

Multistate-CANVAS enables the *de novo* design of enzymes for complex multistep C-C bond-forming reactions

Jean Roch Kouassi

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

© Jean Roch Kouassi, Ottawa, Canada, 2026

Abstract

The design of enzymes for reactions lacking natural biocatalysts is a major goal of computational protein design. While computational methods have enabled the creation of *de novo* enzymes for diverse chemical transformations, designing enzymes for complex multistep C-C bond-forming reactions remains a significant challenge. These reactions require a single active site to accommodate multiple substrates and evolving catalytic requirements throughout the reaction coordinate, making them difficult to represent using conventional single-state computational design approaches. To address this challenge, we developed Multistate-CANVAS (MsC), a computational workflow for the *de novo* design of enzymes for multistep C-C bond-forming reactions. As a proof of concept, MsC was applied to the Friedel-Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde, a multistep C-C bond-forming reaction. Experimental characterization demonstrated that multiple variants could be successfully expressed, purified, folded, and exhibited measurable catalytic activity. The most active variant, FCS_1, achieved a 254-fold enhancement in conversion relative to the uncatalyzed background reaction with a catalytic efficiency of $0.01 \text{ M}^{-1} \text{ s}^{-1}$. The successful generation of active Friedel-Crafts alkylases suggests the feasibility of extending computational enzyme design to complex multistep C-C bond-forming reactions. More broadly, MsC provides a framework for expanding the scope of *de novo* enzyme design toward increasingly challenging new-to-nature chemistry.

Acknowledgments

I would like to thank everyone who contributed to this work and supported me throughout my graduate studies.

First and foremost, I thank my Lord and Saviour Jesus Christ for the strength, perseverance, and purpose He has provided throughout this journey.

I am deeply grateful to my supervisor, Dr. Roberto A. Chica, for his guidance, mentorship, and support throughout my undergraduate and graduate studies, and for the scientific freedom and encouragement that made this work possible. His mentorship has helped me grow not only as a scientist, but as an individual.

I thank Helena Giramé Rizzo and Dr. Marc Garcia-Borràs at the University of Girona for the DFT calculations that provided the mechanistic foundation of this work, and Prof. Anthony Green at the Manchester Institute of Biotechnology for synthesizing the product standard used in the activity assays.

Thank you to all past and present members of the Chica laboratory, with special mention to Serena Hunt, Benjamin Smith, Rui Guo, Cindy Klaus, and Amir Bunyat-zada, for their support and the many discussions that made this experience both productive and enjoyable.

Finally, I thank my parents, especially my father, Jean-Marc Kouassi, and my mother, Yohou Françoise, for their counsel and financial support; my fiancée, Marie Anièle Loko, for her emotional support and patience throughout this journey; and my friends for their unwavering encouragement and support. This work would not have been possible without you.

Table of contents

Chapter 1.	Introduction.....	1
1.1	Enzymes.....	1
1.2	Approaches in <i>de novo</i> enzyme design.....	1
1.2.1	Theozyme based design.....	2
1.2.2	Motif scaffolding.....	4
1.2.3	CANVAS.....	6
1.3	Achievements and evolution of computational enzyme design.....	7
1.3.1	First example of computational <i>de novo</i> enzyme design.....	9
1.3.2	Theozyme-based design.....	10
1.3.3	Motif scaffolding.....	13
1.4	Towards the <i>de novo</i> design of complex C-C bond-forming enzymes.....	15
1.4.1	Multistep C-C bond-forming reactions.....	15
1.4.2	Multistate enzyme design.....	17
1.4.3	Hybrid scaffold generation.....	21
1.5	The Friedel-Crafts alkylation reaction.....	22
1.6	Thesis objectives.....	26
Chapter 2.	Computational design and experimental validation of <i>de novo</i> enzymes for the Friedel-Crafts alkylation.....	28
2.1	Statement of Contribution.....	28
2.2	Acknowledgements.....	28
2.3	Introduction.....	30
2.4	Characterization of the catalytic mechanism.....	32
2.4.1	Mechanistic framework.....	32
2.4.2	Theozyme construction and QM validation.....	33
2.4.3	Stepwise mechanism.....	36
2.5	Multistate CANVAS (MsC).....	40
2.5.1	Catalytic motif assembly.....	41
2.5.2	Scaffold generation.....	43
2.5.3	Theozyme placement.....	45

2.5.4	Multistate sequence design	46
2.6	Results	50
2.6.1	Structural characterization of designed enzymes.....	50
2.6.2	Functional characterization of the designed enzymes	55
2.6.3	Steady state kinetics	63
2.6.4	Molecular dynamics analysis.....	64
2.7	Discussion	67
2.8	Methods.....	76
2.8.1	Computational Enzyme Design	76
2.8.2	Catalytic motif assembly.....	76
2.8.3	Lid generation	77
2.8.4	Ensemble generation	77
2.8.5	Theozyme placement	78
2.8.6	Active-site repacking	78
2.8.7	Multistate design.....	78
2.8.8	Analysis and filtering.....	79
2.8.9	Surface redesign.....	79
2.8.10	Protein expression and purification	80
2.8.11	Circular dichroism spectroscopy and thermal stability measurements.....	81
2.8.12	Conversion assays	81
2.8.13	pH profile determination.....	82
2.8.14	Conversion time course.....	82
2.8.15	Knockouts conversion assays	83
2.8.16	Steady-state kinetics.....	83
2.8.17	Molecular dynamics (MD) simulations	84
2.9	Summary	85
Chapter 3.	Conclusions and perspectives	87
3.1	Validation of Designed Activity.....	87
3.2	Structural validation and conformational sampling.....	88
3.3	Experimental validation of the catalytic mechanism	89
3.4	Evaluation of enantioselectivity.....	90

3.5	Alternative assessment of catalytic competency.....	90
	References.....	92

List of tables

Table 1. Multistate active-site design filtering metrics.	49
Table 2. Catalytic properties of <i>de novo</i> designed C-C bond-forming enzymes.	75
Supplementary table 1. Amino acid sequences of the FCS variants.....	103
Supplementary table 2. Purification yields.	105
Supplementary table 3. Geometric constraints used to define catalytic contacts during computational design.	106
Supplementary table 4. Amino-acid and positions designed during active-site repacking.	107
Supplementary table 5. Active-site energies of the FCS variants across transition states.	109
Supplementary table 6. Triad filtering parameters for the FCS variants.	110
Supplementary table 7. Boltz-2 filtering parameters for the FCS variants using the substrate-bound complexes.	111
Supplementary table 8. Boltz-2 filtering parameters for the FCS variants using the product-bound complexes.	112
Supplementary table 9. Boltz-2 filtering parameters and cutoff values.....	113

List of figures

Figure 1.1. Theozyme-based design.	2
Figure 1.2. Motif scaffolding.	4
Figure 1.3. CANVAS.	6
Figure 1.4. Reactions catalyzed by computationally designed <i>de novo</i> enzymes.	8
Figure 1.5. Comparison of Morita–Baylis–Hillman enzymes designed using MenD (blue) and SSD (green).	21
Figure 1.6. The Friedel-Crafts alkylation of indoles.	23
Figure 1.7. Overview of thesis objectives.	26
Figure 2.1. Iminium mediated catalysis of retro aldol (a) and Friedel-Crafts alkylation (b) reactions.	32
Figure 2.2. QM calculated theozymes for the Friedel-Crafts alkylation between 2-methylindole and <i>trans</i> -cinnamaldehyde.	35
Figure 2.3. Proposed reaction mechanism for the Friedel-Crafts alkylation of 2-methylindole with <i>trans</i> -cinnamaldehyde.	37
Figure 2.4. Multistate CANVAS workflow for <i>de novo</i> enzyme design.	40
Figure 2.5. Preparation of a TIM-barrel scaffold for catalytic motif incorporation.	43
Figure 2.6. Boltz-2 models of selected variants.	49
Figure 2.7. SDS-PAGE analysis of purified enzyme variants.	52
Figure 2.8. Structural characterization of designed FCS variants.	53
Figure 2.9. Conversion assay.	56
Figure 2.10. FCS_1 structure.	57
Figure 2.11. Functional characterization of FCS_1.	58
Figure 2.12. Characterization of FCS_1 knockouts.	62
Figure 2.13. Molecular dynamics simulations of FCS variants in the absence and presence of the product.	66
Figure 2.14. Alternate proposed mechanism for the Friedel-Crafts alkylation of 2-methylindole with <i>trans</i> -cinnamaldehyde.	74
Figure 3.1. Crystal structures for enhanced conformational sampling.	88

