product release. While *de novo* TIM barrels have been successfully designed, their minimalistic architectures lack these essential structural elements. In this work, we present CANVAS, a computational framework for *de novo* enzyme design that introduces a structural lid into a minimal *de novo* TIM barrel to anchor catalytic residues and form an active-site pocket. To evaluate this workflow, we design TIM barrel enzymes with active sites for the Kemp elimination reaction. Four out of nine designs showed measurable activity, with the most active reaching a catalytic efficiency of $21,000 \text{ M}^{-1} \text{ s}^{-1}$ at pH 10. Crystal structures of this variant bound to a transition-state analogue confirmed the accuracy of the designed lid and identified a shift in the designed lid between bound and unbound forms, reminiscent of rearrangements in natural TIM-barrel enzymes. Using the X-ray structure of a lower-activity variant ($19 \text{ M}^{-1} \text{ s}^{-1}$), we applied ensemble-based design to optimize its active site, increasing catalytic efficiency to $32,000 \text{ M}^{-1} \text{ s}^{-1}$. This optimized enzyme also demonstrated high regioselective preference for the target substrate. These results demonstrate that *de novo* TIM barrels can be customized with bespoke substrate binding pockets supporting efficient catalysis, establishing a platform for building *de novo* enzymes from minimal protein scaffolds. We propose strategies to push catalytic efficiency further by expanding CANVAS and combining it with complementary approaches, and we describe the next frontiers for *de novo* enzyme design.

Acknowledgements

First, I would like to thank my supervisor, Dr. Roberto Chica. The guidance I have received throughout my undergraduate and masters projects has been invaluable.

I would also like to thank our project collaborators from the Höcker lab, in particular Birte Höcker and Julian Beck. This was a true collaborative project, and I greatly enjoyed our long-distance discussions.

To the current and past members of the Chica Lab, I am thankful for our friendship. You are what makes being a graduate student enjoyable. In particular, I would like to thank Niayesh Zarifi for the mentorship I received during my honours project and teaching me all I know about enzyme design.

Finally, I would like to thank my parents, David and Denise, my brother Robbie, and my girlfriend Alyssa for all the support throughout all my years as a student. Thank you for your belief in me and for sitting through many hours of presentations and posters on this project. Hopefully new projects with many more presentations to come!

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	viii
List of Figures	ix
List of Abbreviations	xii
Chapter 1. Introduction.....	1
1.1 <i>De novo</i> enzyme design	1
1.2 The first <i>de novo</i> enzyme	1
1.3 Traditional <i>de novo</i> enzyme design methods	3
1.3.1 Theozyme creation	4
1.3.2 Theozyme placement	4
1.3.3 Active-site packing	6
1.4 <i>De novo</i> enzymes designed with traditional methods	7
1.4.1 <i>De novo</i> Kemp eliminase	8
1.4.2 <i>De novo</i> retro-aldolase	9
1.4.3 <i>De novo</i> ester hydrolase	10
1.4.4 <i>De novo</i> Diels-Alderase	11
1.4.5 <i>De novo</i> Morita-Baylis-Hillmanase	12
1.5 Limitations of traditional <i>de novo</i> enzyme design	13
1.5.1 Neglect of protein stability and solubility	14
1.5.2 Conformational sampling inaccuracies	15
1.5.3 Search limitations and scoring inaccuracies	16
1.5.4 Oversimplification of enzyme mechanisms	16

1.6 Current <i>de novo</i> enzyme design methods	17
1.6.1 Employing pre-existing <i>de novo</i> scaffolds	18
1.6.2 Family-wide hallucination and fragment assembly.....	19
1.6.3 Crystallographic fragment screening.....	20
1.6.4 Structure generation around a catalytic array	21
1.6.5 Sequence-based methods for <i>de novo</i> enzyme design.....	27
1.7 Computational optimization of <i>de novo</i> enzymes	27
1.7.1 Ensemble-based design	28
1.7.2 Stability-guided sequence space refinement	28
1.7.3 Iterative <i>in silico</i> optimization.....	29
1.8 Lingering challenges motivate alternate <i>de novo</i> enzyme design approaches	30
1.8.1 The TIM barrel fold in enzymes.....	31
1.8.2 Functionalizing <i>de novo</i> TIM barrels	34
1.9 Thesis objectives: customizing minimal TIM barrels to design enzymes.....	35
Chapter 2. Customizing the structure of minimal TIM barrels to craft efficient <i>de novo</i>	
enzymes	37
Statement of contribution	37
Acknowledgements	37
2.1 Introduction	38
2.2 The CANVAS approach	39
2.3 Applying CANVAS to minimal TIM barrels.....	40
2.3.1 Minimal TIM barrel selection	40
2.3.2 Kemp eliminase design with CANVAS	41
2.4 Results	51
2.4.1 Experimental characterization of variants 1–9.....	51
2.4.2 Rescuing expression and stability with scaffold redesign.....	55

2.4.3 Confirming the lid and active-site features	64
2.4.4 Improving the efficiency of KempTIM4 by ensemble-based design	72
2.4.5 Molecular dynamics of KempTIMs	75
2.5 Discussion.....	79
2.6 Methods	81
2.6.1 Reagents and solutions.	81
2.6.2 Circular permutation of NovoTIM6-SB with inpainting	81
2.6.3 Computational enzyme design	82
2.6.4 Theozyme placement.....	84
2.6.5 Lid design	86
2.6.6 Active-site packing	86
2.6.7 Computational library design	87
2.6.8 Design filtering	88
2.6.9 Sequence optimization of variants 1–5 for enhanced stability	89
2.6.10 Ensemble-based design	89
2.6.11 Cloning and generation of constructs.	92
2.6.12 Protein expression and purification.	92
2.6.13 Far-UV CD spectroscopy	94
2.6.14 Size-exclusion chromatography-multi angle light scattering	95
2.6.15 Steady-state kinetics	95
2.6.16 Crystallization and structure determination.	96
2.6.17 Size-exclusion chromatography small-angle X-ray scattering (SEC-SAXS)	97
2.6.18 Molecular dynamics simulations	98
Chapter 3. Conclusions and outlook	100
3.1 Summary.....	100
3.1.1 Methodological insights from KempTIM characterization.....	102
3.2 Opportunities for CANVAS designs	104

3.2.1 Combining CANVAS with complementary design approaches.....	104
3.2.2 CANVAS with non-TIM barrel scaffolds.....	107
3.2.3 CANVAS in the design of multiple functionally relevant states.....	109
3.3 Outstanding challenges for <i>de novo</i> enzyme design.....	112
References.....	115

List of Tables

Table 2.1. Amino-acid sequences.....	47
Table 2.2. Post-purification culture yields.	54
Table 2.3. Kinetic parameters of KempTIMs.	59
Table 2.4. Kinetic parameters of KempTIM1 and KempTIM4b at various pH values.	63
Table 2.5. Crystallographic data collection and refinement statistics.....	65
Table 2.6. Amino-acid positions for theozyme placement during ensemble-based design	73
Table 2.7. Geometric constraints used to define catalytic contacts during computational design.	83
Table 2.8. Amino-acid positions for theozyme placement.....	84
Table 2.9. Geometric definitions for generation of transition-state poses off the side chain of catalytic base and H-bond donor.....	85
Table 2.10. Design filtering.	88
Table 2.11. Templates generated for ensemble-based design of KempTIM4.	90
Table 2.12. Triad model filtering during ensemble-based design.	91
Table 2.13. Filtering using Boltz-2 models during ensemble-based design.	91

List of Figures

Figure 1.1. Ester hydrolysis.	2
Figure 1.2. The traditional <i>de novo</i> enzyme design workflow.	3
Figure 1.3. Theozyme placement used in <i>de novo</i> enzyme design.	5
Figure 1.4. The generic computational protein design algorithm used in active-site packing.	6
Figure 1.5. Catalytic efficiencies of <i>de novo</i> enzymes designed with traditional methods.	7
Figure 1.6. Kemp elimination.	8
Figure 1.7. Retro-aldol reaction.	10
Figure 1.8. Diels-Alder cycloaddition.	12
Figure 1.9. Morita-Baylis-Hillman reaction.	12
Figure 1.10. Efficiencies achieved through experimental optimization of <i>de novo</i> enzymes designed with traditional approaches.	14
Figure 1.11. Efficiencies achieved by design, computational redesign, and experimental optimization of <i>de novo</i> enzymes.	18
Figure 1.12. Denoising diffusion probabilistic models (DDPMs) for protein structure generation.	22
Figure 1.13. Message-passing neural networks for protein sequence generation.	24
Figure 1.14. <i>De novo</i> enzyme design by generating a scaffold around a theozyme.	26
Figure 1.15. Structures of <i>de novo</i> enzymes created by catalytic motif scaffolding.	31
Figure 1.16. Natural and artificial TIM barrels.	32
Figure 1.17. Overview of thesis objectives.	36
Figure 2.1. The general CANVAS workflow.	39
Figure 2.2. The Kemp elimination reaction of 5-nitrobenzisoxazole.	40

Figure 2.3. Circular permutation of NovoTIM6-SB.	41
Figure 2.4. Structure and sequence filtering throughout the CANVAS pipeline.	42
Figure 2.5. Structures and topologies of minimal TIM barrels and CANVAS-designed variants.	45
Figure 2.6. Active sites of designed enzymes.	46
Figure 2.7. AlphaFold2 models.	49
Figure 2.8. Structural similarity of CANVAS-designed variants to natural proteins.	50
Figure 2.9. SDS-PAGE of expressible variants.	52
Figure 2.10. Biophysical characterization of expressible variants.	53
Figure 2.11. Variant 4 kinetic characterization.	55
Figure 2.12. SDS-PAGE of KempTIMs.	56
Figure 2.13. Biophysical characterization of expressible KempTIMs.	57
Figure 2.14. Michaelis–Menten plots of KempTIMs.	60
Figure 2.15. Characterization of KempTIM knockout mutants.	61
Figure 2.16. pH-rate profiles of KempTIM1 and KempTIM4b.	62
Figure 2.17. Crystal structures of KempTIM1 and KempTIM4.	66
Figure 2.18. Comparison of KempTIM4 experimental SAXS data with theoretical curves.	67
Figure 2.19. KempTIM1 crystal structures with and without 6-nitrobenzotriazole (6NBT).	69
Figure 2.20. Tautomeric forms of 6-nitrobenzotriazole.	70
Figure 2.21. Active-site pocket sizes.	71
Figure 2.22. Ensemble-based enzyme design pipeline.	72
Figure 2.23. Boltz-2 predictions of ensemble-based designs.	74
Figure 2.24. Backbone flexibility of select KempTIMs.	76
Figure 2.25. Catalytic contacts assessed by molecular dynamics.	78

Figure 2.26. Molecular dynamics-derived catalytic competency moderately correlates with experimental efficiency.	79
Figure 2.27. Atom definitions used in theozyme placement and design filtering.....	83
Figure 3.1. Efficiencies of <i>de novo</i> enzymes.	102
Figure 3.2. Proposed second iteration of the CANVAS workflow.	106
Figure 3.3. Generation of extensions on non-TIM barrel minimal <i>de novo</i> scaffolds.....	108
Figure 3.4. Using CANVAS to design enzymes with multiple functionally relevant states.	111

List of Abbreviations

5NBZ	5-nitrobenzisoxazole
6NBT	6-nitrobenzotriazole
6NBZ	6-nitrobenzisoxazole
CANVAS	Customizing amino-acid networks for virtual active-site scaffolding
CD	Circular dichroism
DDPM	Denoising diffusion probabilistic model
dRI	Differential refractive index
EC	Enzyme Commission
IPTG	Isopropyl- β -thiogalactoside
KT	KempTIM
LB	Lysogeny broth
MALS	Multi-angle light scattering
MBH	Morita-Baylis-Hillman
NMR	Nuclear magnetic resonance
NTF2	Nuclear transport factor 2
PDB	Protein Data Bank
PNPA	Para-nitrophenyl acetate
QM	Quantum mechanics
RIF	Rotamer interaction field
SAXS	Small-angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
TIM	Triose phosphate isomerase
TS	Transition state

Chapter 1. Introduction

1.1 *De novo* enzyme design

Enzymes exhibit unmatched catalytic efficiency, offering exceptional substrate selectivity alongside precise stereochemical control over product formation. They achieve this remarkable efficiency and precision while remaining environmentally friendly and operating under ambient conditions. These properties make enzymes attractive catalysts for academic research and industrial applications. Directed evolution¹ and rational design² have long been used to improve the efficiency, specificity, and stability of natural enzymes for target applications²⁻⁴. However, these methods require a baseline starting activity to screen/select upon, and their ability to achieve entirely novel function is limited by the vast protein sequence space. A general computational method to design *de novo* enzymes that exhibit high efficiency and specificity under defined conditions could unlock chemical processes that are currently inaccessible with existing catalysts.

1.2 The first *de novo* enzyme

The first enzyme computationally designed to transform a predefined substrate into a target product in a scaffold absent of activity was PZD2, an enzyme designed by Bolon and Mayo in 2001 to catalyze the hydrolysis of the para-nitrophenyl acetate (PNPA) ester into para-nitrophenol and acetate (Figure 1.1a)⁵. According to transition state (TS) theory^{6,7}, the rate of a reaction can be increased by reducing the energy barrier required to reach the least stable TS. In the absence of a TS for their target reaction, the design of PZD2 involved modelling a high-energy nucleophile–para-nitrobenzoyl tetrahedral intermediate from the proposed reaction mechanism. A histidine residue was selected to be the nucleophile, and an *E. coli* thioredoxin scaffold devoid of PNPA

hydrolase activity was scanned for positions to place the histidine and create an active-site pocket. To do this, the tetrahedral intermediate was tested at each position of the scaffold and amino acids in the surrounding region were allowed to sample various rotameric isomers (rotamers). If no suitable rotamers of the wild-type residue were identified, the residue was mutated to alanine instead. PZD2 showed a k_{cat} of 0.00046 s^{-1} and $k_{\text{cat}}/K_{\text{M}}$ of $2.7 \text{ M}^{-1} \text{ s}^{-1}$, orders of magnitude below the median natural enzyme turnover number (14 s^{-1}) and catalytic efficiency ($125,000 \text{ M}^{-1} \text{ s}^{-1}$)⁸. The use of a reaction intermediate instead of a TS coupled with the lack of an explicit packing step led to sub-optimal active-site complementarity⁹. Despite the low activity, this was the first study to demonstrate that computational methods could be used to introduce novel catalytic activity into a protein scaffold devoid of the target function.

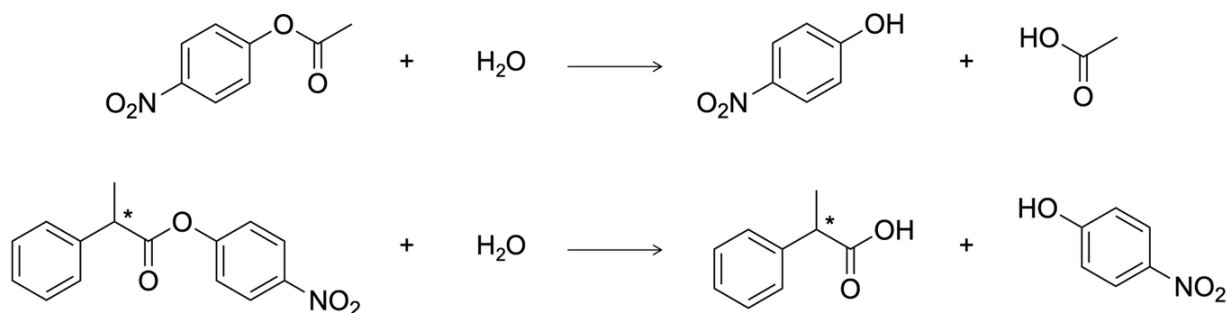


Figure 1.1. Ester hydrolysis. (a) Hydrolysis of the para-nitrophenyl ester into para-nitrophenol and acetate. (b) Hydrolysis of the para-nitrophenyl-2-phenylpropanoate ester into 2-phenylpropanoate and para-nitrophenol.

1.3 Traditional *de novo* enzyme design methods

Following the foundational work by Bolon and Mayo, the next challenge was creating a broadly applicable method for designing enzymes from scratch. *De novo* enzyme design adopts a structure-based approach, predicated on the close conceptual link between three-dimensional structure and function. This approach allows for the strategic selection of amino acid positions in a target protein scaffold to be optimized for TS stabilization. Consequently, the sequence space of designed libraries is enriched for the structural features necessary for catalysis³. In practice, early *de novo* enzyme design campaigns achieved this enrichment through computational workflows consisting of three general steps (Figure 1.2) centered around fitting an active site into an existing protein scaffold: (1) theozyme creation, (2) theozyme placement, (3) active site-packing.

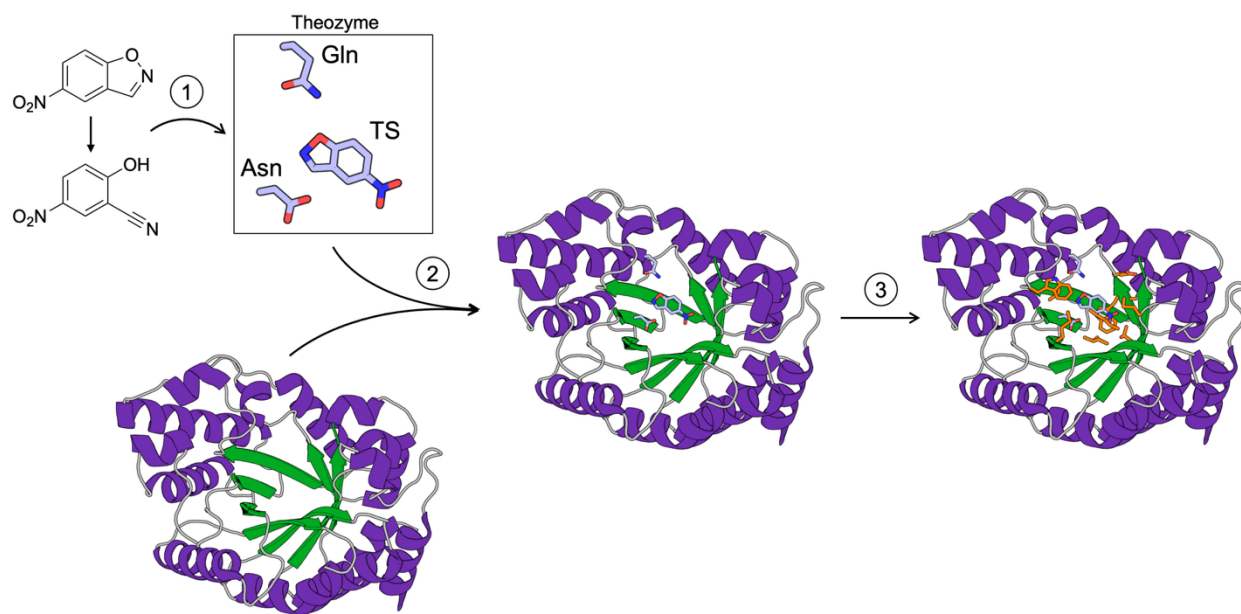


Figure 1.2. The traditional *de novo* enzyme design workflow. (1) Theozyme creation. A theozyme is determined from the transition state (TS) of the reaction mechanism using chemical intuition and quantum mechanical methods. (2) Theozyme placement. The theozyme is placed into a natural protein backbone. (3) Active-site packing. The sequence of the active site is optimized for TS stabilization.

1.3.1 Theozyme creation

Enzymes catalyze reactions by binding tightly to the TS and reducing the free energy barrier to this unstable intermediate^{10,11}. To create a protein capable of this, the idea of a theozyme, an array of functional groups about a TS, geometrically suitable for TS stabilization, was introduced¹². Starting from a proposed reaction mechanism, the theozyme components are derived from experimentally determined active sites or chemical intuition, and precise geometry is calculated using quantum mechanical (QM) approaches such as Hartree–Fock theory, density functional theory, and semi-empirical methods¹². The resulting theozyme contains the computationally optimized TS surrounded by functional groups of amino-acid side chains, representing catalytic residues oriented for TS stabilization. In downstream design protocols, the theozyme is represented by geometric constraints between select TS and catalytic residue atoms.

1.3.2 Theozyme placement

Once the theozyme is calculated, a protein scaffold that can accommodate its precise spatial constraints is identified. In this step, termed theozyme placement, one or more protein templates are explored for backbone positions to house the catalytic residues in the theozyme geometry. These templates, typically high-resolution protein structures derived from X-ray crystallography, with existing ligand-binding pockets and documented high expression in *E. coli*¹³, are retrieved from the Protein Data Bank (PDB)¹⁴. Theozyme placement is done with either a targeted ligand placement¹⁵ or geometric matching algorithm¹⁶ (Figure 1.3). Both methods involve pruning side chains within a cavity of the protein template and sampling various catalytic residue side-chain conformations from a rotamer^{17,18} or conformer¹⁹ library at user-specified positions in the cavity. The TS is built based on torsional, rotational, and translational degrees of freedom with respect to the position of a catalytic residue side-chain functional group. In targeted ligand placement, the

TS pose is generated from a single catalytic residue, often chosen by importance in the reaction mechanism. From the set of TS poses, each conformation of the other catalytic residues is evaluated for the desired catalytic contact. In geometric matching, TS poses are generated from each theozyme residue, and TS poses that overlap for all residues are selected. Both procedures find catalytic residue positions in the protein scaffold that are compatible with the QM-optimized TS and functional group geometry.

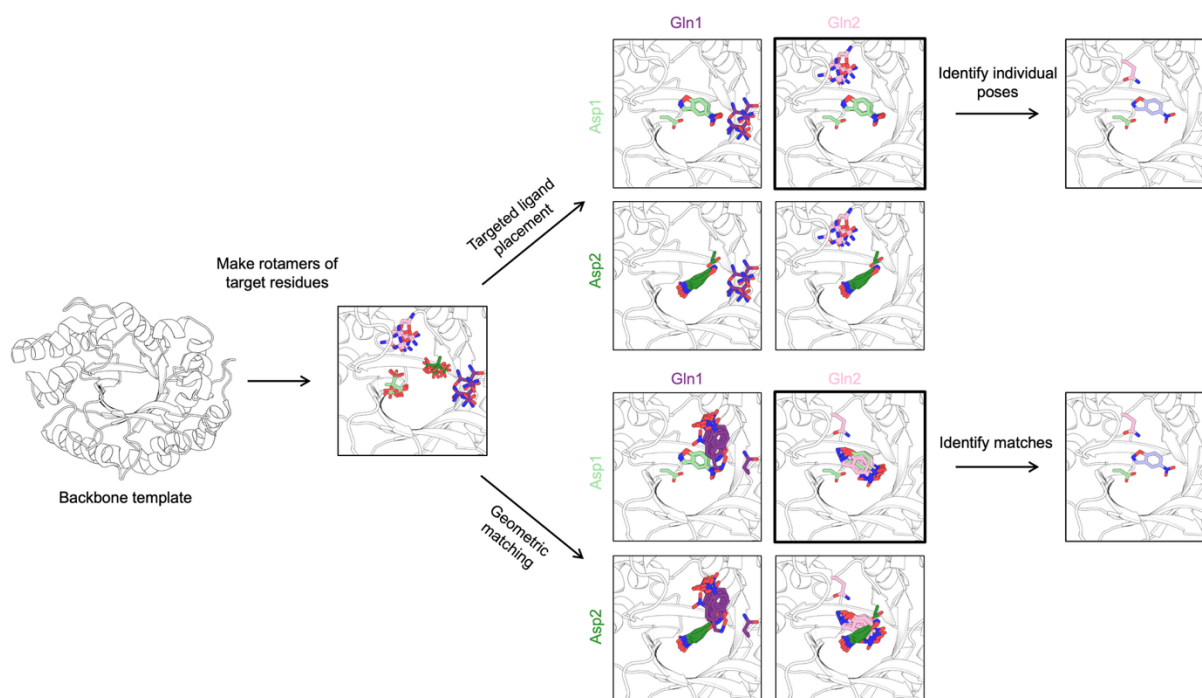


Figure 1.3. Theozyme placement used in *de novo* enzyme design. Starting from a backbone template, rotamers of all catalytic residues are generated. In this case, catalytic Asp and Gln residues are sampled at two positions each. In targeted ligand placement (top row), TS poses are built from each Asp rotamer, and Gln rotamers that can produce the desired contacts are selected, yielding the final placed theozyme. In geometric matching (bottom row), TS poses are built from all rotamers of each catalytic residue. Asp and Gln rotamers where TS poses match are selected, yielding the final placed theozyme. In this case, both targeted ligand placement and geometric matching produce the same result.

1.3.3 Active-site packing

After positions for the catalytic residues have been identified and the theozyme has been fit into the active site, surrounding amino acids are optimized for TS complementarity. Algorithms for this active-site packing step are largely adopted from structure-based computational protein design. These algorithms typically involve placing a library of rotamers of each allowed amino acid type at all design positions, calculating all inter- and intra-residue energies, ranking sequences based on the stability of their rotameric configurations, and identifying the lowest energy sequences that satisfy the design objective⁹. Following theozyme placement, active site-packing proceeds through these steps, with the design objective of maximizing non-bonding interactions with the TS (Figure 1.4). The resulting sequences can be further filtered or directly experimentally characterized.

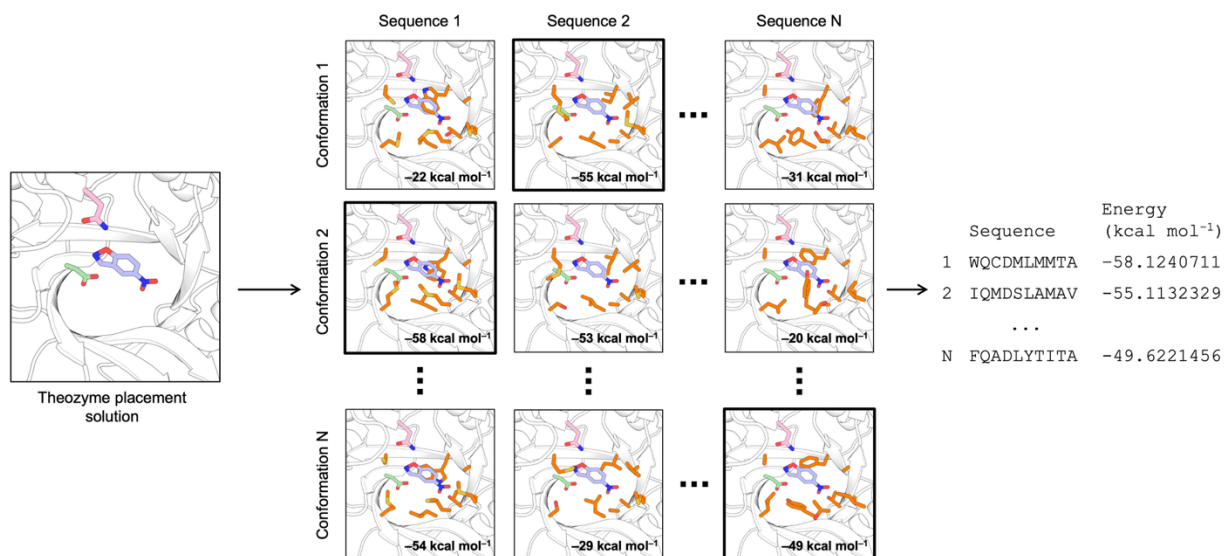


Figure 1.4. The generic computational protein design algorithm used in active-site packing. From a successful theozyme placement, user-specified positions in the backbone template are targeted for placement of side chain rotamer libraries of each allowed amino-acid type. The stability (energy) of each conformation is computed using a selected potential energy function. A search algorithm explores rotamer and sequence space to identify optimal sequences that yield the lowest energy values, producing a list of ranked sequences along with their structures. In this example, the single lowest energy rotameric configuration of each sequence is identified. However, if specified, a ranked list of rotameric configurations, rather than sequences, can be output instead.

1.4 *De novo* enzymes designed with traditional methods

The first campaign by Bolon and Mayo in 2001 was a critical proof-of-concept for *de novo* enzyme design. Since then, there have been many advancements in broadly applicable methods for *de novo* enzyme design. From 2008 to 2013, enzymes have been engineered for the ester hydrolysis, Kemp elimination, retro-aldol, Diels-Alder, and Morita-Baylis-Hillman (MBH) reactions (Figure 1.5), which have become model transformations for *de novo* enzyme design.

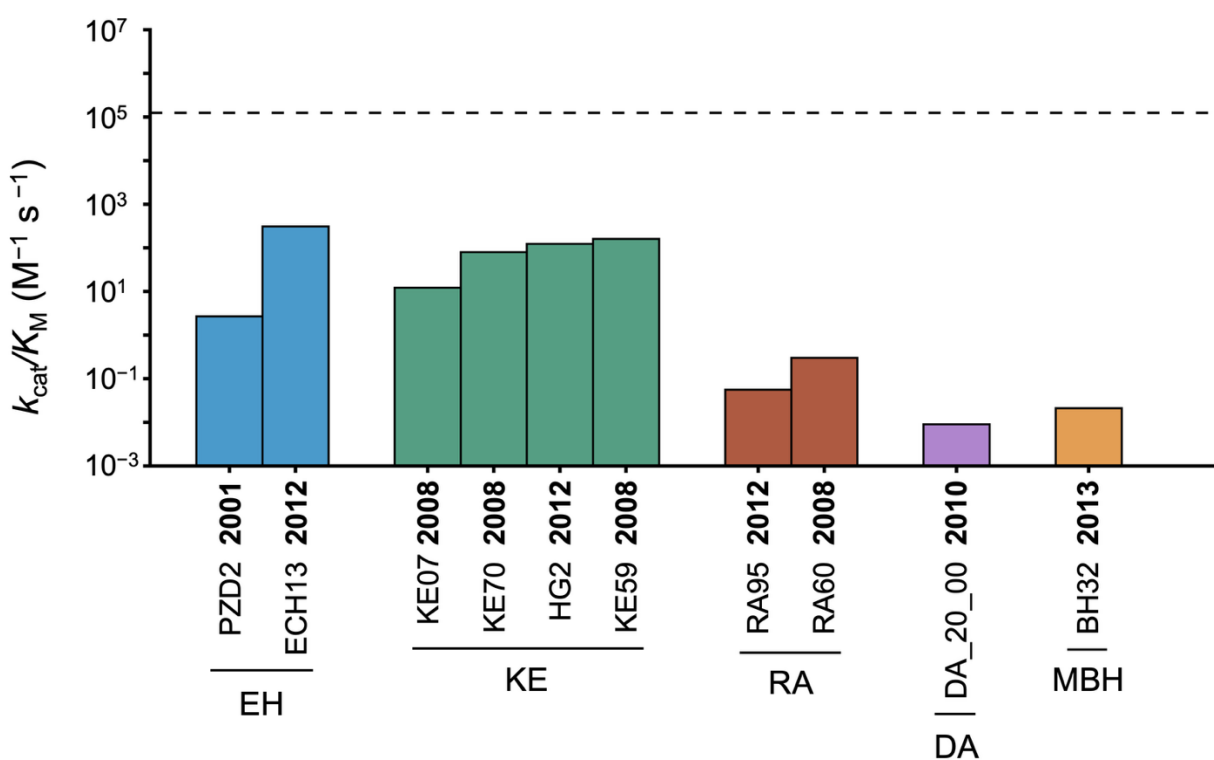


Figure 1.5. Catalytic efficiencies of *de novo* enzymes designed with traditional methods. Enzymes are grouped by reaction type and ordered by catalytic efficiency, with their name and publication year listed below each column^{5,13,20–25}. The dashed line indicates the median $k_{\text{cat}}/K_{\text{M}}$ of natural enzymes ($125,000 \text{ M}^{-1} \text{s}^{-1}$)⁸. Reaction types EH, KE, RA, DA, and MBH, corresponding to ester hydrolysis, Kemp elimination, retro-aldol condensation, Diels–Alder cycloaddition, and the Morita–Baylis–Hillman reaction, are shown in blue, green, brown, purple, and orange, respectively. For bimolecular reactions, $k_{\text{cat}}/K_{\text{M}}$ was calculated using the K_{M} for 4-carboxybenzyl trans-1,3-butadiene-1-carbamate or 4-nitrobenzaldehyde for DA_20_00 or BH32, respectively. Catalytic efficiency for BH32 was taken from Crawshaw *et al.*²⁶.

1.4.1 *De novo* Kemp eliminase

The Kemp elimination is a ring-opening reaction involving the deprotonation of an isoxazole by a general base, proceeding through a single TS to form a hydroxybenzonitrile product (Figure 1.6)²⁷. The simplicity of this reaction makes it widely used as a benchmark for *de novo* enzyme design. The first *de novo* Kemp eliminase was designed by R  thlisberger *et al.*¹³ in 2008, targeting 5-nitrobenzisoxazole as a substrate. Rather than modelling a reaction intermediate, they used QM-derived theozymes designed to stabilize the single TS. They attempted multiple theozyme configurations, with a glutamate or histidine base, a phenylalanine or tryptophan π -stacking residue, and a lysine or serine to stabilize the developing negative charge on the isoxazolic oxygen. Theozyme placement was done using the geometric matching¹⁶ algorithm (Figure 1.3) to search for suitable catalytic residue positions on dozens of high-resolution, experimentally derived protein scaffolds. Explicit active-site packing was done (Figure 1.4) to find a sequence that maximized TS complementarity, and designs were filtered based on deviation from theozyme geometry and TS binding energy. Eight of 59 designs tested were found to be active, with the highest efficiency reaching $160 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 1.5). Thus, R  thlisberger *et al.* demonstrated the first general approach for the computational design of enzymes for single-step reactions from calculated active site geometry.

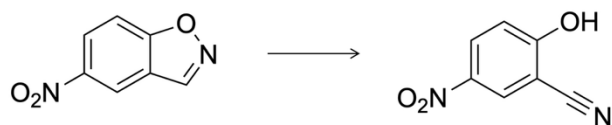


Figure 1.6. Kemp elimination. Kemp elimination of 5-nitrobenzisoxazole yielding 2-hydroxy-5-nitrobenzonitrile.

Also using a computed TS, Privett *et al.*²¹ applied the targeted ligand placement¹⁵ method to the *de novo* design of a Kemp eliminase. Targeting a natural xylanase as the protein scaffold, they identified positions in its active site that could support the catalytic residues: a glutamate or aspartate general base, an aromatic π -stacking residue, and a serine, threonine, or tyrosine residue to hydrogen-bond with the isoxazolic oxygen. The active site was then packed as described previously. However, their initial design, HG1, was inactive despite retaining the overall fold of the parent xylanase. To understand where the design failed, Privett *et al.* crystallized HG1 and found a highly solvent-exposed active site as well as considerable inaccuracy in active-site side-chain positions. Furthermore, molecular dynamics simulations from this structure revealed high active-site conformational heterogeneity. To exclude bulk solvent and generate a more hydrophobic active site, targeted ligand placement was performed in a deeper pocket of the xylanase scaffold. Following active site-packing, the variant HG2 was obtained with an efficiency of $120 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 1.5). To decrease heterogeneity, an active-site serine was mutated to a threonine, producing HG3, with an efficiency of $430 \text{ M}^{-1} \text{ s}^{-1}$, demonstrating the power in iterative design and that an initial “shotgun” approach to identify active variants is unnecessary.

1.4.2 *De novo* retro-aldolase

De novo enzymes were also designed to catalyze the retro-aldol condensation, a bond-breaking reaction where a β -hydroxy carbonyl compound decomposes into two smaller carbonyl compounds (Figure 1.7), proceeding through six steps. The multistep character of this reaction makes the design of a *de novo* retro-aldolase conceptually challenging compared to a Kemp eliminase. To do this, Jiang *et al.*²³ calculated a theozyme for each TS of the decomposition of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone to 6-methoxy-2-naphthaldehyde and acetone (Figure 1.7) and superimposed them to create a composite theozyme. Again, multiple sets of

catalytic residues were used in the absence of a known ideal catalytic motif, all containing a catalytic lysine to perform enamine catalysis through a Schiff base intermediate. The geometric matching algorithm was used to explore protein scaffolds for positions that could accommodate the catalytic residues, and the active site was packed around the composite TS. Out of 72 designs tested, 32 showed retro-aldolase activity. Despite having many active variants, the most active design, RA60, achieved an efficiency of $0.7 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 1.5) and rate enhancement of 10^4 , an order of magnitude lower than the rate enhancement achieved for the Kemp elimination¹³, demonstrating the challenges in designing novel enzymes for multistep reactions.

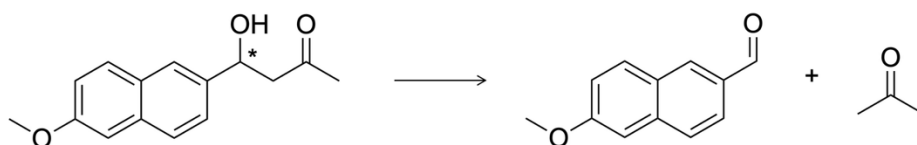


Figure 1.7. Retro-aldol reaction. Retro-aldol condensation of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone yielding 6-methoxy-2-naphthaldehyde and acetone.

In another design campaign, Althoff *et al.*²² used a more thorough design protocol with additional filtering steps to identify preorganized variants. Testing 42 new designs yielded 33 new retro-aldolases, a higher success rate than the campaign by Jiang *et al.* but with no improvement in activity.

1.4.3 *De novo* ester hydrolase

De novo design of an ester hydrolase was attempted again, this time with more complete active-site design targeting a different substrate (Figure 1.1b). In their TS, Richter *et al.*²⁰ included elements often found in natural hydrolytic enzymes^{28,29} such as a cysteine to perform a nucleophilic attack on the ester, a histidine for proton shuffling, and an explicit oxyanion hole consisting of a backbone amide, water molecule, or side-chain functional group to stabilize the

intermediate oxyanion. The design ECH13 showed an efficiency of $310 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 1.5), two orders of magnitude higher than that of PZD2⁵. Despite the improvement by including these additional active-site features, the efficiency remains three orders of magnitude lower than that of natural enzymes with similar catalytic motifs that hydrolyze similar substrates³⁰.

1.4.4 *De novo* Diels-Alderase

The first *de novo* enzymes catalyzing a bimolecular reaction were those designed by Siegel *et al.*²⁴ for the Diels-Alder reaction (Figure 1.8). The theozymes calculated for this reaction contained a glutamine or asparagine to accept a hydrogen bond from the diene carbamate, and a serine, threonine, or tyrosine to donate a hydrogen bond to the dienophile amide oxygen. A composite ligand was constructed similarly to Jiang *et al.*²³ but containing the TS along with the diene and dienophile substrates and product. Design was performed using geometric matching and active site-packing as done previously, yielding 84 designs, only two of which were active. The most active variant, DA_20_00, reached an efficiency far lower than previous *de novo* enzymes (Figure 1.5). Both the low success rate and low activity demonstrate the challenge in designing *de novo* enzymes for even single-step bimolecular reactions. Leveraging the concept of iterative redesign introduced by Privett *et al.*²¹, Siegel *et al.* improved the efficiency of DA_20_00 by two orders of magnitude through molecular dynamics-guided prediction of point mutations enhancing TS binding, followed by identification and combination of beneficial mutations. This improved variant showed high stereospecificity with an enantiomeric excess of 97% for the target product.

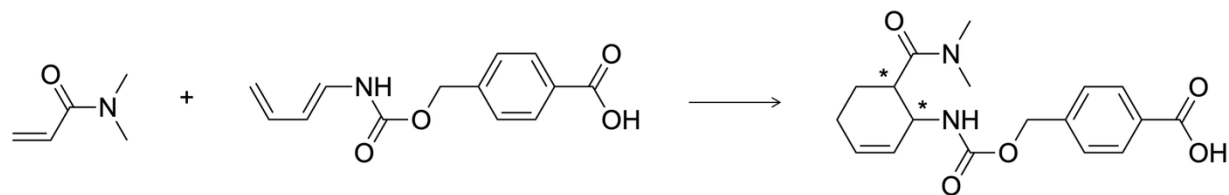


Figure 1.8. Diels-Alder cycloaddition. Diels-Alder cycloaddition of 4-carboxybenzyl-trans-1,3-butadiene-1-carbamate and N,N-dimethylacrylamide producing a substituted cyclohexene.