Supplementary figure 1. Designed active sites.	101
Supplementary figure 2. Standard curve for quantification of the Friedel–Crafts alkylation product.	102

List of abbreviations

AEI	Acyl-enzyme intermediate
CANVAS	Computational design framework
CD	Circular dichroism
CPD	Computational protein design
DA	Diels–Alder
DFT	Density functional theory
DMF	Dimethylformamide
FCA	Friedel–Crafts alkylation
FPLC	Fast protein liquid chromatography
HPLC	High-performance liquid chromatography
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β -D-1-thiogalactopyranoside
IRC	Intrinsic reaction coordinate
k_{cat}	Catalytic turnover number
K_{M}	Michaelis constant

k_{cat}/K_M	Catalytic efficiency
LB	Lysogeny broth
MBH	Morita–Baylis–Hillman
MD	Molecular dynamics
MM	Molecular mechanics
MsC	Multistate-CANVAS
MW	Molecular weight
Ni-NTA	Nickel–nitrilotriacetic acid
NPT	Isothermal–isobaric ensemble (constant number, pressure, and temperature)
NVT	Canonical ensemble (constant number, volume, and temperature)
OD600	Optical density at 600 nm
PDB	Protein Data Bank
pLDDT	Predicted Local Distance Difference Test
PLP	Pyridoxal-5'-phosphate
PNPA	<i>p</i> -Nitrophenyl acetate
QM	Quantum mechanics

QM/MM	Quantum mechanics/molecular mechanics
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
SASA	Solvent-accessible surface area
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
TIM	Triose phosphate isomerase
T _m	Melting temperature
TS	Transition state
TTN	Total turnover number

Chapter 1. Introduction

1.1 Enzymes

Enzymes are highly efficient biological catalysts that enable reaction rate accelerations up to 26 orders of magnitude.¹⁻³ Their specificity, selectivity, and mild reaction conditions have led to their widespread use across various industrial sectors.⁴ Unfortunately, the broader implementation of enzymes in industrial processes is limited by the finite repertoire of naturally occurring catalysts. In particular, many industrially relevant transformations, especially new-to-nature reactions, have no naturally evolved enzymatic counterpart.^{4,5} While protein engineering techniques have enabled the creation of new functions through rational design and/or directed evolution, they remain constrained by the enormous size of protein sequence space, making identification of optimal variants challenging.⁴ Thus, one of the major goals of computational protein engineering is to develop reliable computational workflows capable of designing efficient *de novo* enzymes for any desired chemical reaction.⁶

1.2 Approaches in *de novo* enzyme design

De novo enzyme design is the use of computational methods to create enzymes from scratch. Traditionally, this has been achieved by combining computational protein design (CPD) methodologies with theoretical models of chemical reactivity, placing a model of the catalytic machinery into an inert protein scaffold; an approach known as theozyme-based design.^{6,7} Advances in machine learning have enabled the inverse strategy, motif scaffolding, in which the scaffold is generated *de novo* around a predefined catalytic motif.^{8,9} More recently, a hybrid approach known as CANVAS has emerged and bridges

these two strategies, fixing part of the scaffold while generating the remainder *de novo* around the unaccommodated catalytic residues.¹⁰

1.2.1 Theozyme based design

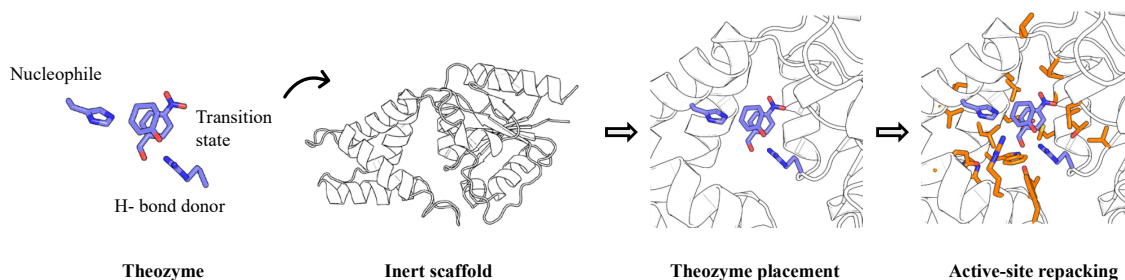


Figure 1.1. Theozyme-based design. The theozyme, comprising the transition state and catalytic residues, is incorporated into an inert protein scaffold by theozyme placement. Active site residues are optimized through repacking to reinforce transition-state stabilization and active-site packing.

Theozyme-based enzyme design is grounded in transition-state theory, which holds that reaction rates are governed by the free energy difference between the ground-state reactants and the transition state, the highest-energy point along the reaction coordinate.¹¹ Accordingly, the first step in the design process is the construction of a theoretical enzyme, or theozyme, a minimal quantum mechanical model comprising the transition state together with the catalytic residues arranged to optimally stabilize it. The theozyme¹² defines the geometric and chemical constraints required for catalysis and serves as the basis for subsequent active-site construction. It is then incorporated into an inert protein scaffold through a process termed theozyme placement, in which computational protein design algorithms, such as ligand placement¹³ or matching¹⁴ identify backbone positions capable of accommodating the theozyme while preserving the catalytic geometry. Both approaches systematically sample catalytic sidechain rotamers and transition state poses under user-defined geometric constraints. In ligand placement, rotamers of the first catalytic residue

are sampled; for each rotamer satisfying the first catalytic contact, a transition state pose is generated, and the remaining residues are then sampled to identify conformations that maintain all contacts simultaneously. In matching, each catalytic residue is sampled independently, and overlapping transition state poses compatible with rotamers from all residues are identified to yield a common geometry. Since theozyme placement is concerned primarily with satisfying the geometric constraints of the catalytic motif, it does not optimize the surrounding protein environment. Consequently, a subsequent active-site repacking step is performed to redesign neighboring residues while preserving the catalytic geometry.¹⁵ There, non-catalytic residues surrounding the transition state are redesigned through an algorithmic search of a user-defined sequence space, simultaneously sampling rotameric conformations. Guided by energy functions and geometric constraints, this identifies residue configurations that preserve catalytic geometry while further stabilizing the transition state through favorable non-bonding interactions. Designs are ranked by computed energy scores, and top candidates are selected for experimental validation. Theozyme-based design methods have enabled the successful creation of *de novo* enzymes for reactions including Kemp elimination, retro-aldol cleavage, Diels–Alder cycloaddition, and ester hydrolysis, establishing transition-state stabilization as a viable strategy for *de novo* enzyme design.^{16–19}

1.2.2 Motif scaffolding

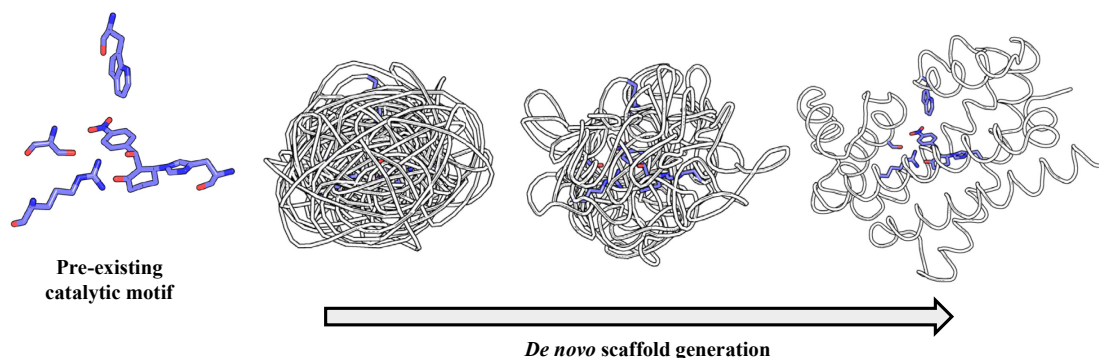


Figure 1.2. Motif scaffolding. A pre-existing catalytic motif is fixed as a geometric constraint while a compatible protein scaffold is generated *de novo* around it.

An alternative strategy for *de novo* enzyme design is motif scaffolding, which reverses the classical design paradigm. Rather than placing a catalytic motif into a pre-existing protein scaffold, a protein backbone is generated *de novo* around a fixed catalytic motif.^{8,20} The starting point is a catalytic motif derived from the active site of a known enzyme. This motif is held fixed as a geometric constraint throughout the design process, and the scaffold is built around it rather than the motif being fitted into the scaffold. Backbone generation is performed using deep learning-based generative models such as RFDiffusion.²¹ Two broad implementations have been developed, differing in how the catalytic motif is represented prior to scaffold generation. In the first,⁸ the atom-level catalytic array is converted into a backbone-level representation through a preprocessing step in which each catalytic residue is embedded into a short backbone fragment compatible with its preferred rotamer conformation. Compatible, non-clashing fragment assemblies are identified and ranked to generate a library of candidate motif geometries, which is then provided as input to the generative model. In the second implementation, the preprocessing step is bypassed entirely. The atom-level coordinates of the catalytic functional groups are provided directly

as input, and the generative model simultaneously infers backbone topology, rotamer conformations, and sequence positions during scaffold generation. This eliminates the need for combinatorial enumeration and enables the scaffolding of more complex active sites with multiple discontinuous catalytic residue segments.²⁰ Following backbone generation, sequences are fitted to the scaffold using inverse folding models such as ProteinMPNN²² or LigandMPNN²³, conditioned on both the backbone geometry and the coordinates of the catalytic motif to ensure compatibility with the active site environment. The resulting sequences are passed to structure prediction models such as AlphaFold2²⁴, Chai-1²⁵, and ESMFold²⁶, and the predicted structures are ranked and filtered using a combination of computational metrics to identify the best candidates for experimental characterization. By generating protein backbones *de novo* around predefined catalytic motifs, motif scaffolding eliminates the dependence on compatible pre-existing protein scaffolds. This substantially expands the structural design space accessible to computational enzyme design and enables catalytic motifs to be accommodated within tailor-made protein architectures. Consequently, motif scaffolding has broadened the scope of *de novo* enzyme design and accelerated the development of functional enzyme scaffolds.

1.2.3 CANVAS

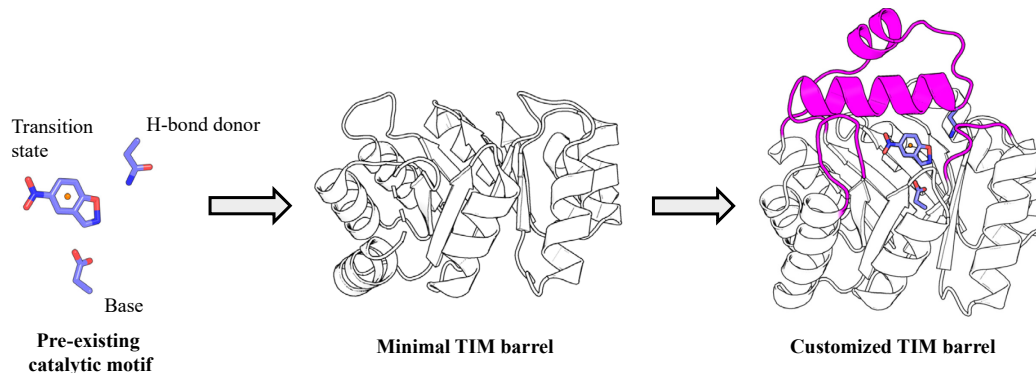


Figure 1.3. CANVAS. A theozyme is partially matched into a pre-existing scaffold, while the remaining catalytic residues are accommodated through *de novo* backbone generation.

More recently, Beck et al.¹⁰ introduced CANVAS, a hybrid computational workflow that combines elements of theozyme-based design and motif scaffolding. In theozyme-based design, the entire scaffold is fixed, and the catalytic machinery is placed within it. In motif scaffolding, the entire scaffold is generated *de novo* around a fixed motif. CANVAS occupies a middle ground: part of the scaffold is fixed, providing a stable structural foundation, while the remaining portion is generated *de novo* around catalytic residues that cannot be accommodated by the existing structure. The workflow begins with theozyme placement onto a pre-existing scaffold, positioning the catalytic residues that can be incorporated into the available structure. Then, using deep learning-based generative models, a structural addition is generated *de novo* around the residues unaccommodated by the existing scaffold. The resulting hybrid is then subjected to active site repacking to optimize the surrounding residue environment for transition state stabilization. Designs are subsequently ranked and filtered using a combination of computational metrics before experimental validation. By combining the structural stability of an established scaffold

with the flexibility of *de novo* backbone generation, CANVAS expands the range of catalytic geometries that can be explored beyond what is accessible to either approach alone.

Collectively, theozyme-based design, motif scaffolding, and CANVAS represent the principal computational approaches currently available for *de novo* enzyme design. Their successful application to an expanding range of chemical transformations has demonstrated the feasibility of engineering catalytic activity *de novo*. The following section reviews the major achievements of these methodologies across representative reaction classes.

1.3 Achievements and evolution of computational enzyme design

Over the past two decades, computational enzyme design has been applied across a chemically diverse range of reactions, from transformations with natural enzymatic counterparts to chemistries with no known biological precedent. Successes spanning Kemp eliminases,^{10,16} retro-aldolases,^{8,17} ester hydrolases,^{20,27} Diels-Alderase and Morita-Baylis-Hilman^{8,28} catalysts have demonstrated that active sites computationally installed onto protein scaffolds can confer genuine catalytic activity. This capacity to engineer catalysis *de novo* was first established by a single proof-of-concept study, which demonstrated that computational protein design could be used to engineer catalytic activity into an inert scaffold.