1.4.5 *De novo* Morita-Baylis-Hillmanase

The MBH reaction is a multistep bimolecular reaction, providing a stringent test for *de novo* enzyme design methods. Bjelic *et al.*²⁵ targeted the combination of 2-cyclohexen-1-one and 4-nitrobenzaldehyde (Figure 1.9), an MBH reaction with three TS. They designed a composite ligand combining the second TS and second intermediate. The theozyme included this composite ligand, a cysteine or histidine nucleophile, a lysine (for cysteine) or aspartate/glutamate (for histidine) to stabilize the nucleophile, an aromatic residue to π -stack with 4-nitrobenzaldehyde, a water molecule for proton shuffling, and backbone amide or polar side chains to form an oxyanion hole. Highlighting the challenge of this reaction, of the 48 enzymes tested, only two were active, and again the efficiency was below that of previous *de novo* designs and far below that of natural enzymes (Figure 1.5).

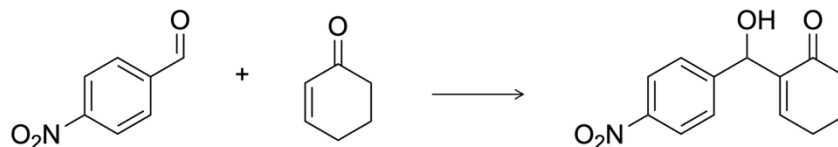


Figure 1.9. Morita-Baylis-Hillman reaction. Morita-Baylis-Hillman reaction between 4-nitrobenzaldehyde and 2-cyclohexenone yielding 2-(hydroxy(4-nitrophenyl)methyl)cyclohex-2-en-1-one.

1.5 Limitations of traditional *de novo* enzyme design

Traditional *de novo* enzyme design has yielded only modest efficiency, far below that of natural enzymes (Figure 1.5). Comparison with random libraries of identical amino-acid composition demonstrated that while design methods enrich for active enzymes, increasing hit rates from less than 1% to 50%, the level of efficiency achieved by random libraries is similar, and in some cases random libraries produced the highest efficiency variants³¹. Consequently, traditional *de novo* enzyme design presented a platform for creating baseline activity by enriching libraries, while downstream optimization remained necessary to improve efficiency. *De novo*-designed Kemp eliminases^{32–35}, retro-aldolases^{36,37}, Diels-Alderase³⁸, and MBHases²⁶ have all been optimized multiple orders of magnitude by directed evolution, leading to efficiencies approaching those of the median natural enzyme (Figure 1.10). However, designing enzymes with these efficiencies from scratch remains an important goal, as it would both provide superior biocatalysts and confirm our understanding of enzyme function. Achieving this requires addressing several inaccuracies and approximations in the traditional workflow such as the neglect of protein stability, inaccuracies in conformational sampling, limitations to search algorithms and scoring functions, and the oversimplification of enzyme mechanisms.

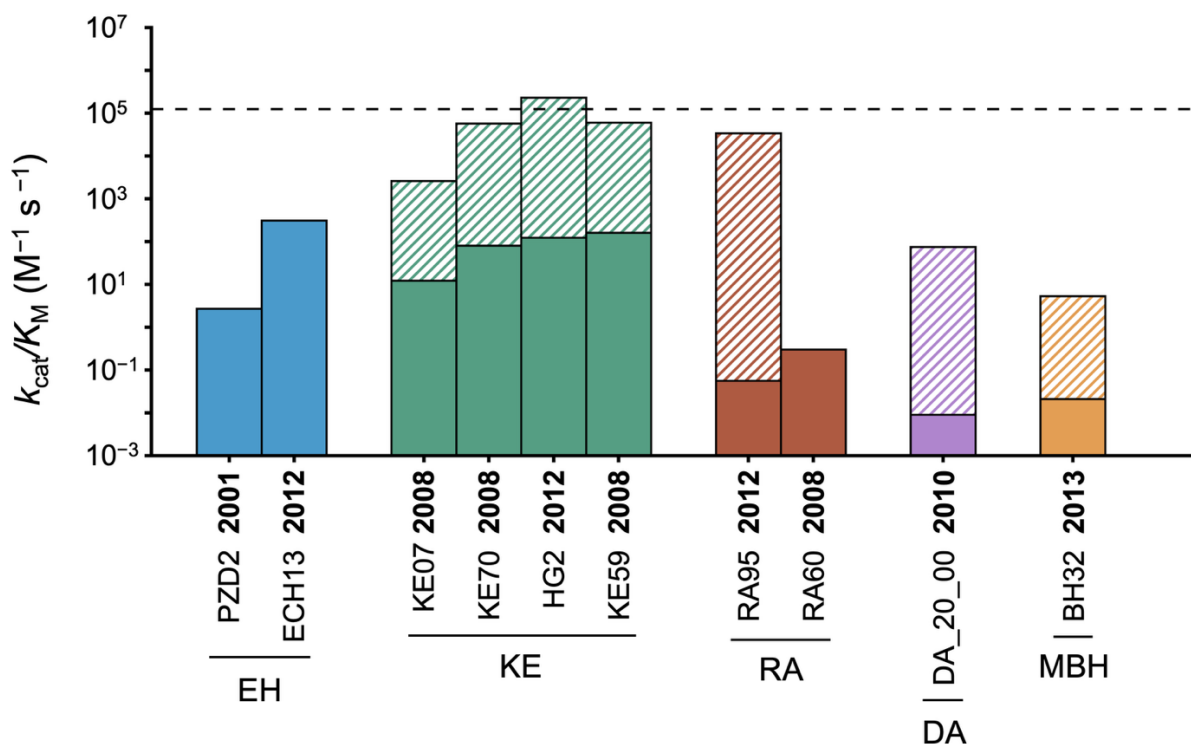


Figure 1.10. Efficiencies achieved through experimental optimization of *de novo* enzymes designed with traditional approaches. Enzymes are grouped by reaction type and ordered by catalytic efficiency, with their name and publication year listed below each column. Efficiencies of initial *de novo* designs^{5,13,20–25} are shown as solid-coloured bars, while those for experimentally optimized variants^{24,26,32–35} are shown as hashed bars. The dashed line indicates the median k_{cat}/K_M of natural enzymes ($125,000 M^{-1} s^{-1}$)⁸. Reaction types EH, KE, RA, DA, and MBH, corresponding to ester hydrolysis, Kemp elimination, retro-aldol condensation, Diels–Alder cycloaddition, and the Morita–Baylis–Hillman reaction, are shown in blue, green, brown, purple, and orange, respectively. For bimolecular reactions, k_{cat}/K_M was calculated using the K_M for 4-carboxybenzyl trans-1,3-butadiene-1-carbamate or 4-nitrobenzaldehyde for DA_20_00 or BH32, respectively. Catalytic efficiency for BH32 was taken from Crawshaw *et al.*²⁶.

1.5.1 Neglect of protein stability and solubility

Despite the ability of traditional workflows to improve success rates compared to random, success rates are still relatively low compared to what is desired. A significant contribution to the low hit rates of computational enzyme design is the disregard for protein stability and solubility. Natural proteins are only marginally stable³⁹, and thus mutations to natural protein scaffolds accumulated during design are likely to result in misfolding, insolubility, and poor expression^{20,24}. Thus, many promising variants in *de novo* enzyme design campaigns do not show soluble expression, which greatly reduces success rates^{24,40}. In some cases, starting from thermostable

protein scaffolds may improve stability^{23,40}, but this greatly limits the number of structures that can be used for design. Instead, the low solubility success rate is compensated by testing many variants in a “shotgun” approach.

1.5.2 Conformational sampling inaccuracies

Catalytic efficiency is highly sensitive to the positioning of catalytic residues. A deviation from ideal geometry of just 1 Å can lead to a 10⁴-fold decrease in activity, and catalysis by natural enzymes involves precision of 0.1 Å or better^{34,41}. During theozyme placement, precision is poor due to the discrete nature of rotamer^{17,18} and conformer¹⁹ libraries. Furthermore, the crystallographic protein structures used to derive side chain conformations in either library type have resolutions of 1 Å or worse¹⁸. This limits the search for side chain conformations that satisfy the theozyme requirements during design. Scaffolds used for design are single, fixed structures derived from natural proteins. These templates are highly evolved for their target functions and may not be perfectly suited to the activity of interest. Furthermore, this fixed-backbone approximation ignores the flexibility and motion of proteins in solution. Productive catalytic contacts often cannot be made in all backbone conformations, and high-energy states may be functionally essential^{42–44}. Combining fixed, natural enzyme backbones with sequence optimization using discrete side chains leads to failure in finding ideal catalytic residue positions. Indeed, perfect matches were not obtained in the theozyme placement steps of early *de novo* enzyme design campaigns⁴⁰. This results in the use of tolerance values, allowing deviation from the ideal geometry. Furthermore, mutations acquired during design may alter the backbone conformation and protein dynamics in unpredictable ways, representing another challenge not considered during design.

1.5.3 Search limitations and scoring inaccuracies

An exhaustive search through all 20 amino acids at all active site positions of a *de novo* enzyme is impossible, even when the sequence space is refined based on knowledge of the optimal reaction conditions. To navigate this, sequence optimization requires random sampling algorithms such as simulated annealing⁴⁵ to identify sequence minima, limiting the sequence space that can be explored^{46,47}. This is compounded by inaccuracies in sequence scoring. Potential energy functions used in theozyme placement and active-site packing inaccurately treat non-bonded interactions^{9,48,49}. In particular, long-range electrostatics are not well accounted for, despite the major contribution of electrostatic effects to catalysis⁵⁰.

1.5.4 Oversimplification of enzyme mechanisms

The modest activity seen in *de novo* enzyme design is partly attributable to the reductionist view of the catalytic cycle. For single-step reactions, design focuses on the sole TS of the reaction, while for multistep reactions the rate-limiting TS or a composite TS is modelled. However, this neglects substrate binding, conformational exchange throughout the reaction, and product release. These steps, each of which can be rate-limiting^{51–55}, are coordinated by natural enzymes for efficient catalysis^{56–62}. In multistep reactions, the focus on a single or composite TS does not consider the continuous path taken through each intermediate and fails to design preorganization at each TS²⁸. The optimal active site arrangement for one TS may be suboptimal for others⁶³. Furthermore, the role of solvent is largely ignored; while water is excluded from active sites to minimize the desolvation barrier of substrate binding²¹, it may also play a role in accelerating steps such as product release⁶⁴. Mechanistic oversights are exacerbated by the inaccuracies present in theozyme creation. Approximations in QM methods⁶⁵ and subjective selection of catalytic residues

limit this fundamental step of enzyme design. Consequently, directed evolution often finds superior catalytic arrangements^{34,36,37}, even if multiple are evaluated during the design process²³.

1.6 Current *de novo* enzyme design methods

Recent *de novo* enzyme design campaigns attempt to address failures of traditional approaches, particularly protein stability and conformational sampling. Novel protein scaffolds have been adapted to design enzymes, providing simple, highly stable templates compared to the highly specialized scaffolds of natural enzymes. Alternative approaches have continued to harness the complex topologies of natural scaffolds by diversifying their backbones prior to theozyme placement. Although these approaches have found success, the most attention has been given to the *de novo* design of protein backbones tailored to the theozyme. Driven by recent improvements in machine learning, a considerable number of *de novo* enzyme design campaigns have used generative models to design enzyme backbones and sequences entirely from scratch. In this way, the backbone can be crafted around the ideal theozyme geometry rather than having to search for a suitable pocket. These approaches have led to multiple orders of magnitude improved efficiency compared to traditional methods (Figure 1.11).

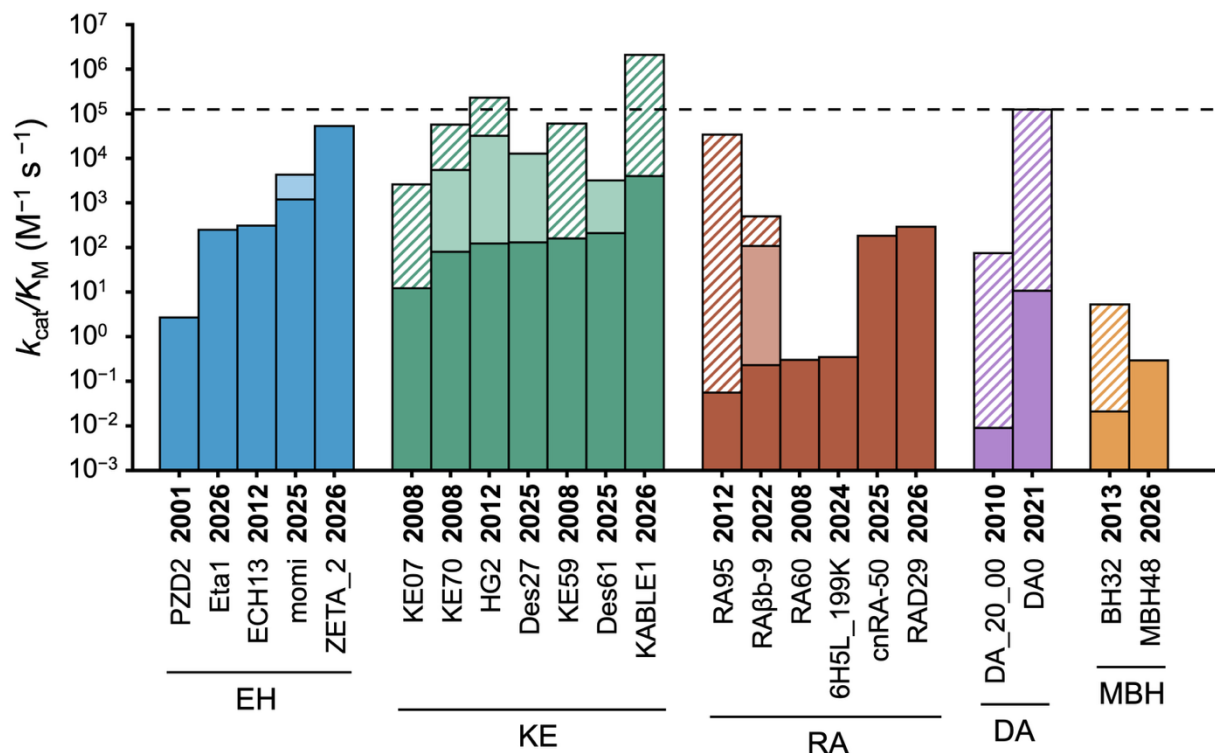


Figure 1.11. Efficiencies achieved by design, computational redesign, and experimental optimization of *de novo* enzymes. Enzymes are grouped by reaction type and ordered by catalytic efficiency, with their name and publication year listed below each column. Efficiencies of initial *de novo* designs^{5,13,20–25,66–74} are shown as solid-coloured bars, those for computationally redesigned variants^{67,69,71,75} are shown as lighter-coloured bars, and those for experimentally optimized variants^{24,26,32–35,70,71,74} are shown as hashed bars. The dashed line indicates the median k_{cat}/K_M of natural enzymes ($125,000 M^{-1} s^{-1}$)⁸. Reaction types EH, KE, RA, DA, and MBH, corresponding to ester hydrolysis, Kemp elimination, retro-aldol condensation, Diels–Alder cycloaddition, and the Morita–Baylis–Hillman reaction, are shown in blue, green, brown, purple, and orange, respectively. For bimolecular reactions, k_{cat}/K_M was calculated using the K_M for 4-carboxybenzyl trans-1,3-butadiene-1-carbamate, azachalcone, or 4-nitrobenzaldehyde for DA_20_00, DA0, or BH32 and MBH48, respectively. Catalytic efficiency for BH32 was taken from Crawshaw *et al.*²⁶.

1.6.1 Employing pre-existing *de novo* scaffolds

De novo α -helical proteins are structurally well-understood^{76,77} and have many beneficial properties for industrial biocatalysis, including a small size⁷⁸, modular architecture⁷⁹, elevated thermostability⁸⁰, strong organic solvent tolerance⁸¹, and high designability^{77,82}. Thus, these proteins may address the lack of soluble expression and stability that plagued previous design campaigns. Enzymes for the retro-aldol reaction⁸³, Diels–Alder reaction⁷⁴, sulfoxidation⁸⁴, cyclopropane synthesis^{85,86}, Si–H insertion⁸⁶, photocatalysis⁸⁷, olefin metathesis⁸⁸, and metal-

dependent hydrolysis^{89,90} have been designed starting from *de novo* helical bundles or other structurally simple helical scaffolds. These scaffolds can even be truncated and rebuilt to accommodate larger substrates⁸⁶ or improve substrate specificity⁹⁰. Furthermore, this approach has been extended to *de novo* proteins adopting other folds such as the β -barrel⁷¹. Designed retroaldolases (RA β b-9⁷¹ and 6H5L_199K⁸³) and Diels-Alderases (DA0⁷⁴) displayed equal or better efficiency than those designed by traditional methods (Figure 1.11), but required a fraction of the experimental effort. Thus, adapting a *de novo* scaffold presents a useful approach to yield a stable and modestly efficient enzyme.

1.6.2 Family-wide hallucination and fragment assembly

For many reactions, a suitable binding pocket may not be found in any existing scaffold, natural or artificial. However, many sub-optimal pockets may be able to satisfy catalytic contacts with minor backbone adjustments. For example, proteins adopting the nuclear transport factor 2 (NTF2)-like fold possess well-defined pockets with potential for high TS interactions, yet their backbones may not align perfectly with a target theozyme geometry. To expand the structural diversity in such folds, a family-wide hallucination approach⁹¹ can be harnessed, where loop and variable region sequences are randomized, followed by iterative structure prediction and further sequence diversification. A rotamer interaction field (RIF) is then docked into this library of diversified scaffolds. RIFs consist of a set of residue side chains making favorable interactions with an ensemble of ligand conformations⁹². The RIF approach is more exhaustive and less computationally expensive than either geometric matching or targeted ligand placement. Retroaldolase cnRA-50 designed with this approach demonstrated an efficiency over two orders of magnitude above variants created with traditional approaches (Figure 1.11)⁷², although 320 variants were tested. Similarly, luciferases⁹³ designed with this approach reached efficiencies of

1,400 M⁻¹ s⁻¹, but this required testing 7648 variants to identify only three active luciferases. Furthermore, this efficiency came primarily from excellent productive binding of the substrate (K_M = 18 μ M), rather than high turnover.

In a related approach, Kemp eliminases were designed by modular combinatorial assembly of fragments from thousands of homologous natural TIM barrel scaffolds⁶⁹. A method termed PROSS⁹⁴ was employed to identify stabilizing mutations for these new scaffolds, followed by theozyme placement and active-site packing. Filtering was performed with a “fuzzy-logic” multi-objective function to simultaneously achieve low active-site energy, high catalytic residue desolvation, and accurate theozyme geometry. Despite the added complexity of the method, the most active of 73 designs, Des27 and Des61, showed efficiency comparable to what was achieved by traditional methods (Figure 1.11).

1.6.3 Crystallographic fragment screening

A conceptually interesting approach was proposed by Chen *et al.*⁷⁰, where a crystallographic fragment screening method is used to determine a suitable active-site location in a protein scaffold prior to *de novo* enzyme design. To design an enzyme with this approach, small-molecule fragments resembling the TS of a target reaction are soaked into crystals of a generalist binding protein. Binding fragments are identified by electron density difference maps, and the TS analogues are modelled in the protein. These models can then be used for *de novo* enzyme design by finding positions where the catalytic residues can make the required contacts, followed by active-site packing. Chen *et al.* used this approach to design KABLE1, a Kemp eliminase with an efficiency (3,800 M⁻¹ s⁻¹) higher than that of previous designed Kemp eliminases with just 10 designs tested (Figure 1.11)⁷⁰. Site-saturation mutagenesis and nuclear magnetic resonance (NMR)-guided directed evolution⁹⁵ were employed to improve this efficiency to 2,500,000 M⁻¹

s^{-1} , an order of magnitude beyond that of the median natural enzyme⁸. While other approaches rationally place active sites²¹, this step is implicit in the crystallographic fragment screening method. Instead, active site location is guided by existing TS analogue binding affinity in the target scaffold, providing a clear and experimentally supported starting point for design.

1.6.4 Structure generation around a catalytic array

Despite the success of adapting *de novo* proteins or diversified scaffolds, backbones created entirely from scratch could more generally support any arbitrary catalytic array. Recently, many protein structure generative models have been developed^{66,96–117}. Of these, denoising diffusion probabilistic models (DDPMs)¹¹⁸ have been the most widely applied to *de novo* enzyme design. DDPMs, widely used in image and video generation¹¹⁸, are a class of generative models that create new samples by learning to reverse random noise added to training examples. Training these models is done in two phases (Figure 1.12a): the diffusion process and the reverse diffusion process. In the diffusion process, training examples are iteratively corrupted with reversible random noise. In the reverse diffusion process, a neural network is trained to reverse the noise added at each step. Thus, DDPMs capture a target distribution by learning how to map a random distribution to a dataset of interest¹¹⁹. Once the model is trained, it can create new samples that are similar to examples found in the original dataset. For protein structure generation, these models learn to recover orientations and Cartesian coordinates of residues or atoms corrupted with noise (Figure 1.12a). To then generate a new protein structure, random residue coordinates and orientations are initialized, and the DDPM makes a prediction. Residue coordinates are updated with a noised version of this structure, which is then used as input for the next prediction, eventually converging after many iterations (Figure 1.12b). Thus, a well-trained DDPM can generate novel protein backbones from random noise¹²⁰.

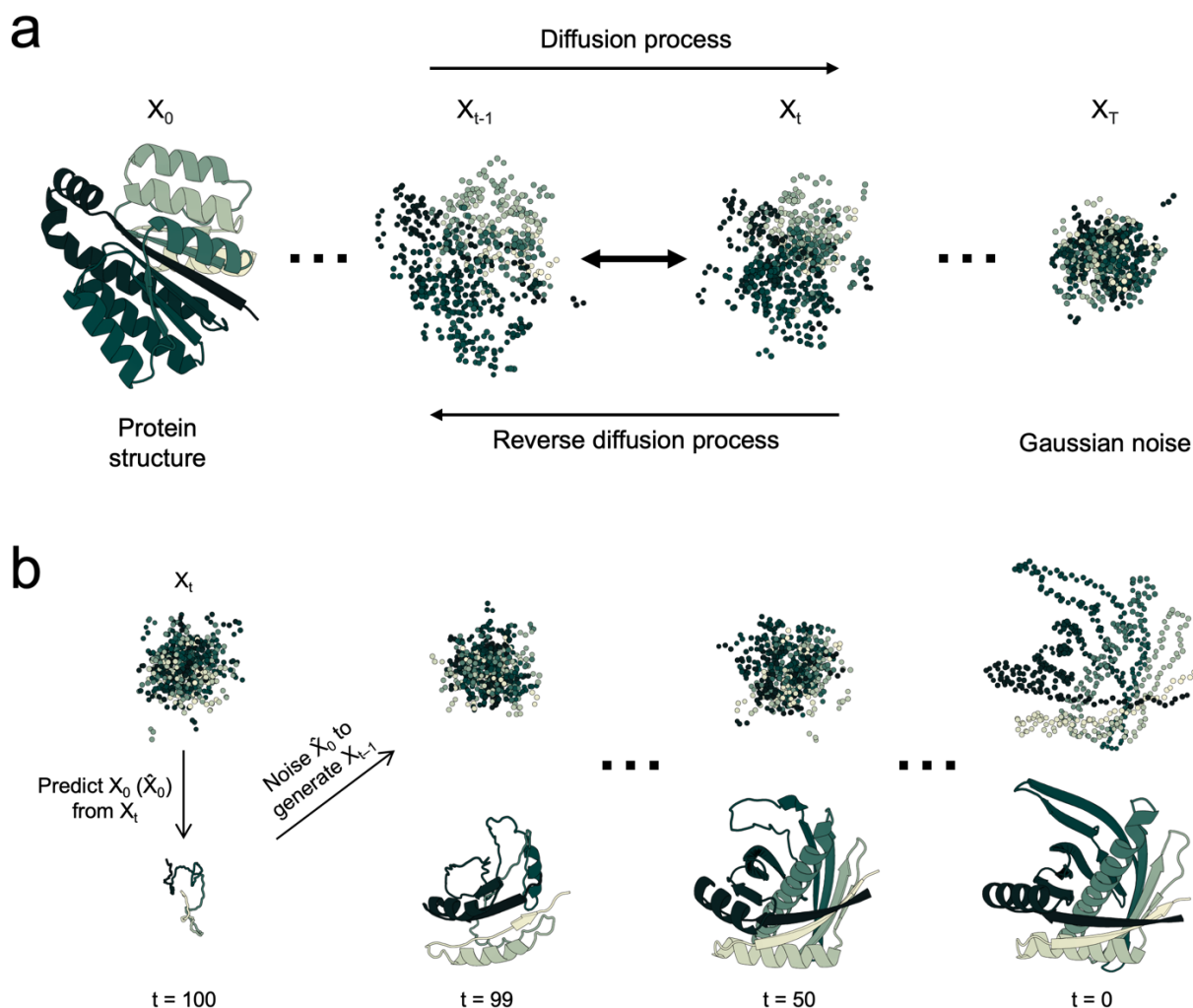


Figure 1.12. Denoising diffusion probabilistic models (DDPMs) for protein structure generation. Forward and reverse diffusion processes in denoising diffusion probabilistic models for protein structure generation. In the forward process, random noise is iteratively added to the training examples over T steps, referred to as timesteps. Timestep X_T represents the protein structure once completely corrupted with noise. In the reverse process, a neural network is trained to predict a protein structure, X_{t-1} , from a noised structure of X_t at each timestep. Three dots indicate states separated by multiple timesteps, while arrows indicate states separated by a single timestep. (b) Generating a novel protein structure with a DDPM. Random coordinates are initialized to generate the first noisy input (X_t), and the model makes a prediction ($\hat{X}_0$). This prediction is noised to generate X_{t-1} , and a structure is predicted again. While little structural detail is apparent in the noised samples (X_t) of early timesteps, the predictions ($\hat{X}_0$) become closer to realistic protein structures as the timestep progresses. Figures adapted from Chu *et al.*¹¹⁹ and Watson *et al.*¹¹¹.

Structure generative models have found many applications in protein design with the help of structure-conditioned sequence generative models. To generate sequences given a protein backbone, these models sample from a learned amino-acid probability distribution for each target position. This sampling is done autoregressively, so predictions at previously decoded positions can inform the next positions¹²⁰. Of the sequence generative models that have been developed^{100,121–139}, models adopting the message-passing neural network (MPNN)¹⁴⁰ framework have been the most widely applied to *de novo* enzyme design. These models represent a protein as an attributed graph, where node features capture local geometry such as backbone dihedral angles, while edge features encode relative inter-residue relationships including backbone atom distances, spatial orientations, and a positional encoding⁹⁷. Amino acid identities are predicted one position at a time in a sequential (N- to C-terminal)⁹⁷ or random¹³⁸ order (Figure 1.13). The computational efficiency of these models, coupled with the fact that their sequences are predicted to fold into the target structures with high confidence¹³⁸, has led to their widespread adoption over physics-based methods when designing *de novo* backbones. Thus, structure and sequence generative models are highly complementary; the former generates novel backbones, while the latter produces sequences that encode these backbones. While many structure generative models can create all protein atoms¹⁰⁷, designing with a sequence generative model often yields improved *in silico* success⁹⁸.

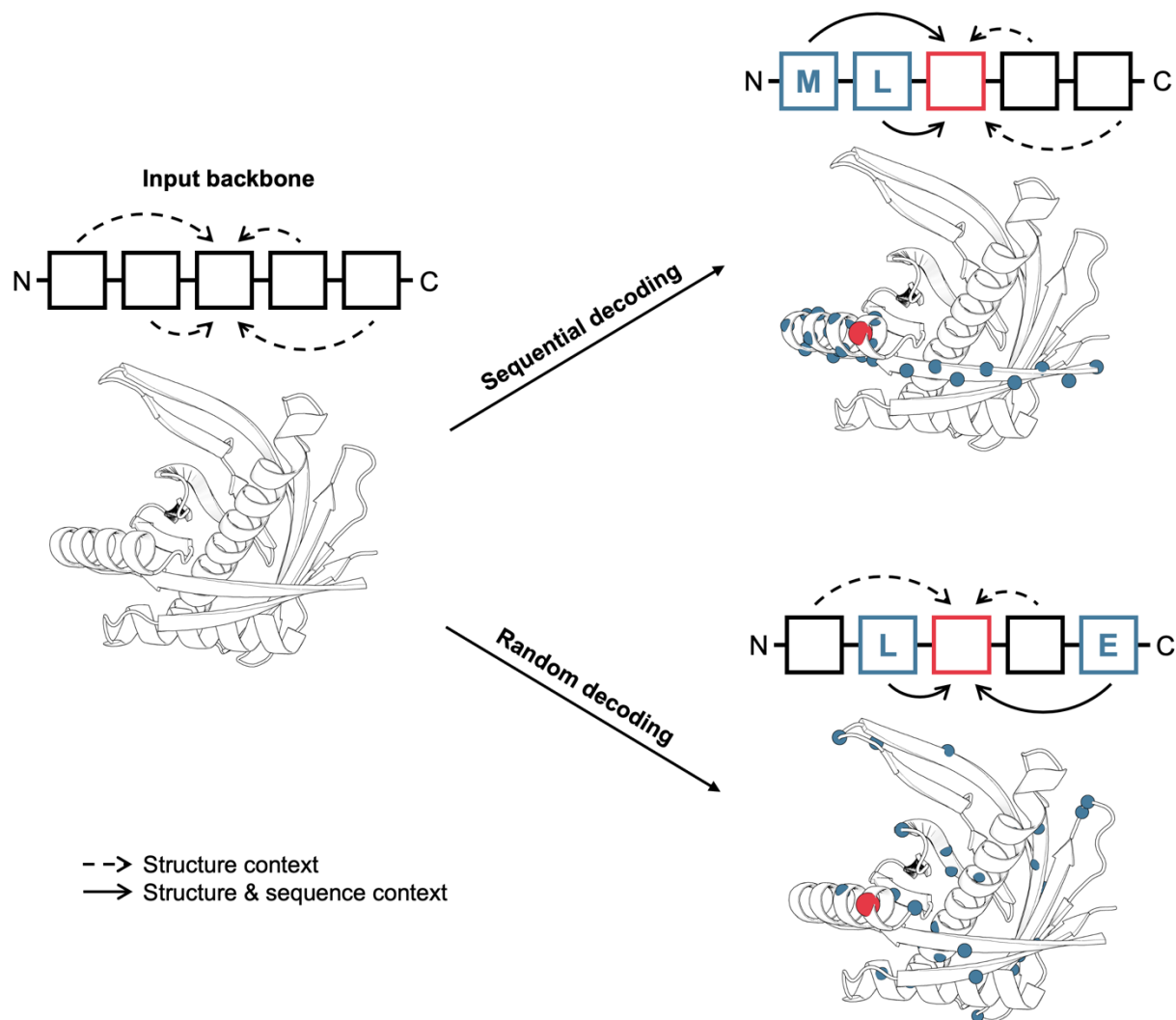


Figure 1.13. Message-passing neural networks for protein sequence generation. When a backbone is passed as input, node and edge features are extracted from each position, creating an internal representation of this 3D structure. The sequence is then predicted either sequentially from N- to C-terminus (top row) or randomly (bottom row). Due to the autoregressive nature, the first amino-acid prediction uses structure context alone. Subsequent predictions then use previously decoded amino acids, along with the structure, as context. In the diagram and protein structure, the current position is shown in red, while previously decoded amino-acid positions are shown in blue. Figure reproduced from Ingraham *et al.*⁹⁷.

These workflows have found applications in *de novo* enzyme design, where, instead of searching for a suitable scaffold to accommodate the theozyme, a protein backbone can be created around the ideal theozyme geometry. To do this, catalytic residues are fixed in space, and the protein is generated around them, known as catalytic motif scaffolding. Generation of custom

backbones forgoes the need for a theozyme placement step, and thus the theozyme geometry does not require tolerance values. The use of a natural protein scaffold with a single, fixed backbone conformation no longer limits the ability to target the ideal catalytic geometry. These *de novo* proteins are also highly soluble and stable¹¹¹, addressing another limitation of traditional *de novo* enzyme design campaigns.

Generative models have enabled the emergence of novel workflows for *de novo* enzyme design, broadly proceeding through the steps outlined in Figure 1.14. Following theozyme creation and catalytic motif scaffolding, sequences predicted to be compatible with the backbone are produced. To filter these sequences, a structure prediction tool such as AlphaFold2¹⁴¹ or ESMFold¹⁴² is employed, and high-confidence predictions that are similar to the generated structure can be used as input for active-site packing. Central to this workflow is the catalytic motif scaffolding step, which dictates how catalytic residues are positioned within the generated scaffold. Early structure generative models required that the catalytic residue backbones of an input motif occupy pre-determined locations in space^{107,111}. An inverted rotamer tree can be used to sample multiple backbone locations relative to the fixed side chains. However, for a many-residue theozyme, exhaustive structure generation from all inverted rotamer combinations becomes computationally intractable, and many rotamer combinations will clash anyway when scaffolded. To overcome this, Braun *et al.*⁷³ and Lauko *et al.*⁶⁷ proposed grafting idealized helices onto each inverted rotamer prior to structure generation. These larger input motifs can be filtered early for sterically incompatible rotamer combinations, and motif scaffolding success rates greatly improve. Recent models, however, do not require specification of the input motif backbone location and can generate scaffolds from functional group alone^{66,98,117}.

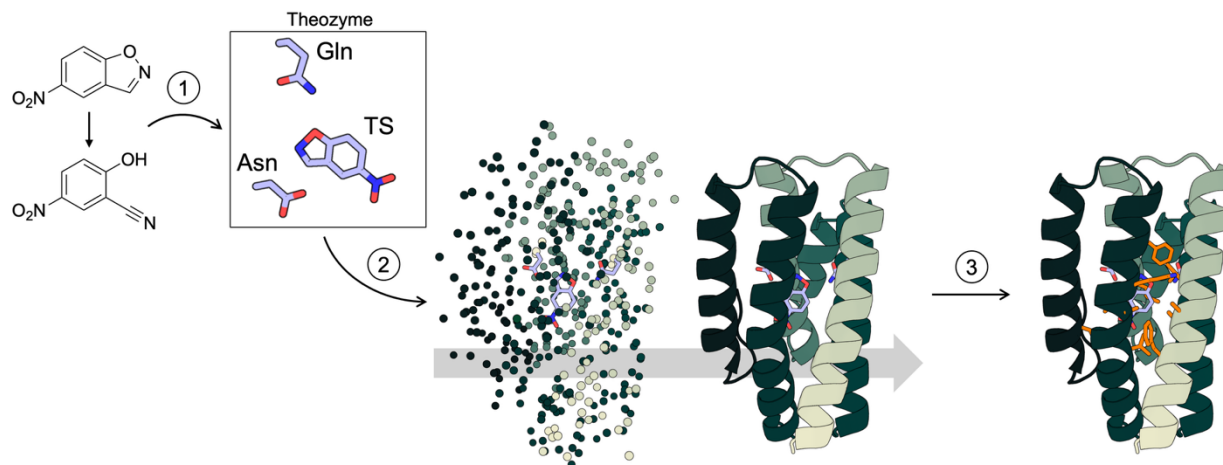


Figure 1.14. *De novo* enzyme design by generating a scaffold around a theozyme. (1) Theozyme creation. A theozyme is determined from the transition state (TS) of the reaction mechanism using chemical intuition and quantum mechanical methods. (2) Catalytic motif scaffolding. A protein scaffold is generated around the theozyme in a process termed catalytic motif scaffolding. (3) Active-site packing. The sequence of the active site is optimized for TS stabilization.

The use of predicted structures for active-site packing rather than those determined experimentally can introduce inaccuracy. To mitigate this, improvements to the structure generation and active-site packing steps have been implemented. It has been noted that the ability of designed sequences to encode the target backbone geometry improves through iterative sequence design and structure prediction¹⁴³. Thus, Braun *et al.*⁷³ employed a backbone refinement protocol integrating this concept. In this iterative approach, predicted structures are relaxed under geometric constraints towards the input catalytic motif, followed by sequence design and subsequent structure prediction. Alternatively, novel structures can be diversified to improve structure prediction success. In the approach by Kim *et al.*⁶⁸, generated structures sample loop extensions of various lengths, expanding backbone diversity. Once high-confidence predicted structures are obtained, active-site packing is achieved by designing sequences with generative models which are evaluated with physics-based rotamer optimization^{67,68,73}. Through these advancements, structure generation around a catalytic array has produced retro-aldolases^{66,73}, MBHases⁷³, and ester hydrolases^{67,68,98} that demonstrate efficiency orders of magnitude higher

than enzymes created by traditional *de novo* enzyme design for the same reactions (Figure 1.11). Notably, ester hydrolase ZETA_2 achieves efficiency approaching that of the median natural enzyme⁶⁸. Furthermore, generating structures around cofactors has yielded novel enzymes for cyclopropane synthesis, B–H insertion, and C(*sp*³)–H alkylation¹¹⁴.

1.6.5 Sequence-based methods for *de novo* enzyme design

The conceptual connection between active-site geometry and enzyme function has resulted in successful structure-based approaches for *de novo* enzyme design. However, purely sequence-based approaches have also been applied to enzyme design. By conditioning or fine-tuning on functional classifications or enzyme families, language models can produce a novel sequence for a target reaction type^{120,144}. Using these models, a variety of enzymes have been designed that catalyze natural reactions, achieving activity near their most efficient natural counterparts^{145–148}. Although this approach can generate efficient enzymes for certain functional classifications, they struggle to generate enzymes for predefined chemical transformations in enzyme classes with broad substrate scopes¹⁴⁶, and thus are not yet applied to the design of enzymes for novel reactions.

1.7 Computational optimization of *de novo* enzymes

As for traditional design, experimental optimization remains the most powerful platform to achieve high catalytic activity^{70,93}. However, directed evolution is resource-intensive, may involve time-consuming screening methods, and struggles to navigate rugged fitness landscapes¹⁴⁹. Consequently, various approaches have been developed to improve enzymes without experimental optimization. These methods largely build upon the concept of iterative design introduced by Privett *et al.*²¹, which seeks to investigate and correct the failures of first-round designs.

1.7.1 Ensemble-based design

Traditional *de novo* design makes use of a single, fixed backbone which may not represent catalytically competent sub-states for a given reaction. An effective computational approach for enzyme optimization that addresses the fixed-backbone approximation is ensemble-based design. This approach generates a structural ensemble from a low-activity enzyme and performs enzyme design calculations from each ensemble member, identifying those where optimal catalytic contacts can be made. In an early example of this approach, the design model of a low-activity enzyme was used to create an artificial ensemble¹⁵⁰ on which TS-stabilizing mutations were identified, increasing activity 3-fold³³. It was more recently demonstrated that a crystallographic ensemble of designed Kemp eliminase HG3 could be used to recapitulate the active site of a highly evolved variant¹⁵¹. Leveraging experimentally derived ensembles in this case provided access to design templates that parallel the catalytically competent sub-states sampled during directed evolution^{151,152}. Further demonstrating this point, ensemble-based design with crystallographic ensembles was used to improve the activity of multiple *de novo* Kemp eliminases over 100-fold (Figure 1.11)⁷⁵. The success of experimentally derived ensembles for this approach enables enzyme engineers to create an enzyme, determine an experimental structure to identify failures of the design, and then apply ensemble-based design to address these failures.