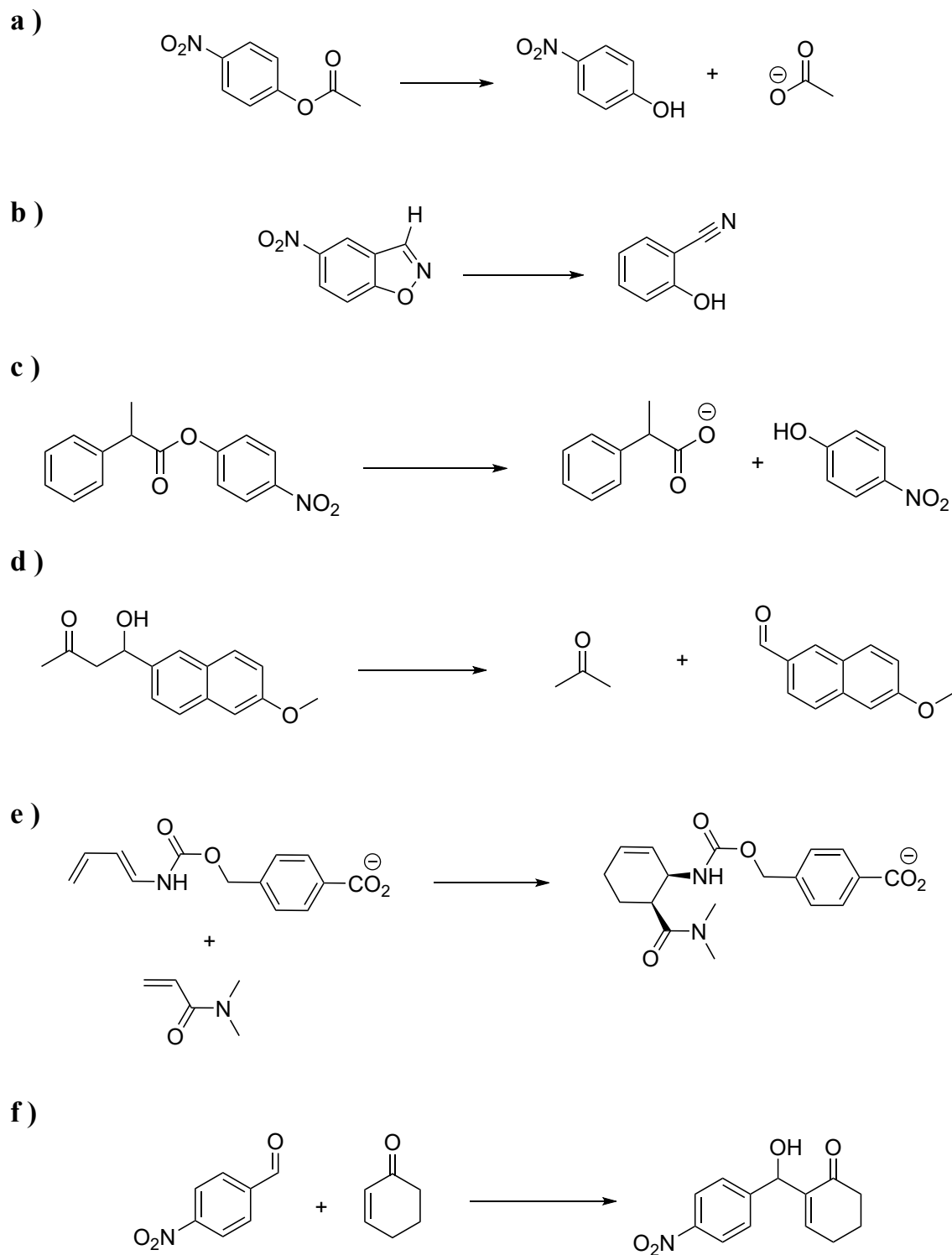


Figure 1.4. Reactions catalyzed by computationally designed *de novo* enzymes. a) Hydrolysis of *p*-nitrophenyl acetate into *p*-nitrophenol and acetate. b) Kemp elimination of 5-nitrobenzisoxazole to cyanophenol. c) Ester hydrolysis of *p*-nitrophenyl-2-phenylpropanoate to 2-phenylpropanoate and *p*-nitrophenol. d) Retro-aldol cleavage of 4-

hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) to yield 6-methoxy-2-naphthaldehyde (6-MNA) and acetone. e) Diels–Alder [4 + 2] cycloaddition of 4-carboxy-benzyl-trans-1,3-butadiene-1 carbamate on N,N-dimethylacrylamide. f) Morita–Baylis–Hillman reaction between 4-nitrobenzaldehyde and 2-cyclohexenone.

1.3.1 First example of computational *de novo* enzyme design

The first demonstration of *de novo* enzyme design to engineer a biocatalyst for a defined organic reaction was reported by Bolon and Mayo in 2001.²⁷ Using the nucleophilic histidine-mediated hydrolysis of p-nitrophenyl acetate (PNPA) to p-nitrophenol and acetate as a model reaction, the authors introduced esterase-like activity into the catalytically inert 108-residue *E. coli* thioredoxin scaffold. At the time, formal theozyme placement methodologies had not yet been developed, and the computational tools required to explicitly model and position transition state geometries within protein scaffolds were unavailable. A simplified "superrotamer" strategy was therefore adopted: rather than explicitly stabilizing the true transition state, rotamers of the high-energy tetrahedral intermediate were placed at each backbone position of thioredoxin to identify sites where a catalytic histidine could be geometrically positioned for nucleophilic attack. The surrounding residues were then repacked computationally using the ORBIT software²⁹ to accommodate substrate binding, effectively using the high-energy intermediate as a transition state proxy. An active-site scan across all 94 non-glycine, non-proline positions yielded ten candidate designs, of which the top two, PZD1 and PZD2, were expressed and experimentally characterized. The more active variant, PZD2, displayed saturation kinetics with a k_{cat}/K_M of $2.7 \text{ M}^{-1} \text{ s}^{-1}$, well below that of natural enzymes, which typically operate in the range of $10^5 \text{ M}^{-1} \text{ s}^{-1}$.³⁰ While the study was a landmark proof of concept, the modest rate enhancement achieved by PZD2 exposed the core limitation of the approach:

approximating the transition state with a high-energy intermediate, rather than modeling the transition state itself, is insufficient to achieve the precise electrostatic and geometric complementarity needed for efficient catalysis. The subsequent development of explicit theozyme placement methodologies directly addressed this shortcoming, enabling the design of active sites with far greater transition state complementarity, as described in the following sections.

1.3.2 Theozyme-based design

The introduction of theozyme-based design enabled catalytic architectures to be constructed directly from the geometric requirements of a reaction transition state. One of the first major demonstrations of this approach was the computational design of Kemp eliminases by R  thlisberger *et al.*¹⁶ Of 59 designs tested across 17 scaffolds, the most active design, KE59, achieved a $k_{\text{cat}}/K_{\text{M}}$ of 163 M⁻¹ s⁻¹ and a rate enhancement of 2.5×10^5 . Subsequent incorporation of structural information and molecular dynamics simulations by Privett *et al.*³¹ produced HG-3, which reached a $k_{\text{cat}}/K_{\text{M}}$ of 425 M⁻¹ s⁻¹ and a rate enhancement of 5.9×10^5 . These studies established that theozyme-based design could successfully generate substantial catalytic activity when the catalytic requirements could be represented by a relatively simple arrangement of functional groups around a single transition state. The approach was subsequently extended to reactions requiring more elaborate catalytic arrangements. Richter and Blomberg *et al.*¹⁹ applied theozyme-based design to ester hydrolysis by incorporating a Cys–His catalytic dyad together with an oxyanion hole. Three strategies for oxyanion stabilization were investigated, of which only the use of backbone NH groups produced active designs. Following optimization, the most active variant achieved a $k_{\text{cat}}/K_{\text{M}}$ of 400 M⁻¹ s⁻¹. However, structural characterization

revealed that the intended Cys–His dyad was not properly formed in any of the designs, illustrating the difficulty of reproducing more complex networks of geometrically constrained catalytic interactions within existing protein scaffolds. Theozyme-based design was also successfully extended to bimolecular bond-forming chemistry. Siegel and Zanghellini *et al.*¹⁸ designed a Diels–Alderase requiring the simultaneous positioning of two substrates and two catalytic residues. Of 84 designs experimentally tested, two displayed activity, and subsequent active-site optimization produced DA_20_10, which generated exclusively the (3*R*,4*S*) endo product with approximately 97% stereoselectivity. This demonstrated that computational enzyme design could extend beyond unimolecular bond-breaking reactions to reactions requiring simultaneous organization of multiple substrates and stereochemical control. Nevertheless, increasing the number of substrates, catalytic residues and geometrically constrained interactions that must be simultaneously accommodated increases the difficulty of identifying protein environments compatible with the desired catalytic architecture.