1.7.2 Stability-guided sequence space refinement

Both traditional and current approaches to active-site packing rely on search algorithms involving random sampling to traverse vast sequence space. One method for computational enzyme optimization that addresses this is FuncLib¹⁵³, which applies strict pruning of the sequence space for enzyme design. FuncLib performs a phylogenetic analysis to identify evolutionarily tolerated mutations and removes predicted destabilizing mutations. The narrowed sequence space

can be exhaustively enumerated, ranked by calculated energy, and clustered to identify a diverse set of active-site sequences predicted to not destabilize the structure. FuncLib has been applied to enhancing promiscuous activity in natural enzymes¹⁵³ and target activity in *de novo* enzymes^{69,154}. For example, this approach was used to improve the efficiency of *de novo* Kemp eliminase Des27 by nearly two orders of magnitude (Figure 1.11)⁶⁹, with a single serendipitous mutation further improving this efficiency to match that of the median natural enzyme⁸. Although phylogenetic input may be less relevant for *de novo* enzyme active sites, FuncLib is effective with the predicted stability filter alone⁶⁹, demonstrating the power of exhaustively enumerating through a refined sequence space.

1.7.3 Iterative *in silico* optimization

The abundance of high-throughput machine learning-based models for protein design has enabled workflows that optimize enzymes through cycles of design and prediction in a manner similar to the backbone refinement protocol by Braun *et al.*⁷³. In an approach termed AI.zymes¹⁵⁵, the structure of a low-activity variant is first used as input to generate a library of mutants. Physics-based packing is used for active-site design, and sequence generative models design the rest of the protein, applying a bias to the input sequence. The resulting library is scored based on user-specified metrics, including stability, electrostatics, and ligand interactions. Promising mutants are identified through multi-objective Boltzmann selection, and predicted structures of these variants are used as input for redesign, adding new variants to a growing pool until a maximum number is reached. Top-performing variants can be further filtered and experimentally characterized. This approach has been used to optimize the promiscuous Kemp eliminase activity of a natural ketosteroid isomerase, increasing efficiency 8-fold while improving stability and eliminating oligomeric forms of the enzyme¹⁵⁵.

1.8 Lingering challenges motivate alternate *de novo* enzyme design approaches

Recent approaches to *de novo* enzyme design nicely address multiple limitations of traditional *de novo* design. The custom generation of structures around a catalytic motif helps tailor backbone geometry directly to the needs of the theozyme. Family-wide hallucination and ensemble-based design engage the fixed backbone approximation, while the use of *de novo* scaffolds confronts the protein stability challenge. However, both methods face distinct challenges depending on the origin of the input backbones. Pre-existing natural and *de novo* protein backbones are not tailored to the reaction of interest. In contrast, machine learning-generated structures can be crafted to the theozyme and are highly stable, but are largely α -helical and lack resemblance to any natural proteins (Figure 1.15)^{67,68,73}. These helical backbones often perform well during *in silico* filtering¹⁵⁶ due to their large sequence space, or designability¹⁵⁷. However, the tightly packed, regular secondary structure of these scaffolds that imparts stability and designability may also stifle conformational flexibility and the subtle rearrangements that facilitate catalysis¹⁵⁸. A middle ground, the generation of bespoke structures that retain familiar enzyme folds, offers a promising alternative for *de novo* enzyme design.

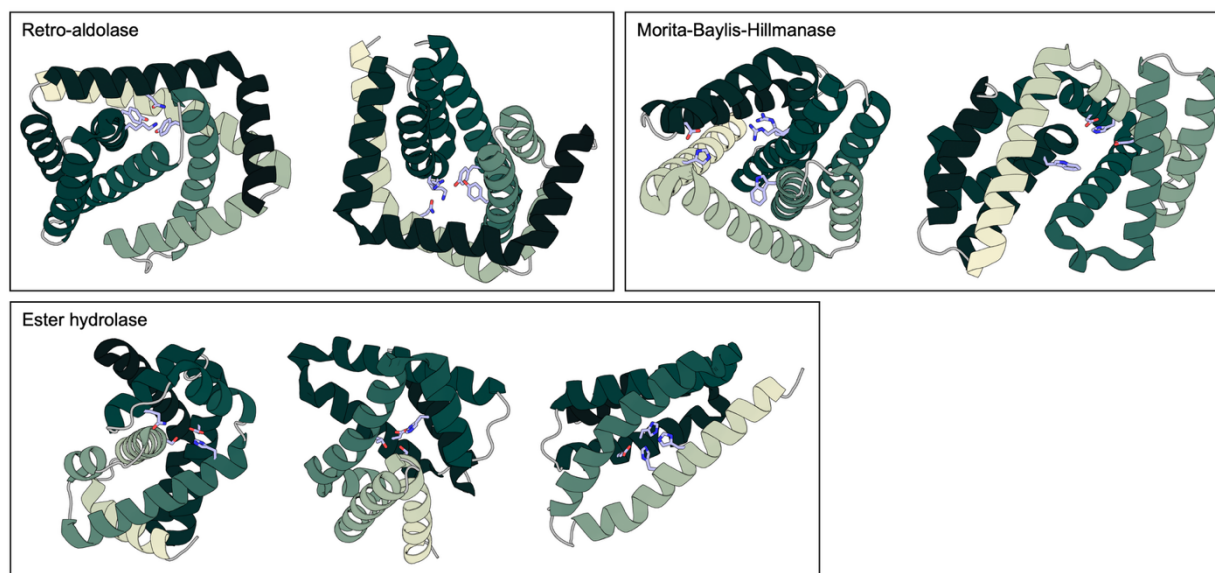


Figure 1.15. Structures of *de novo* enzymes created by catalytic motif scaffolding. These *de novo* retro-aldolase, Morita-Baylis-Hillmanase, and ester hydrolase scaffolds are small and largely α -helical.

1.8.1 The TIM barrel fold in enzymes

The TIM barrel presents a compelling fold for *de novo* enzyme design. First identified in triose phosphate isomerase^{159,160}, it is one of the most common folds in natural enzymes¹⁶¹ and supports catalysis at rates approaching the diffusion limit^{162,163}. The canonical TIM barrel contains a central closed β -barrel comprised of eight β -strands, surrounded by eight α -helices arranged in repeating β/α units (Figure 1.16a). The core of the β -barrel can be divided into four layers, with four side chains pointing towards the center in each layer¹⁶⁴. Enzymes with this fold are highly varied, catalyzing entirely unrelated reactions. The TIM barrel is the only single-domain fold that supports function across five of the seven Enzyme Commission (EC) classes¹⁶⁵. Their functional breadth is nearly twice that of the next mostly enzymatically diverse fold¹⁶⁵. These functions are typically related to molecular or energy metabolism, and thus they are the most abundant protein fold in all genomes¹⁶⁶. Therefore, TIM barrels present a unique versatility not observed in any other enzyme fold.

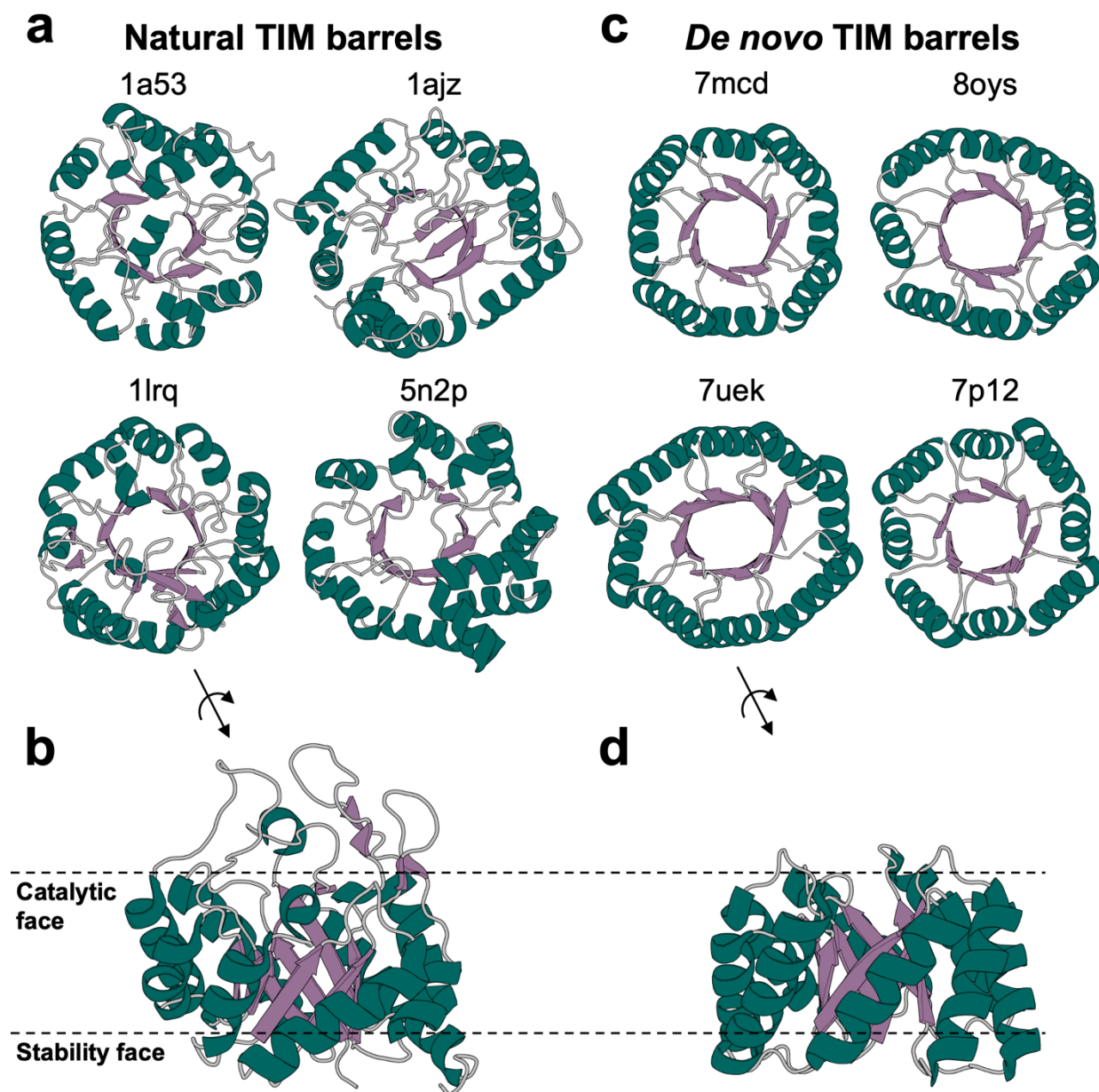


Figure 1.16. Natural and artificial TIM barrels. Crystal structures of natural TIM barrels^{167–170} (a), which feature extended loops and other insertions on their catalytic face (b). In contrast, *de novo* TIM barrels^{121,171–173} (c) have short and rigid $\beta\alpha$ -loops (d). PDB IDs are shown above each structure.

The unique versatility of the TIM barrel arises from its architecture. All TIM barrel enzymes have their active sites towards the C-terminal end of the closed β -sheet. Here, about 80% of the catalytic residues are housed on $\beta\alpha$ -loops¹⁷⁴, in regions unimportant for stability. This spatial separation of the TIM barrel core (stability face) and active-site $\beta\alpha$ -loops (catalytic face) (Figure 1.16b) allows function-enhancing mutations to be explored without compromising the structure, a property associated with evolvability of protein folds^{174,175}. Accordingly, the $\beta\alpha$ -loops are highly modular and mutable, often featuring extended loops and other insertions (Figure 1.16b) that facilitate substrate binding and passage along the reaction coordinate^{164,176,177}. In triose phosphate isomerase, a TIM-barrel enzyme which displays diffusion-limited efficiency, each $\beta\alpha$ -loop extension plays a distinct role in catalysis¹⁷⁵. These roles include substrate interactions, electrostatic catalysis, and acid-base catalysis; loops not directly involved in catalysis additionally facilitate the conformational change to an active, closed form. For efficient catalysis, this conformational change occurs with rapid, microsecond loop movement induced by substrate binding^{178,179}. Alternatively, TIM-barrel enzymes can orchestrate the interdependent motion of multiple loops through extreme conformational changes on the millisecond timescale to bind larger substrates¹⁸⁰. Thus, extended $\beta\alpha$ -loop structures demonstrate dynamic motions on multiple timescales^{177,179–181}, supporting efficient catalysis and the emergence of new activity¹⁸².

Many traditional campaigns have favored the TIM barrel fold as a scaffold for enzyme construction^{13,21–23}. In the design of Kemp eliminases by R  thlisberger *et al.*, all successful designs adopted the TIM barrel fold, despite it comprising only 25% of the input scaffold set¹³. This enrichment was due to the high density of backbone positions with side chains oriented toward the barrel interior, ideal for placing catalytic residues and maximizing non-bonding interactions with the TS¹³. The separation of stability and catalysis in the TIM barrel enables the accumulation of

beneficial active-site mutations during design with minimal impact on the overall fold. This feature likely contributes to the success of directed evolution from *de novo* enzymes adopting this fold (Figure 1.10)^{33–37,183}.

1.8.2 Functionalizing *de novo* TIM barrels

The modular and mutable active-site features of TIM barrels, combined with their catalogue of diversity, makes them particularly appealing for protein design. Towards this goal, artificial TIM barrels have been created from scratch by targeting the canonical topology^{121,171–173,176,184}, a simple four-fold repeated $\beta\alpha\beta\alpha$ unit. The first *de novo* TIM barrel was created by determining optimal lengths for each secondary structural element and designing the sequence of this topology using physics-based calculations¹⁸⁴. *De novo* TIM barrels are structurally simple (Figure 1.16c) compared to their natural counterparts, containing only the canonical eightfold $\beta\alpha$ repeat with short, rigid $\beta\alpha$ -loops on their catalytic face (Figure 1.16d); this minimalism makes them highly stable but functionally inert, lacking extended loops and other insertions that can form an enclosed binding pocket¹⁷⁷. However, their excessive stability and high mutability, particularly at $\beta\alpha$ -loops and inward-facing β -strand positions¹⁷², make *de novo* TIM barrels excellent candidates for functionalization.

Headway toward function has involved extending $\beta\alpha$ -loops with long α -helices^{185–187}, forming a central cavity that could be modified downstream¹⁷⁷. Unfortunately, these cavities lack definition and remain too solvent exposed to support the microenvironments required for complex function like catalysis. Entire domains have been inserted into *de novo* TIM barrel $\beta\alpha$ -loops to create metal-binding sites at the domain interface¹⁸⁸, imparting photocatalytic activity¹⁸⁹. Metal-binding sites have also been installed on structurally diverse, generative-model derived TIM barrels to create metalloenzymes¹⁹⁰. Although these approaches achieve enzymatic function, the

resulting TIM barrels lack defined substrate binding pockets and instead largely rely on the intrinsic reactivity of the metal cofactor. Broadly applicable strategies to form custom active-site pockets for pre-defined chemical transformations are required to endow *de novo* TIM barrels with true enzymatic function. Rather than relying on the stepwise addition of structural extensions and catalytic machinery, such approaches could directly tailor these extensions to the target reaction geometry.

1.9 Thesis objectives: customizing minimal TIM barrels to design enzymes

Chapter 1 summarized traditional *de novo* enzyme design methods, the enzymes designed using these methods, and challenges that prompted novel approaches. These novel approaches have improved catalytic efficiency, but not to the level of laboratory evolved enzymes. While many strategies repurpose existing scaffolds, others aim to generate new scaffolds from scratch. However, a hybrid approach uniting the creation of custom backbones with the versatility and modularity of TIM barrels could be a valuable alternative. This thesis describes a workflow, termed CANVAS (Customizing Amino-acid Networks for Virtual Active-site Scaffolding), for introducing enzymatic activity into a minimal TIM barrel by adding a structural lid (Figure 1.17). The workflow will be evaluated through the design of Kemp eliminase enzymes, a model reaction for *de novo* enzyme design (Figure 1.6). Chapter 2 outlines the workflow and presents the computational and experimental characterization of CANVAS-designed enzymes, with special attention to their structural and functional features. Chapter 3 contrasts these designs with those from recent approaches, explores how existing methods could complement CANVAS, and pinpoints opportunities for CANVAS to advance *de novo* enzyme design.

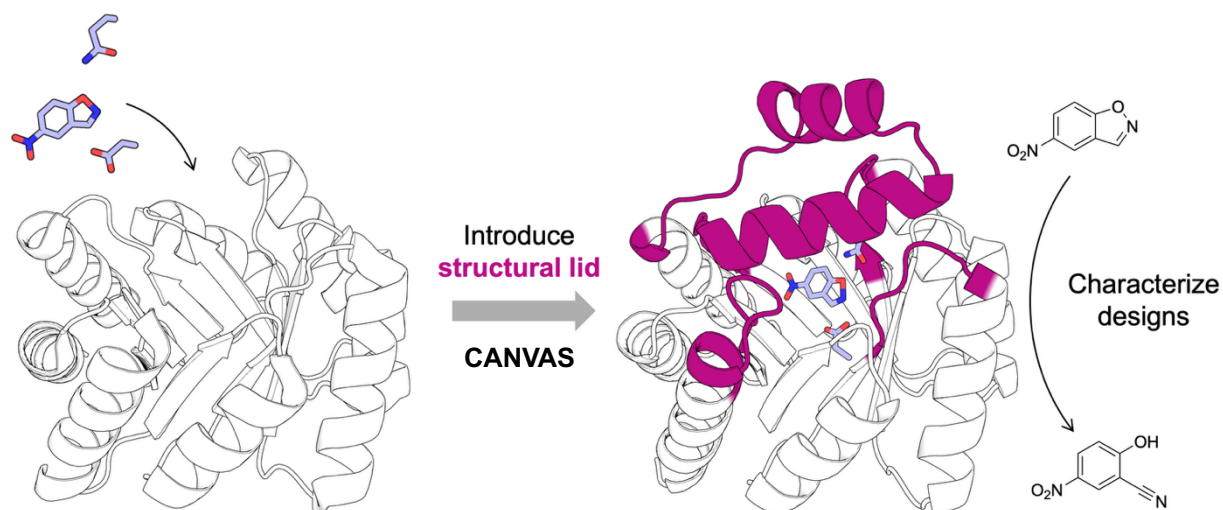


Figure 1.17. Overview of thesis objectives. We aim to create a framework, termed CANVAS, to introduce a structural lid into a minimal *de novo* TIM barrel, creating a custom active site and scaffolding a catalytic residue. To evaluate this, we target the model Kemp elimination reaction. The designed Kemp eliminase enzymes will be experimentally characterized to assess activity and validate the lid architecture.

Chapter 2. Customizing the structure of minimal TIM barrels to craft efficient *de novo* enzymes

Statement of contribution

Benjamin J. Smith, Niayesh Zarifi, Julian Beck, and Emily Freund performed the *de novo* design of Kemp eliminases using CANVAS. Julian Beck and Emily Freund performed the circular permutation of 7OSV. Benjamin J. Smith, Julian Beck, Emily Freund, Niayesh Zarifi, and Jan Hartmann experimentally characterized the initial Kemp eliminase variants. Julian Beck and Mark Kriegel crystallized KempTIM4 and KempTIM1, respectively. Benjamin J. Smith performed the ensemble-based redesign of KempTIM4. Benjamin J. Smith, Mark Kriegel, and Ahana G. Harsha experimentally characterized the ensemble-based redesigns of KempTIM4. Benjamin J. Smith performed the molecular dynamics simulations. Roberto A. Chica and Birte Höcker conceived the project. This chapter was written by Benjamin J. Smith, Roberto A. Chica, Birte Höcker, and Julian Beck.

Acknowledgements

Benjamin J. Smith was the recipient of a Collaborative Research and Training Experience (CREATE) program scholarship and a Canadian Graduate Research Scholarship–Masters (CGRS-M) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

2.1 Introduction

Enzymes are nature's most efficient catalysts, with properties that make them attractive for research and industry. The ability to design enzymes for any target reaction is therefore highly desired. Current *de novo* enzyme design strategies typically place active sites into diversified natural scaffolds^{69,72,93} or build α -helical scaffolds around catalytic motifs to create pockets with high transition-state (TS) complementarity^{67,68,73,114}. Despite some success, the former approach does not cater the backbone geometry to the specific reaction requirements. Concurrently, the highly stable, regular secondary structures from the latter approach may impede structural rearrangements that facilitate substrate binding and passage along the reaction coordinate, critical for the high efficiency of natural enzymes¹⁸². A more robust strategy could entail generating bespoke structures that retain familiar but versatile folds.

The TIM barrel is one of the most functionally diverse enzyme folds. It exhibits the greatest breadth of catalytic function of any single-domain fold, while supporting efficiency near the diffusion limit^{165,166,191,192}. While idealized *de novo* TIM barrels have been successfully designed^{121,171–173,176,184}, none display enzymatic function due to their minimalistic architecture, which lacks structural elements required for catalysis such as cavities, pockets, and extended loops. Efforts to address these deficiencies have led to the introduction of secondary structural elements to create rudimentary pockets^{185–187}. However, these pockets are too solvent exposed and undefined to support the microenvironments required for catalysis. Therefore, novel strategies are needed to craft custom active sites and endow *de novo* TIM barrels with true enzymatic function.

2.2 The CANVAS approach

Here, we introduce CANVAS (Customizing Amino-acid Networks for Virtual Active-site Scaffolding), a computational workflow for transforming minimal proteins such as *de novo* TIM barrels into functional enzymes (Figure 2.1). The process begins with a minimal TIM-barrel template and a theozyme, which is a computationally derived, idealized model of an active site with functional groups arranged to stabilize the transition state of a reaction¹². Using protein design software¹⁹³, the theozyme is placed onto the TIM-barrel template. Catalytic residues that cannot be accommodated are positioned in the empty space above the catalytic face. These residues can serve as anchors for constructing a custom lid to form the active-site pocket, which is created using a structure-generative model. Lid sequences are designed with a structure-conditioned sequence generative model and filtered using predicted structures. In this iteration of CANVAS, RFdiffusion¹¹¹, ProteinMPNN¹³⁸, and AlphaFold2¹⁴¹ were selected as the respective structure generative, sequence generative, and structure prediction models, owing to the experimental validation of this workflow^{111,194}. However, other tools with expanded functionality could substitute in future iterations. Following lid design, the active site sequence is optimized to build a pocket complementary to the TS.

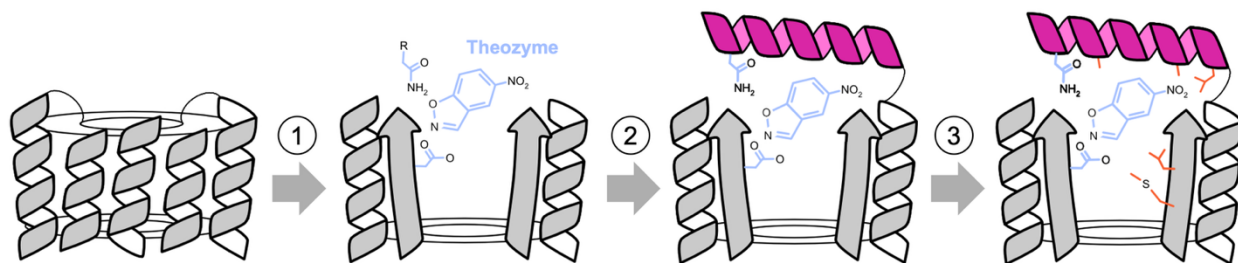


Figure 2.1. The general CANVAS workflow. CANVAS follows three general steps: (1) placement of the theozyme (blue) onto the minimal TIM-barrel scaffold using a single catalytic residue as an anchor; (2) construction of a custom lid to form the active-site pocket and anchor additional catalytic residues; (3) active-site packing to further stabilize the transition state.

2.3 Applying CANVAS to minimal TIM barrels

We applied CANVAS to the Kemp elimination, a model organic transformation commonly used to benchmark *de novo* enzyme design methods^{13,21,69,75}. This reaction involves the deprotonation of a benzisoxazole by a base and delocalization of the resulting negative charge, producing a hydroxybenzonitrile compound through ring opening¹⁹⁵. The theozyme for the transformation of 5-nitrobenzisoxazole to 2-hydroxy-5-nitrobenzonitrile (Figure 2.2) features an Asp as catalytic base and an Asn or Gln as H-bond donor to stabilize the binding of the substrate and properly orient it for reaction with the base²¹. The Kemp elimination further benefits from a hydrophobic environment¹⁹⁵. Therefore, we set out to introduce a lid onto a minimal TIM barrel that could form a hydrophobic pocket to house the active site.

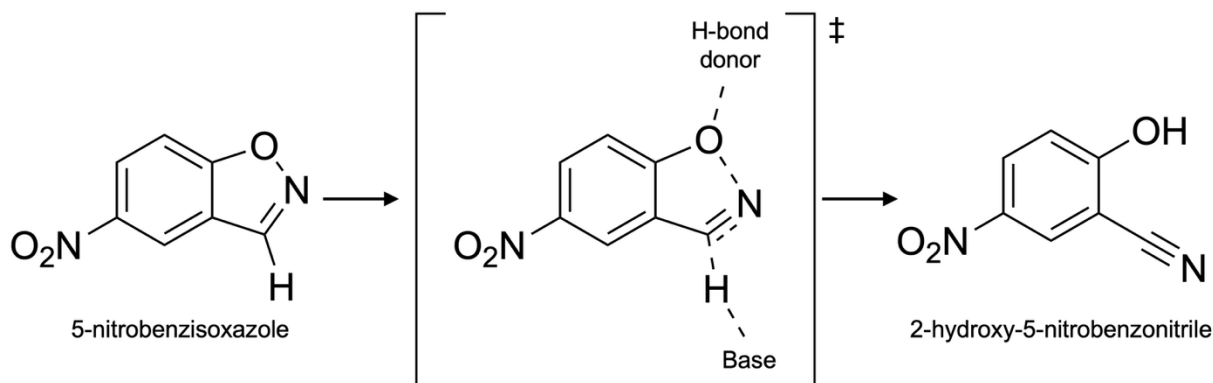


Figure 2.2. The Kemp elimination reaction of 5-nitrobenzisoxazole. The unimolecular reaction proceeds through a single transition state which can be stabilized by a catalytic base and an H-bond donor.

2.3.1 Minimal TIM barrel selection

We used the crystal structures of two *de novo* TIM barrels (PDB IDs: 7OSV¹⁷³ and 7MCD¹²¹) as templates. These TIM barrels are minimal and highly stable, representing ideal starting points for the introduction of a structural lid. Since 7OSV has its N- and C-termini on the catalytic face of the TIM barrel, we applied an *in silico* circular permutation based on the inpainting

method¹⁹⁶ to generate two modified structures named NT6-CP1 and NT6-CP2 that contain termini on the stability face (Figure 2.3).

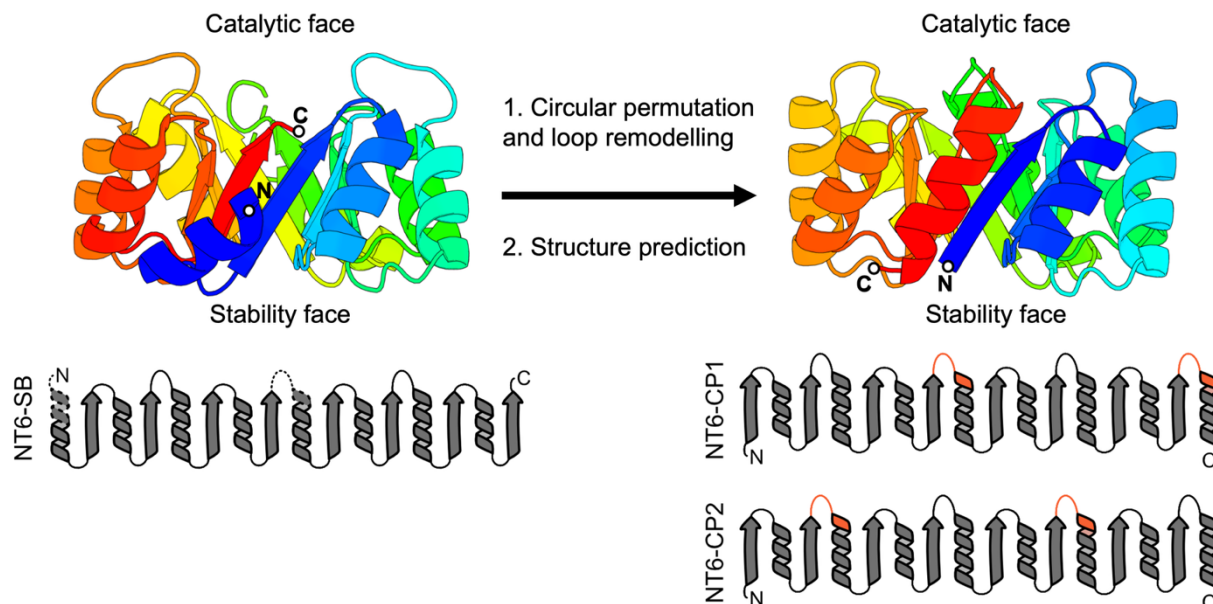


Figure 2.3. Circular permutation of NovoTIM6-SB. Circular permutation and loop remodeling were used to relocate the N- and C-termini of NovoTIM6-SB (left, NT6-SB, PDB ID: 7OSV) from the catalytic face to the stability face. After structure prediction with AlphaFold2, variants NT6-CP1 and NT6-CP2 (right) were generated. Protein structures are colored from blue (N-terminus) to red (C-terminus). Topology diagrams show that the first few residues of NT6-SB and the fourth $\beta\alpha$ -loop were deleted, indicated by dashed outlines. Orange segments in the resulting TIM barrels represent the remodelled or inserted loops required for circular permutation.

2.3.2 Kemp eliminase design with CANVAS

Starting from these three minimal TIM barrel structures, we used CANVAS to generate enzymes for the Kemp elimination reaction. First, the catalytic base was placed computationally, targeting positions in the first two β -barrel layers¹⁹¹ closest to the stability face. TS poses were built from the catalytic base¹⁵, and positions for the base where the TS was placed with its isoxazolic oxygen pointed out of the catalytic face were selected (Figure 2.4). The catalytic H-bond donor is then built from each TS pose. The C_α of this second residue is thus positioned in the empty space above the TIM barrel catalytic face, to be used as an “anchor” for lid generation.

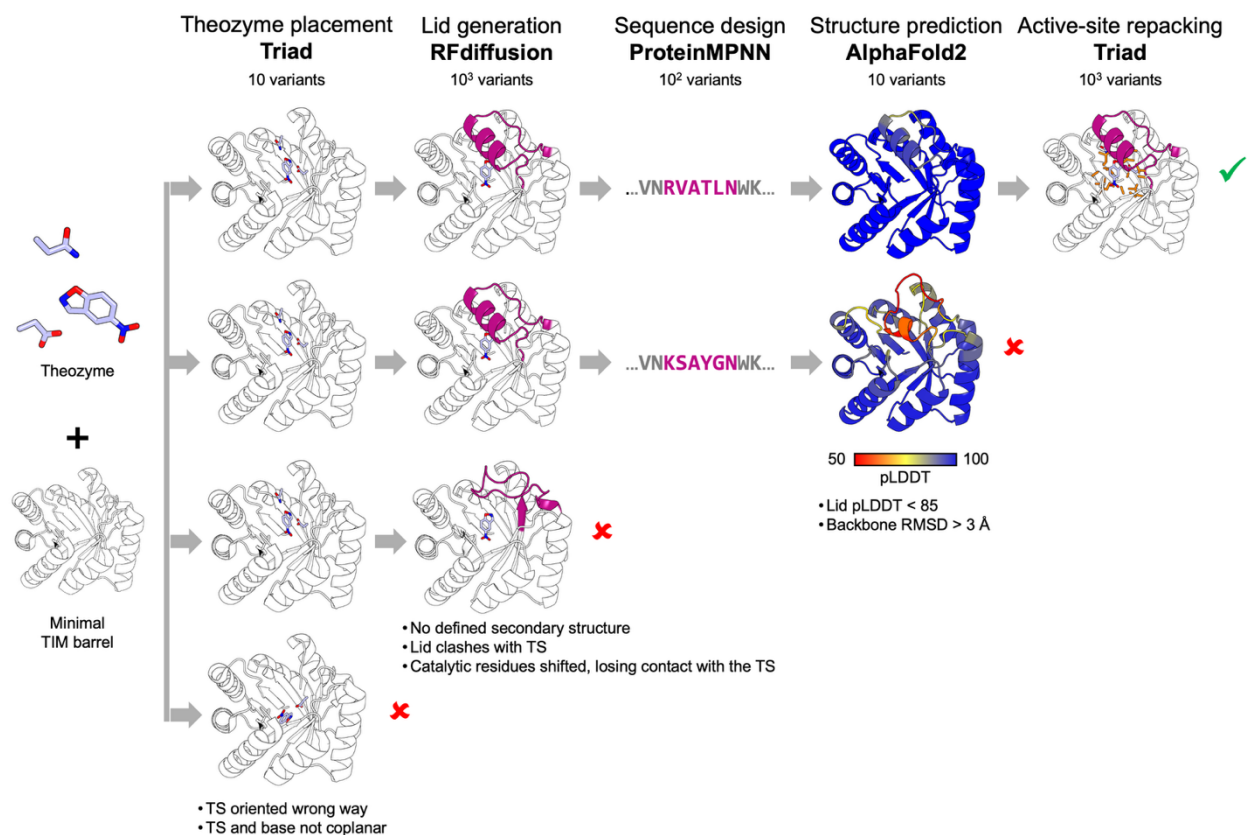


Figure 2.4. Structure and sequence filtering throughout the CANVAS pipeline. Step 1: a catalytic residue from the theozyme is placed onto a minimal TIM barrel scaffold, and the transition state (TS) is positioned from its side chain using Triad. The second catalytic residue is built from the TS, positioning its α -carbon in the empty space above the TIM barrel catalytic face. Additional residues can be added inside the barrel or in the empty space above using similar steps. Step 2: a lid composed of protein fragments of desired length and secondary structure is generated with RFdiffusion to anchor catalytic residues located above the TIM barrel face. Step 3: lid sequences are designed with ProteinMPNN while maintaining catalytic residue identity. Step 4: designed structures are predicted with AlphaFold2. Step 5: the amino-acid sequence and rotameric configuration of the active site are optimized with Triad to maximize TS packing and catalytic contact geometry. Resulting sequences are then filtered using key enzyme design criteria. At each step, structures failing filtering criteria are rejected. Approximate number of structures/sequences generated at each step is indicated (# variants).

Lids were created on the theozyme-containing TIM barrels using motif scaffolding in RFdiffusion¹¹¹. For each target $\beta\alpha$ -loop, a fragment of 10–30 residues was generated from the last β -strand residue to the H-bond donor and back to the first residue of the subsequent α -helix. Other $\beta\alpha$ -loops were sometimes elongated by 3–5 residues to increase contacts with the primary fragment. Because RFdiffusion inference begins with random noise, multiple trajectories were run from each input structure. Diffused fragments with a defined secondary structure that retained the required catalytic contacts were selected (Figure 2.4), and four lid sequences were designed with ProteinMPNN¹³⁸ for each scaffold. Sequences were filtered using AlphaFold2¹⁴¹ predictions; those that yielded high-confidence structures similar to the RFdiffusion model were carried forward (Figure 2.4). Lid structures remaining after filtering included α -helices, two-stranded β -sheets, and helix-turn-helix motifs.

Following generation of lid structures, the active sites of each scaffold were packed for TS stabilization. Active-site positions were allowed to sample rotamers of hydrophobic and aromatic residues, also allowing the ProteinMPNN-proposed residue at lid positions. To select the most promising sequences for experimental characterization, they were filtered using various criteria chosen to target structural features associated with efficient catalysis such as solvent-accessible surface area of the TS, active-site preorganization, and catalytic contact geometry⁷⁵.

Filtering yielded nine variants featuring three distinct lids (Figure 2.5a), one from each minimal TIM barrel: variants 1–2 (NT6-CP2), variants 3–5 (NT6-CP1), and variants 6–9 (7MCD). The lid diffused onto 7MCD consists of a single 26-residue fragment introduced in $\beta\alpha$ -loop 7 (Figure 2.5b). In contrast, the designed lids in NT6-CP1 and NT6-CP2 contain a major 33- or 39-residue fragment introduced into $\beta\alpha$ -loops 2 or 6, respectively, to scaffold the catalytic H-bond donor, as well as three or four elongated loops. For all three structures, the major fragment is

largely α -helical, reflecting the designability of helices compared to other secondary structures generated by RFdiffusion^{111,156}. Each variant features a unique active-site configuration (Figure 2.6) and sequence (Table 2.1), which are predicted with high confidence by AlphaFold2 (Figure 2.7). Structural similarity searches with FoldSeek¹⁹⁷ confirm that while these variants preserve the conventional TIM barrel fold of natural enzymes, they possess novel catalytic face architectures (Figure 2.8) which are tailored to the Kemp elimination, a reaction not catalyzed by the native activity of a natural enzyme.

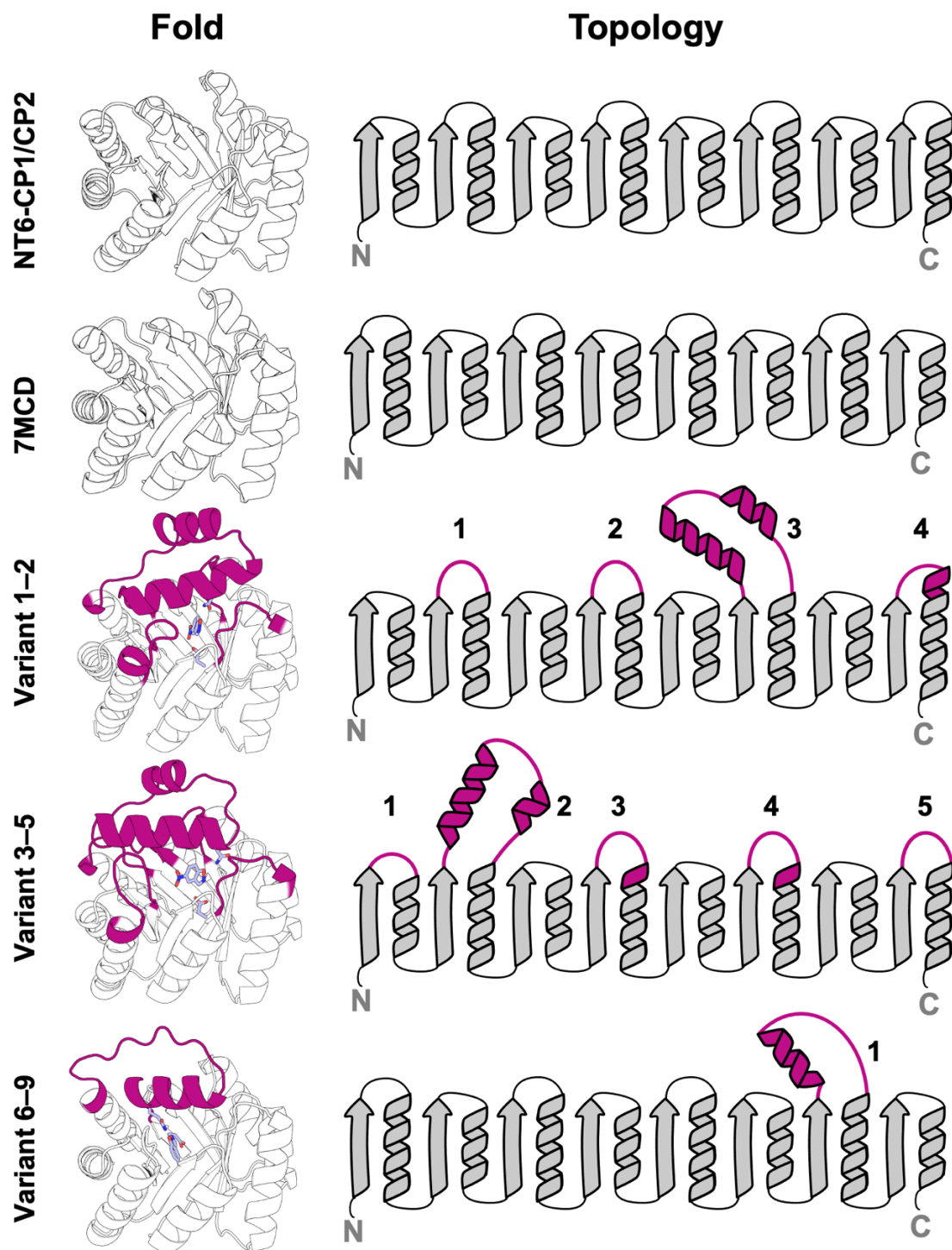


Figure 2.5. Structures and topologies of minimal TIM barrels and CANVAS-designed variants. Structures (a) and topologies (b) are shown for various TIM barrels. The designed lids (magenta) of variants 1–5 consist of four or five inserted fragments, including a long helix-turn-helix and three or four extended loops. The lid of variants 6–9 contains a single inserted helix-loop fragment. The transition state and catalytic residues of the designed variants are shown as light blue sticks.