The extension of computational enzyme design to multistep reactions introduced the additional challenge of accommodating the catalytic requirements of multiple intermediates and transition states. Jiang and Althoff *et al.*¹⁷ addressed this problem in the design of enzymes for the retro-aldol decomposition of methodol by constructing a composite transition-state representation incorporating key states along the reaction coordinate. Four alternative catalytic motifs were evaluated, with designs containing a hydrophobic environment surrounding the catalytic lysine and an explicit water molecule proving most successful. The most active initial design, RA61, displayed a catalytic efficiency of 0.74 M⁻¹ s⁻¹ and a rate enhancement of 2.3×10^4 . Subsequent computational

design generated the RA95 lineage,³² which was further improved through directed evolution by Giger *et al.*³² to produce RA95.5-8, with a $k_{\text{cat}}/K_{\text{M}}$ of 850 M⁻¹ s⁻¹. These studies demonstrated that computational enzyme design could be extended to multistep reactions and that catalytic requirements associated with several mechanistic states could be incorporated into the design process. The composite state strategy was subsequently applied to the more complex Morita–Baylis–Hillman reaction, which combines several of the challenges encountered individually in earlier design efforts. The MBH reaction is bimolecular, involves C–C bond formation, proceeds through three covalent intermediates and four transition states, and has no known natural enzymatic counterpart. Bjelic *et al.*²⁸ used a composite representation incorporating structures derived from different points along the reaction coordinate and searched for existing scaffolds capable of accommodating the resulting catalytic architecture. Of 48 designs tested, only BH25 and BH32 displayed measurable activity, with rate enhancements of 27- and 19-fold, respectively. Although modest, these activities demonstrated that computational design could produce catalysts for a complex, multistep, new-to-nature bond-forming reaction.

Together, these studies established theozyme-based design as an effective strategy for generating catalytic activity from predefined catalytic geometries and demonstrated, through composite representations, that requirements from multiple mechanistic steps can be incorporated into a single design target. However, this approach becomes increasingly challenging as catalytic motifs grow in size and complexity, particularly because of its dependence on pre-existing scaffolds. Moreover, composite representations do not treat individual mechanistic states independently, limiting their ability to account for changing catalytic requirements throughout complex multistep reactions. Thus, the design of

increasingly complex catalytic architectures and their optimization across multiple distinct mechanistic states remain important challenges.

More recently, Beck *et al.*¹⁰ applied the CANVAS methodology to the Kemp elimination, demonstrating for the first time that an enzymatically active *de novo* enzyme could be constructed from a minimal TIM barrel scaffold. Using two *de novo* TIM barrel crystal structures as templates, CANVAS was employed to introduce a custom structural lid around a theozyme comprising an Asp general base and an Asn hydrogen bond donor, generating nine KempTIM variants with distinct lid architectures and active site configurations. Of these, one active design, KempTIM4, was identified, achieving a k_{cat}/K_M of $11 \text{ M}^{-1} \text{ s}^{-1}$ comparable to first-generation theozyme-based Kemp eliminases such as KE07. Together, these successive demonstrations on the Kemp elimination benchmark illustrate how advances in computational design methodology, from theozyme-based design to CANVAS, have progressively expanded the range of achievable scaffolds and active site architectures, establishing a growing toolkit for the creation of novel biocatalysts.

1.3.3 Motif scaffolding

More recent motif-scaffolding approaches have substantially expanded the structural space available for accommodating catalytic motifs and have produced notable improvements over earlier fixed-scaffold design strategies. This is particularly evident for ester hydrolysis. Lauko and Pellock *et al.*²⁰, designed proteins around a complete Ser–His–Asp catalytic triad together with a two-component oxyanion hole, an architecture that had proven difficult to accommodate using earlier fixed-scaffold approaches. The most active

design, momi120-103, achieved a k_{cat}/K_M of $2.2 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$, representing a several-hundred-fold improvement over the earlier designs of Richter and Blomberg *et al.*¹⁹. These results demonstrate how increasing the structural freedom available during scaffolding can enable more elaborate catalytic arrangements to be incorporated into designed proteins. Similar improvements have been observed for multistep reactions. Braun *et al.*⁸ applied motif scaffolding to the retro-aldol reaction using the well-characterized catalytic tetrad of evolved retro-aldolases as the input catalytic array. Of only 35 expressed designs, 32 exhibited activity above background. RAD29 achieved a catalytic efficiency of $290 \text{ M}^{-1} \text{ s}^{-1}$, approaching that of the extensively evolved RA95.5-5, while RAD35 displayed an enantiomeric excess of 99% for the (*R*) product. These activities were obtained without directed evolution, demonstrating the ability of motif-scaffolding approaches to generate active proteins efficiently when an effective catalytic arrangement is already available. The same methodology was subsequently applied to the more demanding MBH reaction. Using two catalytic arrays derived from previously characterized MBHases, they generated 63 designs, more than 93% of which displayed measurable activity above background. The most active designs, MBH18 and MBH48, achieved k_{cat} values of 7.7×10^{-3} and 0.025 min^{-1} , respectively. This represents a substantial improvement in design success rate compared with the earlier MBH study, in which only two of 48 designs displayed measurable activity.²⁸ Nevertheless, despite the improved success rate, the catalytic rates remained modest relative to those obtained for simpler transformations, further illustrating the difficulty of designing efficient catalysts as reaction complexity increases.

Together, these studies demonstrate that motif scaffolding can overcome some of the structural constraints associated with placing catalytic arrangements into pre-existing

scaffolds and can substantially increase the probability of obtaining active designs. However, the successful applications described above relied on catalytic motifs that could be specified a priori, either from established catalytic arrangements or from previously characterized and optimized enzymes. As catalytic motifs increase in size and complexity, their definition and incorporation into a protein structure become increasingly challenging. This limitation is particularly relevant for complex multistep reactions, where different catalytic interactions may be required at different stages of the reaction coordinate. Consequently, the challenge is no longer solely to generate a protein capable of accommodating a predefined catalytic motif, but also to define a sufficiently complex catalytic architecture that captures the requirements of the complete reaction within a common active site.

1.4 Towards the *de novo* design of complex C-C bond-forming enzymes

1.4.1 Multistep C-C bond-forming reactions

While *de novo* enzyme design has enabled the successful creation of biocatalysts for a broad range of chemical transformations, the *de novo* design of enzymes for complex multistep C-C bond-forming reactions remains an open challenge^{33–35}. Although computational methodologies have progressively expanded from unimolecular single-step reactions to multistep and bimolecular transformations, no computational framework has yet been specifically developed to address the combined requirements of complex multistep C-C bond-forming reactions. These reactions present several levels of mechanistic complexity arising from the combination of multiple reacting species, successive chemical transformations, and substantial changes in molecular structure and electronic character throughout the catalytic cycle. For example, the Morita-Baylis-

Hillman reaction investigated by Bjelic *et al.*²⁸ involves two substrates and proceeds through two distinct catalyst-substrate intermediates and four transition states²⁸. As reaction mechanisms become more elaborate, the active site must accommodate substantial structural and electronic changes as the reacting species progress from substrates, through chemically distinct intermediates and transition states, to products whose structures differ markedly from either reactant. Consequently, the catalytic requirements of the active site evolve throughout the reaction coordinate. Distinct catalytic interactions become necessary at different stages of the mechanism, and catalytic residues that play a central role during one step may contribute little or not at all during another. The active site must therefore simultaneously maintain productive binding of multiple reacting species, support the different chemical transformations occurring throughout the catalytic cycle, and accommodate the structural and conformational rearrangements associated with C-C bond formation, catalyst regeneration, and product release.