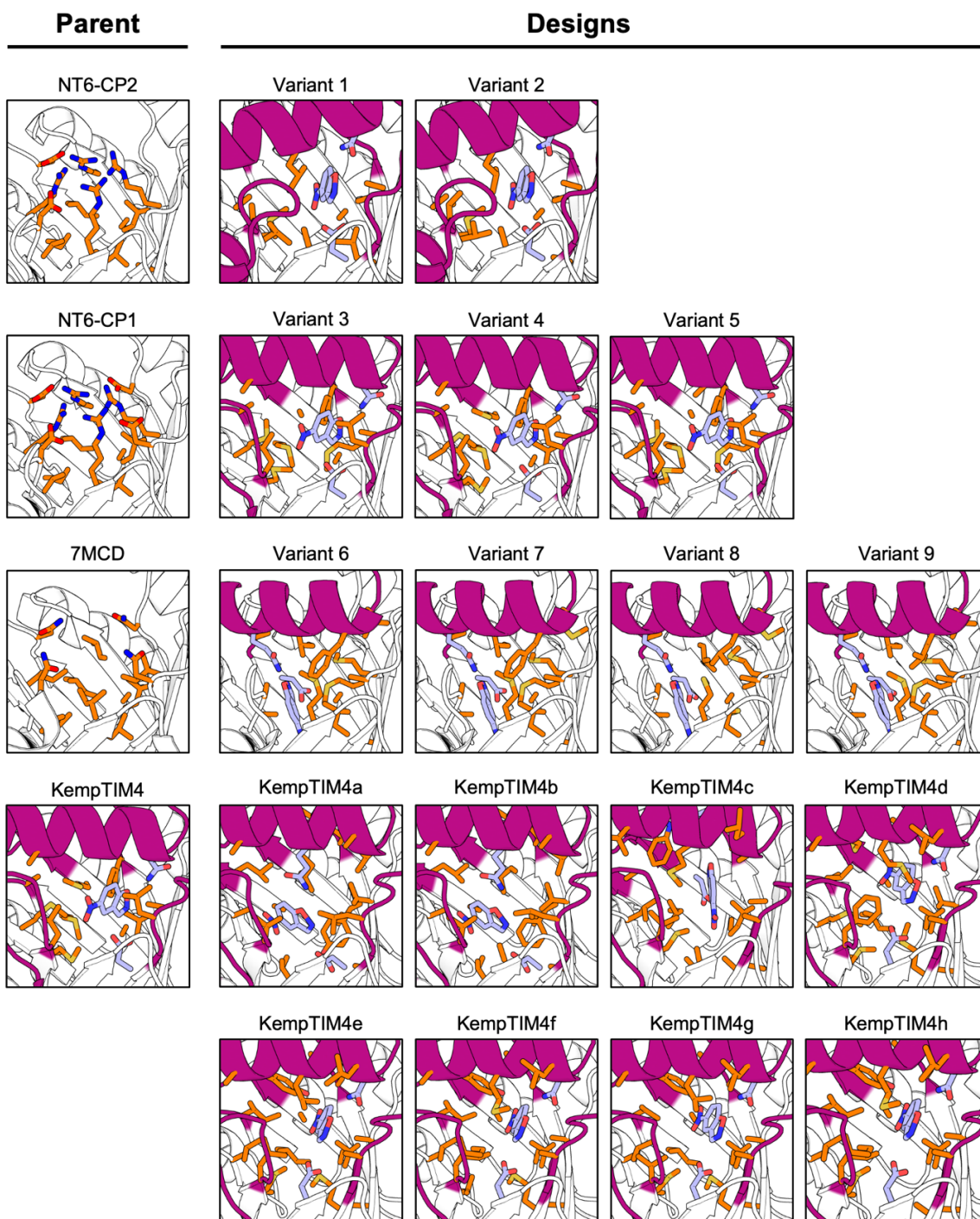


Figure 2.6. Active sites of designed enzymes. Residues optimized by Triad during active-site repacking of KempTIM variants, along with the corresponding residues on the parent minimal TIM barrel, are shown in orange. Lids introduced by RFDiffusion are shown in magenta. The catalytic base, catalytic hydrogen bond donor, and transition state are depicted as light blue sticks.

Table 2.1. Amino-acid sequences.

Enzyme	# residues	MW ^a (kDa)	Sequence ^b
NT6-CP1	188	22.1	RIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKETAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRR
NT6-CP2	188	22.1	EIAYRSDDWRDLQEALKKGADILIVDATDPEEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLEHHHHHH
7MCD	191	21.2	DILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAAELKKAEGAGADILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAAELKKAEGAGADILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAAELKKAEGAGLEHHHHHH
Variant 1	225	26.3	EIAYASDDWRDLQEALKKGADILVNVNPEEAWKQVEILRRLGAKRIYASLRWRDLQEALKKGADILAVPLVDKDEAWKQVEILRRLGAKETAYGSDWRDLQEALKKGADILGVASNTPRVRLVLRNLETLYSREEAERIVEEKIKLNDPEEAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADILLVAADDES EESKDEAWKQVEILRRLEHHHHHH
Variant 2	225	26.3	EIAYASDDWRDLQEALKKGADILVNVNPEEAWKQVEILRRLGAKRIYASLRWRDLQEALKKGADILAVPLVDKDEAWKQVEILRRLGAKETAYGSDWRDLQEALKKGADILGVASNTPRVRLVLRNLETLYSREEAERIVEEKIKLNDPEEAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADILLVAADDES EESKDEAWKQVEILRRLEHHHHHH
Variant 3	227	25.6	RIAYASLASRWDLQEALKKGADILMVALARDAASVRALAEGAAGNGDPRSVAEELLALLRPTDKDEAWKQVEILRRLGAKETAYMSDDWRDLQEALKKGADILMVLAASSALSPEEAWKQVEILRRLGAKRIAYGSGDWRDLQEALKKGADILVGLGAKGKDEAWKQVEILRRLGAKETAYASFWRDLQEALKKGADILMVFAVLDAERALKQVEILRRLEHHHHHH
Variant 4	227	25.6	RIAYMSLASRWDLQEALKKGADILMVALARDAASVRALAEGAAGNGDPRSVAEELLALLRPTDKDEAWKQVEILRRLGAKETAYMSDDWRDLQEALKKGADILMVLAASSALSPEEAWKQVEILRRLGAKRIAYGSGDWRDLQEALKKGADILVGLGAKGKDEAWKQVEILRRLGAKETAYASFWRDLQEALKKGADILVFAVLDAERALKQVEILRRLEHHHHHH
Variant 5	227	25.6	RIAYASLASRWDLQEALKKGADILMVALARDAASVRALAEAAAGNGDPRSVAEELLALLRPTDKDEAWKQVEILRRLGAKETAYMSDDWRDLQEALKKGADILMVLAASSALSPEEAWKQVEILRRLGAKRIAYGSGDWRDLQEALKKGADILVGLGAKGKDEAWKQVEILRRLGAKETAYASFWRDLQEALKKGADILMVFAVLDAERALKQVEILRRLEHHHHHH
Variant 6	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEGAGADILLVAPDDFEKGVEEVKELKRHGAKIIAYMSKSAEELKKAEGAGADILMVYPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAEGAGADILDVANAEQEKIASKFLGRKTKVKIEENDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKAEGAGWHHHHHH
Variant 7	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEGAGADILLVAPDDFEKGVEEVKELKRHGAKIIAYMSKSAEELKKAEGAGADILMVFPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAEGAGADILDVANAEQEKIASKFMGRKTKVKIEENDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKAEGAGWHHHHHH
Variant 8	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEGAGADILMVAPDDFEKGVEEVKELKRHGAKIIAYMSKSAEELKKAEGAGADILAVFPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAEGAGADILDVANAEQEKIASKLMGRKTKVKIEENDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKAEGAGWHHHHHH
Variant 9	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEGAGADILLVAPDDFEKGVEEVKELKRHGAKIIAYMSKSAEELKKAEGAGADILMVYPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAEGAGADILDVANAEQEKIASKLLGRKTKVKIEENDFEKGVEEVKELKRHGAKIIAYGSKSAEELKKAEGAGWHHHHHH
KempTIM1	235	25.2	HHHHHHENLYFQSGSGEITAYASGDADHLLAAREAGADILVDVDVDPARALAQVRLRAAGARRILYASLRVDDLAALEAGADILAVPDLDDHDAALAQIRALKAGAREITAYGSPDADHLLAAREAGADILGVAADTPAVRELWLRNLRQVFSEEEAREILDRRIHLPDPAEAEIQVRLRAAGAKRIMFASDDVDHLLAAAKRAGADILLVAEAGGSAEARAALAQVRLKALW
KempTIM2	235	25.6	HHHHHHENLYFQSGSGEITAFASGDVDHLLAAMEAGADILDVNVNPDPAAGLEQVRLRAAGAKRIAYASLRVEDLLAALKAGADILAVPDYDHERELRIIRELKRAGAKEIAYGSSDADHLLAARRAGADILGVAADTPAVRELWLRNLRQVFSEEEARRILDERIHLPPPAEAEIQVRLRAAGAKRIMYASDDVDHLLAAAKEAGADILLVVEAGGSAEALERALAQVRLKALW
KempTIM3	238	25.3	HHHHHHENLYFQSGSGRIYAAALASDLSLVEALKLGADILMVALMADAAAVRAMAEGHLNANGDPRSVAELEALLRPTDLDALAAVRELKALGAKETAFMSHDVDHLIRAMEAGADILMVLESSATSVAEALA

			QVRLKAAGAKRISFGSGDV AHLKAAMEAGADIL <u>DVLERHGLD</u> VALAQIRELKAAGAKEIAFASLDP DHLRLAREEGADILMV <u>FGATDP</u> PARALATVRYLRAQAW
KempTIM4	233	25.0	HHHHHHENLYFQSRIAYM <u>ALASDLD</u> SLVEALKLGADILMV <u>ALMADAAAVRAMAEG</u> <u>LHANGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFM SHD VDH LIRAMEAGADILMV <u>LESSATSVEA</u> AALAQVR RLKAAGAKRISFGSGDV AHLKAAMEAGADIL <u>DVLERHGLD</u> VALAQIRELKAAGAKEIAFASLDPDHL LRAREEGADILV <u>FGATDP</u> PARALATVRYLRAW
KempTIM5	238	25.1	HHHHHHENLYFQSGSGRIALA <u>AALAADE</u> DTLLRALELGADILMV <u>ALLPDAAAVRAAAIAAHANGDPRS</u> <u>VEELEALLRPTD</u> VDAALAVVRLKAAGAKEIAFM SHD VDH LIAAMEAGADILMV <u>LASVATSVEA</u> AALAE QVRLKAAGAKRIAFSGSDVDE LIAAMEAGADIL <u>DVLEARGLD</u> VALAQIRRLKEAGAKEIAFASFDA AHLLEARRAGADILMV <u>FALLDP</u> DRALAIVEELRAAAW
KempTIM4a	233	24.8	HHHHHHENLYFQSRIAYI <u>ALASDLD</u> SLVEALKLGADILAV <u>LLMADAAAVRALAEQLHAIGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFASHD VDH LIRAMEAGADILAV <u>VESSATSVEA</u> AALAQVR RLKAAGAKRILFSGSDV AHLKAAMEAGADIL <u>DVLA RHGLD</u> VALAQIRELKAAGAKEIAFVSLDPDHL LRAREEGADILV <u>VGATDP</u> PARALATVRYLRAW
KempTIM4b	233	24.8	HHHHHHENLYFQSRIAYI <u>ALASDLD</u> SLVEALKLGADILAV <u>LLMADAAAVRALAEQLHAIGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFASHD VDH LIRAMEAGADILAV <u>VESSATSVEA</u> AALAQVR RLKAAGAKRILFSGSDV AHLKAAMEAGADIL <u>DVLA RHGLD</u> VALAQIRELKAAGAKEIAFVSVDPDHL LRAREEGADILV <u>LGATDP</u> PARALATVRYLRAW
KempTIM4c	233	25.0	HHHHHHENLYFQSRIAYD <u>ALASDLD</u> SLVEALKLGADILLV <u>ALMADAAAVRVWAEMQHLAGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFLSHD VDH LIRAMEAGADILMV <u>VEVSATSVEA</u> AALAQVR RLKAAGAKRIMFSGSDV AHLKAAMEAGADILAVA <u>ERHGLD</u> VALAQIRELKAAGAKEIAFASMDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4d	233	25.1	HHHHHHENLYFQSRIAYA <u>ALASDLD</u> SLVEALKLGADILIV <u>LLMADAAAVRAMAEMLHLNGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFM SHD VDH LIRAMEAGADILMV <u>FELSATSVEA</u> AALAQVR RLKAAGAKRIAFDSGDV AHLKAAMEAGADILMV <u>VAERHGLD</u> VALAQIRELKAAGAKEIAFVSFDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4e	233	25.0	HHHHHHENLYFQSRIAYA <u>ALASDLD</u> SLVEALKLGADILIV <u>ILMADAAAVRAYAELLHLNGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFLSHD VDH LIRAMEAGADILMV <u>LESSATSVEA</u> AALAQVR RLKAAGAKRIAFDSGDV AHLKAAMEAGADILMV <u>VAERHGLD</u> VALAQIRELKAAGAKEIAFASFDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4f	233	25.0	HHHHHHENLYFQSRIAYA <u>ALASDLD</u> SLVEALKLGADILLV <u>ILMADAAAVRAYAEMLHLNGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFLSHD VDH LIRAMEAGADILLV <u>LESSATSVEA</u> AALAQVR RLKAAGAKRISFDSGDV AHLKAAMEAGADILMV <u>VAERHGLD</u> VALAQIRELKAAGAKEIAFASFDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4g	233	24.9	HHHHHHENLYFQSRIAYA <u>ALASDLD</u> SLVEALKLGADILIVV <u>LMADAAAVRAFAELLHLNGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFLSHD VDH LIRAMEAGADILMV <u>VESSATSVEA</u> AALAQVR RLKAAGAKRIAFDSGDV AHLKAAMEAGADILMV <u>VAERHGLD</u> VALAQIRELKAAGAKEIAFASFDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4h	233	24.9	HHHHHHENLYFQSRIAYA <u>ALASDLD</u> SLVEALKLGADILLV <u>ILMADAAAVRAYAEMLHLNGDPRSVAE</u> <u>LEALLRPTDLD</u> DALA AVRELKALGAKEIAFLSHD VDH LIRAMEAGADILMV <u>LESSATSVEA</u> AALAQVR RLKAAGAKRISFDSGDV AHLKAAMEAGADILAVA <u>ERHGLD</u> VALAQIRELKAAGAKEIAFASFDPDHL LRAREEGADILV <u>AGATDP</u> PARALATVRYLRAW
KempTIM4-gsg ^c	236	25.1	HHHHHHENLYFQSGSGRIAYM <u>ALASDLD</u> SLVEALKLGADILMV <u>ALMADAAAVRAMAEG</u> <u>LHANGDPRS</u> <u>VAELEALLRPTDLD</u> DALA AVRELKALGAKEIAFM SHD VDH LIRAMEAGADILMV <u>LESSATSVEA</u> AALAE QVRLKAAGAKRISFGSGDV AHLKAAMEAGADIL <u>DVLERHGLD</u> VALAQIRELKAAGAKEIAFASLDP DHLRLAREEGADILV <u>FGATDP</u> PARALATVRYLRAW

^a Molecular weight

^b Catalytic and lid residues are underlined and bolded, respectively.

^c This enzyme contains an additional GSG linker between the His-tag and first TIM barrel residue.

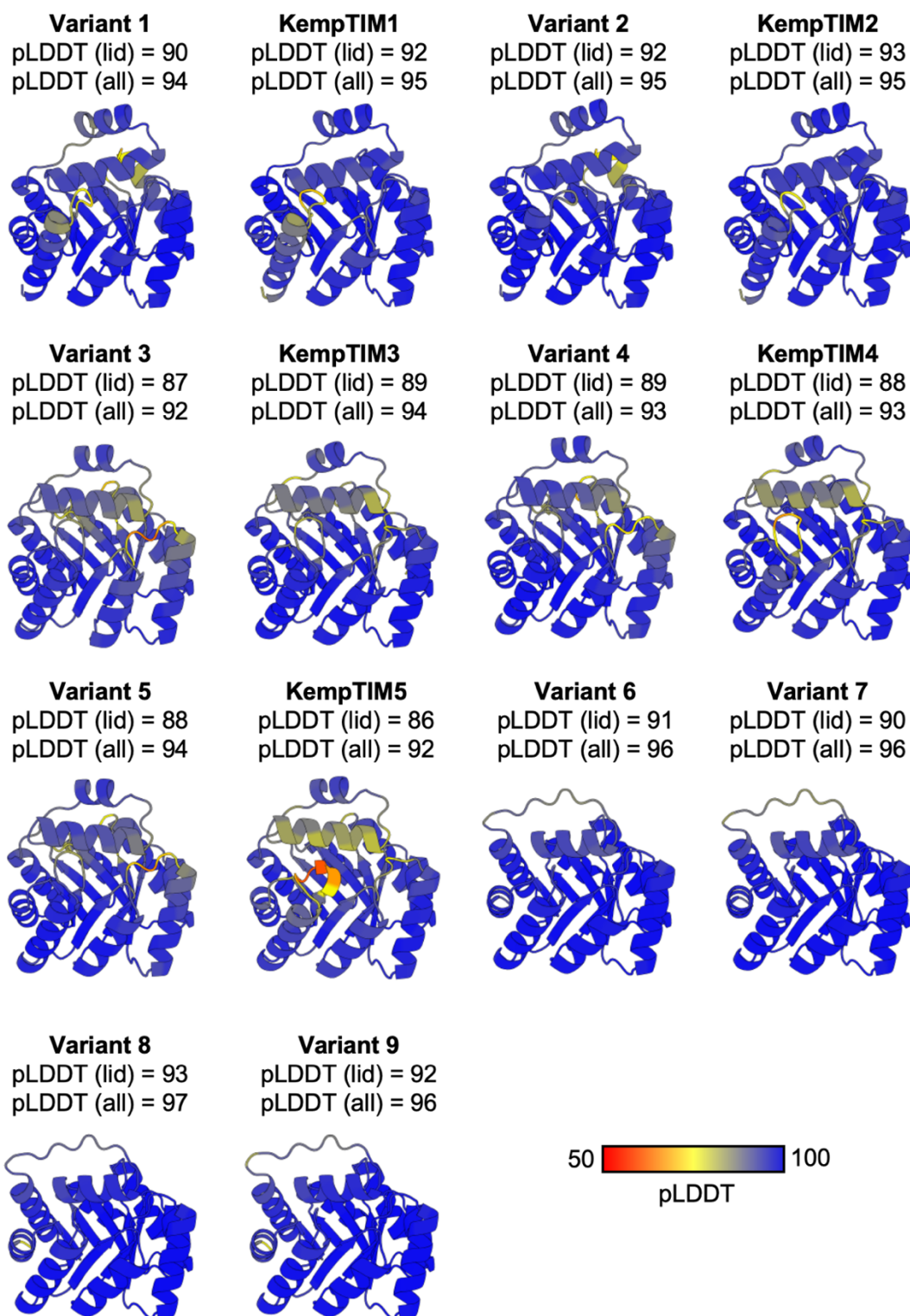


Figure 2.7. AlphaFold2 models. Predicted structures of variants are coloured by confidence (pLDDT), with the average pLDDT of lid residues and the entire structure listed above each variant.

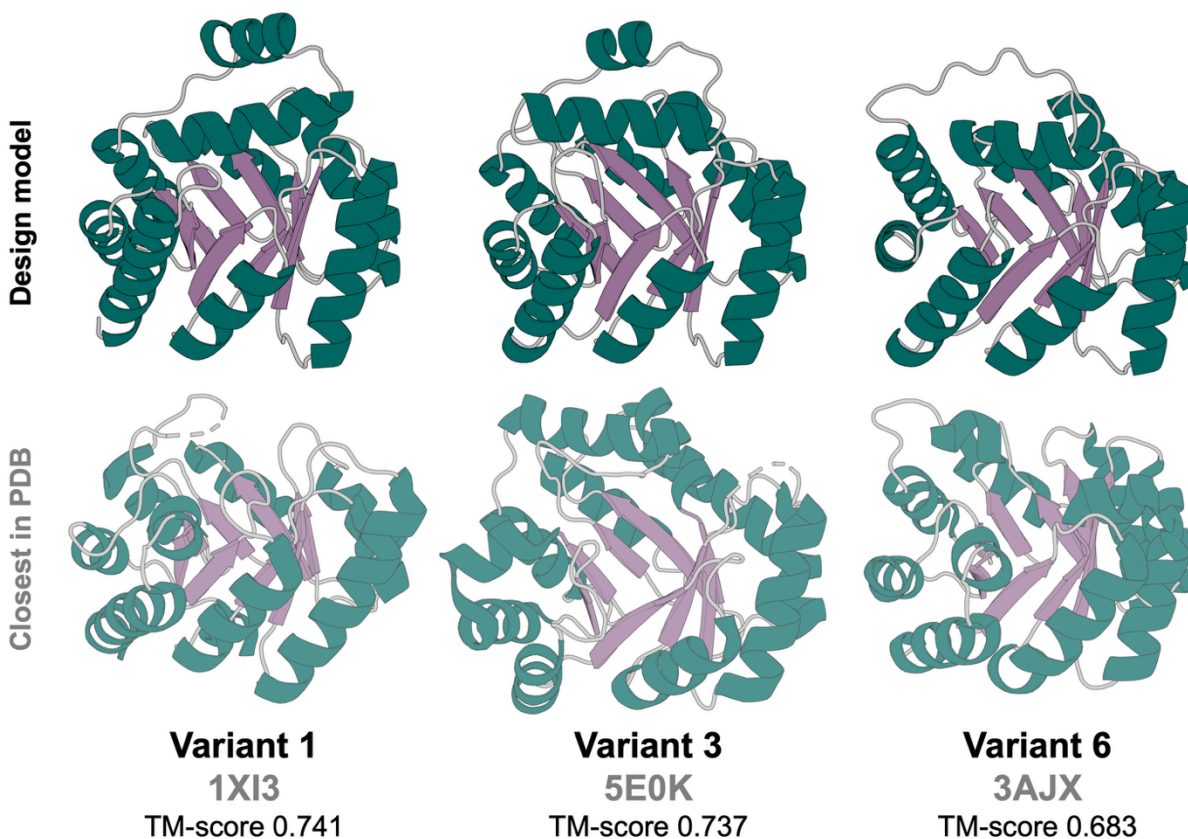


Figure 2.8. Structural similarity of CANVAS-designed variants to natural proteins. FoldSeek¹⁹⁷ was used to search representative design models for each lid structure against the PDB100 database, and the most similar natural proteins by TM-score are shown. PDB IDs^{198,199} and TM-scores for these proteins are shown below the name of the variant that was used as the query. These natural proteins adopt similar overall topologies, particularly in the TIM-barrel core, but contain different lid fragments. All three are natural TIM barrel enzymes.

2.4 Results

2.4.1 Experimental characterization of variants 1–9

Of the nine designed variants, only variants 4 and 6–9 showed soluble expression and could be purified to homogeneity (Figure 2.9, Figure 2.10a). Those derived from NT6-CP2 (variants 1–2) and most derived from NT6-CP1 (variants 3 and 5) could not be expressed. Notably, the NT6-CP1 template itself also failed to express, in contrast to NT6-CP2 and 7MCD. Variants 4 and 6–9 were purified as predominantly monomeric species (Figure 2.10b) with yields of 3–16 mg L⁻¹ (Table 2.2), although variant 4 showed a slight tendency to form dimers. Circular dichroism (CD) spectroscopy of variants 4 and 6–9 revealed a predominantly α secondary structure content (Figure 2.10c), confirming that the variants were not unfolded. However, these variants showed greater 208/222 peak ratios than the NT6-CP2 and 7MCD templates, possibly indicating increased random coil content. Additionally, these variants were destabilized by 21–34 °C relative to the parent TIM barrels (Figure 2.10d). Only variant 4 showed detectable activity (Figure 2.11); however, detailed characterization was challenged by low protein yields and a tendency to aggregate.

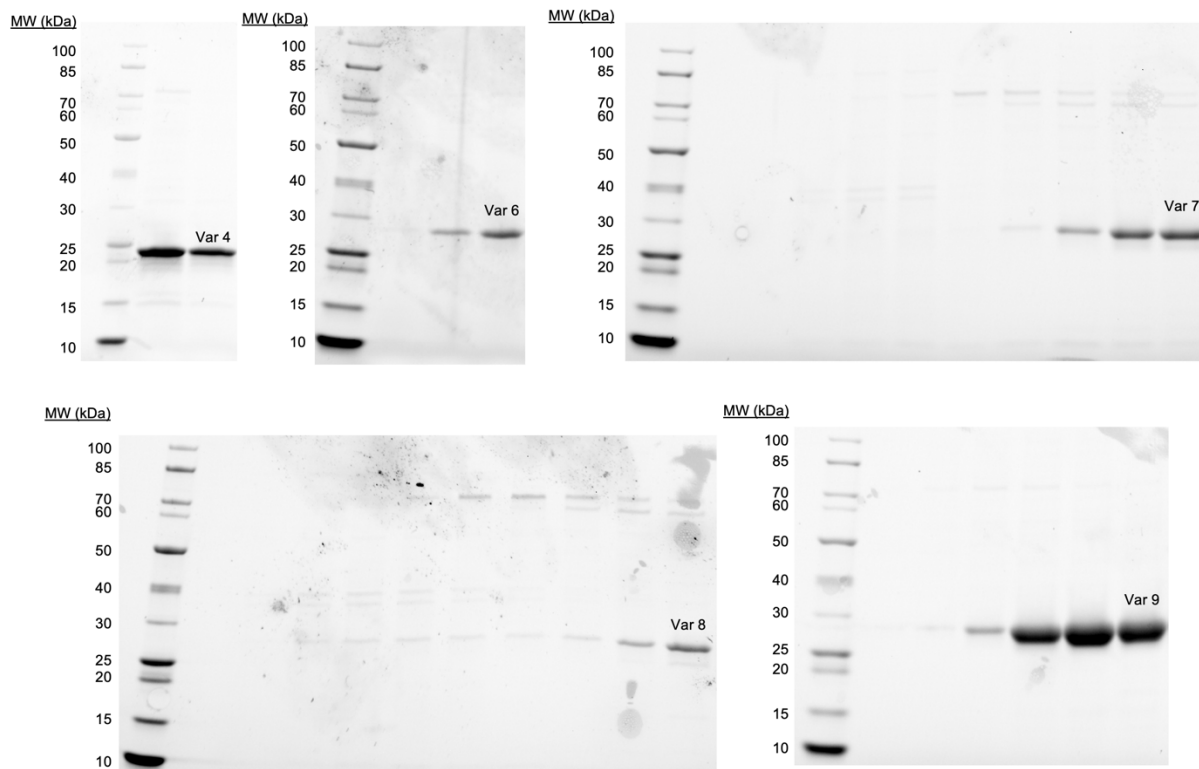


Figure 2.9. SDS-PAGE of expressible variants. The well containing the relevant variant (Var) is labeled. Mini-PROTEAN TGX Stain-Free Precast gels (Biorad) were loaded with each sample. The first column of each gel was loaded with an unstained protein standard (NEB). MW: molecular weight.

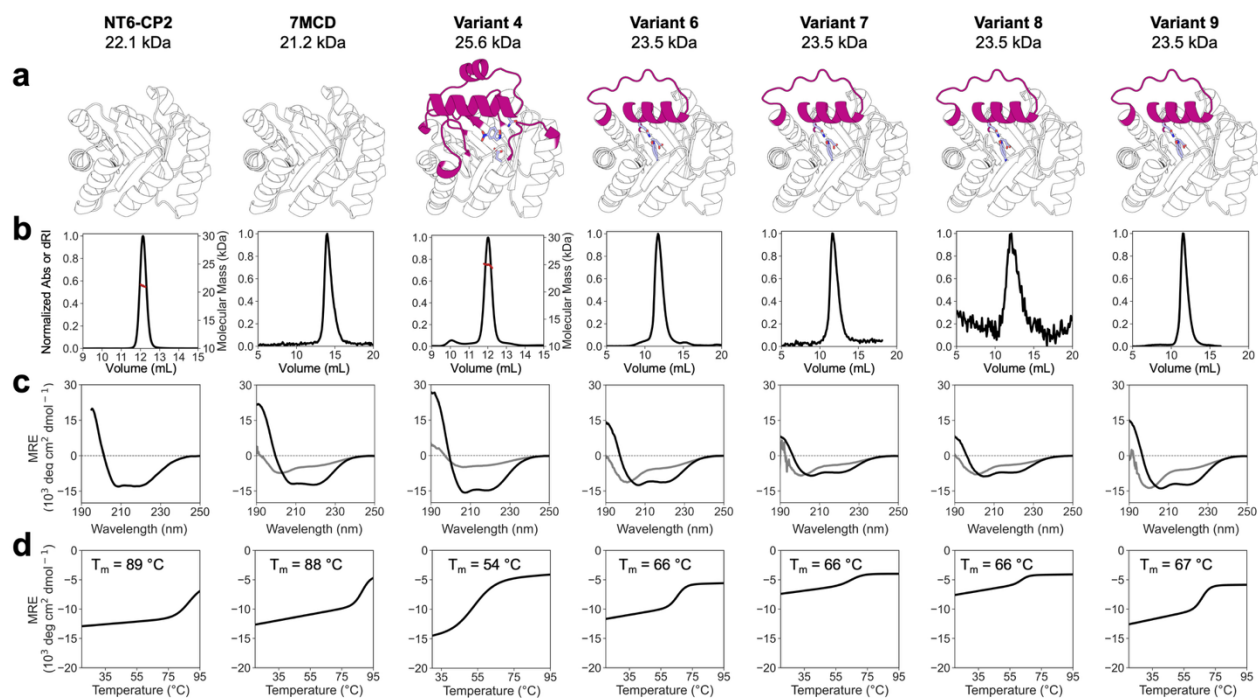


Figure 2.10. Biophysical characterization of expressible variants. (a) Models of minimal TIM barrels and CANVAS-designed variants, with the generated lids shown in magenta and the theozyme as sticks. (b) SEC-MALS (NT6-CP2, variant 4) and SEC (7MCD, variants 6–9) indicate that the proteins are predominantly monomeric. However, variant 4 shows a slight tendency to dimerize. The red line indicates the measured molecular mass of the species. Abs: absorbance. dRI: differential refractive index. (c) Circular dichroism spectra reveal a predominantly α -helical secondary structure but with a more pronounced peak at 208 nm compared to the minimal TIM barrels. Secondary scans at 20 °C and 95 °C are shown in black and grey, respectively. The 95 °C scan is not shown for NT6-CP2. MRE: mean residue ellipticity. (d) Melting curves demonstrate that variants 4 and 6–9 are substantially destabilized compared to their parent minimal TIM barrels. T_m : melting temperature.

Table 2.2. Post-purification culture yields.

Enzyme	Yield (mg L ⁻¹) ^a
NT6-CP1	–
NT6-CP2	70 ± 10
7MCD	31 ± 3
Variant 1	–
Variant 2	–
Variant 3	–
Variant 4	3 ± 1
Variant 5	–
Variant 6	12 ± 6
Variant 7	8
Variant 8	9
Variant 9	16 ± 7
KempTIM1	17 ± 9
KempTIM2	–
KempTIM3	59
KempTIM4	30 ± 10
KempTIM5	7
KempTIM4a	7 ± 5
KempTIM4b	30 ± 20
KempTIM4c	40 ± 10
KempTIM4d	30 ± 20
KempTIM4e	30 ± 20
KempTIM4f	40 ± 20
KempTIM4g	28 ± 1
KempTIM4h	30 ± 20

^a Data represent the average measurements from 1 to 17 independent protein batches, with error bars indicating the standard deviation. NT6-CP1, Variants 1, 2, 3 and 5, and KempTIM2 could not be expressed solubly.

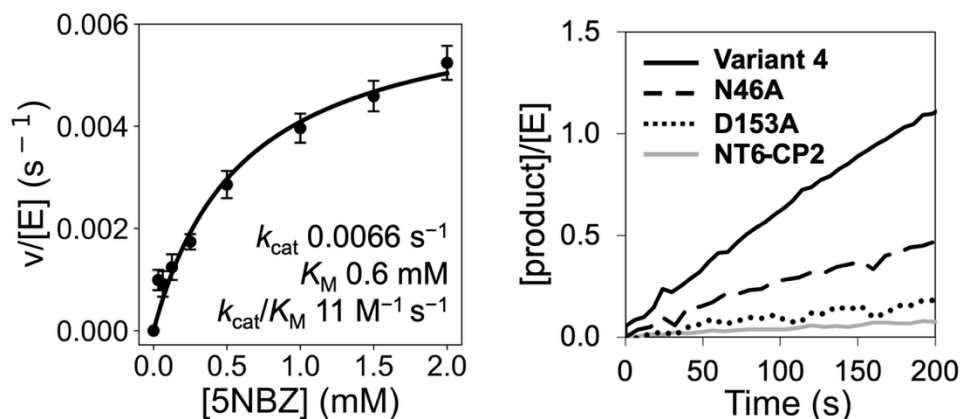


Figure 2.11. Variant 4 kinetic characterization. Michaelis–Menten plot (left) of normalized initial rates as a function of 5-nitrobenzoxazole (5NBZ) concentration. Data represent the average of 15 replicate measurements from five independent protein batches (mean $\pm$ SEM). Reaction progress curves (right) at 2 mM 5NBZ demonstrate the contributions of designed catalytic residues D153 and N46 to catalysis.

2.4.2 Rescuing expression and stability with scaffold redesign

Due to their lack of expression and poor stability, we endeavoured to redesign the scaffolds of variants 1–5. ProteinMPNN has been previously applied to rescue failed *de novo* designs¹³⁸ and stabilize natural enzymes²⁰⁰. Thus, we used ProteinMPNN to redesign all distal positions while preserving the Triad-designed active site, yielding five new variants henceforth termed KempTIMs. Expression and solubility were restored in all but one of the KempTIMs (Figure 2.12). The four expressible KempTIMs (KempTIM1, KempTIM3–5; Figure 2.13a) were purified as monomeric species (Figure 2.13b) at yields of 7–59 mg L⁻¹ (Table 2.2). In contrast with the initial variants, CD spectroscopy revealed a stronger α -helical signal with a greater 222/208 ratio than the template, NT6-CP2 (Figure 2.13c). Furthermore, all KempTIMs displayed melting temperatures above 70 °C, and several remained folded until 90 °C (Figure 2.13d). Notably, AlphaFold2 confidence was equivalent between KempTIM1–5 and variants 1–5, not indicative of the improved stability and expression of these new designs (Figure 2.7).

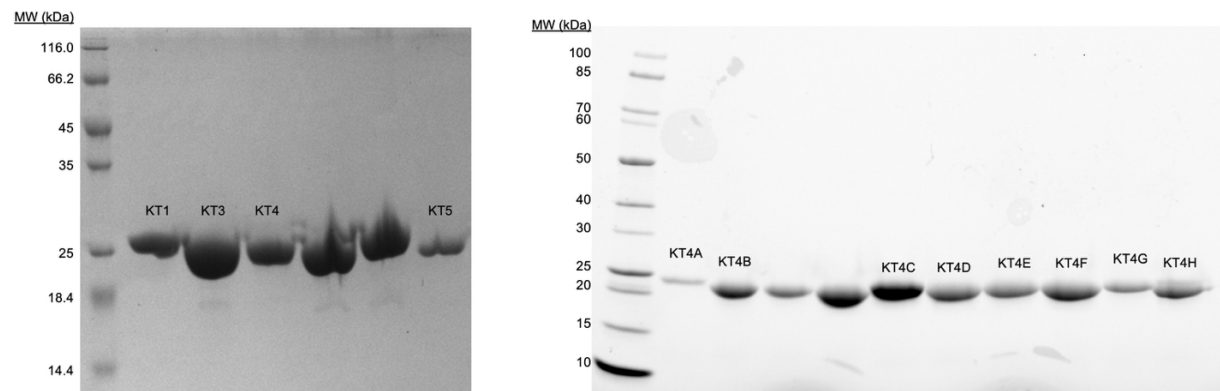


Figure 2.12. SDS-PAGE of KempTIMs. The well containing the relevant KempTIM (KT) is labeled. Mini-PROTEAN TGX Stain-Free Precast gels (Biorad) were loaded with each sample. The first column of each gel was loaded with a pEqGOLD Protein Marker I (VWR, left) or an unstained protein standard (NEB, right). MW: molecular weight.

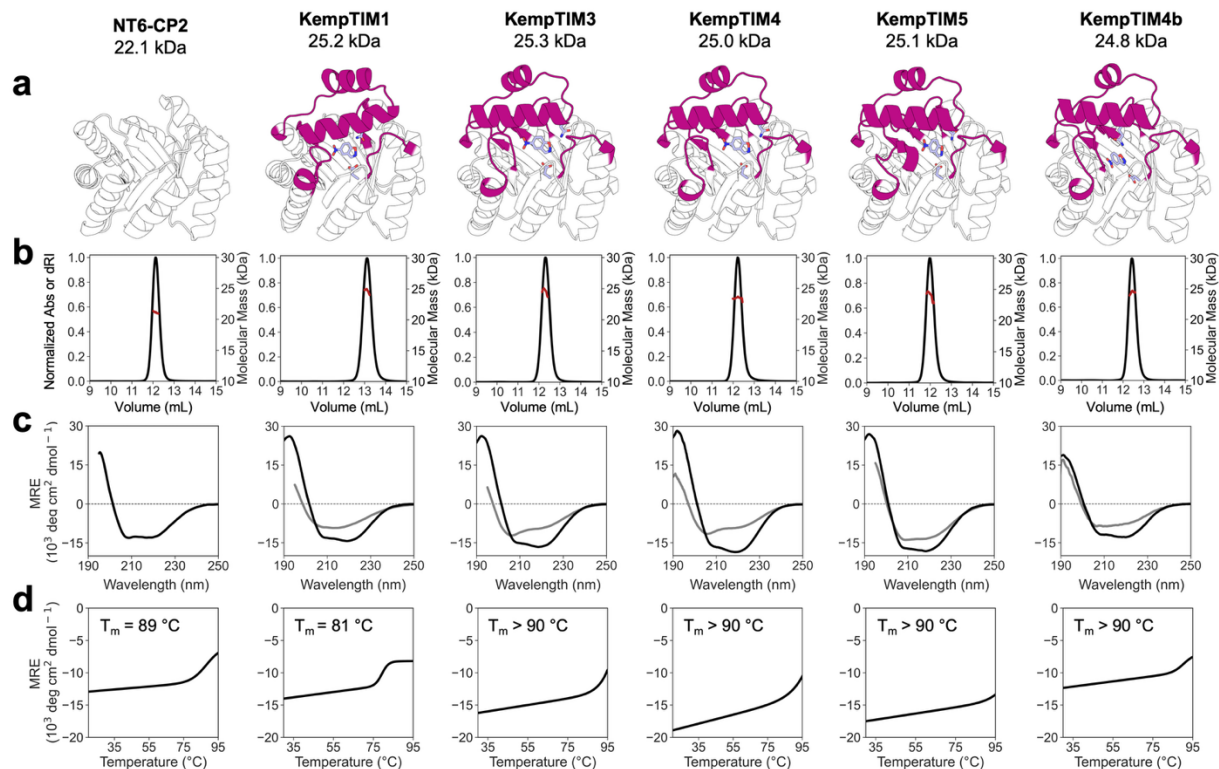


Figure 2.13. Biophysical characterization of expressible KempTIMs. (a) Models of minimal TIM barrel NT6-CP2 and KempTIMs, with the generated lids shown in magenta and the theozyme as sticks. (b) SEC-MALS indicates that the proteins are predominantly monomeric. The red line indicates the measured molecular mass of the species. Abs: absorbance. dRI: differential refractive index. (c) Circular dichroism spectra reveal a predominantly α -helical secondary structure, but with a more pronounced peak at 222 nm compared to NT6-CP2. Secondary scans at 20 °C and 95 °C are shown in black and grey, respectively. The 95 °C scan is not shown for NT6-CP2. MRE: mean residue ellipticity. (d) Melting curves demonstrate greater thermostability for designed KempTIMs than for the template NT6-CP2. T_m : melting temperature.

Kinetic analysis revealed activity for all expressed KempTIMs (Table 2.3, Figure 2.14a). The most active design, KempTIM1, exhibited a catalytic efficiency of $1,400 \text{ M}^{-1} \text{ s}^{-1}$ and turnover number of 4.2 s^{-1} at pH 7. Mutation of the catalytic base abolished activity (Figure 2.15a), whereas mutation of the H-bond donor decreased k_{cat} sixfold (Figure 2.15a,b, Table 2.3) without disrupting the secondary structure content (Figure 2.15c), confirming the intended functional roles of the designed residues. A pH-rate profile (Figure 2.16, Table 2.4) revealed increasing activity at alkaline pH, with a pKa of approximately 9, consistent with the catalytic base Asp residing in a hydrophobic environment. At pH 10, KempTIM1 showed a $k_{\text{cat}}/K_{\text{M}}$ of $21,000 \text{ M}^{-1} \text{ s}^{-1}$ and k_{cat} of 13.7 s^{-1} (Figure 2.14b, Table 2.3).

Table 2.3. Kinetic parameters of KempTIMs.

Enzyme	Base	H-bond donor	pH	Substrate	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M ($\text{M}^{-1} \text{s}^{-1}$)
<i>CANVAS</i>							
KempTIM1	D24	N131	7	5NBZ	4.2 ± 0.2	3.1 ± 0.2	$1,400 \pm 100$
	D24	N131	10	5NBZ	13.7 ± 0.5	0.66 ± 0.06	$21,000 \pm 2,000$
	D24	N131	7	6NBZ	5.6 ± 0.9	2.2 ± 0.5	$2,500 \pm 700$
	D24	N131	10	6NBZ	27 ± 3	0.6 ± 0.1	$45,000 \pm 9,000$
	D24A	N131	7	5NBZ	N.D.	N.D.	N.D.
	D24	N131A	7	5NBZ	0.65 ± 0.02	2.6 ± 0.1	250 ± 10
KempTIM3	D153	N46	7	5NBZ	0.0176 ± 0.0004	2.11 ± 0.08	8.3 ± 0.4
KempTIM4	D153	N46	7	5NBZ	N.D.	N.D.	7.01 ± 0.08
	D153	N46	10	5NBZ	N.D.	N.D.	18.6 ± 0.4
	D153A	N46	7	5NBZ	N.D.	N.D.	N.D.
	D153	N46A	7	5NBZ	N.D.	N.D.	6.5 ± 0.1
KempTIM5	D153	N46	7	5NBZ	N.D.	N.D.	1.60 ± 0.02
Variant 4	D153	N46	7	5NBZ	0.0066 ± 0.0006	0.6 ± 0.1	11 ± 3
<i>Ensemble-based design</i>							
KempTIM4a	D153	Q42	7	5NBZ	0.55 ± 0.01	1.18 ± 0.06	470 ± 20
KempTIM4b	D153	Q42	7	5NBZ	1.21 ± 0.04	0.39 ± 0.04	$3,100 \pm 300$
	D153	Q42	10	5NBZ	26.3 ± 0.9	0.83 ± 0.06	$32,000 \pm 3,000$
	D153	Q42	7	6NBZ	0.139 ± 0.008	0.22 ± 0.04	600 ± 100
	D153	Q42	10	6NBZ	1.02 ± 0.02	0.27 ± 0.02	$3,800 \pm 300$
	D153A	Q42	7	5NBZ	N.D.	N.D.	N.D.
	D153	Q42A	7	5NBZ	0.46 ± 0.03	0.7 ± 0.1	640 ± 90
KempTIM4c	D5	Q43	7	5NBZ	0.18 ± 0.01	2.6 ± 0.2	69 ± 6
KempTIM4d	D134	N46	7	5NBZ	0.61 ± 0.01	0.89 ± 0.04	690 ± 30
KempTIM4e	D134	N46	7	5NBZ	0.15 ± 0.01	2.8 ± 0.2	54 ± 5
KempTIM4f	D134	N46	7	5NBZ	N.D.	N.D.	19 ± 2
KempTIM4g	D134	N46	7	5NBZ	0.42 ± 0.06	3.1 ± 0.7	140 ± 30
KempTIM4h	D134	N46	7	5NBZ	N.D.	N.D.	9.7 ± 0.4

5NBZ and 6NBZ are 5-nitrobenzoxazole and 6-nitrobenzoxazole, respectively. k_{cat} and K_M were calculated by fitting the data to the Michaelis–Menten model $v_0 = k_{\text{cat}}[\text{E}_0][\text{S}]/(K_M + [\text{S}])$. Errors of nonlinear regression fitting, which represent the absolute measure of the typical distance that each data point falls from the regression line, are provided. 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 or 1 mM for 5NBZ or 6NBZ, respectively), which is the substrate’s solubility limit. In those cases, catalytic efficiencies (k_{cat}/K_M) were calculated from the slope of the linear portion ($[\text{S}] \ll K_M$) of the Michaelis–Menten model ($v = (k_{\text{cat}}/K_M)[\text{E}_0][\text{S}]$). $n = 1$ independent protein batch for KempTIM3 and KempTIM5. $n = 2$ independent protein batches for KempTIM1 and all ensemble-based designs. $n = 4$ independent protein batches for KempTIM4 and variant 4. At least two technical replicates were performed for each independent batch. Kinetic assays were carried out in 50 mM sodium phosphate containing 100 mM NaCl and 10% MeOH (pH 7) or 50 mM CHES containing 100 mM NaCl and 10% MeOH (pH 10).

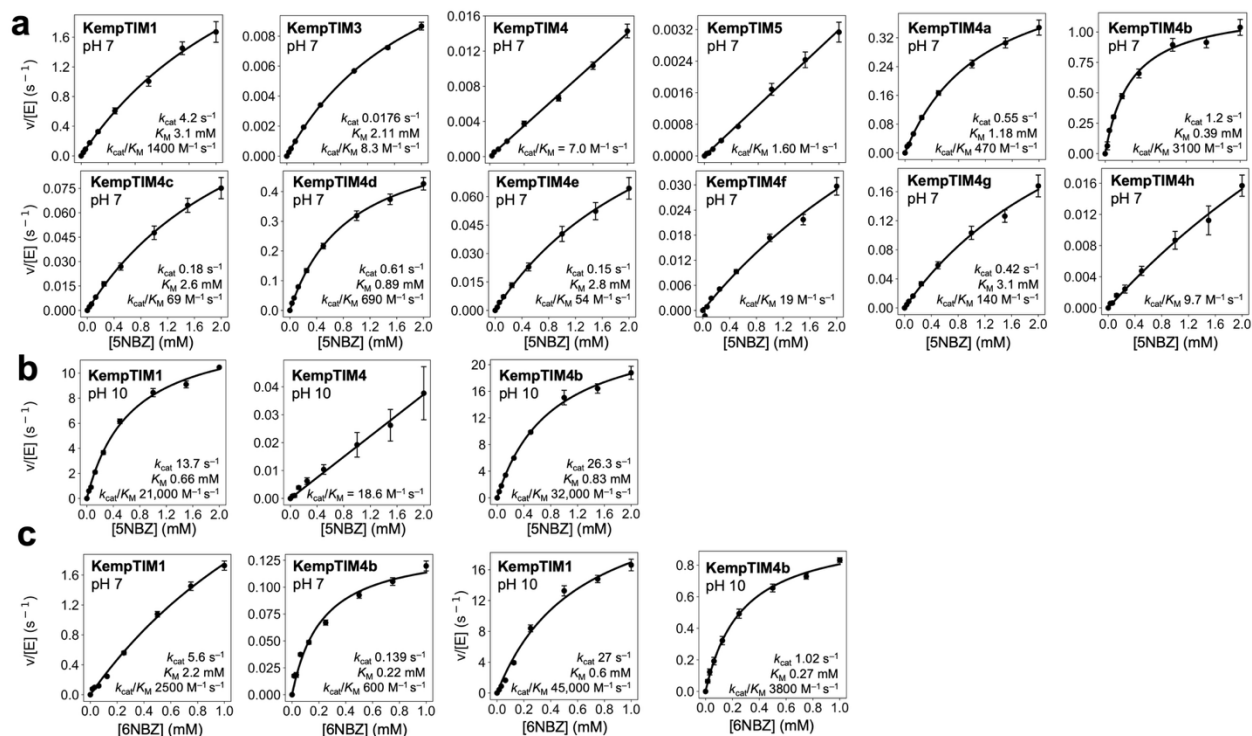


Figure 2.14. Michaelis–Menten plots of KempTIMs. Michaelis–Menten plots show normalized initial rates as a function of 5-nitrobenzoxazole (5NBZ, panels a and b) or 6-nitrobenzoxazole (6NBZ, panel c) concentration. Kinetic assays were carried out in 50 mM sodium phosphate containing 100 mM NaCl and 10% MeOH (pH 7, panel a) or 50 mM CHES containing 100 mM NaCl and 10% MeOH (pH 10, panel b). Data represent the mean $\pm$ SEM (or $\pm$ s.d. when $n=1$ protein batch) of ≥ 3 replicate measurements from ≥ 1 independent protein batches.

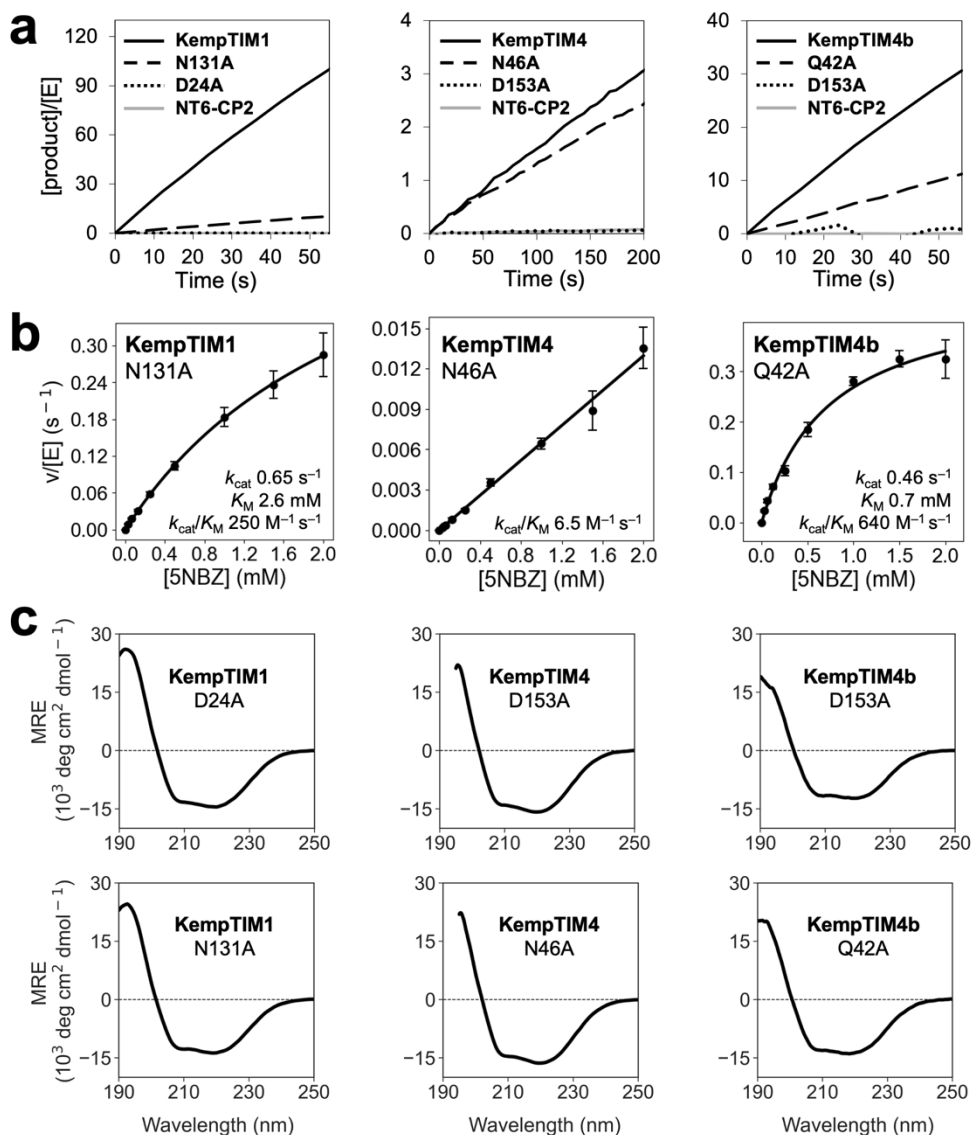


Figure 2.15. Characterization of KempTIM knockout mutants. (a) Reaction progress curves at 2 mM 5NBZ illustrating the contributions of the designed catalytic residues to overall catalysis. (b) Michaelis–Menten plots for variants lacking H-bond donor residues. Data represent the mean $\pm$ s.d. of ≥ 3 replicate measurements from one independent protein batch. Kinetic assays were carried out in 50 mM sodium phosphate containing 100 mM NaCl and 10% MeOH (pH 7). (c) Circular dichroism spectra demonstrate that mutation of catalytic residues does not change the secondary structure content. MRE: mean residue ellipticity.

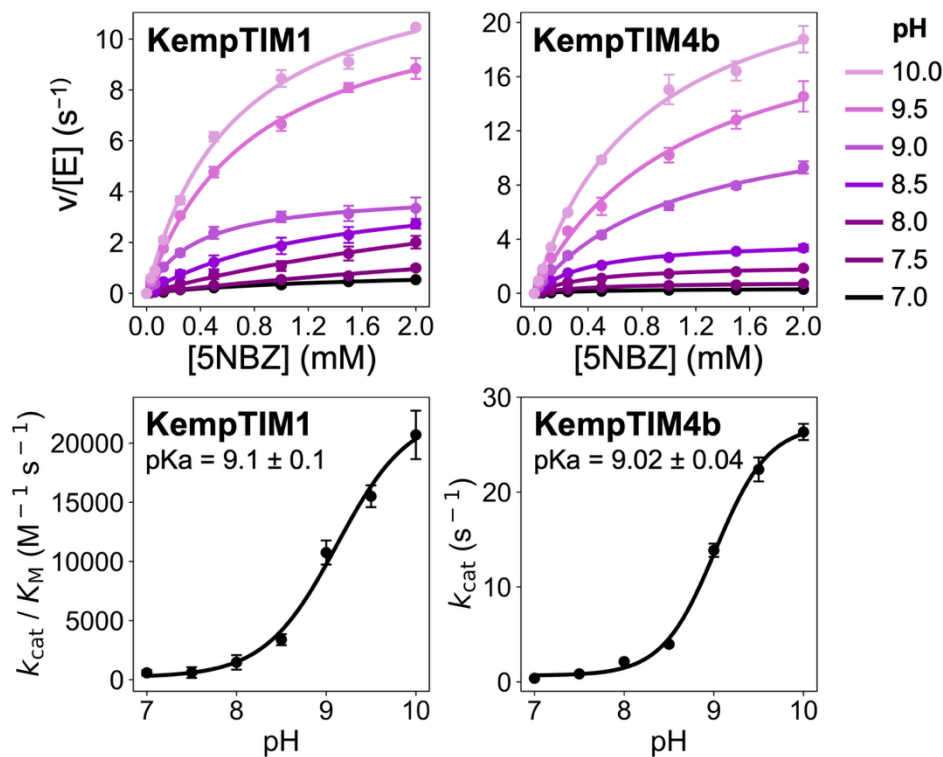


Figure 2.16. pH-rate profiles of KempTIM1 and KempTIM4b. Michaelis–Menten plots (top) at varying pH show normalized initial rates as a function of 5-nitrobenzisoxazole (5NBZ) concentration. pH-rate profiles (bottom) reveal a single inflection near pH 9, with greatest enzymatic activity observed at pH ≥ 10 . Kinetic assays were performed in 50 mM Tris buffer containing 100 mM NaCl and 10% MeOH (pH 7.0–8.5) or in 50 mM CHES buffer containing 100 mM NaCl and 10% MeOH (pH 9.0–10.0). Data represent mean $\pm$ SEM from ≥ 2 independent protein batches each with ≥ 2 replicate measurements.

Table 2.4. Kinetic parameters of KempTIM1 and KempTIM4b at various pH values.

Enzyme	pH	k_{cat} (s^{-1})	K_{M} (mM)	$k_{\text{cat}}/K_{\text{M}}$ ($\text{M}^{-1} \text{s}^{-1}$)
KempTIM1	7.0	4.2 ± 0.2	3.1 ± 0.2	1400 ± 100
	7.5	N.D.	N.D.	700 ± 50
	8.0	N.D.	N.D.	1700 ± 100
	8.5	4.4 ± 0.3	1.3 ± 0.2	3400 ± 500
	9.0	4.0 ± 0.1	0.37 ± 0.03	$11,000 \pm 1000$
	9.5	12.3 ± 0.3	0.79 ± 0.04	$15,600 \pm 900$
	10.0	13.7 ± 0.5	0.66 ± 0.06	$21,000 \pm 2000$
KempTIM4b	7.0	1.21 ± 0.04	0.39 ± 0.04	3100 ± 300
	7.5	0.84 ± 0.03	0.42 ± 0.04	2000 ± 200
	8.0	2.14 ± 0.08	0.43 ± 0.05	5000 ± 600
	8.5	4.0 ± 0.1	0.43 ± 0.03	9300 ± 700
	9.0	13.9 ± 0.7	1.1 ± 0.1	$13,000 \pm 2000$
	9.5	22 ± 1	1.1 ± 0.1	$20,000 \pm 3000$
	10.0	26.3 ± 0.9	0.84 ± 0.06	$32,000 \pm 3000$

k_{cat} and K_{M} using 5-nitrobenzisoxazole as substrate were calculated by fitting the data to the Michaelis–Menten model $v_0 = k_{\text{cat}}[E_0][S]/(K_{\text{M}} + [S])$. Errors of nonlinear regression fitting are provided. 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), which is the substrate’s solubility limit. In those cases, catalytic efficiencies ($k_{\text{cat}}/K_{\text{M}}$) were calculated from the slope of the linear portion ($[S] \ll K_{\text{M}}$) of the Michaelis–Menten model ($v = (k_{\text{cat}}/K_{\text{M}})[E_0][S]$). $n = 2$ independent protein batches. At least two technical replicates were performed for each independent batch. Kinetic assays were carried out in 50 mM sodium phosphate containing 100 mM NaCl and 10% MeOH (pH 7.0), 50 mM Tris containing 100 mM NaCl and 10% MeOH (pH 7.5–8.5), or 50 mM CHES with 100 mM NaCl and 10% MeOH (pH 9.0–10.0).

2.4.3 Confirming the lid and active-site features

To assess structural accuracy of the CANVAS-designed lids and their associated active sites, we solved crystal structures (Table 2.5) of the most efficient design, KempTIM1 ($21,000 \text{ M}^{-1} \text{ s}^{-1}$), and a lower-efficiency variant, KempTIM4 ($19 \text{ M}^{-1} \text{ s}^{-1}$), with a different lid. In both enzymes, the designed lids fold as intended, with the helix–turn–helix motif and elongated $\beta\alpha$ -loops closely matching the design models, with lid C_α RMSD of $0.9 \text{ \AA}$ and $1.3 \text{ \AA}$ for KempTIM1 and KempTIM4, respectively (Figure 2.17a,b). In both cases, however, the lid is slightly displaced relative to the TIM-barrel core, shifting the C_α position of the catalytic H-bond donor by $1.0 \text{ \AA}$ in KempTIM1 and $2.0 \text{ \AA}$ in KempTIM4 relative to the design models (Figure 2.17c,d). Because the lid region of KempTIM4 showed extensive crystal contacts, we probed its solution structure using small-angle X-ray scattering (SAXS). Comparison of different structural models with the SAXS data indicates that the lid adopts a conformation in solution similar to that observed in the crystal structure (Figure 2.18).

Table 2.5. Crystallographic data collection and refinement statistics.

	KempTIM1 + 6NBT	KempTIM1 – 6NBT	KempTIM4 – 6NBT
PDB ID	9TZD	9U01	9QKX
Data collection ^a			
Beamline	BESSY BL14.1	BESSY BL14.1	BESSY BL14.1
Wavelength [Å]	0.9184	0.9184	0.9184
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 3 ₁ 2 1
Unit cell [Å, °]	a = 41.22 b = 67.27 c = 70.92 α = β = γ = 90	a = 41.17 b = 66.69 c = 71.14 α = β = γ = 90	a = b = 59.8 c = 120.69 α = β = 90 γ = 120
Resolution range [Å]	33.63-1.2 (1.23-1.2)	33.35-1.25 (1.28-1.25)	47.59-2.30 (2.53-2.30)
Unique reflections	61955 (4030)	54927 (3588)	11469 (2772)
Multiplicity	13.0 (12.4)	12.6 (13.2)	19.5 (20.4)
Completeness [%]	99.22 (98.01)	99.98 (99.94)	98.65 (98.30)
R-meas [%]	0.07119 (2.28)	0.07025 (2.128)	0.4812 (2.946)
<I/σI>	16.97 (1.09)	14.09 (1.03)	8.24 (1.58)
CC _{1/2}	1 (0.468)	1 (0.573)	0.995 (0.404)
CC*	1 (0.798)	1 (0.854)	0.999 (0.759)
Wilson B-factor [Å ²]	15.18	17.30	33.87
Refinement			
R _{work} /R _{free} [%]	15.55 / 17.73	15.70 / 17.74	25.47 / 29.96
<u>No. of atoms (non-H)</u>			
macromolecules	1752	1754	1640
ligands	36	16	42
solvent	191	185	52
<u>RMSD from ideal geometry</u>			
bonds [Å]	0.007	0.007	0.001
angles [°]	1.03	0.91	0.310
<u>Ramachandran statistics</u>			
favoured [%]	100.00	99.08	98.63
outliers [%]	0.00	0.00	0.00
Clashscore	5.24	0.83	4.96
<u>Average B [Å²]</u>			
macromolecules	23.02	27.50	42.19
ligands	21.72	26.15	41.79
solvent	30.25	38.16	59.28
	33.56	39.30	41.22

^a Statistics for the highest-resolution shell are shown in parentheses.

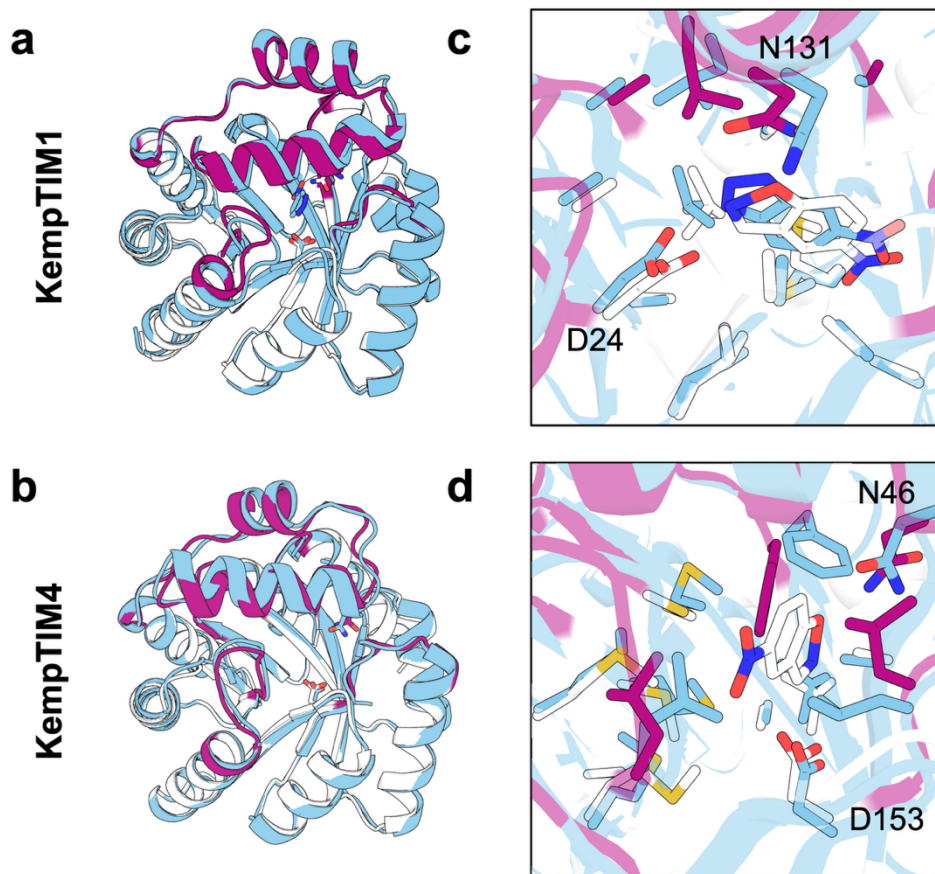


Figure 2.17. Crystal structures of KempTIM1 and KempTIM4. Experimental structures are shown in blue and design models in white (minimal TIM barrel) and magenta (lid). (a) The crystal structure of KempTIM1 closely matches the design model, with backbone C_{α} root-mean-square deviations (RMSDs) of 0.68 Å for the full structure and 0.88 Å for the lid. (b) The crystal structure of KempTIM4 also shows close backbone agreement with the design model (backbone C_{α} RMSD = 0.86 Å for the full structure and 1.34 Å for the lid). (c) Active-site residues within the minimal TIM barrel of KempTIM1 more closely match their designed conformations than those in the lid, consistent with an ~ 1 Å displacement of the lid relative to the TIM-barrel core. The co-crystallized transition-state (TS) analogue 6-nitrobenzotriazole adopts a binding pose similar to that in the design model, enabling hydrogen bonding to both the catalytic base D24 and the H-bond donor N131, despite N131 adopting a different rotamer in the crystal structure. (d) Active-site residues of KempTIM4 show substantial divergence from the design model, potentially reflecting the absence of a bound TS analogue. The catalytic base D153 adopts a similar rotamer in both structures, whereas the catalytic H-bond donor N46 adopts a rotamer incompatible with TS stabilization, likely caused by the ~ 2 Å displacement of the lid relative to the TIM-barrel core.

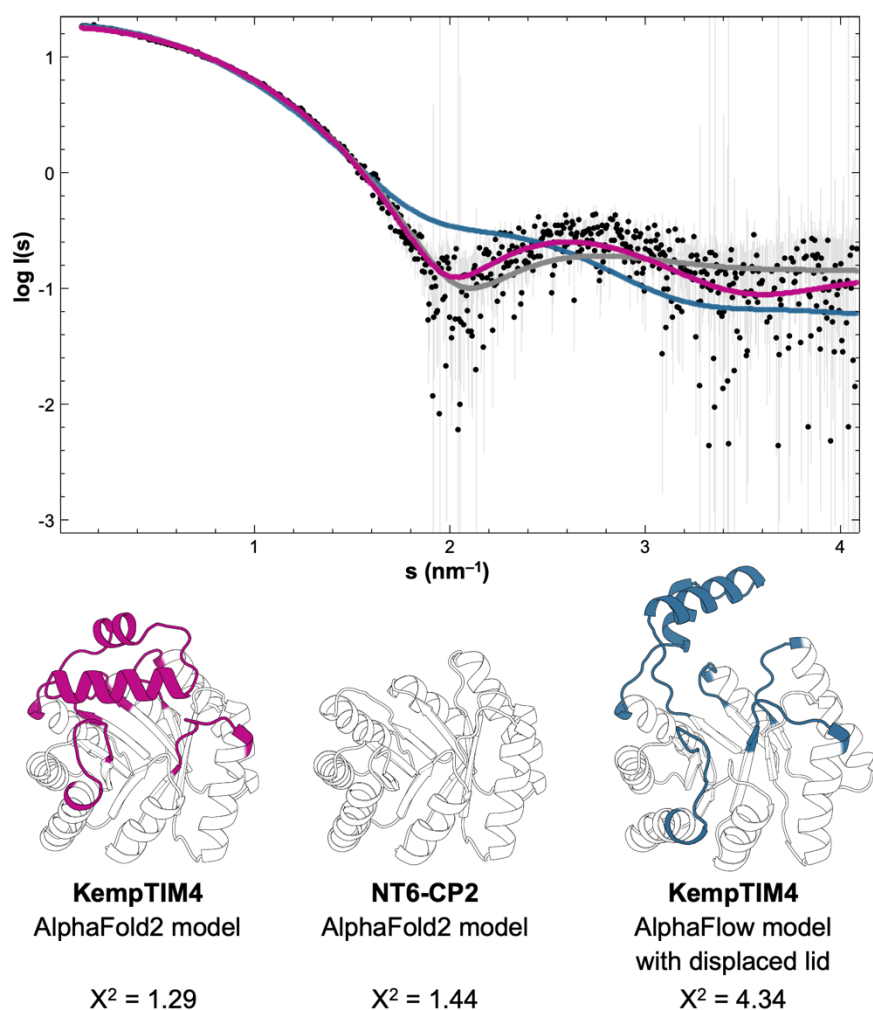


Figure 2.18. Comparison of KempTIM4 experimental SAXS data with theoretical curves. Experimental scattering data for KempTIM4 (black) align more closely with the calculated scattering curve from the KempTIM4 AlphaFold2¹⁴¹ model (magenta) than with models of the minimal TIM barrel NT6-CP2 (grey) or a KempTIM4 model with improperly positioned lid generated using AlphaFlow²⁰¹ (teal). The chi-value below each model represents the goodness of fit between the calculated scattering curves and the experimental data, where values closer to 1.0 indicate a better fit.

KempTIM1 was crystallized in the presence of the TS analogue 6-nitrobenzotriazole (6NBT), and we observed clear electron density for this ligand in the active site (Figure 2.19a). In this co-crystal structure, the designed catalytic base D24 adopts a rotamer similar to that of the design model, enabling formation of a hydrogen bond with the bound TS analogue despite the ligand being slightly rotated relative to its designed orientation (Figure 2.17c). This productive conformation of D24 is stabilized by an ordered water molecule (Figure 2.19a). In contrast, the H-bond donor N131 adopts an alternative rotamer yet still forms a hydrogen bond with 6NBT, consistent with the observed sixfold decrease in catalytic efficiency following mutation of this residue (Figure 2.15, Table 2.3). The structure of KempTIM1 was also determined without bound TS analogue. Notably, N131 exhibits a C_{α} displacement of 1.1 Å relative to the 6NBT-bound structure (Figure 2.19b), placing it closer to its location in the design model. This backbone shift indicates a lid rearrangement upon substrate binding, a behavior commonly observed in natural enzymes²⁰². This conformational adjustment is accompanied by a rotameric change in the neighbouring residue R130, whose side chain reorients toward the binding pocket in the co-crystal structure and forms a water-mediated hydrogen bond with the TS analogue (Figure 2.19c).

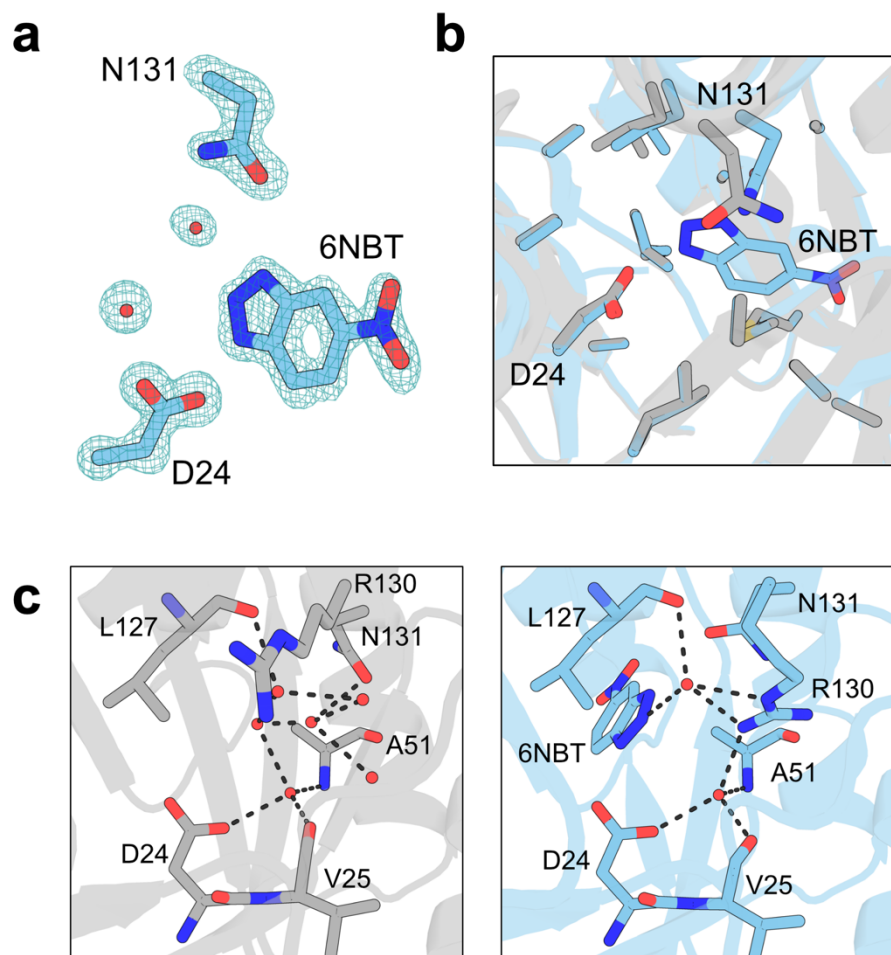


Figure 2.19. KempTIM1 crystal structures with and without 6-nitrobenzotriazole (6NBT). (a) Polder map of 6NBT, catalytic residues, and interacting water molecules contoured at $\pm 6\sigma$. Calculated map shows a well-defined density for all displayed atoms. (b) Comparison of the active sites in the 6NBT-bound (blue) and unbound (grey) structures shows that residues within the minimal TIM barrel adopt identical side-chain rotamers in both states. In contrast, the lid undergoes a subtle but concerted shift upon ligand binding, resulting in an altered C_α position and a corresponding rotameric change for the catalytic H-bond donor N131. This lid displacement indicates localized structural rearrangements associated with 6NBT binding. (c) Water networks in the active site of KempTIM1. In the unbound structure (grey, left panel), six ordered water molecules are present owing to the absence of bound ligand and a rotamer of R130 that places its side chain away from the active site. In the co-crystal structure (blue, right panel), two ordered water molecules are observed: one hydrogen bonds to the catalytic base D24 and the other to 6NBT. Both waters also interact with R130, which adopts a conformation that positions its guanidinium group within the active-site pocket.

The KempTIM1 co-crystal structure further shows a deviation of the TS analogue nitro group. Notably, the catalytic residues form hydrogen bonds that better accommodate a tautomeric form of 6NBT (Figure 2.17c, Figure 2.19, Figure 2.20) corresponding to the 6-nitrobenzotriazole substrate rather than the 5-nitro regioisomer targeted in the design. To determine whether this binding preference translates into altered regioselectivity, we measured kinetic parameters using 6-nitrobenzotriazole as substrate. KempTIM1 exhibits an approximately twofold higher k_{cat}/K_M ($45,000 \text{ M}^{-1} \text{ s}^{-1}$) with 6-nitrobenzotriazole compared to the 5-nitro regioisomer (Figure 2.14c, Table 2.3). This lack of regioselective substrate preference is consistent with the relatively large active-site pocket of KempTIM1, which could accommodate both regioisomers (Figure 2.21).

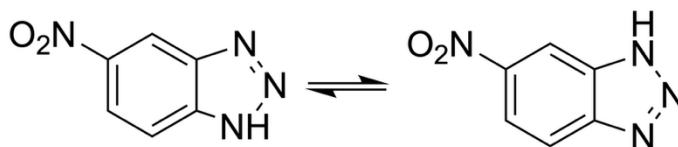


Figure 2.20. Tautomeric forms of 6-nitrobenzotriazole. The transition-state analogue 6-nitrobenzotriazole tautomerizes in water, which allows it to bind to enzymes that react with either 5- or 6-nitrobenzotriazole.

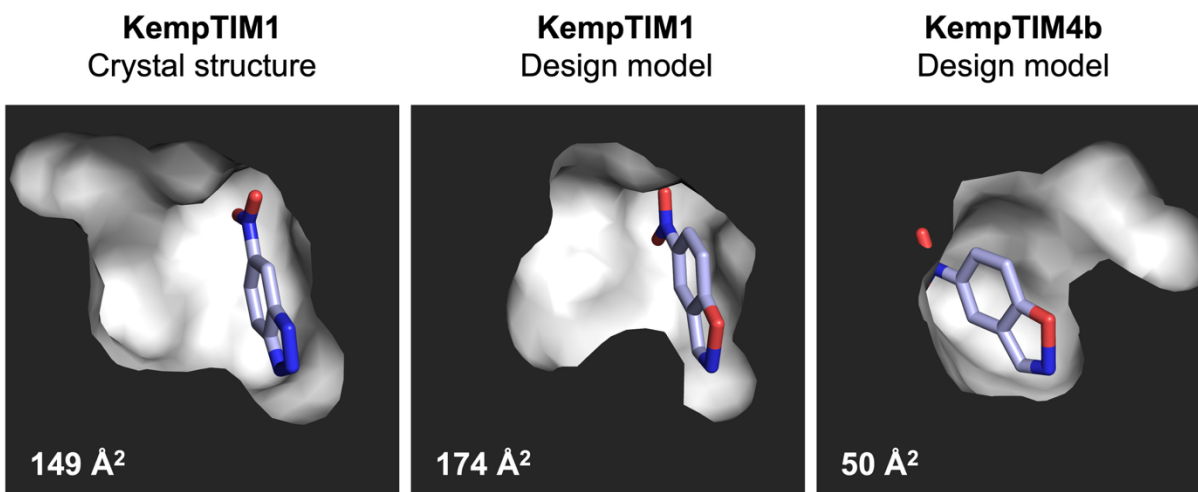


Figure 2.21. Active-site pocket sizes. Both the design model and experimental structure of KempTIM1 show relatively large binding pockets. In contrast, the design model of KempTIM4b contains a more tightly packed pocket. Solvent-accessible surface area values for the transition state are shown in the bottom-left corners of each panel.

Despite crystallizing the lower-activity design KempTIM4 in the presence of 6NBT, no electron density for this ligand was observed in the active site, indicating weak binding under the crystallization conditions. Instead, the observed density was best interpreted as a bound glycerol molecule from the cryoprotectant. Furthermore, most KempTIM4 active-site residues adopt conformations that differ from those intended to stabilize the TS, indicating limited preorganization (Figure 2.17d). While the catalytic base D153 adopts the designed rotamer, the H-bond donor N46 adopts a conformation incompatible with TS stabilization. Consistent with this structural observation, mutation of N46 to alanine results in only a minor decrease in catalytic activity (Figure 2.15, Table 2.3). Together, these structural features account for the far lower catalytic efficiency of KempTIM4 relative to KempTIM1 and offer clear targets for optimization.

2.4.4 Improving the efficiency of KempTIM4 by ensemble-based design

To correct active-site deficiencies observed in KempTIM4, we applied ensemble-based design^{75,151} (Figure 2.22). As described in Chapter 1, this approach uses an ensemble of protein backbones as design templates, revealing conformational substates that better accommodate the TS and its catalytic interactions⁷⁵.