Although similarly complex multistep mechanisms occur in other classes of enzymatic reactions, C-C bond-forming transformations are of particular interest because they enable the construction of new carbon frameworks from simpler molecular building blocks. C-C bond formation is fundamental to organic synthesis and provides a means of assembling the molecular frameworks of pharmaceuticals, natural products, agrochemicals, and other complex organic molecules^{36,37}. Complex C-C bond-forming reactions therefore represent valuable targets for expanding the synthetic scope of *de novo* enzyme design. While naturally evolved enzymes satisfy the diverse requirements of such reactions through highly optimized active-site architectures and conformational flexibility, reproducing this level of catalytic organization through computational design remains challenging.

To date, computational enzyme design methodologies have largely been developed around the optimization of a single catalytic model during the design process, typically consisting of an idealized transition-state representation surrounded by catalytic functional groups positioned to promote the targeted chemical transformation. Although this paradigm has enabled the successful design of enzymes for a broad range of chemical transformations, extending it to complex multistep C-C bond-forming reactions requires consideration of a broader set of catalytic and structural requirements throughout the catalytic cycle. Developing computational methodologies capable of translating these requirements into active-site architectures that support the complete reaction therefore represents an important step toward extending *de novo* enzyme design to increasingly complex and synthetically valuable chemical transformations.

1.4.2 Multistate enzyme design

Current computational enzyme design methodologies have predominantly followed a single-state design (SSD) paradigm, in which the active site is optimized to stabilize a single transition state or a composite state. This approach has formed the basis of most *de novo* enzyme design efforts to date, including theozyme-based design strategies. In single-state design, the rate-limiting transition state is typically selected as the target for optimization, on the grounds that stabilizing this state will have the greatest impact on the overall reaction rate. This strategy has enabled the successful design of enzymes for a broad range of chemical transformations. However, multistep reactions proceed through a sequence of chemically distinct intermediates and transition states, each contributing to the overall activation free energy of the catalytic cycle. Consequently, optimizing an active site

for a single catalytic state does not explicitly consider the catalytic requirements of the remaining reaction pathway.

From a computational standpoint, single-state design operates by evaluating a scoring function, typically a molecular mechanics or knowledge-based energy function, against a fixed structural representation of the chosen catalytic state. Sequence selection is then carried out by identifying amino acid combinations that minimize this energy, effectively asking which active-site residues best stabilize that one catalytic geometry. The fundamental limitation of this approach for multistep reactions is that the geometric and chemical requirements of the active site differ between states. Residues selected to optimize interactions with one transition state may be incompatible with or actively disfavor the geometries required by other states along the reaction coordinate. Because these conflicts are never evaluated during design, they can only be discovered experimentally, after the fact.

Multistate design (MSD) is a general computational strategy in which a single protein sequence is optimized simultaneously across multiple structural states, rather than against a single fixed representation.³⁸ In this framework, sequence selection is driven not by a single energy evaluation but by a combined fitness function that integrates contributions from all included states, such that the selected sequence must perform well across all of them simultaneously. MSD has been applied across a range of protein engineering problems, including the design of sequences capable of adopting two distinct conformations,³⁹ the engineering of binding specificity through the simultaneous inclusion of desired and undesired interaction states,⁴⁰ and the improvement of protein stability predictions through the incorporation of backbone conformational ensembles as design

states.⁴¹ This framework was subsequently extended to enzyme substrate specificity by St-Jacques *et al.*,⁴² who developed a multistate methodology termed multichemical state analysis to optimize enzyme sequences across large structural ensembles for productive binding of multiple target substrates, demonstrating that MSD could be applied to engineer the chemical recognition properties of enzyme active sites.

The application of MSD to enzyme catalysis represents a natural extension of this framework to the problem of multistep reaction mechanisms. Rather than optimizing substrate binding across multiple chemical identities, multistate enzyme design optimizes catalytic competence across multiple points along a single reaction coordinate. To address this, our laboratory developed Multistate Enzyme Design (MenD), an unpublished methodology that extends the MSD framework to the *de novo* design of enzyme active sites for multistep reactions. In MenD, each transition state or catalytic intermediate along the reaction pathway is represented as an independent structural state, and sequence selection is performed such that the resulting active-site architecture must satisfy the geometric and chemical requirements of all included states at once. This is achieved using a sum fitness function, in which the energies computed across all state ensembles are summed into a single fitness score for each candidate sequence, which is subsequently optimized using a Monte Carlo search algorithm. To further ensure that selected sequences remain energetically favorable independently of their interactions with the transition-state models, each transition-state-containing structure is paired with a corresponding transition-state-free structure, both of which contribute to the combined fitness score. A sequence that stabilizes one state at the expense of another will therefore score poorly, driving the optimization toward solutions broadly compatible with the full reaction coordinate. This

represents a fundamental departure from single-state design: whereas SSD asks what sequence best accommodates one catalytic geometry, MenD asks what sequence is most compatible with all of them.

Furthermore, naturally evolved enzymes catalyze complete reaction mechanisms rather than isolated chemical steps. It has therefore been hypothesized that their remarkable catalytic efficiencies arise, at least in part, from their ability to stabilize multiple catalytic states along the reaction coordinate. Although this hypothesis has not been conclusively established, it provides a compelling conceptual framework for the computational design of enzymes for complex multistep reactions and lends further motivation to the development of multistate design strategies such as MenD.

Motivated by this idea, MenD was first applied in our laboratory to the de novo design of Morita–Baylis–Hillman enzymes, where ten enzymes designed using a multistate approach were compared against ten counterparts generated using conventional single-state design under identical assay conditions. MenD-derived enzymes consistently displayed higher catalytic activity than their SSD counterparts, as measured by HPLC (unpublished data, Chica laboratory). This initial result suggests that simultaneously accommodating multiple catalytic states during sequence design may confer a catalytic advantage for multistep reactions and raises the possibility that MenD could prove particularly powerful for the design of enzymes operating through complex, multi-state mechanisms.

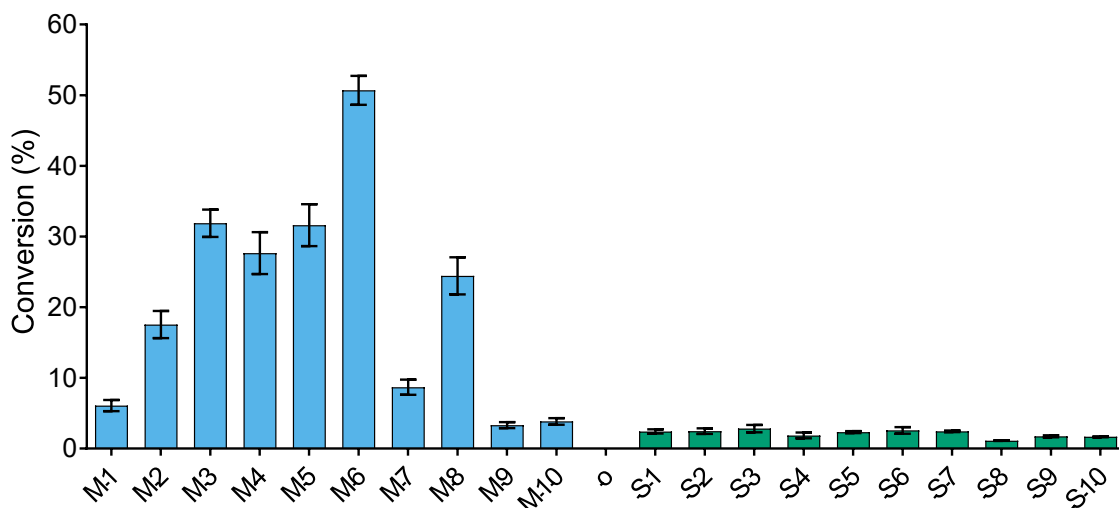


Figure 1.5. Comparison of Morita–Baylis–Hillman enzymes designed using Mend (blue) and SSD (green). Biotransformations were performed with 3 mM 2-cyclohexen-1-one, 0.6 mM 4-nitrobenzaldehyde, and 18 μ M enzyme (3 mol%) at 30 °C in phosphate-buffered saline (pH 7.4) containing 3% (v/v) acetonitrile, and analyzed after 22 h by HPLC. Error bars represent the experimental uncertainty associated with the measured conversion.