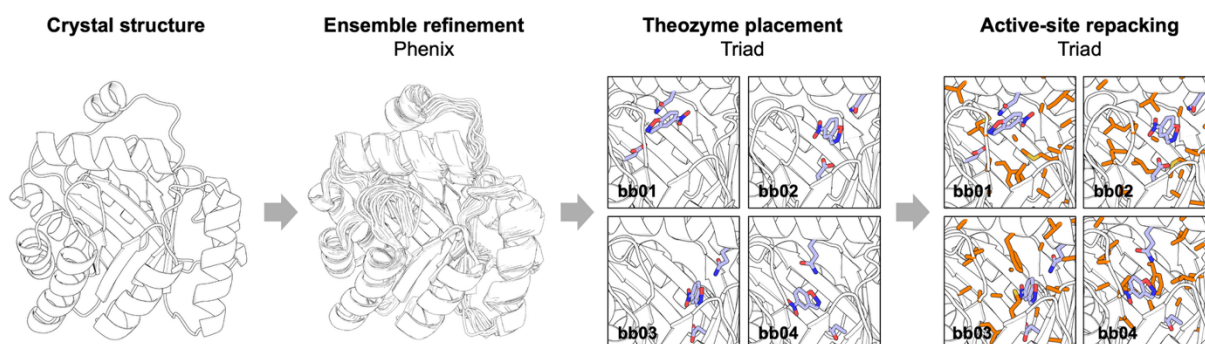


Figure 2.22. Ensemble-based enzyme design pipeline. A conformational ensemble is generated from the crystal structure of a low-activity enzyme using ensemble refinement⁴² in the Phenix software⁴³. Theozymes (blue sticks) containing an Asp catalytic base and either Asn or Gln as hydrogen-bond donors are then placed on each backbone template from the ensemble using the protein design software Triad. Thus, distinct catalytic motifs can be obtained on each backbone (bb01–bb04). Active sites (orange sticks) are subsequently packed around the transition state by optimizing the identity and side-chain conformations of designed residues, thereby maximizing favorable nonbonding interactions. Final designs are then filtered using structural metrics designed to reproduce mechanisms known to enhance enzyme catalysis.

For KempTIM4 optimization, a 60-member structural ensemble was generated by crystallographic ensemble refinement²⁰³. In this method, molecular dynamics explores protein conformations while restrained by electron density, providing an ensemble concordant with experimental data. Using each template of this ensemble for design, the catalytic base (Asp) was allowed to sample various positions within the β -barrel (Table 2.6). The H-bond donor (Asn or Gln) was allowed to sample β -barrel and lid positions (Table 2.6), but the target geometry could only be achieved with this residue on the designed lid. Following theozyme placement, the active site was packed using a sequence space of aromatic and hydrophobic residues as well as the wild-

type KempTIM4 residue, further refined by LigandMPNN¹³¹ residue probabilities at each position. The design models were filtered for the target catalytic contacts and active-site accessibility. To confirm the intended TS binding site, 6NBT-bound structures of each sequence were predicted using the protein-ligand co-folding model Boltz-2²⁰⁴ (Figure 2.21). This filtering pipeline yielded eight redesigned variants (KempTIM4a–h) comprising one of three distinct catalytic motifs (Figure 2.6, Table 2.1), none of which reproduced the motif found in the parent KempTIM4.

Table 2.6. Amino-acid positions for theozyme placement during ensemble-based design

Input structure	Residues for catalytic base placement	Residues for catalytic H-bond donor placement	Residues mutated to alanine
KempTIM4 crystallographic ensemble	M5, M26, M85, M104, G134, D153, A181, V200	M5, M26, M39, G42, L43, A45, N46, M85, F202	A3, M5, M26, A28, A38, M39, G42, L43, A45, N46, A83, M85, M104, L106, S108, S132, G134, D153, L155, E156, R157, A179, A181, L183, V200, F202

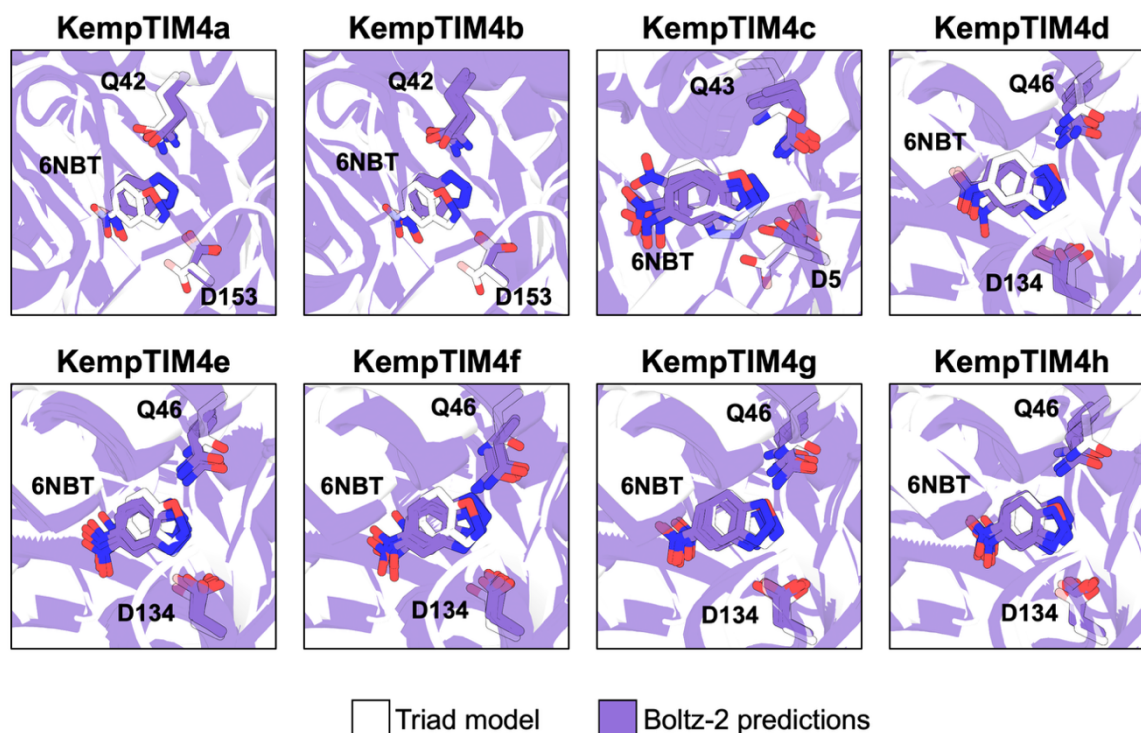


Figure 2.23. Boltz-2 predictions of ensemble-based designs. Five Boltz-2-predicted structures (purple) of KempTIM4 variants bound to the transition-state analogue 6-nitrobenzotriazole (6NBT) are overlaid on the Triad design model (white). The predicted 6NBT binding poses show strong agreement with the design model, with RMSDs $< 1 \text{ \AA}$ in all cases. Catalytic contacts between designed residues and 6NBT are preserved, even when the catalytic residue adopts a different side-chain rotamer in the Boltz-2 models compared with the Triad model.

All second-round designs expressed solubly (Figure 2.12, Table 2.2) and exhibited increased catalytic efficiency relative to KempTIM4 at pH 7, with two variants displaying improvements of at least two orders of magnitude (Figure 2.14a, Table 2.3). The most active design, KempTIM4b, showed similar biophysical properties to KempTIM4 (Figure 2.13), but achieved a >400-fold improvement in k_{cat}/K_M ($3100 \text{ M}^{-1} \text{ s}^{-1}$) at pH 7, surpassing even the most active first-round design KempTIM1 (Figure 2.14a, Table 2.3). Mutation of the catalytic base of KempTIM4b abolished activity (Figure 2.15a), whereas mutation of the H-bond donor reduced catalytic efficiency approximately fivefold (Figure 2.15a,b). This sharply contrasts with the parent KempTIM4, in which mutation of the H-bond donor had little effect. Like KempTIM1, KempTIM4b also showed enhanced activity at alkaline pH, with a pK_a around 9 (Figure 2.16, Table 2.4). At pH 10, k_{cat} increased to 26 s^{-1} and k_{cat}/K_M improved >1600-fold to $32,000 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 2.14b, Table 2.3).

In contrast with KempTIM1, KempTIM4b reacted more efficiently with the intended substrate 5-nitrobenzisoxazole than with 6-nitrobenzisoxazole, exhibiting 5- and 8-fold higher k_{cat}/K_M at pH 7 and 10, respectively (Figure 2.14c, Table 2.3). Consistent with this regioselective substrate preference, the active-site pocket of the KempTIM4b design model shows tighter packing compared to the structure and design model of KempTIM1 (Figure 2.21), reflecting the enhanced TS complementarity afforded by ensemble-based design⁷⁵.

2.4.5 Molecular dynamics of KempTIMs

To probe the origins of activity differences among the designed enzymes, we performed molecular dynamics simulations (10 replicates of 100 ns) of select KempTIMs, using their design models as starting structures. Most regions of these designs were somewhat rigid (Figure 2.24). The most flexible regions tended to be the major lid fragment or loop extensions near the active

sites of the KempTIMs. Notably, the highly active design KempTIM1 demonstrated the greatest major lid fragment rigidity of any designs, suggesting that increased rigidity of this fragment, which contains the catalytic H-bond donor, may help stabilize the substrate and maintain catalytically productive geometry. In general, the most flexible regions, such as the extended loops and C-terminal ends of the major fragment α -helix (Figure 2.24), also displayed the lowest AlphaFold2 confidence (Figure 2.7).

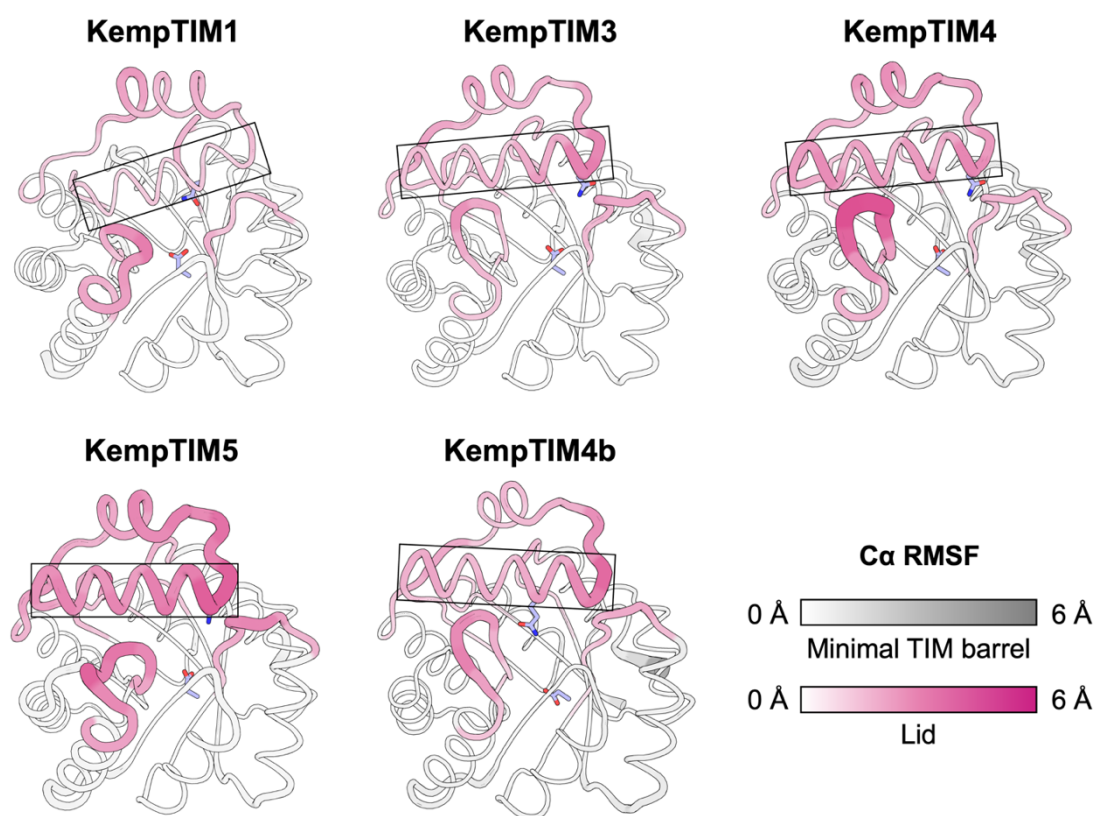


Figure 2.24. Backbone flexibility of select KempTIMs. C α root-mean-square fluctuations (RMSF), shown as putty plots, indicate that the lids are consistently more flexible than the minimal TIM-barrel core across all designs. Within the lids, loop regions display the highest mobility, whereas the α -helix anchoring the designed catalytic H-bond donor (boxed) remains comparatively rigid. This helix is especially rigid in the most active first-round design, KempTIM1. Catalytic residues are shown as blue sticks.

To evaluate the ability of designed active sites to maintain catalytically competent poses, we ran simulations (10 replicates of 100 ns) of all active KempTIMs bound to the TS analogue 6NBT. For each trajectory snapshot, we calculated both the distance and angle of the two designed catalytic contacts. The most active variants displayed narrow distributions centred on ideal values³⁴ (Figure 2.25), whereas lower-activity variants showed broadened distributions for one or both contacts. The H-bond donor interaction was generally the least consistently formed, in agreement with the observed lid displacement in the crystal structures (Figure 2.17c,d) as well as the mutational studies (Figure 2.15). We calculated a catalytic competency score⁷⁵, defined as the percent of simulation snapshots in which both catalytic contacts satisfied geometric criteria (Figure 2.26a) expected for efficient catalysis. Higher-activity variants like KempTIM1 and KempTIM4b scored better than lower-activity variants like KempTIM4 (Figure 2.26b), and catalytic competency scores correlated moderately with measured $k_{\text{cat}}/K_{\text{M}}$ values across the variant set (Pearson R = 0.76, Figure 2.26c). These findings indicate that this metric may serve as a useful in-silico filter for prioritizing future enzyme designs prior to experimental characterization.

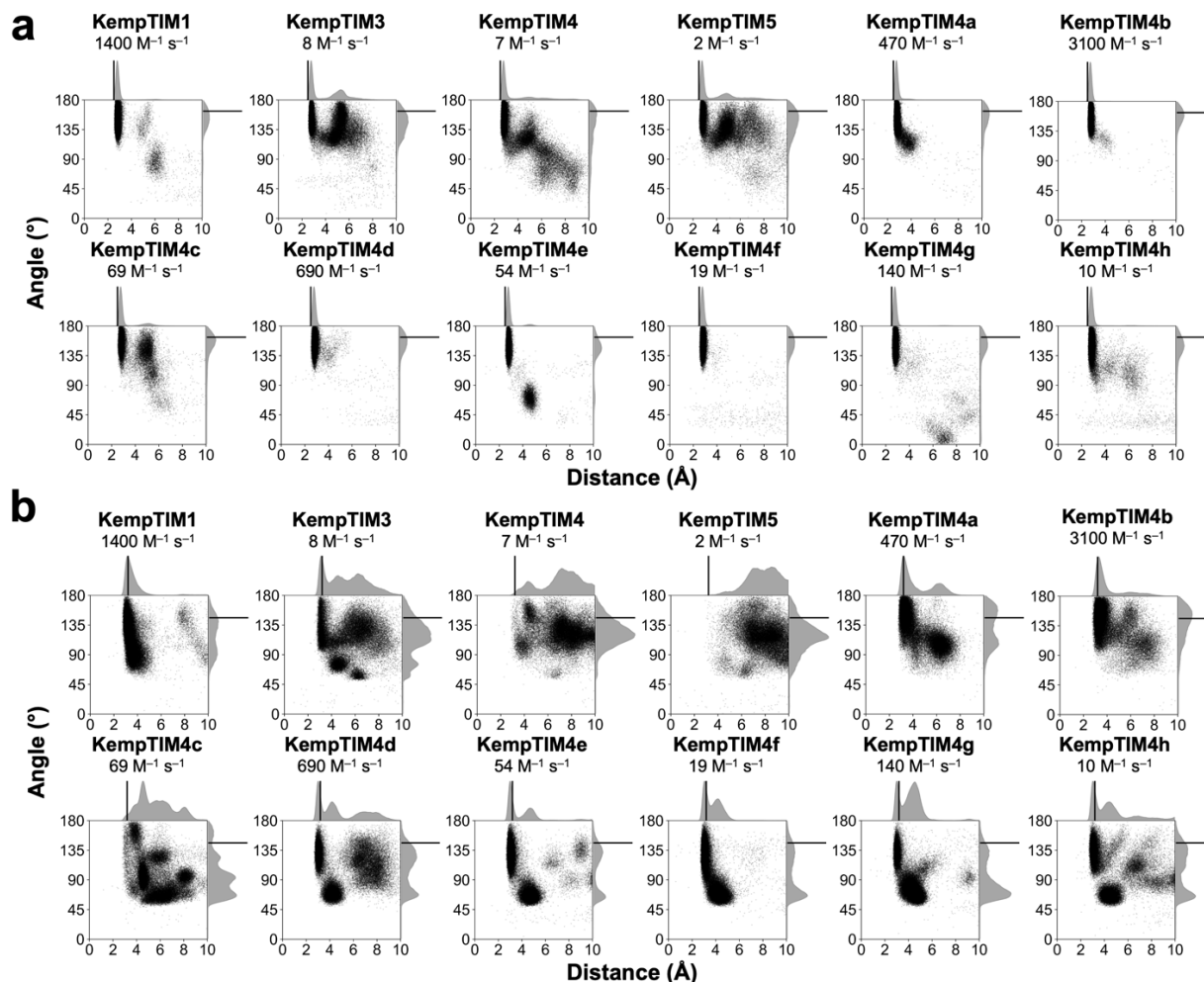


Figure 2.25. Catalytic contacts assessed by molecular dynamics. Scatter plots of angle versus distance for the base (top) and H-bond donor (bottom) contacts, compiled from ten 100-ns MD trajectories of KempTIM variants bound to the transition-state analogue 6-nitrobenzotriazole. For reference, the base/H-bond donor distances and angles in the crystal structure of HG3.17 (PDB ID: 5RGE)¹⁵¹, 2.5/3.2 Å and 163.0/146.1°, respectively, are shown as solid lines, representing ideal values associated with the high catalytic efficiency of this extensively evolved Kemp eliminase. Definitions of catalytic distances and angles between 6NBT and the designed catalytic residues are provided in Figure 2.26a. Catalytic efficiencies at pH 7 are listed for each variant.

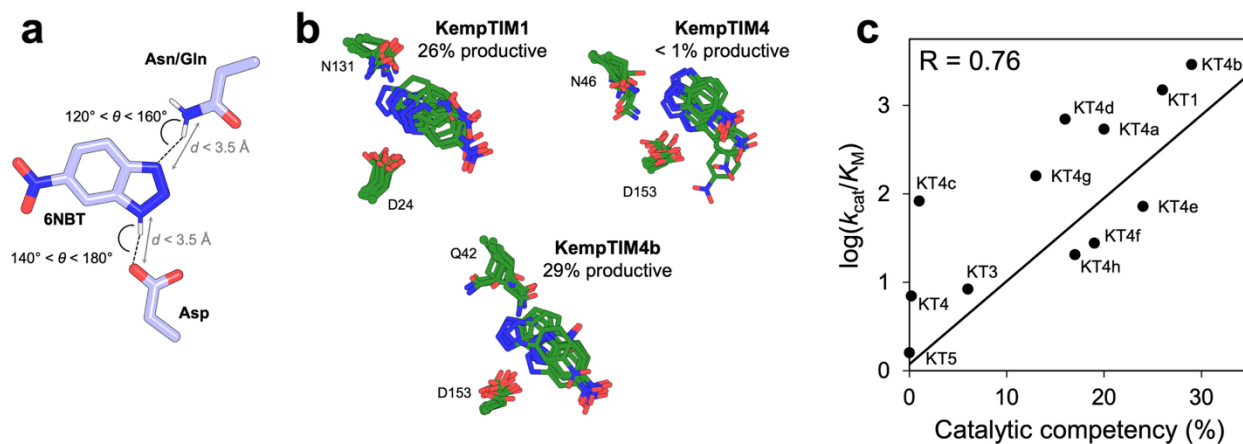


Figure 2.26. Molecular dynamics-derived catalytic competency moderately correlates with experimental efficiency. (a) Geometric definitions of the catalytic contacts between 6-nitrobenzotriazole (6NBT) and the designed general base and H-bond donor. The catalytic angle (θ) and heavy-atom distance (d) thresholds defining competent contacts are indicated. (b) Representative high-efficiency variants KempTIM1 and KempTIM4b show a higher catalytic competency score, defined as the fraction of molecular dynamics snapshots in which both catalytic contacts simultaneously satisfy the geometric criteria. In contrast, the lower-efficiency KempTIM4 shows a lower score. (c) Correlation between the catalytic-competency score and experimental catalytic efficiencies. KempTIM (KT) variants with higher catalytic competency display higher activities ($R = 0.76$), with KempTIM1 and KempTIM4b ranking the highest.

2.5 Discussion

Natural TIM barrels use lids to bind substrates and form active sites conducive to catalysis¹⁶⁴. In contrast, *de novo* TIM barrels lack such lids and are functionless. Using the CANVAS approach, we demonstrate that extended loops and custom lid fragments can transform these inert scaffolds into efficient enzymes. By designing various lids to simultaneously anchor a catalytic residue and form a substrate binding pocket, we created a series of *de novo* Kemp eliminases. One of these, KempTIM1, ranks among the most active *de novo* Kemp eliminases whose active site was obtained from a single round of computational design. Its catalytic efficiency is greater than that of first-round Kemp eliminases designed in scaffolds derived from natural TIM barrels^{13,21} or recombined TIM-barrel fragments⁶⁹, despite testing fewer variants. Thus, equipping minimal TIM barrels with custom lids provides an effective approach to *de novo* enzyme design.

Importantly, CANVAS designs are optimizable. Using a crystallographic ensemble^{75,151} of KempTIM4 as a template for design, we increased catalytic efficiency by more than two orders of magnitude. The catalytic efficiency and substrate selectivity of KempTIM4b demonstrate that *de novo* TIM barrels equipped with CANVAS-designed lids can host fully functional, selective active sites and provide a platform for further optimization toward even more powerful catalysts. Further activity improvements might be achieved using experimental optimization such as NMR-guided directed evolution⁹⁵, which has previously uncovered distal mutations that boost Kemp elimination rates in *de novo* enzymes to values approaching the diffusion limit²⁰⁵.

CANVAS enables the modular customization of lids to satisfy the geometric and chemical requirements of many reactions. This design versatility becomes more critical as the number of catalytic residues grows, a regime where design using pre-existing backbones becomes increasingly difficult. While recent methods address this problem by building entirely new folds around catalytic motifs^{67,68,73} yielding functional *de novo* enzymes, the resulting proteins often bear little resemblance to natural enzymes. Furthermore, the stable, regular structures generated by these methods may not be amenable to the structural rearrangements necessary for efficient catalysis^{164,182}. In contrast, minor rearrangements can be recapitulated in CANVAS-designed enzymes, demonstrated by the shift in lid conformation in KempTIM1 upon TS analogue binding. CANVAS therefore offers a complementary strategy, retaining a familiar, well-behaved fold while creating novel architecture tailored to the essential catalytic features.

In the context of TIM barrels, this approach maintains a fold that is inherently versatile and widely used in nature across diverse chemistries¹⁶⁵. The β -barrel provides a stable base to house the active site, while the lid offers a tuneable element for controlling substrate binding and anchoring any catalytic residues. Previous enzyme design efforts have sought to harness this

versatility by repurposing natural TIM barrels^{13,21–23} or by recombining TIM-barrel fragments in silico⁶⁹, approaches that restrict design space to pre-existing architectures. Instead, CANVAS facilitates the design of bespoke lids that are restricted only by the catalytic demands. As a result, minimal TIM barrels can be employed as blank canvases for enzyme construction. We anticipate that this approach will unlock the full catalytic potential of designed TIM-barrel proteins for diverse chemical transformations.

2.6 Methods

2.6.1 Reagents and solutions.

All experiments were conducted using analytical-grade chemicals, with solutions prepared using double-distilled water.

2.6.2 Circular permutation of NovoTIM6-SB with inpainting

Circular permutation of NovoTIM6-SB (PDB ID: 7OSV¹⁷³) was performed using the inpainting method described by Wang *et al.*²⁰⁶ The first residue of either the first or third β -strand was selected as the new N-terminus, with the corresponding $\beta\alpha$ -loop deleted (Figure 2.3). A new connection for the corresponding $\beta\alpha$ -loop was modeled by inpainting 5 to 10 residues. Unresolved residues in the crystal structure were remodeled, keeping intact the positions and identities of the remaining residues. For each new connection, 100 designs were modeled, followed by relaxation and scoring using the Rosetta protein design software²⁰⁷. Structures of each design were predicted using AlphaFold2¹⁴¹ with Model 4 weights. Designs with a pLDDT value above 90 and a Rosetta score below 3 Rosetta energy units per residue were kept. One design from each connection was selected for experimental characterization (Table 2.1). The proteins were named NT6-CP1 (N-terminus on the first β -strand) and NT6-CP2 (N-terminus on the third β -strand).

2.6.3 Computational enzyme design

All calculations were performed with the Triad protein design software¹⁹³ (Protabit, Pasadena, CA, USA). Rotamer optimization was performed using a Monte Carlo with simulated annealing search algorithm. Input structures were prepared for Triad calculations via the addH.py application within Triad. The Kemp elimination transition state (TS) structure was built using parameters from Privett *et al.*²¹. To provide sidechain conformations, a backbone-independent conformer library (bda-bbind_1.0)²⁰⁸ was used for theozyme placement, and a backbone-independent rotamer library (bbind02.May.e2)¹⁸ with expansions of ± 1 standard deviation around χ_1 and χ_2 was used for active-site repacking. Energies were calculated using a modified version of the Phoenix energy function^{193,209} consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field²¹⁰ 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 Å²¹¹, and an electrostatic term modeled using Coulomb’s law with a distance-dependent dielectric of 10. An energy benefit of –100 kcal mol⁻¹ was applied when TS-side-chain interactions satisfied catalytic contact geometries (Figure 2.27, Table 2.7), as described by Lassila *et al.*¹⁵.

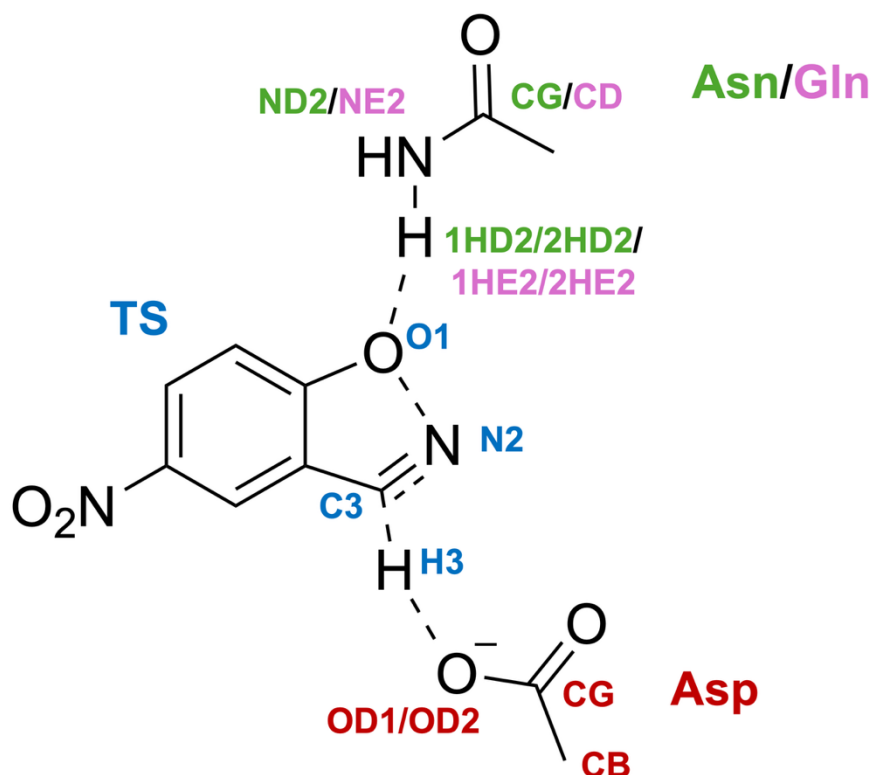


Figure 2.27. Atom definitions used in theozyme placement and design filtering. Relevant atoms of the catalytic asparagine/glutamine, aspartate, and transition state are shown in green/pink, red, and blue, respectively.

Table 2.7. 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
Base	Asp	Distance	OD1 or OD2	H3			1.0	1.6
		Angle	CG	OD1 or OD2	H3		109	131
		Angle	OD1 or OD2	H3	C3		159	180
		Torsion	CB	CG	OD1 or OD2	H3	-21, 159	21, 201
H-bond donor	Asn	Distance	1HD2 or 2HD2	O1			1.2	2.3
		Angle	ND2	1HD2 or 2HD2	O1		145	157
		Angle	1HD2 or 2HD2	O1	N2		120	140
		Torsion	1HD2 or 2HD2	O1	N2	C3	-20, 160	20, 200
	Gln	Distance	1HE2 or 2HE2	O1			1.2	2.3
		Angle	NE2	1HE2 or 2HE2	O1		145	157
		Angle	1HE2 or 2HE2	O1	N2		120	140
		Torsion	1HE2 or 2HE2	O1	N2	C3	-20, 160	20, 200

^a Atoms of the catalytic asparagine/glutamine, aspartate, and transition state are coloured in green/pink, red, and blue, respectively.

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

2.6.4 Theozyme placement

AlphaFold2¹⁴¹ predictions of NT6-CP1 and NT6-CP2, along with the 7MCD crystal structure, were used as backbone templates for theozyme placement. Amino-acid residues with a C_{α} – C_{β} vector oriented towards the β -barrel interior and located toward the C-terminus of each β -strand were selected as positions for introduction of the catalytic base (Asp). Neighbouring residues were mutated to alanine to avoid steric clashes with the TS (Table 2.8). TS poses were built using the contact geometries (Figure 2.27) listed in Table 2.9, and those where the isoxazolic oxygen pointed toward the catalytic face of the TIM barrel and remained coplanar with the catalytic base were selected (Figure 2.4). In parallel, the catalytic H-bond donor (Asn or Gln) was placed on a Gly-Asn/Gln-Gly tripeptide scaffold, and TS poses were built using H-bond donor side-chain contacts specified in Table 2.9. The resulting TS was superimposed with the TS from the catalytic base calculation to create the full theozyme, positioning the catalytic base within the minimal TIM barrel and the H-bond donor in the empty space above the catalytic face. The Gly residues from the tripeptide were then deleted, preserving only the Asn/Gln residue, whose alpha carbon atom served as an anchor for the designed lid.

Table 2.8. Amino-acid positions for theozyme placement.

Input structure	Residues for catalytic base placement ^a	Residues mutated to alanine ^a
7MCD	I4, I31, I50, I77, I96, I123, I142, I169	N6, I31, N52, I77, N98, I123, N144, I169
NT6-CP1	R5, I24, R51, I70, R97, I116, R143, I162	A3, R5, I24, D26, A49, R51, I70, D72, A95, R97, I116, D118, A141, R143, I162, D164
NT6-CP2	R5, I24, R51, I70, R97, I116, R143, I162	A3, R5, I24, D26, A49, R51, I70, D72, A95, R97, I116, D118, A141, R143, I162, D164

^a Residue numbering based on input structure

Table 2.9. Geometric definitions for generation of transition-state poses off the side chain of catalytic base and H-bond donor.

Catalytic 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
Asn	Distance	1HD2 or 2HD2	O1			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	ND2	1HD2 or 2HD2	O1		145, 151, 157
	Angle	1HD2 or 1HD2	O1	N2		120, 135, 140
	Torsion	CG	ND2	1HD1 or 2HD2	O1	120, 135, 140
	Torsion	ND2	1HD2 or 2HD2	O1	N2	-121, -116, -111
	Torsion	1HD2 or 2HD2	O1	N2	C3	160, 180, 200
Gln	Distance	1HE2 or 2HE2	O1			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	NE2	1HE2 or 2HE2	O1		145, 151, 157
	Angle	1HE2 or 2HE2	O1	N2		120, 135, 140
	Torsion	CD	NE2	1HE2 or 2HE2	O1	120, 135, 140
	Torsion	NE2	1HE2 or 2HE2	O1	N2	-121, -116, -111
	Torsion	1HE2 or 2HE2	O1	N2	C3	160, 180, 200

^a Atoms of the catalytic asparagine/glutamine, aspartate, and transition state are coloured in green/pink, red, and blue, respectively.

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

2.6.5 Lid design

For all theozyme placements, peptide connections between the catalytic H-bond donor residue and the minimal TIM barrel structures were designed using RFdiffusion¹¹¹. For each position of the H-bond donor alpha carbon, various $\beta\alpha$ -loops were selected as insertion points for a primary peptide fragment, and 10–30 residues were sampled to connect the catalytic residue while preserving the TIM-barrel topology. Additional $\beta\alpha$ -loops were sometimes elongated to improve interactions with the primary fragment (Figure 2.5b). For each inserted primary fragment, 100–200 structures were generated. Primary fragments lacking secondary structure, causing steric clashes with the TS, or displacing the H-bond donor from its theozyme position were discarded (Figure 2.4). The sequences of the filtered fragments were designed using ProteinMPNN¹³⁸ with a temperature factor of 0.1, without modifying the minimal TIM barrel sequence. Cysteine and methionine residues were excluded during sequence design. Structures of designed sequences were predicted using AlphaFold2¹⁴¹ or ColabFold²¹² with all five model weights, and those displaying an average lid pLDDT value > 85 and backbone RMSD $< 3 \text{ \AA}$ relative to the RFdiffusion model were selected for active-site packing.

2.6.6 Active-site packing

Predicted models generated above were used for a theozyme placement step, as described previously, to verify that contacts between catalytic residues and the TS could be formed on these scaffolds. Active-site repacking calculations were then performed on the structures with this new theozyme placement. In these calculations, the TS was translated by $\pm 0.4 \text{ \AA}$ along each Cartesian coordinate in $0.2\text{-}\text{\AA}$ steps and rotated about all three axes (origin at the TS geometric center) in 5° increments over a 10° range (clockwise and counterclockwise). This resulted in a total combinatorial search space of 15,625 possible poses. Residues near the catalytic amino acids and

TS were designated as design positions. During repacking, these positions were allowed to sample rotamers of various amino acids, favoring hydrophobic and aromatic residues, while the identities of the catalytic residues remained fixed. At active-site lid positions, the residue introduced by ProteinMPNN was also included in the sequence space. Gly was only allowed if no other residue could be sterically accommodated.

2.6.7 Computational library design

After repacking, the CLEARSS computational library design algorithm²¹³ was used to generate a combinatorial library comprising the most favorable amino acids predicted by Triad at each designed active-site position. In this method, libraries of a specific size configuration are generated from a pre-scored list of sequences using the highest probability set of amino acids at each position based on the sum of their Boltzmann weights, used to convert raw energy values into statistical weights. Using the list of energy-ranked sequences from active-site repacking as input, libraries of 192 sequences were generated. Rotamer configurations for each sequence in the library were optimized using the cleanSequences.py application within Triad to find the lowest-energy conformation of each sequence on its respective backbone, generating “cleaned” structures. To compare energies, the energy difference between each “cleaned” structure and a corresponding all-Gly structure in the absence of the TS was calculated. These energies are reported in Table 2.10. Structures were cleaned with and without the TS to evaluate preorganization, as described previously⁷⁵.

Table 2.10. Design filtering.

Enzyme	Base dihedral (°)	H-bond dihedral (°)	Energy (kcal mol ⁻¹) ^a	SASA (Å ²) ^b	# residues not preorganized ^c	Tunnel present ^d
Variant 1	-42.5	25.7	-39.4	174	5	Yes
Variant 2	-7.3	25.7	-36.3	188	6	Yes
Variant 3	27.6	-156.2	-10.4	107	6	Yes
Variant 4	27.6	-156.2	-2.3	107	5	Yes
Variant 5	27.6	-156.2	-10.4	107	6	Yes
Variant 6	-1	28.8	-115.54	87	2	Yes
Variant 7	-1	28.8	-111.84	87	2	Yes
Variant 8	-0.7	44.2	-108.87	100	3	Yes
Variant 9	-13.3	29.6	-107.41	91	2	Yes

^a Energies are calculated by subtracting the energy of an all-Gly structure as described in the methods under *Active-site repacking*.

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

^c Preorganized residues are those predicted to adopt the same rotamer in the presence and absence of the transition state. Variant 1–2, Variant 3–5, and Variant 6–9 had 14, 17, and 18 residues that were optimized during active-site repacking, respectively.

^d The presence of a tunnel was determined by Caver 3.0 using a minimum cutoff bottleneck radius of 0.9 Å.

2.6.8 Design filtering

Energies of the TS-bound and unbound “cleaned” structures were calculated using Triad. Dihedral angles of catalytic residues relative to the TS, as well as the solvent-accessible surface area (SASA) of the TS, were analyzed using PyMOL (v2.3.0, Schrödinger, LLC). Bottleneck radii of active-site entrances were evaluated using CAVER 3.0²¹⁴, and the number of residues mutated to glycine or methionine was counted. Designs were filtered based on the following criteria (Table 2.10): energy difference between bound and unbound structures below 0 kcal mol⁻¹, catalytic residue dihedral angles outside 50° to 130° and -50° to -130°, SASA of TS between 80 and 200 Å², < 6 Met and < 4 Gly, bottleneck radius of active-site entrance > 0.9 Å, and < 7 non-preorganized residues. Designs passing all criteria were used for structure prediction with AlphaFold2¹⁴¹ or ColabFold²¹², where designs with a backbone RMSD < 3 Å relative to the structure cleaned by Triad, overall pLDDT > 90, and lid pLDDT > 85 were chosen for experimental characterization (Table 2.1).

2.6.9 Sequence optimization of variants 1–5 for enhanced stability

Sequence optimization of variants 1–5 was performed using ProteinMPNN with LigandMPNN weights¹³¹. Amino-acid identities of all residues designed during active-site repacking with Triad were retained. Four sequences were generated using the predicted structure of each variant as input, omitting cysteine. The generated sequences were used for structure prediction with ColabFold²¹² using all five model weights. Predictions were filtered based on their all-atom RMSD to the input model. Theozyme placement was carried out using ColabFold structures, as previously described, to ensure that catalytic contacts were preserved. A single design for each variant was selected for experimental characterization (named KempTIM1–5, Table 2.1).

2.6.10 Ensemble-based design

Two ensembles of backbone templates were generated from the KempTIM4 crystal structure (PDB ID: 9QKX) using the ensemble refinement protocol²⁰³ implemented in the PHENIX macromolecular structure determination software²¹⁵. 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 addition of riding hydrogens, parallel ensemble refinement simulations were performed using various combinations of the parameters pTLS (0.6, 0.8, 0.9, 1.0), τ_x (0.5, 1.5, 2.0) and wx-ray (2.5, 5.0, 10.0), where pTLS 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 wx-ray is the coupled bath 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. Two ensembles were combined to form a 60-member ensemble (Table 2.11).

Table 2.11. Templates generated for ensemble-based design of KempTIM4.

Ensemble	MD sampling parameters			Before ensemble refinement		After ensemble refinement		No. structures
	ρ_{TLS}	τ_x	$w_{\text{x-ray}}$	R_{work}	R_{free}	R_{work}	R_{free}	
1	0.8	1.5	5	0.2547	0.2996	0.2343	0.2782	20
2	0.8	2.0	10	0.2547	0.2996	0.2165	0.2808	40

Theozyme placement was performed on each ensemble member as described previously, except with the use of the bbind02.May.e2 rotamer library. Positions for the base (Asp) and H-bond donor (Asn or Gln) as well as neighboring positions mutated to Ala are indicated in Table 2.6. LigandMPNN was used to predict active-site mutations, allowing only hydrophobic and aromatic residues. Amino acids with a mean probability $>1\%$ across ten batches were selected as the sequence space for active-site repacking in Triad, performed as described above. Triad models (Figure 2.6) and Boltz-2 predictions (Figure 2.23) were used to filter the designs (Table 2.12, Table 2.13). Designs were retained only if Boltz-2 models reproduced the intended catalytic contacts and the predicted ligand binding pose deviated by $< 1.0 \text{ \AA}$ root-mean-square deviation from the corresponding Triad model.

Table 2.12. Triad model filtering during ensemble-based design.

Metric	Residue	Type	Atom 1 ^a	Atom 2 ^a	Atom 3 ^a	Atom 4 ^a	Min ^b	Max ^b
Base	Asp	Distance	OD1 or OD2	H3				1.6
		Angle	CG	OD1 or OD2	H3		92	132
		Angle	OD1 or OD2	H3	C3		139	179
		Torsion	CB	CG	OD1 or OD2	H3	−20, 160	20, 200
		Torsion	CG	OD1 or OD2	H3	C3	−20, 160	20, 200
		Torsion	OD1 or OD2	H3	C3	N2	−30, 150	30, 210
H-bond donor	Asn	Distance	1HD2 or 2HD2	O1				2.3
		Angle	ND2	1HD2 or 2HD2	O1		131	171
		Angle	1HD2 or 2HD2	O1	N2		110	150
		Torsion	CB	CG	1HD2 or 2HD2	O1	−20, 160	20, 200
		Torsion	CG	1HD2 or 2HD2	O1	N2	−20, 160	20, 200
		Torsion	1HD2 or 2HD2	O1	N2	C3	−20, 160	20, 200
	Gln	Distance	1HE2 or 2HE2	O1				2.3
		Angle	NE2	1HE2 or 2HE2	O1		131	171
		Angle	1HE2 or 2HE2	O1	N2		110	150
		Torsion	CG	CD	1HE2 or 2HE2	O1	−20, 160	20, 200
		Torsion	CD	1HE2 or 2HE2	O1	N2	−20, 160	20, 200
		Torsion	1HE2 or 2HE2	O1	N2	C3	−20, 160	20, 200
		Bottleneck radius ^c						1.0
		Radius						

^a Atoms of the catalytic asparagine/glutamine, aspartate, and transition state are coloured in green/pink, red, and blue, respectively.

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

^c Determined with Caver 3.0 using a minimum cutoff bottleneck radius of 0.9 Å, only for predictions containing a tunnel

Table 2.13. Filtering using Boltz-2 models during ensemble-based design.

Metric	Type	Atom 1 ^a	Atom 2 ^a	Min ^b	Max ^b
Base distance	Distance	OD1 or OD2	N3	–	3.3
H-bond donor	Distance	ND2 or NE2	N1	–	3.3
Triad vs Boltz-2 transition-state analogue RMSD ^c	RMSD	–	–	–	1.0
Bound vs unbound active-site RMSD ^d	RMSD	–	–	–	1.0
Bottleneck radius ^e	Radius	–	–	1.2	–
Tunnel length ^e	Distance	–	–	–	15

^a Atoms of the catalytic asparagine/glutamine, aspartate, and transition state analogue 6NBT are coloured in green/pink, red, and blue, respectively.

^b All measurements given in Å.

^c RMSD between the atoms of the benzene rings of the transition-state analogue in the Boltz2 model and transition-state in the Triad model.

^d RMSD between the sidechains of active site residues in the Boltz2 transition-state analogue-bound and -unbound models

^e Determined with Caver 3.0 using a minimum cutoff bottleneck radius of 0.9 Å, only for predictions containing a tunnel

2.6.11 Cloning and generation of constructs.

Genes for all proteins were ordered as codon-optimized fragments flanked by restriction sites for NdeI and XhoI from Twist Bioscience (South San Francisco, CA, USA). After digestion with NdeI and XhoI, the gene fragment was ligated into pET21b(+) for variants 1–5 and KempTIM1–5, or pET29b(+) for variants 6–9 and KempTIM4a–h. Catalytic knockout mutations were introduced by a modified QuikChange PCR protocol utilizing KAPA Polymerase followed by DpnI digestion. *E. coli* Top10 or BL21 (DE3) cells were transformed with ligated vector or reaction mixture and plated on Lysogeny Broth (LB) agar plates containing 100 $\mu\text{g mL}^{-1}$ ampicillin or 50 $\mu\text{g mL}^{-1}$ kanamycin as selection markers. Single colonies were used to inoculate 5 mL LB supplemented with 100 $\mu\text{g mL}^{-1}$ ampicillin or 50 $\mu\text{g mL}^{-1}$ kanamycin. After overnight growth (37 °C, 250 rpm), cells were harvested by centrifugation and DNA was isolated using NucleoSpin Plasmid EasyPure-Kit (Macherey-Nagel) according to the manufacturer's protocol. Vector assembly and introduction of mutations were validated by sequencing (Eurofins Genomics) using standard T7 primers.

2.6.12 Protein expression and purification.

E. coli BL21 (DE3) cells (Novagen) were transformed with plasmid, plated on agar plates containing antibiotic, and incubated overnight at 37 °C. Single colonies were picked to inoculate precultures and incubated at 30–37 °C overnight. On the next day, 1 L LB supplemented with antibiotic was inoculated with 5–10 mL of preculture and incubated at 37 °C until OD600 reached a value between 0.6 and 0.8. Overexpression was induced by adding isopropyl- β -thiogalactoside (IPTG) to a final concentration of 0.1 mM for variants 1–5, KempTIM1–5, and KempTIM4a–h or 1 mM for variants 6–9 followed by overnight incubation at 16–20 °C with shaking. Cells were

harvested by centrifugation and pellets were either frozen at $-20\text{ }^{\circ}\text{C}$ or used directly for purification.

Two different purification schemes were employed. In the first scheme, cell pellets were resuspended in 35 mL buffer A (50 mM sodium phosphate pH 7.0, 100 mM NaCl, 10 mM imidazole). Resuspended cells were lysed by sonication (Branson Ultrasonic Sonifier 250, output 4, duty cycle 40%, 3×3 min) and centrifuged. The supernatant was loaded onto a HisTrapHP column (5 mL, Cytiva Life Science) equilibrated with buffer A and coupled to an ÄKTApure system (Cytiva Life Science). After washing with 10 column volumes of buffer A, the protein was eluted with a linear gradient over 20 CV to 60 % buffer B (50 mM sodium phosphate pH 7.0, 100 mM NaCl, 500 mM imidazole). For crystallisation of KempTIM1 and KempTIM4-gsg, which contains an additional GSG tripeptide linker between its His-tag and first TIM barrel residue (Table 2.1), an additional overnight TEV-cleavage step was performed during a dialysis against buffer C (50 mM of sodium phosphate pH 7.0, 100 mM NaCl). The dialyzed protein was loaded onto a HisTrapHP column (5 mL, Cytiva Life Science) equilibrated with buffer C and coupled to an ÄKTApure system (Cytiva Life Science) and the flowthrough was collected. Fractions containing the protein were pooled, concentrated with a centrifugal concentrator, and loaded onto a HiLoad 26/600 Superdex 75 preparative grade column (Cytiva Life Sciences) preequilibrated in buffer C and eluted with 1 column volume buffer C.

In the second scheme, cell pellets were resuspended in 8 mL buffer D (5 mM imidazole in 100 mM potassium phosphate buffer at pH 8.0) supplemented with 1 mg mL^{-1} lyophilized lysozyme (MP Biomedicals) and 1 U mL^{-1} nuclease (Merck Millipore). Cells were homogenized using an Avestin EmulsiFlex-B15 cell disruptor and centrifuged. Due to their high thermostability, greater purity for KempTIM4a–h could be achieved through heating the lysate at $70\text{ }^{\circ}\text{C}$ for 1 hour

prior to centrifugation. Proteins were purified from the lysate by immobilized metal affinity chromatography using Ni-NTA agarose (Qiagen) pre-equilibrated with buffer D in individual Econo-Pac gravity-flow columns (Bio-Rad). Contaminants were washed away using buffer E (10 mM imidazole in 100 mM potassium phosphate buffer pH 8.0) and then 20 mM imidazole in the same buffer. Proteins were eluted with 5 mL of buffer F (250 mM imidazole in 100 mM potassium phosphate buffer pH 8.0). Proteins were further subjected to gel filtration in buffer C using an ENrich SEC 70 size-exclusion chromatography column (Bio-Rad). Purity was confirmed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and fractions containing the protein of interest were pooled. Protein concentration was determined spectrophotometrically using the absorption at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPASy ProtParam tool (<https://web.expasy.org/protparam/>).

2.6.13 Far-UV CD spectroscopy

CD spectra were collected with a Jasco J-710 or J-815 spectrometer. Spectra were recorded using a 1-mm quartz cuvette at 20 °C with scanning speed 10 or 100 nm min⁻¹, bandwidth 1 nm, response time 1 s. Experiments were performed in 20 mM sodium phosphate buffer (pH 7.0), or 50 mM sodium phosphate buffer (pH 8.0) supplemented with 100 mM sodium chloride, using a protein concentration of approximately 0.2 mg mL⁻¹. For each protein, 3 or 10 spectra were accumulated and averaged. For data normalization, a buffer spectrum was subtracted and the signal was converted to mean residue molar ellipticity using $[\theta_{\text{MRE}}] = \theta / (n \times d \times c)$, where n is the number of residues in the protein, θ the collected ellipticity in mdeg, d the path length in mm, and c the protein concentration in M. Thermal denaturation assays were performed by heating the samples from 20 to 95 °C at a rate of 1 °C per minute and ellipticity at 222 nm was measured every 1 °C. Melting temperature (T_m) was determined following the protocol by Greenfield²¹⁶.

2.6.14 Size-exclusion chromatography-multi angle light scattering

Size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) measurements were performed as described in Beck *et al.*¹⁸⁷ For all measurements, buffer C, an injection volume of 50 μL , and a protein concentration of 2 mg mL^{-1} were used.

2.6.15 Steady-state kinetics

Assays were performed in sodium phosphate (pH 7.0), Tris (pH 7.0–8.5), or CHES (pH 9.0–10.0) buffers, all at 50 mM with 100 mM NaCl and 10% MeOH in a Spark (Tecan) or Synergy H1 (Biotek) plate reader. Enzyme concentrations varied from 0.2 to 80 μM . Reactions (200 μL final volume) were initiated by addition of varying concentrations (0.015–2 mM) of 5-nitrobenzisoazole (abcr or AAblocks) or 6-nitrobenzisoazole (AAblocks) dissolved in methanol (final methanol concentration 10%). Product formation was monitored at 380 nm ($\epsilon = 15,800 \text{ M}^{-1} \text{ cm}^{-1}$)^{21,195} or 400 nm ($\epsilon = 2,870 \text{ M}^{-1} \text{ cm}^{-1}$)^{183,195} for 5-nitrobenzisoazole and 6-nitrobenzisoazole, respectively, in individual wells of a 96-well plate (Nunc or Greiner Bio-One). Path lengths for each well were calculated ratiometrically using the difference in absorbance at 900 and 975 nm. Control measurements without protein were conducted for each substrate concentration and subtracted from the enzyme measurements. Linear phases of the kinetic traces were used to measure initial reaction rates. If the enzyme showed saturation within the substrate concentration used, the data were fitted to the Michaelis–Menten model using ($v_0 = (v_{\text{max}} \times [\text{S}]) / (K_{\text{M}} + [\text{S}])$) where v_0 is the initial velocity rate, v_{max} is the maximal velocity rate, $[\text{S}]$ is the free substrate concentration, K_{M} is the Michaelis constant. The catalytic constant k_{cat} was calculated with $k_{\text{cat}} = v_{\text{max}} / [\text{E}]$ where $[\text{E}]$ is the enzyme concentration. If saturation was not achieved within the substrate's solubility limit, a linear equation ($v_0 = k_{\text{cat}}/K_{\text{M}} \times [\text{S}]$) was used. For the pH-rate profile, $k_{\text{cat}}/K_{\text{M}}$ or k_{cat} values as a function of pH were fit to a sigmoid curve defined by $k_{\text{cat}}/K_{\text{M}} =$

$A / (1 + e^{(-k \times (pH - pKa))}) + C$ for KempTIM1 or $k_{cat} = A / (1 + e^{(-k \times (pH - pKa))}) + C$ for KempTIM4b

where A is the upper asymptote, C is the lower asymptote, k is the steepness, and pKa is the point of inflection. The point of inflection represents the approximate pKa of the catalytic base Asp.

2.6.16 Crystallization and structure determination.

KempTIM4-gsg and KempTIM1 were purified using the first purification scheme, with the final size-exclusion chromatography step performed in 20 mM HEPES buffer pH 8.0 supplemented with 20 mM NaCl. Crystallization screens were set up using the sitting-drop vapor diffusion method with a Phoenix pipetting robot (Art Robbins Instruments) and commercially available sparse-matrix screens (NeXtal) in 96-well sitting-drop plates (3-drop Intelli-Plates, Art Robbins Instruments). Protein, reservoir solutions were mixed in ratios of 1:1, 2:1, and 1:2 and crystals were grown at 293 K. The protein concentration of KempTIM4-gsg was 12 mg mL⁻¹, supplemented with 5 mM 6NBT (5% DMSO), and diffracting crystals were obtained after 40 days in 4 M sodium formate. Cryoprotection was achieved by adding glycerol to a final concentration of 25%.

Crystals for KempTIM1 (8.75 mg mL⁻¹ protein) without TS analogue were obtained after 48 h in droplets containing 0.2 M ammonium acetate, 0.1 M Bis-Tris pH 5.5, and 25% w/v PEG 3350 as reservoir solution. For the TS analogue-bound structure of KempTIM1 (8.75 mg mL⁻¹ protein), 6NBT was added as a solid in excess and the solution was incubated for 1 h and subsequently centrifuged. These crystals were observed after 24 h in droplets with 0.1 M sodium acetate pH 4.5, 0.1 M Bis-Tris pH 5.5, 25% w/v PEG 3350 as reservoir. The KempTIM1 crystals were cryo-protected by adding 20% ethylene glycol to the corresponding reservoir solution. All crystals were mounted using cryo-loops on SPINE standard bases and flash-cooled in liquid nitrogen.

Diffraction data were collected at 100 K at beamline BL 14.1 with a Pilatus3 S 6M detector at the BESSY II synchrotron (Helmholtz-Zentrum Berlin). Data processing was performed using X-ray Detector Software APP3 (XDSAPP3)²¹⁷ with XDS²¹⁸, and data quality was assessed using phenix.xtriage²¹⁵. Phases were solved by molecular replacement using AlphaFold2 predictions as search models with Phaser²¹⁹. The resulting models were manually rebuilt using Coot²²⁰ and refined with phenix.refine²¹⁵ in an iterative process.

The structure of KempTIM4-gsg was refined to a resolution of 2.3 Å. Despite crystallization with 6NBT, no electron density corresponding to the transition-state analogue was observed in the active site; instead, the density was best interpreted as a bound glycerol molecule from the cryoprotectant. The structures of KempTIM1 apo and co-crystallized with 6NBT were refined to resolutions of 1.25 Å and 1.20 Å, respectively. The high resolution of the co-crystal structure and the well-defined electron density of 6NBT in the active site enabled conclusive placement of the TS analogue. Additionally, there is also a diffuse electron density, which cannot be interpreted unambiguously, indicating weak binding of solvent molecules. However, a putative ethylene glycol molecule was placed to account for the residual density, as it was used as cryoprotectant. Refinement statistics and crystallographic data are shown in Table 2.5. Coordinates and structure factors were validated and deposited in the PDB under accession codes 9TZD, 9U01 and 9QKX for KempTIM1 with 6NBT, KempTIM1 without 6NBT, and KempTIM4, respectively.

2.6.17 Size-exclusion chromatography small-angle X-ray scattering (SEC-SAXS)

SEC-SAXS measurements were conducted at the BioSAXS beamline BM29 of the ESRF in Grenoble, France. A Superdex 75 Increase 10/300 GL column (Cytiva Life Sciences) was used with a flow rate of 0.8 mL min⁻¹ in buffer C. KempTIM4-gsg was measured at a concentration of 5 mg mL⁻¹, with an injection volume of 100 µL. Data processing and analysis were performed

using ATSAS 3.2.1^{221,222}. For analysis, AlphaFold2 models of KempTIM4 and NT6-CP2, as well as an AlphaFlow²⁰¹ model of KempTIM4 with a displaced lid obtained using the base molecular dynamics weights were used. For each of these models a theoretical scattering curve was generated and fitted to the experimental data using CRY SOL²²³.

2.6.18 Molecular dynamics simulations

The Amber 2020 software²²⁴ (<http://ambermd.org/>) with the ff19SB protein force field²²⁵, gaff2 ligand force field²²⁶, and OPC water model²²⁷ was used for all simulations. A cutoff of 10 Å was applied to electrostatics modeled using the particle mesh Ewald method²²⁸. All trajectories were run with a time step of 2 fs. Models of each design generated using Triad were prepared for simulations, which were run for each enzyme in the presence and absence of bound TS analogue 6NBT. Parameters for 6NBT were generated using the Antechamber package²²⁹. Hydrogen atoms were added using Reduce²³⁰, and the prepared structures were solvated with OPC water in a truncated octahedral box with periodic boundary conditions where the distance between the protein surface and the box edges was set to 10 Å. The additions2 algorithm in Amber was used to place counterions to neutralize the system. The system was minimized (solvent, protein, then whole system) using the steepest descent method with positional restraints of 500 kcal mol⁻¹ Å⁻². The system was then heated to a temperature of 300 K over 240 ps while restraints were gradually removed, followed by equilibration under an NPT ensemble at 300 K and 1 bar for 10 ns and subsequent equilibration under an NVT ensemble at 300 K for another 10 ns. Constant temperature and pressure were achieved using the Langevin thermostat²³¹ and Berendsen barostat²³², respectively. Following equilibration, ten replicas of 100-ns production simulations were run. For analysis, 1000 snapshots separated by 100 ps each were extracted from the production trajectory.

Catalytic interactions and RMSF values were extracted using CPPTRAJ and PYTRAJ, respectively²³³.

Chapter 3. Conclusions and outlook

3.1 Summary

Despite the immense potential held by the creation of custom biocatalyst for any chemical transformation, existing *de novo* enzymes suffer from low turnover rates. Chapter 1 reviewed traditional and recent design strategies, highlighting their strengths and limitations. A major constraint on the activity achieved by these approaches is the reliance on natural proteins as templates for design; theozyme placement on these scaffolds strains the precise geometry required for catalysis, forcing sub-optimal catalytic contacts. To address this, recent approaches have either diversified and recombined natural protein templates^{69,72,93} or generated *de novo* scaffolds to fit catalytic groups^{66–68,73,98}. The *de novo* structure generation approach is motivated by the observation that pre-existing scaffolds, no matter how diversified, remain incompatible with the theozyme because the backbone is not tailored to exact transition state (TS) and residue geometry⁶⁷. Conversely, in their current form, *de novo* backbones lack the structural complexity that supports efficient catalysis, which can be accessed through the diversification and recombination of natural backbones^{69,158}. Both perspectives address valid shortcomings in the opposing methodology. Thus, a hybrid approach that captures the precision of *de novo* design with the versatile architecture of natural enzymes could be beneficial.

In Chapter 2, we described CANVAS (Customizing Amino-acid Networks for Virtual Active-site Scaffolding), an approach to design *de novo* TIM-barrel enzymes by integrating reaction-specific lids customized to the geometry of the theozyme. This approach enables the creation of enzymes with backbones tailored for catalytic requirements while preserving the structural complexity necessary for the diverse catalytic function of the TIM barrel fold. CANVAS-

designed enzymes exhibit structural similarity to natural TIM-barrel enzymes, but the lid architectures deviate from the catalytic-face extended loops and other structural segments found in nature. Still, the crystal structures of CANVAS-designed KempTIM1 reveal a lid conformational change induced by TS analogue binding, reminiscent of rearrangements observed in natural TIM-barrel enzymes¹⁶⁴. Furthermore, the elevated pKa of the catalytic base in designed KempTIM1 and KempTIM4b indicates the successful creation of a hydrophobic active-site environment, demonstrating that these engineered lids can form custom active sites tailored to specific reaction requirements.

KempTIM1 achieves high efficiency ($21,000 \text{ M}^{-1} \text{ s}^{-1}$) and turnover (14 s^{-1}) for a *de novo* enzyme, comparable to or exceeding many Kemp eliminases optimized through directed evolution (Figure 3.1)^{32,33,35}. In contrast with KempTIM1, other recent *de novo* enzymes achieve high efficiency primarily through productive substrate binding rather than turnover rate^{67,68,93,190}. Notably, the turnover rate of KempTIM1 is on the same order of magnitude as that of a recent *de novo* Kemp eliminase obtained through two rounds of computational design and one serendipitous point mutant⁶⁹. CANVAS-designed enzymes can be optimized using computational approaches, as ensemble-based design produced KempTIM4b ($32,000 \text{ M}^{-1} \text{ s}^{-1}$) from a lower-efficiency precursor ($19 \text{ M}^{-1} \text{ s}^{-1}$). This represents a greater fold-improvement than achieved by the directed evolution of the similar-efficiency KE07 ($12 \text{ M}^{-1} \text{ s}^{-1}$) to KE07-10/11G ($2,600 \text{ M}^{-1} \text{ s}^{-1}$). Despite this success, neither KempTIM1 nor KempTIM4b match the efficiency of the most extensively evolved Kemp eliminases^{34,70} or the median natural enzyme (Figure 3.1)⁸. Pushing the boundaries of *de novo* enzyme design will require the explicit consideration of complex enzyme features that are ignored by existing design protocols.

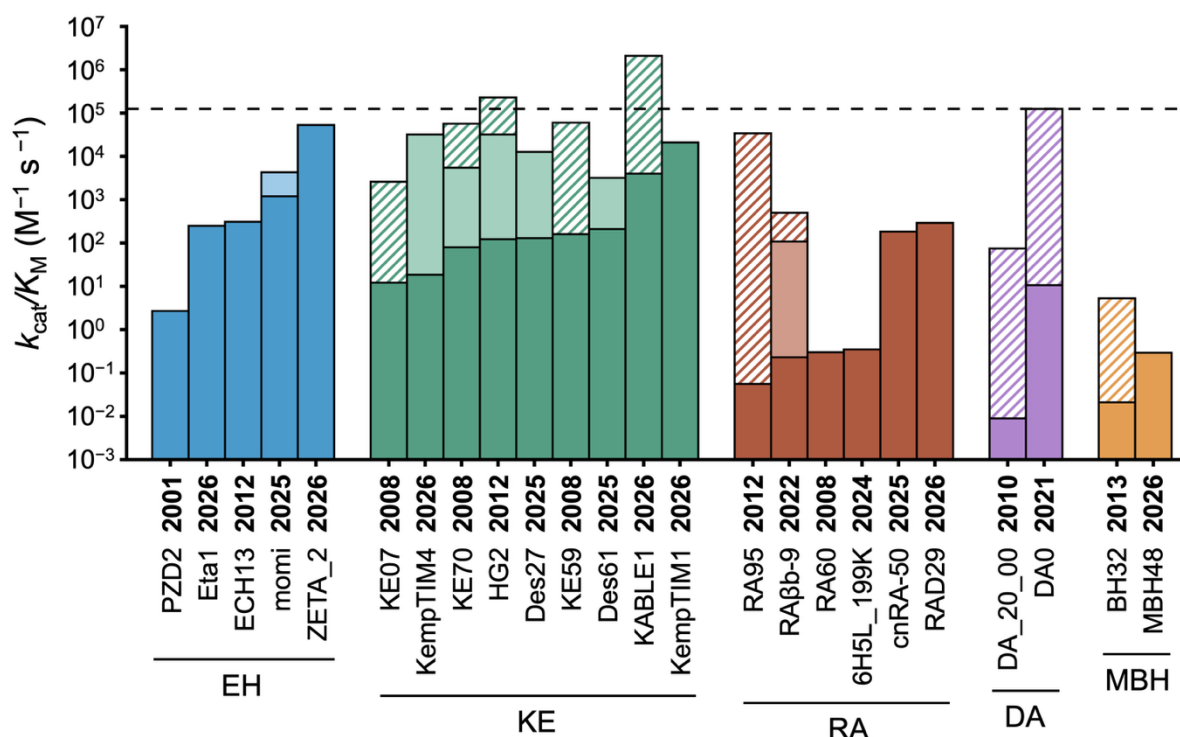


Figure 3.1. Efficiencies of *de novo* enzymes. Enzymes are grouped by reaction type and ordered by catalytic efficiency, with its name and publication year listed below each column. Efficiencies of initial *de novo* designs^{5,13,20–25,66–74} are shown as solid-coloured bars, those for computationally redesigned variants^{67,69,71,75} are shown as lighter-coloured bars, and those for experimentally optimized variants^{24,26,32–35,70,71,74} are shown as hashed bars. The dashed line indicates the median k_{cat}/K_M of natural enzymes ($125,000 M^{-1} s^{-1}$)⁸. Reaction types EH, KE, RA, DA, and MBH, corresponding to ester hydrolysis, Kemp elimination, retro-aldol condensation, Diels–Alder cycloaddition, and the Morita–Baylis–Hillman reaction, are shown in blue, green, brown, purple, and orange, respectively. For bimolecular reactions, k_{cat}/K_M was calculated using the K_M for 4-carboxybenzyl trans-1,3-butadiene-1-carbamate, azachalcone, or 4-nitrobenzaldehyde for DA_20_00, DA0, or BH32 and MBH48, respectively. Catalytic efficiency for BH32 was taken from Crawshaw *et al.*²⁶.

3.1.1 Methodological insights from KempTIM characterization

Characterization of the KempTIMs highlights specific features and metrics that could be integrated into the CANVAS design protocol. Notably, the concerted design of the lid and minimal TIM barrel sequence was critical for ensuring high expression and stability. Design of the entire scaffold sequence, rather than the lid alone, following lid generation could be readily incorporated into the CANVAS workflow. Given its success for improving the activity of KempTIM4, integrating ensemble-based design directly into CANVAS could be beneficial. This could be done

by employing predicted ensembles^{201,234} rather than ensembles derived experimentally or from computationally expensive molecular dynamics simulations. Although the accuracy of methods for predicting multiple conformational states of proteins remains sub-optimal²³⁵, the increased diversity may assist in identifying templates that better accommodate the theozyme. When combined with the catalytic competency metric to enrich high-activity designs, these refinements could bolster experimental success.

Analysis of flexibility in these designs reveals the structural determinants of effective lids. In KempTIM1, the most active first-round design, the major lid fragment housing the H-bond donor is notably rigid. While such lid rigidity could hinder substrate entry or product release, the efficiency of this enzyme compared to KempTIM3–5 suggests otherwise. Also noted was the flexibility of the extended $\beta\alpha$ -loops near the active site, originally intended to improve interactions with the major lid fragment, suggesting a potential division of labor between these diffused fragments. Future work could involve kinetic solvent viscosity effects assays to investigate the role of these loops in substrate binding and product release²³⁶, potentially informing a novel filtering metric that selects rigid major lid fragments but flexible extended loops, based on H-bonding density or other criteria. This could enrich for designs like KempTIM1, where catalytic residues are precisely anchored, without compromising conformational dynamics. The observation that low-confidence regions of predicted KempTIM structures overlap with flexible regions agrees with recent assessments of AlphaFold2 confidence²³⁷ and may inform this filtering. However, employing prediction confidence as a filter requires deconvoluting genuine flexibility from structural uncertainty, a challenge exacerbated by the tendency of generative models to optimize for more designable secondary structures at the expense of loop quality¹⁵⁶. To address this,

structure prediction confidence could provide an initial screen, followed by short molecular dynamics simulations to assess flexibility.

3.2 Opportunities for CANVAS designs

3.2.1 Combining CANVAS with complementary design approaches

CANVAS presents an extensible framework that treats minimal TIM barrels as blank canvases to be customized for target reaction geometry. The TIM barrel fold is highly evolvable due to the spatial separation of stability and function¹⁷⁴; its core is highly ordered and rigid, while the catalytic face is highly customizable and mutable. In contrast, backbones generated from scratch around catalytic motifs lack the structural complexity required for this separation¹⁵⁸. CANVAS leverages the highly stable cores of *de novo* TIM barrels to retain stability while allowing for more complex reaction-specific architecture. This structural partitioning renders CANVAS-designed enzymes highly amenable to downstream optimization. As observed for KempTIM4, these enzymes can be optimized using computational methods. However, experimental optimization has still yielded greater success in improving the efficiency of *de novo* enzymes⁷⁰. Thus, approaches such as NMR-guided directed evolution⁹⁵ could enhance the catalytic power of CANVAS. The favorable initial activity afforded by CANVAS through the customization of a minimal TIM barrel, a highly evolvable fold, could be leveraged by laboratory evolution to create even more efficient *de novo* enzymes.

An advantage of CANVAS is its conceptual modularity, making it compatible with alternative computational design and optimization approaches. While the simplicity of the initial workflow assisted in evaluating the core concept, CANVAS can be readily combined with other advanced design strategies. In particular, future iterations could harness structure diversification⁶⁸

and iterative backbone refinement⁷³ protocols developed for enzyme design from generated scaffolds. Following lid generation, lids or loop extensions could be diversified by hallucination^{91,206}, a method shown to improve structural diversity and *in silico* sequence design success rates⁶⁸. These structures could also be coupled with the fragment recombination approach to generate even more lid structural diversity^{69,238}. In its current iteration, the CANVAS workflow eliminates a considerable portion of scaffolds due to poor designability and thus poor structure prediction success. The backbone refinement protocol by Braun *et al.*⁷³ could resolve this, improving the designability of generated scaffolds through iterative structure prediction and sequence design.

Various computational tools have been developed that could improve CANVAS. Current tools such as RFdiffusion¹¹¹, ProteinMPNN¹³⁸, and AlphaFold2¹⁴¹ for lid design and validation could be substituted for tools with expanded functionality. An alternative tool for structure generation, RFdiffusion3⁹⁸, explicitly models ligands and can scaffold residue motifs at the functional group level. Additionally, the structure generation tool Protpardelle-1c¹¹⁷ has been shown to generate a more diverse set of designable protein backbones compared to RFdiffusion¹⁵⁶. Protpardelle-1c can create TIM barrels with elevated designability compared to other approaches, a characteristic highly desirable for CANVAS. With this, a larger library of *de novo* TIM barrels could be used as input to CANVAS, increasing the diversity of starting points to the workflow. Additionally, sequence generative models such as Caliby^{124,127} have demonstrated superior success in designing sequences that accurately capture the target backbone when their structures are predicted. To validate designed sequences in the context of a ligand, co-folding models such as Boltz-2²⁰⁴ and AlphaFold3²³⁹, as well as models that predict an ensemble of protein-ligand interactions such as PLACER⁷², have been developed. These models could help validate the

designed TS binding pose and enrich for preorganized designs through predictions made in the presence and absence of ligand. The use of advanced computational tools and the combination of CANVAS with complementary design strategies suggests a second iteration of the workflow (Figure 3.2).

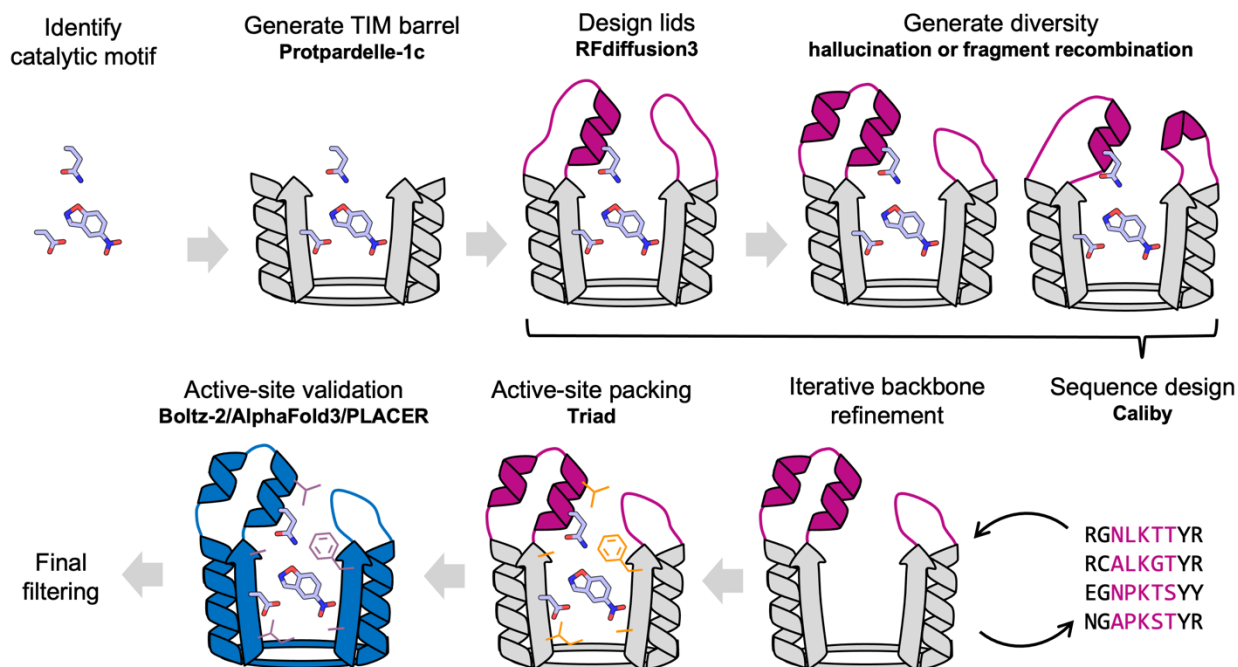


Figure 3.2. Proposed second iteration of the CANVAS workflow. Highly designable TIM barrels generated by Protpardelle-1c could be used as input for design. Subsequently, RFdiffusion3 could be used to construct lid fragments with high TS complementarity. These scaffolds can be diversified through hallucination of the loops or fragment recombination, and sequences of this diversified scaffold library can be produced with Caliby, designing the entire backbone. The iterative backbone refinement protocol could be used to increase structure prediction confidence. As before, active site optimization would be done with physics-based programs such as Triad. Protein-ligand co-folding models could validate the designed active sites, followed by any final filtering, including the identified catalytic competency metric.

3.2.2 CANVAS with non-TIM barrel scaffolds

Many other protein folds have been created *de novo* that lack complex structural elements for catalysis¹¹⁹. Although the power behind CANVAS is evident in the TIM barrel fold, these other *de novo* proteins could also be used as input structures (Figure 3.3). Demonstrating this concept, truncating and re-building fragments of *de novo* helical bundles to modify active sites has demonstrated success in the design of *de novo* cyclopropane synthases⁸⁶. Even natural protein scaffolds could be adapted to specific reactions with CANVAS. Often, *de novo* enzyme design campaigns fail due to the inability to place one or more residues of the theozyme into natural scaffolds, which have structural elements specific to their evolved function⁶⁷. CANVAS could be applied to amend this; for example, if a subset of the theozyme could be successfully placed into a pocket, the structure could be truncated and rebuilt with CANVAS to scaffold any remaining catalytic residues. Therefore, the concept behind CANVAS holds potential for many applications using non-TIM-barrel backbones.

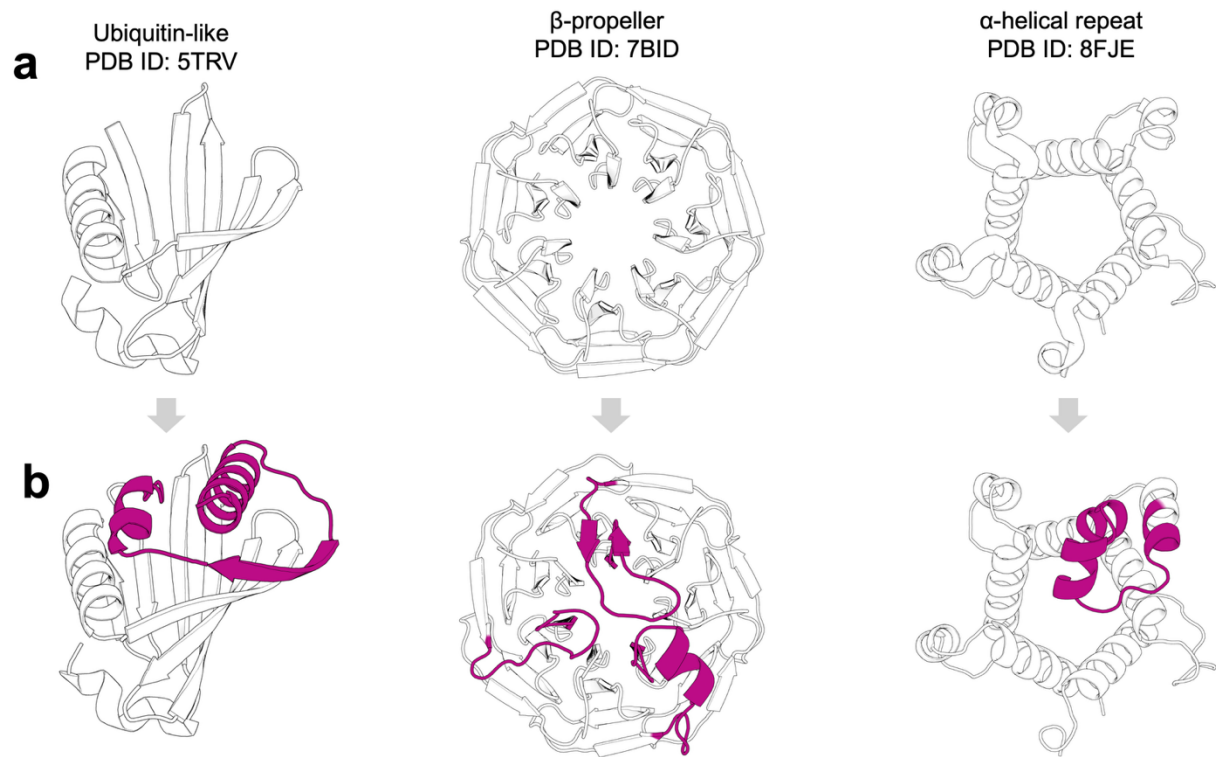


Figure 3.3. Generation of extensions on non-TIM barrel minimal *de novo* scaffolds. Minimal proteins (a) adopting ubiquitin-like²⁴⁰, β -propeller²⁴¹, and α -helical repeat²⁴² folds, as well as many possible others, (b) could be customized with CANVAS.

3.2.3 CANVAS in the design of multiple functionally relevant states

Internal motions and large conformational changes that facilitate substrate binding, passage along the reaction coordinate, and product release are essential for the high efficiency of natural enzymes^{62,243}. The exchange between specific states and the timescale of this exchange are often governed by amino acids at positions distal to the active site, which are not considered in *de novo* enzyme design; directed evolution has been required to sample these activity enhancing distal mutations^{244,236}. Developing computational strategies to design this remains a frontier for creating *de novo* enzymes capable of adopting multiple conformational states. While methods exist to shift the conformational equilibrium of natural protein backbones^{245,246}, they require prior knowledge of the functionally relevant states. For a novel reaction, these conformational states are likely unknown or may deviate greatly from those of natural protein scaffolds. The complete remodeling of the complex backbone dynamics intrinsic to natural protein scaffolds may be challenging. Consequently, conformational dynamics remain largely unexplored in *de novo* enzyme design. Enzyme structures generated entirely from scratch around a catalytic motif may be too rigid and highly ordered to easily accommodate large conformational changes. In contrast, the flexible loops of TIM barrels naturally facilitate substrate binding and the transition to a catalytically active, closed state^{161,175,180}.