Although these findings remain preliminary, they support the hypothesis that explicitly considering multiple catalytic states during enzyme design may provide a more suitable framework for engineering enzymes catalyzing multistep reactions. Extending this concept to increasingly demanding reaction mechanisms therefore represents a logical next step in the development of computational enzyme design methodologies.

1.4.3 Hybrid scaffold generation

Accommodating a complex catalytic architecture within a protein scaffold presents a major challenge in *de novo* enzyme design. The geometric requirements imposed by catalytic motifs optimized across multiple reaction states may not be readily satisfied by existing protein folds, limiting the range of active-site architectures that can be explored. Frameworks such as CANVAS may provide a solution to this challenge. By combining

naturally occurring protein folds with *de novo* designed structural elements, CANVAS expands the structural space available for active-site incorporation while preserving the overall architecture of a natural protein scaffold. This ability to introduce additional structural elements expands the range of positions available for the placement of catalytic residues or motifs beyond those provided by the original scaffold. Such capabilities are especially valuable for reactions that lack a naturally evolved enzymatic counterpart, where the catalytic architecture must be established *de novo* and accommodated within a suitable protein scaffold. Collectively, these concepts highlight the key computational requirements for extending *de novo* enzyme design to complex multistep C-C bond-forming reactions. Integrating catalytic design across multiple reaction states with scaffold-generation strategies capable of accommodating the resulting active-site architecture provides a promising direction for addressing this class of transformations. The Friedel-Crafts alkylation of indoles with α,β -unsaturated aldehydes exemplifies of this class of reactions. As a complex multistep C-C bond-forming reaction, it provides a stringent test for computational methodologies aimed at extending *de novo* enzyme design for this reaction class.

1.5 The Friedel-Crafts alkylation reaction

The Friedel-Crafts alkylation⁴³ reaction, discovered in 1877 by Charles Friedel and James Mason Crafts, was developed as a method to introduce alkyl groups onto aromatic rings using a Lewis acid catalyst such as aluminum chloride. Its adaptation to indoles is of particular pharmaceutical significance, as the indole scaffold is a common structural motif in numerous bioactive molecules including neurotransmitters, alkaloids, and therapeutic agents. C3-alkylated indole derivatives are found in compounds with anticancer, antiviral,

antimicrobial, and anti-inflammatory activities, making the enantioselective synthesis of such products a priority in medicinal chemistry and drug discovery.^{44–46}

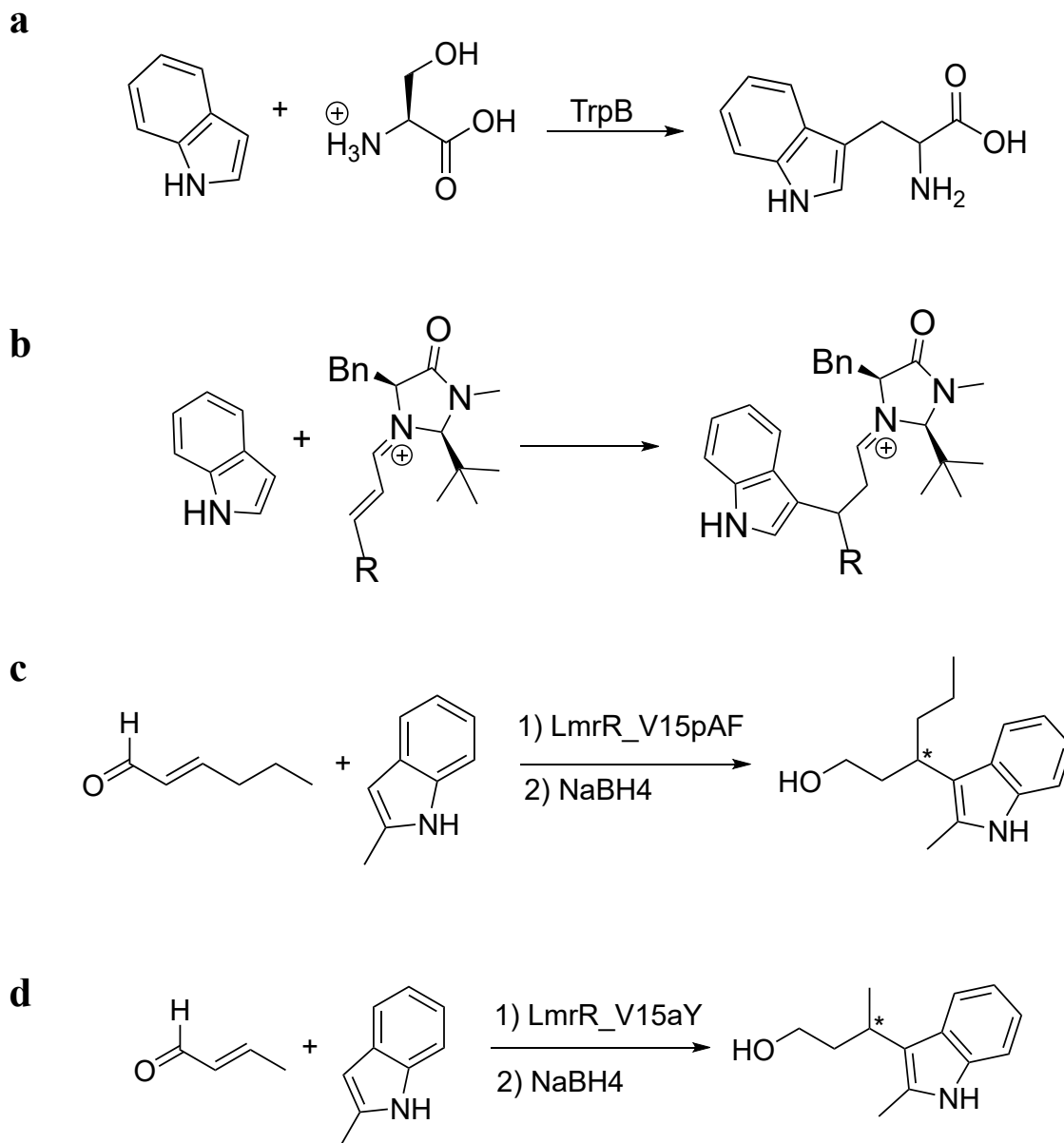


Figure 1.6. The Friedel-Crafts alkylation of indoles. (a) Natural Friedel-Crafts alkylation catalyzed by the tryptophan synthase TrpB. (b) Organocatalytic Friedel-Crafts alkylation mediated by the MacMillan catalyst. (c–d) Artificial Friedel-Crafts alkylation catalyzed by engineered enzymes featuring a non-canonical amino acid.