Although KempTIM1 exhibited a rearrangement in the lid and active-site residues induced by TS analogue binding, these bound and unbound states were not explicitly engineered, and the role of distal mutations in this rearrangement is unknown. Consequently, this could provide a model system for understanding the role of distal mutations in the conformational equilibrium of a *de novo* enzyme that has not been experimentally optimized. KempTIM1 therefore offers a clean platform, without the complex dynamics of natural enzymes, for engineering novel distal

mutations that shift the conformational equilibrium of a *de novo* enzyme. Successfully doing so would establish CANVAS as a platform to design enzymes with multiple functionally relevant states from scratch. Recently, sequence generative models like ProteinMPNN¹³⁸ have been used to stabilize multiple conformational states of *de novo* scaffolds, finding sequences that exchange on functionally relevant timescales²⁴⁷. Integrating this approach with CANVAS lid design could identify sequences that adopt multiple functionally relevant lid conformations (Figure 3.4). For novel reactions, however, a non-trivial challenge will be deriving target conformational landscapes from scratch. Future work should explore lid generation approaches in CANVAS that capture functionally relevant open and closed states, alongside any other backbone conformations that promote passage along the reaction coordinate. This will challenge our ability to precisely model loop conformations and control dynamics that are tightly coupled to function¹⁷⁷.

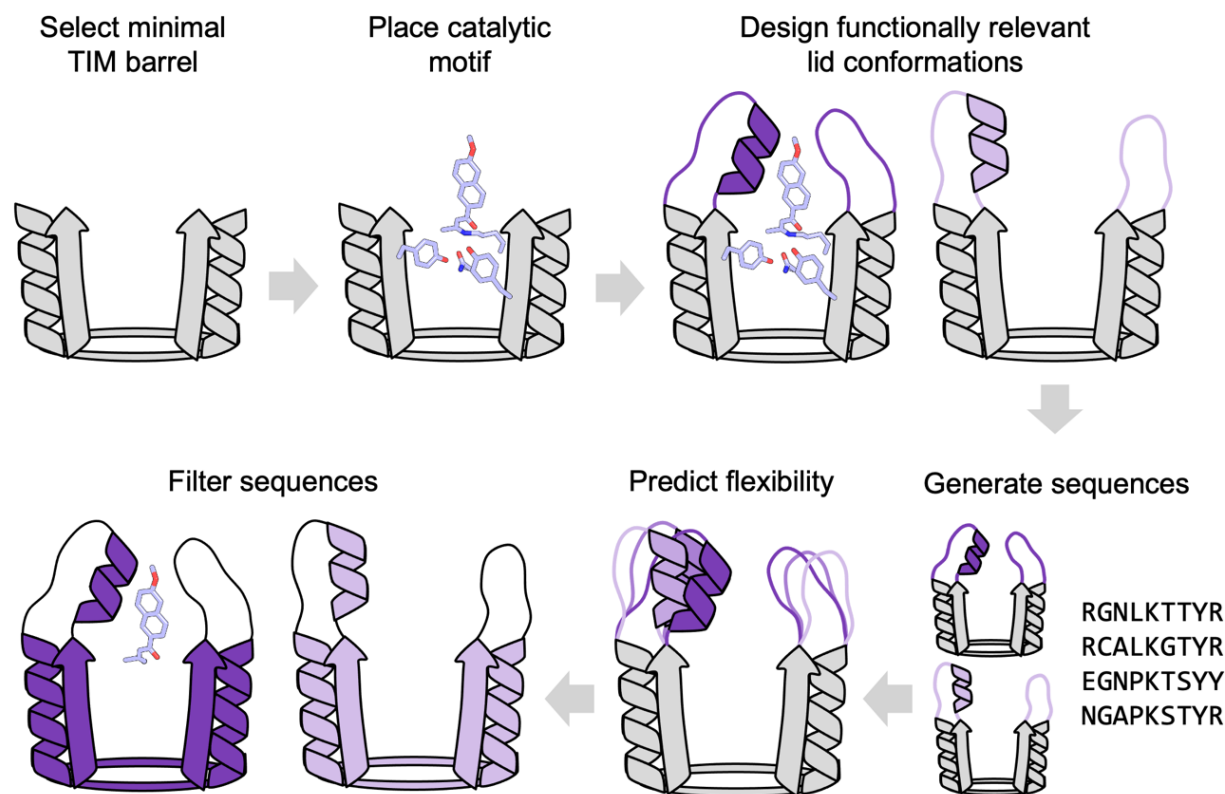


Figure 3.4. Using CANVAS to design enzymes with multiple functionally relevant states. CANVAS could be used to generate multiple lid conformations, representing open and closed enzyme states. Sequences can be produced that are predicted to stabilize both conformations. Machine-learning methods to predict conformational ensembles or molecular dynamics simulations could be used to evaluate the flexibility and identify sequences that adopt the target states. As further validation, protein-ligand co-folding models can be used to identify sequences predicted to be closed in the presence and open in the absence of substrate.

3.3 Outstanding challenges for *de novo* enzyme design

Despite the success of *de novo* enzyme design for developing biocatalysts for model chemical transformations, the efficiencies of designed enzymes remain lower than those optimized through directed evolution and the median natural enzymes (Figure 3.1). Furthermore, *de novo* enzymes with higher efficiencies have been limited to relatively simple and well-understood chemical transformations such as the Kemp elimination²⁰⁵ and ester hydrolysis^{67,68}, or bind cofactors where the intrinsic reactivity comes from the cofactor itself rather than protein side chains^{68,86,114,190}. Computational optimization has further improved efficiencies^{69,75}, but has still been limited to these simple reactions. Thus, despite recent improvements in *de novo* enzyme design, many challenges remain unresolved.

Limited work is done to address mechanistic approximations. Future work should explore designing biocatalysts by considering the entire catalytic cycle, including substrate binding, product release, and progression along the reaction coordinate across all intermediates and TSs^{54,248}. Furthermore, the role of electrostatics is not considered in *de novo* enzyme design, despite its critical role in natural enzyme catalysis⁵⁰. Methods to probe this feature are computationally expensive and require molecular dynamics-derived structural ensembles for analysis²⁴⁴. With increases in computational power and machine learning-guided ensemble prediction accuracy²³⁵, this could serve as a future filter. While many design and filtering methods continue to advance, little progress has been made on theozyme generation methods. As the first and most critical step of *de novo* enzyme design, inaccuracies in theozymes will quickly become activity-limiting for novel reactions. Constructing a theozyme requires rational selection of residues to install around the TS prior to quantum mechanical (QM) geometry optimization¹². Furthermore, despite the sub-angstrom accuracy of QM methods, artifacts can be included in

theozymes due to the gas-phase approximation²⁴⁹. Reflecting this, some calculated theozymes have been proven sub-optimal³⁶, and proof-of-concept *de novo* enzyme design campaigns have been most successful when a laboratory-optimized or natural enzyme catalytic motif is known for the reaction of interest^{67,69,73}. This is true for the present CANVAS campaign as well, as filtering metrics were derived from previous studies of optimized Kemp eliminases^{34,75,151}. *De novo* design of highly efficient enzymes without any knowledge from previous design and optimization campaigns has yet to be accomplished.

The prevalence and success of protein structure prediction models have enabled the current *de novo* enzyme design paradigm, involving the creation of novel scaffolds around a catalytic motif. However, structure prediction methods still do not have perfect accuracy. In the present CANVAS design campaign, the low expression and stability of variants 1–5, possibly due to misfolding, was not reflected by AlphaFold2 confidence. Furthermore, predicted backbone accuracy was sub-optimal for many designs, as demonstrated by KempTIM4. Thus, there remain challenges in using predicted structures as templates for design rather than those derived experimentally. Structure prediction methods continue to improve, but atomic-level accuracy must be obtained for broad applicability to enzyme design. Furthermore, prediction accuracy is limited for proteins with low sequence homology¹⁴² or proteins that adopt multiple conformational states²³⁵.

Engineering activity-enhancing distal mutations also remains elusive to *de novo* enzyme design. For efficient catalysis in natural enzymes, allosteric residues are known to modify the local electric field²⁴⁴, fine-tune backbone conformational dynamics^{236,244}, and promote preorganization either through direct residue networks²⁵⁰ or by prepaying the entropic penalty of active-site positioning²⁵¹. In experimentally optimized *de novo* enzymes, distal residues have been found to

play a role in electrostatic catalysis and modulating protein conformational energy landscapes toward catalytically competent conformations^{236,244}. Methods have been employed to design sequences that stabilize multiple states^{128,246,247}. However, functionally relevant states must be known beforehand; furthermore, this represents the design of conformational equilibrium rather than true dynamics, although exchange on functionally relevant timescales may arise coincidentally²⁴⁷. The design of true enzyme dynamics, accounting for the conformational exchange of pre-specified functionally relevant states on a precise, targeted timescale for catalysis will require considerable effort.

CANVAS provides an important step for *de novo* enzyme design, producing enzymes that retain a familiar and versatile protein fold while customizing the backbone for a target reaction. However, the purely computational design of enzymes in a single, concerted step that achieve natural-level efficiency for entirely novel reactions under pre-specified conditions will be a challenging test of our understanding of enzymatic catalysis. This will require a superior understanding of the principles behind enzyme function and improvements in the accuracy and generalizability of computational methods.

References

1. Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
2. Bornscheuer, U. T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol.* **5**, 137–143 (2001).
3. Lutz, S. Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21**, 734–743 (2010).
4. Vidal, L. S., Isalan, M., Heap, J. T. & Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. *RSC Chem. Biol.* **4**, 271–291 (2023).
5. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* **98**, 14274–14279 (2001).
6. Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **3**, 107–115 (1935).
7. Laidler, K. J. & King, M. C. Development of transition-state theory. *J. Phys. Chem.* **87**, 2657–2664 (1983).
8. Bar-Even, A. *et al.* The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).
9. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Computational Enzyme Design: Successes, Challenges, and Future Directions. in (2018). doi:10.1039/9781788010450-00088.
10. Lienhard, G. E. Enzymatic Catalysis and Transition-State Theory. *Science* **180**, 149–154 (1973).

11. Wolfenden, R. Analog approaches to the structure of the transition state in enzyme reactions. *Acc. Chem. Res.* **5**, 10–18 (1972).
12. Tantillo, D. J., Jiangang, C. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).
13. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
14. Crystallography: Protein Data Bank. *Nature. New Biol.* **233**, 223–223 (1971).
15. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci.* **103**, 16710–16715 (2006).
16. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
17. Dunbrack Jr., R. L. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681 (1997).
18. Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
19. Shetty, R. P., de Bakker, P. I. W., DePristo, M. A. & Blundell, T. L. Advantages of fine-grained side chain conformer libraries. *Protein Eng.* **16**, 963–969 (2003).
20. Richter, F. *et al.* Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).

21. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).
22. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
23. Jiang, L. *et al.* De Novo Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).
24. Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
25. 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).
26. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
27. 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. *J. Am. Chem. Soc.* **118**, 6462–6471 (1996).
28. 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).
29. Simón, L. & Goodman, J. M. Enzyme Catalysis by Hydrogen Bonds: The Balance between Transition State Binding and Substrate Binding in Oxyanion Holes. *J. Org. Chem.* **75**, 1831–1840 (2010).

30. Lowe, G. & Williams, A. Papain-catalysed hydrolysis of some hippuric esters. A new mechanism for papain-catalysed hydrolysis. *Biochem. J.* **96**, 199–204 (1965).
31. Kipnis, Y. & Baker, D. Comparison of designed and randomly generated catalysts for simple chemical reactions. *Protein Sci.* **21**, 1388–1395 (2012).
32. Khersonsky, O. *et al.* Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J. Mol. Biol.* **396**, 1025–1042 (2010).
33. Khersonsky, O. *et al.* Optimization of the *In-Silico*-Designed Kemp Eliminate KE70 by Computational Design and Directed Evolution. *J. Mol. Biol.* **407**, 391–412 (2011).
34. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminate. *Nature* **503**, 418–421 (2013).
35. Khersonsky, O. *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminate KE59. *Proc. Natl. Acad. Sci.* **109**, 10358–10363 (2012).
36. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
37. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
38. Preiswerk, N. *et al.* Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci.* **111**, 8013–8018 (2014).
39. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Bioinforma.* **46**, 105–109 (2002).

40. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).
41. Sigala, P. A. *et al.* Testing Geometrical Discrimination within an Enzyme Active Site: Constrained Hydrogen Bonding in the Ketosteroid Isomerase Oxyanion Hole. *J. Am. Chem. Soc.* **130**, 13696–13708 (2008).
42. Sekhar, A. & Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl. Acad. Sci.* **110**, 12867–12874 (2013).
43. Romero-Rivera, A., Garcia-Borràs, M. & Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **7**, 8524–8532 (2017).
44. Stiller, J. B. *et al.* Structure determination of high-energy states in a dynamic protein ensemble. *Nature* **603**, 528–535 (2022).
45. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science* **220**, 671–680 (1983).
46. Voigt, C. A., Gordon, D. B. & Mayo, S. L. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design 1 Edited by J. Thornton. *J. Mol. Biol.* **299**, 789–803 (2000).
47. Allen, B. D. & Mayo, S. L. Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **27**, 1071–1075 (2006).

48. Alvizo, O. & Mayo, S. L. Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proc. Natl. Acad. Sci.* **105**, 12242–12247 (2008).
49. Leaver-Fay, A. *et al.* Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. in *Methods in Enzymology* vol. 523 109–143 (Elsevier, 2013).
50. Warshel, A. *et al.* Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* **106**, 3210–3235 (2006).
51. Natarajan, S. K. & Sierks, M. R. Identification of Enzyme–Substrate and Enzyme–Product Complexes in the Catalytic Mechanism of Glucoamylase from *Aspergillus awamori*. *Biochemistry* **35**, 15269–15279 (1996).
52. Yao, L., Li, Y., Wu, Y., Liu, A. & Yan, H. Product Release Is Rate-Limiting in the Activation of the Prodrug 5-Fluorocytosine by Yeast Cytosine Deaminase. *Biochemistry* **44**, 5940–5947 (2005).
53. Machado, T. F. G., Gloster, T. M. & Da Silva, R. G. Linear Eyring Plots Conceal a Change in the Rate-Limiting Step in an Enzyme Reaction. *Biochemistry* **57**, 6757–6761 (2018).
54. Kerns, S. J. *et al.* The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124–131 (2015).
55. Cleland, W. W. What limits the rate of an enzyme-catalyzed reaction. *Acc. Chem. Res.* **8**, 145–151 (1975).
56. Menger, F. M. Analysis of ground-state and transition-state effects in enzyme catalysis. *Biochemistry* **31**, 5368–5373 (1992).

57. Ruben, E. A. *et al.* Ground State Destabilization from a Positioned General Base in the Ketosteroid Isomerase Active Site. *Biochemistry* **52**, 1074–1081 (2013).
58. Phillips, R. S. *et al.* Ground-State Destabilization by Phe-448 and Phe-449 Contributes to Tyrosine Phenol-Lyase Catalysis. *ACS Catal.* **6**, 6770–6779 (2016).
59. Biler, M., Crean, R. M., Schweiger, A. K., Kourist, R. & Kamerlin, S. C. L. Ground-State Destabilization by Active-Site Hydrophobicity Controls the Selectivity of a Cofactor-Free Decarboxylase. *J. Am. Chem. Soc.* **142**, 20216–20231 (2020).
60. Narayanan, C., Bernard, D. & Doucet, N. Role of Conformational Motions in Enzyme Function: Selected Methodologies and Case Studies. *Catalysts* **6**, 81 (2016).
61. Torgeson, K. R. *et al.* Conserved conformational dynamics determine enzyme activity. *Sci. Adv.* **8**, eabo5546 (2022).
62. Yon, J. M., Perahia, D. & Ghélis, C. Conformational dynamics and enzyme activity. *Biochimie* **80**, 33–42 (1998).
63. De Raffele, D., Martí, S. & Moliner, V. Understanding the Directed Evolution of De Novo Retro-Aldolases from QM/MM Studies. *ACS Catal.* **10**, 7871–7883 (2020).
64. Kim, J. K. *et al.* Fast product release requires active-site water dynamics in carbonic anhydrase. *Nat. Commun.* **16**, 4404 (2025).
65. DeChancie, J. *et al.* How similar are enzyme active site geometries derived from quantum mechanical theozymes to crystal structures of enzyme-inhibitor complexes? Implications for enzyme design. *Protein Sci.* **16**, 1851–1866 (2007).

66. Ahern, W. *et al.* Atom-level enzyme active site scaffolding using RFdiffusion2. *Nat. Methods* **23**, 96–105 (2026).
67. Lauko, A. *et al.* Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).
68. Kim, D. *et al.* Computational design of metallohydrolases. *Nature* **649**, 246–253 (2026).
69. Listov, D. *et al.* Complete computational design of high-efficiency Kemp elimination enzymes. *Nature* **643**, 1421–1427 (2025).
70. Chen, Y. *et al.* Emergence of specific binding and catalysis from a designed generalist binding protein. *Nat. Chem.* <https://doi.org/10.1038/s41557-026-02125-6> (2026)
doi:10.1038/s41557-026-02125-6.
71. Kipnis, Y. *et al.* Design and optimization of enzymatic activity in a de novo β -barrel scaffold. *Protein Sci.* **31**, e4405 (2022).
72. Anishchenko, I. *et al.* Modeling protein–small molecule conformational ensembles with PLACER. *Proc. Natl. Acad. Sci.* **122**, e2427161122 (2025).
73. Braun, M. *et al.* Computational enzyme design by catalytic motif scaffolding. *Nature* **649**, 237–245 (2026).
74. Basler, S. *et al.* Efficient Lewis acid catalysis of an abiological reaction in a de novo protein scaffold. *Nat. Chem.* **13**, 231–235 (2021).
75. Rakotoharisoa, R. V. *et al.* Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J. Am. Chem. Soc.* **146**, 10001–10013 (2024).

76. Grigoryan, G. & DeGrado, W. F. Probing Designability via a Generalized Model of Helical Bundle Geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
77. Chino, M. *et al.* Artificial Diiron Enzymes with a De Novo Designed Four-Helix Bundle Structure. *Eur. J. Inorg. Chem.* **2015**, 3371–3390 (2015).
78. Regan, L. & DeGrado, W. F. Characterization of a Helical Protein Designed from First Principles. *Science* **241**, 976–978 (1988).
79. Burgess, N. C. *et al.* Modular Design of Self-Assembling Peptide-Based Nanotubes. *J. Am. Chem. Soc.* **137**, 10554–10562 (2015).
80. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
81. Hindson, S. A. *et al.* Rigidifying a De Novo Enzyme Increases Activity and Induces a Negative Activation Heat Capacity. *ACS Catal.* **11**, 11532–11541 (2021).
82. ElGamacy, M. & Hernandez Alvarez, B. Expanding the versatility of natural and de novo designed coiled coils and helical bundles. *Curr. Opin. Struct. Biol.* **68**, 224–234 (2021).
83. Elaily, W. *et al.* Computational design of a thermostable de novo biocatalyst for whole cell biotransformations. 2024.10.07.617055 Preprint at <https://doi.org/10.1101/2024.10.07.617055> (2024).
84. Mann, S. I., Nayak, A., Gassner, G. T., Therien, M. J. & DeGrado, W. F. *De Novo* Design, Solution Characterization, and Crystallographic Structure of an Abiological Mn–Porphyrin-Binding Protein Capable of Stabilizing a Mn(V) Species. *J. Am. Chem. Soc.* **143**, 252–259 (2021).

85. Kalvet, I. *et al.* Design of Heme Enzymes with a Tunable Substrate Binding Pocket Adjacent to an Open Metal Coordination Site. *J. Am. Chem. Soc.* **145**, 14307–14315 (2023).
86. Hou, K. *et al.* De novo design of porphyrin-containing proteins as efficient and stereoselective catalysts. *Science* **388**, 665–670 (2025).
87. Bunzel, H. A. *et al.* Photovoltaic enzymes by design and evolution. Preprint at <https://doi.org/10.1101/2022.12.20.521207> (2022).
88. Zou, Z. *et al.* De novo design and evolution of an artificial metathase for cytoplasmic olefin metathesis. *Nat. Catal.* **8**, 1208–1219 (2025).
89. Zastrow, M. L. & Pecoraro, V. L. Influence of Active Site Location on Catalytic Activity in de Novo-Designed Zinc Metalloenzymes. *J. Am. Chem. Soc.* **135**, 5895–5903 (2013).
90. Hoffnagle, A. M., Srisantitham, S., Neeley, M., Tsai, C.-Y. & Tezcan, F. A. A De Novo Designed Protein with Versatile Metal Binding and Tunable Hydrolytic Activity. *Biochemistry* **64**, 3261–3271 (2025).
91. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
92. Dou, J. *et al.* De novo design of a fluorescence-activating β -barrel. *Nature* **561**, 485–491 (2018).
93. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).

94. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346 (2016).
95. Bhattacharya, S. *et al.* NMR-guided directed evolution. *Nature* **610**, 389–393 (2022).
96. Anand, N., Eguchi, R. & Huang, P.-S. Fully differentiable full-atom protein backbone generation. <https://openreview.net/forum?id=SJxnVL8YOV> (2019).
97. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
98. Butcher, J. *et al.* De novo Design of All-atom Biomolecular Interactions with RFdiffusion3. 2025.09.18.676967 Preprint at <https://doi.org/10.1101/2025.09.18.676967> (2025).
99. Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E. & Ovchinnikov, S. Boltzdesign1: Inverting All-Atom Structure Prediction Model for Generalized Biomolecular Binder Design. 2025.04.06.647261 Preprint at <https://doi.org/10.1101/2025.04.06.647261> (2025).
100. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
101. Geffner, T. *et al.* Proteina: Scaling Flow-based Protein Structure Generative Models. Preprint at <https://doi.org/10.48550/arXiv.2503.00710> (2025).
102. Wang, C. *et al.* Proteus: Exploring Protein Structure Generation for Enhanced Designability and Efficiency. 2024.02.10.579791 Preprint at <https://doi.org/10.1101/2024.02.10.579791> (2024).

103. Lin, Y., Lee, M., Zhang, Z. & AlQuraishi, M. Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2. Preprint at <https://doi.org/10.48550/arXiv.2405.15489> (2024).
104. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
105. Campbell, A., Yim, J., Barzilay, R., Rainforth, T. & Jaakkola, T. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. Preprint at <https://doi.org/10.48550/arXiv.2402.04997> (2024).
106. Wu, K. E. *et al.* Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2024).
107. Chu, A. E. *et al.* An all-atom protein generative model. *Proc. Natl. Acad. Sci.* **121**, e2311500121 (2024).
108. Bose, A. J. *et al.* SE(3)-Stochastic Flow Matching for Protein Backbone Generation. Preprint at <https://doi.org/10.48550/arXiv.2310.02391> (2024).
109. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
110. Lin, Y. & AlQuraishi, M. Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds. Preprint at <https://doi.org/10.48550/arXiv.2301.12485> (2023).
111. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

112. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone generation. Preprint at <https://doi.org/10.48550/arXiv.2302.02277> (2023).
113. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. Preprint at <https://doi.org/10.48550/arXiv.2205.15019> (2022).
114. Rector-Brooks, J. *et al.* General Multimodal Protein Design Enables DNA-Encoding of Chemistry. Preprint at <https://doi.org/10.48550/arXiv.2604.05181> (2026).
115. Jendrusch, M. & Korb, J. O. Efficient protein structure generation with sparse denoising models. *Nat. Mach. Intell.* **7**, 1429–1445 (2025).
116. Jing, B. *et al.* Generating functional and multistate proteins with a multimodal diffusion transformer. Preprint at <https://doi.org/10.1101/2025.09.03.672144> (2025).
117. Lu, T. *et al.* Conditional Protein Structure Generation with Protbardelle-1c. Preprint at <https://doi.org/10.1101/2025.08.18.670959> (2025).
118. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at <https://doi.org/10.48550/arXiv.2006.11239> (2020).
119. Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by de novo protein design. *Nat. Biotechnol.* **42**, 203–215 (2024).
120. Hsu, C., Fannjiang, C. & Listgarten, J. Generative models for protein structures and sequences. *Nat. Biotechnol.* **42**, 196–199 (2024).
121. Anand, N. *et al.* Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).

122. Yi, K., Jamali, K. & Scheres, S. H. All-atom inverse protein folding through discrete flow matching. in (2025).
123. Liu, M. *et al.* ProtInvTree: Deliberate Protein Inverse Folding with Reward-guided Tree Search. Preprint at <https://doi.org/10.48550/arXiv.2506.00925> (2025).
124. Kim, J., Shuai, R. W. & Huang, P.-S. Two-Stage Ligand-Protein Complex Sequence Design with L-Caliby. in (2026).
125. Birnbaum, F. & Keating, A. E. Beyond native sequence recovery: Improved modeling of the sequence-energy landscape of protein structures. 2026.01.14.699067 Preprint at <https://doi.org/10.64898/2026.01.14.699067> (2026).
126. Li, D. *et al.* SurfFold: a unified model for protein inverse folding by integrating surface and structural information. *Bioinformatics* **42**, btaf666 (2026).
127. Shuai, R. W., Lu, T., Bhatti, S., Kouba, P. & Huang, P.-S. Ensemble-conditioned protein sequence design with Caliby. 2025.09.30.679633 Preprint at <https://doi.org/10.1101/2025.09.30.679633> (2025).
128. Abrudan, A. *et al.* Multi-state Protein Design with DynamicMPNN. *arXiv.org* <https://arxiv.org/abs/2507.21938v2> (2025).
129. Tang, X. *et al.* BC-Design: A Biochemistry-Aware Framework for Inverse Protein Design. 2024.10.28.620755 Preprint at <https://doi.org/10.1101/2024.10.28.620755> (2025).
130. Bai, P. *et al.* Mask-prior-guided denoising diffusion improves inverse protein folding. *Nat. Mach. Intell.* **7**, 876–888 (2025).

131. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods* **22**, 717–723 (2025).
132. Shuai, R. W., Widatalla, T., Huang, P.-S. & Hie, B. L. Sidechain conditioning and modeling for full-atom protein sequence design with FAMPNN. 2025.02.13.637498 Preprint at <https://doi.org/10.1101/2025.02.13.637498> (2025).
133. Dieckhaus, H. & Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. 2024.08.20.608844 Preprint at <https://doi.org/10.1101/2024.08.20.608844> (2024).
134. Dieckhaus, H., Brocidiacano, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl. Acad. Sci.* **121**, e2314853121 (2024).
135. Gao, Z. *et al.* KW-Design: Pushing the Limit of Protein Design via Knowledge Refinement. in (2023).
136. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. 2022.04.10.487779 Preprint at <https://doi.org/10.1101/2022.04.10.487779> (2022).
137. Gao, Z., Tan, C., Chacón, P. & Li, S. Z. PiFold: Toward effective and efficient protein inverse folding. Preprint at <https://doi.org/10.48550/arXiv.2209.12643> (2023).
138. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
139. Ren, M., Yu, C., Bu, D. & Zhang, H. Accurate and robust protein sequence design with CarbonDesign. *Nat. Mach. Intell.* **6**, 536–547 (2024).

140. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning* 1263–1272 (PMLR, 2017).
141. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
142. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
143. Goudy, O. J., Nallathambi, A., Kinjo, T., Randolph, N. Z. & Kuhlman, B. In silico evolution of autoinhibitory domains for a PD-L1 antagonist using deep learning models. *Proc. Natl. Acad. Sci.* **120**, e2307371120 (2023).
144. Romero-Romero, S., Lindner, S. & Ferruz, N. Exploring the Protein Sequence Space with Global Generative Models. *Cold Spring Harb. Perspect. Biol.* **15**, a041471 (2023).
145. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
146. Munsamy, G. *et al.* Conditional language models enable the efficient design of proficient enzymes. Preprint at <https://doi.org/10.1101/2024.05.03.592223> (2024).
147. Romero-Romero, S. *et al.* *De novo* design of triosephosphate isomerases using generative language models. Preprint at <https://doi.org/10.1101/2024.11.10.622869> (2024).
148. Ruffolo, J. A. *et al.* Design of highly functional genome editors by modelling CRISPR–Cas sequences. *Nature* **645**, 518–525 (2025).

149. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
150. Davis, I. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **14**, 265–274 (2006).
151. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
152. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).
153. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178-186.e5 (2018).
154. Risso, V. A. *et al.* Enhancing a *de novo* enzyme activity by computationally-focused ultra-low-throughput screening. *Chem. Sci.* **11**, 6134–6148 (2020).
155. Merlicek, L. P. *et al.* AI.zymes: A Modular Platform for Evolutionary Enzyme Design. *Angew. Chem. Int. Ed.* **64**, e202507031 (2025).
156. Lu, T., Liu, M., Chen, Y., Kim, J. & Huang, P.-S. Assessing generative model coverage of protein structures with SHAPES. *Cell Syst.* **16**, 101347 (2025).
157. Li, H., Helling, R., Tang, C. & Wingreen, N. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science* **273**, 666–669 (1996).
158. Listov, D. & Fleishman, S. J. De novo enzyme design: Controlling structure to design function. *Curr. Opin. Struct. Biol.* **98**, 103252 (2026).

159. Banner, D. W. *et al.* Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5Å resolution: using amino acid sequence data. *Nature* **255**, 609–614 (1975).
160. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C. & Wilson, I. A. Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem. Biophys. Res. Commun.* **72**, 146–155 (1976).
161. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
162. Blacklow, S. C., Raines, R. T., Lim, W. A., Zamore, P. D. & Knowles, J. R. Triosephosphate isomerase catalysis is diffusion controlled. *Biochemistry* **27**, 1158–1165 (1988).
163. Radzicka, A. & Wolfenden, R. A Proficient Enzyme. *Science* **267**, 90–93 (1995).
164. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (β α)₈-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).
165. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).
166. Nagano, N., Orengo, C. A. & Thornton, J. M. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *J. Mol. Biol.* **321**, 741–765 (2002).
167. Hennig, M., Darimont, B. D., Jansonius, J. N. & Kirschner, K. The Catalytic Mechanism of Indole-3-glycerol Phosphate Synthase: Crystal Structures of Complexes of the Enzyme

- from *Sulfolobus solfataricus* with Substrate Analogue, Substrate, and Product. *J. Mol. Biol.* **319**, 757–766 (2002).
168. Achari, A. *et al.* Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. *Nat. Struct. Biol.* **4**, 490–497 (1997).
169. Wang, J., Duewel, H. S., Stuckey, J. A., Woodard, R. W. & Gatti, D. L. Function of His185 in *Aquifex aeolicus* 3-Deoxy-d-manno-octulosonate 8-Phosphate Synthase. *J. Mol. Biol.* **324**, 205–214 (2002).
170. Fleming, J. R. *et al.* Evolutionary Morphing of Tryptophan Synthase: Functional Mechanisms for the Enzymatic Channeling of Indole. *J. Mol. Biol.* **430**, 5066–5079 (2018).
171. Goverde, C. A. *et al.* Computational design of soluble and functional membrane protein analogues. *Nature* **631**, 449–458 (2024).
172. Chu, A. E., Fernandez, D., Liu, J., Eguchi, R. R. & Huang, P.-S. De Novo Design of a Highly Stable Ovoid TIM Barrel: Unlocking Pocket Shape towards Functional Design. *BioDesign Res.* **2022**, 9842315 (2022).
173. Kordes, S., Romero-Romero, S., Lutz, L. & Höcker, B. A newly introduced salt bridge cluster improves structural and biophysical properties of *de novo* TIM barrels. *Protein Sci.* **31**, 513–527 (2022).
174. Dellus-Gur, E., Toth-Petroczy, A., Elias, M. & Tawfik, D. S. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *J. Mol. Biol.* **425**, 2609–2621 (2013).

175. Richard, J. P., Zhai, X. & Malabanan, M. M. Reflections on the catalytic power of a TIM-barrel. *Bioorganic Chem.* **57**, 206–212 (2014).
176. Romero-Romero, S. *et al.* The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach. *J. Mol. Biol.* **433**, 167153 (2021).
177. Beck, J. & Romero-Romero, S. Designing *de novo* TIM barrels: insights into stabilization, diversification, and functionalization strategies. *Biochem. Soc. Trans.* **54**, 167–180 (2026).
178. Richard, J. P., Amyes, T. L. & Reyes, A. C. Orotidine 5'-Monophosphate Decarboxylase: Probing the Limits of the *Possible* for Enzyme Catalysis. *Acc. Chem. Res.* **51**, 960–969 (2018).
179. Rozovsky, S. & McDermott, A. E. The time scale of the catalytic loop motion in triosephosphate isomerase11Edited by P. E. Wright. *J. Mol. Biol.* **310**, 259–270 (2001).
180. Romero-Rivera, A., Corbella, M., Parracino, A., Patrick, W. M. & Kamerlin, S. C. L. Complex Loop Dynamics Underpin Activity, Specificity, and Evolvability in the $(\beta\alpha)_8$ Barrel Enzymes of Histidine and Tryptophan Biosynthesis. *JACS Au* **2**, 943–960 (2022).
181. Liao, Q. *et al.* Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. *J. Am. Chem. Soc.* **140**, 15889–15903 (2018).
182. Richard, J. P. Protein Flexibility and Stiffness Enable Efficient Enzymatic Catalysis. *J. Am. Chem. Soc.* **141**, 3320–3331 (2019).
183. 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).

184. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
185. Wiese, J. G., Shanmugaratnam, S. & Höcker, B. Extension of a *de novo* TIM barrel with a rationally designed secondary structure element. *Protein Sci.* **30**, 982–989 (2021).
186. Kordes, S., Beck, J., Shanmugaratnam, S., Flecks, M. & Höcker, B. Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs. *Protein Eng. Des. Sel.* **36**, gzad012 (2023).
187. Beck, J., Shanmugaratnam, S. & Höcker, B. Diversifying *de novo* TIM barrels by hallucination. *Protein Sci.* **33**, e5001 (2024).
188. Caldwell, S. J. *et al.* Tight and specific lanthanide binding in a *de novo* TIM barrel with a large internal cavity designed by symmetric domain fusion. *Proc. Natl. Acad. Sci.* **117**, 30362–30369 (2020).
189. Klein, A. S. *et al.* A De Novo Metalloenzyme for Cerium Photoredox Catalysis. *J. Am. Chem. Soc.* **146**, 25976–25985 (2024).
190. El Nesr, G. *et al.* Zero-shot design of a *de novo* metalloenzyme. Preprint at <https://doi.org/10.64898/2026.04.23.720277> (2026).
191. Höcker, B., Jürgens, C., Wilmanns, M. & Sterner, R. Stability, catalytic versatility and evolution of the ($\beta\alpha$)₈-barrel fold. *Curr. Opin. Biotechnol.* **12**, 376–381 (2001).
192. Wierenga, R. K., Kapetaniou, E. G. & Venkatesan, R. Triosephosphate isomerase: a highly evolved biocatalyst. *Cell. Mol. Life Sci.* **67**, 3961–3982 (2010).

193. Lee, F. S., Anderson, A. G. & Olafson, B. D. Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Eng. Des. Sel.* **36**, gzad013 (2023).
194. Vázquez Torres, S. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).
195. Casey, M. L., Kemp, D. S., Paul, K. G. & Cox, D. D. Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **38**, 2294–2301 (1973).
196. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
197. Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
198. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β -subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci.* **112**, 14599–14604 (2015).
199. Orita, I. *et al.* Crystal structure of 3-hexulose-6-phosphate synthase, a member of the orotidine 5'-monophosphate decarboxylase suprafamily. *Proteins Struct. Funct. Bioinforma.* **78**, 3488–3492 (2010).
200. Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
201. Jing, B., Berger, B. & Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. Preprint at <https://doi.org/10.48550/ARXIV.2402.04845> (2024).

202. Sullivan, S. M. & Holyoak, T. Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proc. Natl. Acad. Sci.* **105**, 13829–13834 (2008).
203. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).
204. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. Preprint at <https://doi.org/10.1101/2025.06.14.659707> (2025).
205. Chen, Y. *et al.* Emergence of specific binding and catalysis from a designed generalist binding protein. 2025.01.30.635804 Preprint at <https://doi.org/10.1101/2025.01.30.635804> (2025).
206. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
207. Leaver-Fay, A. *et al.* Rosetta3. in *Methods in Enzymology* vol. 487 545–574 (Elsevier, 2011).
208. Shetty, R. P., de Bakker, P. I. W., DePristo, M. A. & Blundell, T. L. Advantages of fine-grained side chain conformer libraries. *Protein Eng.* **16**, 963–969 (2003).
209. Chica, R. A., Moore, M. M., Allen, B. D. & Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc. Natl. Acad. Sci.* **107**, 20257–20262 (2010).
210. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).

211. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* **94**, 10172–10177 (1997).
212. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
213. Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc. Natl. Acad. Sci.* **107**, 19838–19843 (2010).
214. Chovancova, E. *et al.* CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **8**, e1002708 (2012).
215. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877 (2019).
216. Greenfield, N. J. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protoc.* **1**, 2527–2535 (2006).
217. Sparta, K. M., Krug, M., Heinemann, U., Mueller, U. & Weiss, M. S. *XDSAPP2.0*. *J. Appl. Crystallogr.* **49**, 1085–1092 (2016).
218. Kabsch, W. *XDS*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
219. McCoy, A. J. *et al.* *Phaser* crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

220. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
221. Hopkins, J. B., Gillilan, R. E. & Skou, S. *BioXTAS RAW*: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.* **50**, 1545–1553 (2017).
222. Manalastas-Cantos, K. *et al.* *ATSAS 3.0*: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343–355 (2021).
223. Svergun, D., Barberato, C. & Koch, M. H. J. *CRY SOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
224. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
225. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).
226. He, X., Man, V. H., Yang, W., Lee, T.-S. & Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J. Chem. Phys.* **153**, 114502 (2020).
227. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
228. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

229. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).
230. 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 1 Edited by J. Thornton. *J. Mol. Biol.* **285**, 1735–1747 (1999).
231. Izaguirre, J. A., Catarella, D. P., Wozniak, J. M. & Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **114**, 2090–2098 (2001).
232. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
233. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
234. Lewis, S. *et al.* Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **389**, eadv9817 (2025).
235. Jing, B., Berger, B. & Jaakkola, T. AI-based Methods for Simulating, Sampling, and Predicting Protein Ensembles. Preprint at <https://doi.org/10.48550/ARXIV.2509.17224> (2025).
236. Zarifi, N. *et al.* Distal mutations enhance catalysis in designed enzymes by facilitating substrate binding and product release. *Nat. Commun.* **16**, 8662 (2025).

237. Vander Meersche, Y., Diharce, J., Gelly, J.-C. & Galochkina, T. Flexibility or uncertainty? A critical assessment of AlphaFold 2 pLDDT. *Structure* **33**, 2157-2163.e2 (2025).
238. Lipsh-Sokolik, R. *et al.* Combinatorial assembly and design of enzymes. *Science* **379**, 195–201 (2023).
239. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
240. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
241. Mylemans, B., Lee, X. Y., Laier, I., Helsen, C. & Voet, A. R. D. Structure and stability of the designer protein WRAP-T and its permutants. *Sci. Rep.* **11**, 18867 (2021).
242. An, L. *et al.* Hallucination of closed repeat proteins containing central pockets. *Nat. Struct. Mol. Biol.* **30**, 1755–1760 (2023).
243. Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838–844 (2007).
244. Hunt, S. E. *et al.* Distal Mutations in a Designed Retro-Aldolase Alter Loop Dynamics to Shift and Accelerate the Rate-Limiting Step. *J. Am. Chem. Soc.* **147**, 30723–30736 (2025).
245. St-Jacques, A. D. *et al.* Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. *Nat. Commun.* **14**, 6058 (2023).
246. Cavanagh, P. E. *et al.* Computational design of conformation-biasing mutations to alter protein functions. *Science* **391**, eadv7953 (2026).

247. Guo, A. B. *et al.* Deep learning-guided design of dynamic proteins. *Science* **388**, eadr7094 (2025).
248. Mak, W. S. & Siegel, J. B. Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
249. Zhang, X. *et al.* Quantum Mechanical Design of Enzyme Active Sites. *J. Org. Chem.* **73**, 889–899 (2008).
250. Deng, J. & Cui, Q. Second-Shell Residues Contribute to Catalysis by Predominately Preorganizing the Apo State in PafA. *J. Am. Chem. Soc.* **145**, 11333–11347 (2023).
251. Wankowicz, S. A. & Fraser, J. S. Advances in uncovering the mechanisms of macromolecular conformational entropy. *Nat. Chem. Biol.* **21**, 623–634 (2025).