Nature has evolved enzymes capable of mediating Friedel-Crafts type alkylation of indoles, the most prominent being tryptophan synthase (TrpB) (Figure 1.6.a), a pyridoxal-5'-phosphate (PLP)-dependent enzyme that condenses L-serine with indole to generate L-tryptophan through a multistep β -replacement mechanism.⁴⁷ Directed evolution has yielded standalone TrpB variants with broad substrate scope, accepting a wide range of indole analogues to produce diverse tryptophan derivatives in high yield and with excellent enantioselectivity. Despite these achievements, TrpB is fundamentally constrained by its PLP-dependent mechanism, limiting its applicability to substrates compatible with the amino-acrylate electrophilic intermediate and precluding its use for broader Friedel-Crafts chemistry beyond the condensation of indoles with amino acid precursors. The first enantioselective organocatalytic Friedel-Crafts alkylation of indoles was reported by MacMillan and coworkers in 2001⁴⁸, who demonstrated that chiral iminium ion catalysis using an imidazolidinone catalyst could mediate the asymmetric alkylation of indoles with α,β -unsaturated aldehydes with high enantioselectivity (Figure 1.6.b). This work established iminium catalysis as a powerful strategy for enantioselective Friedel-Crafts chemistry and inspired the subsequent development of a broad family of organocatalytic systems for this transformation.⁴⁹ However, the MacMillan catalyst presents significant practical limitations. It typically requires relatively high loadings (10–20 mol%) and available imidazolidinone catalyst kits are expensive. This can translate into substantial cost at preparative and industrial scales.⁵⁰ These factors limit their scalability, particularly for large-scale applications. In addition, achieving high enantioselectivity often requires low-temperature conditions, introducing considerable energy demands and operational

complexity.⁴⁸ Finally, the catalyst operates exclusively in organic solvents, restricting its use in aqueous or physiological environments and limiting its compatibility with green chemistry and biologically relevant settings. These limitations of both natural enzymes and organocatalysts motivated efforts to design artificial enzymes for the Friedel-Crafts alkylation of indoles. A particularly successful strategy exploited non-canonical amino acids to introduce catalytic functionality not available in the canonical amino acid repertoire. Leveson-Gower and coworkers⁵¹ demonstrated that incorporating a non-canonical p-aminophenylalanine (pAF) bearing an amine with a low pKa into a protein scaffold enabled iminium mediated Friedel-Crafts catalysis (Figure 1.6.c) with modest enantioselectivity. More recently, Brouwer and coworkers⁵² expanded this approach using an aminotyrosine (aY) (Figure 1.6.d) residue, achieving improved catalytic efficiency. While these designs represent a significant conceptual advance, they rely on non-canonical amino acids that require specialized biosynthetic machinery for incorporation, limiting their accessibility and the ease with which they can be further optimized through conventional computational design. These limitations motivate the *de novo* design of artificial enzymes for the Friedel-Crafts alkylation of indoles built exclusively from canonical amino acids. The Friedel-Crafts alkylation of indoles therefore represents an ideal model reaction for evaluating emerging computational methodologies for the *de novo* design of enzymes catalyzing complex multistep C-C bond-forming reactions.

1.6 Thesis objectives

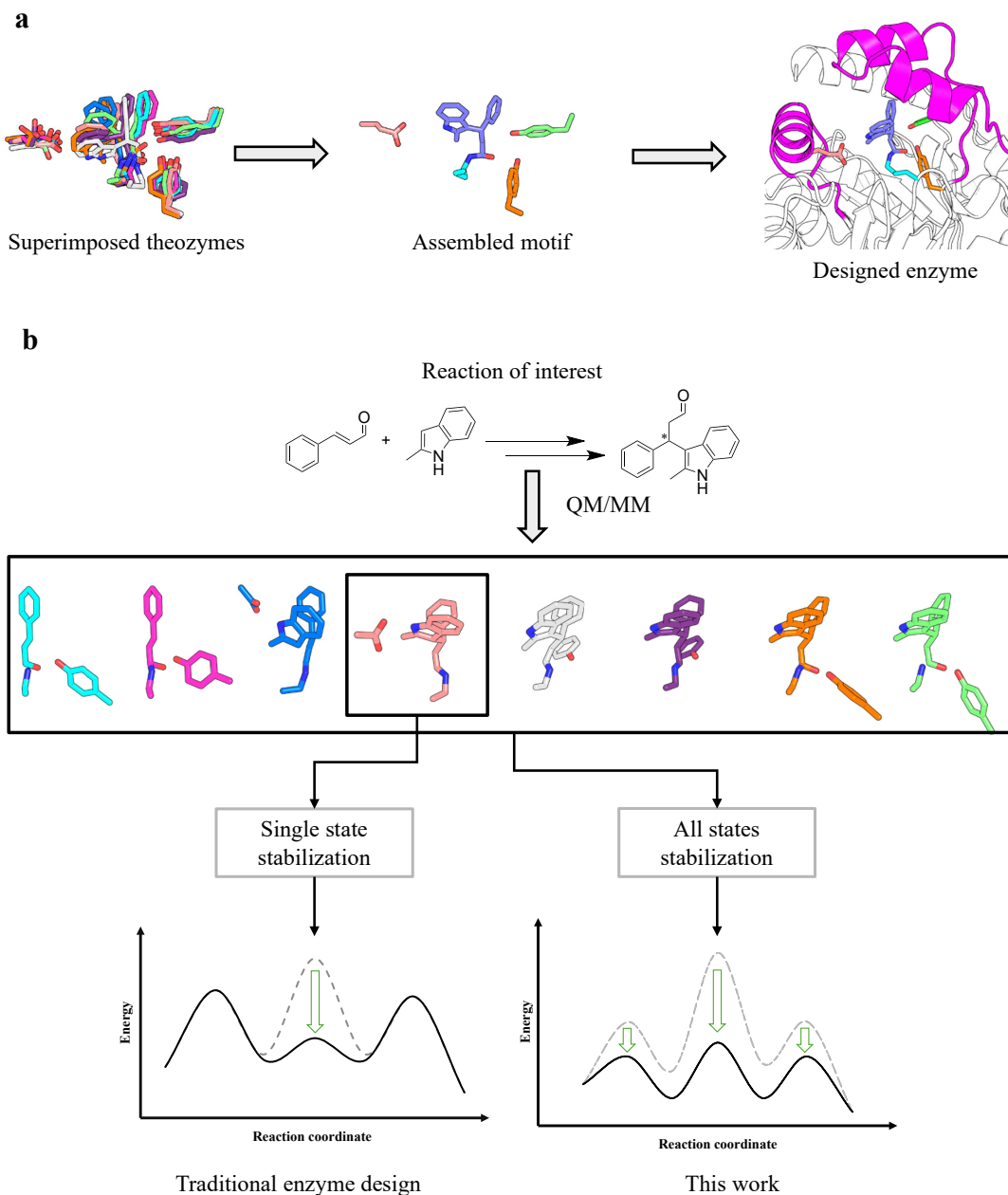


Figure 1.7. Overview of thesis objectives. Chapter 2 describes the development of the Multistate-CANVAS (MsC) computational framework for the design of enzymes for the Friedel-Crafts alkylation reaction. Chapter 3 presents the conclusions drawn from this work, along with perspectives on the limitations of the methodology and future directions for the design of *de novo* enzymes for complex multistep reactions.

Chapter 1 reviewed the principal computational methodologies currently available for *de novo* enzyme design, together with their major achievements. Collectively, these approaches have demonstrated that enzymes can be designed to catalyze a diverse range of chemical transformations. However, extending *de novo* enzyme design to complex multistep C-C bond-forming reactions remains a significant challenge owing to their evolving catalytic requirements and the structural demands they impose on enzyme active sites. To address this challenge, the primary objective of this thesis was to develop Multistate-CANVAS (MsC), a computational framework that integrates multistate computational enzyme design with hybrid scaffold generation for the *de novo* design of enzymes catalyzing complex multistep C-C bond-forming reactions. The workflow was designed to construct catalytic architectures by integrating catalytic information across multiple reaction states and to incorporate these architectures into suitable protein scaffolds through hybrid scaffold generation. The Friedel-Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde was selected as a proof-of-concept reaction to develop and evaluate this methodology.

Chapter 2 describes the development of the MsC workflow together with its application to the computational design and experimental characterization of *de novo* Friedel-Crafts alkylases. Finally, Chapter 3 discusses the strengths, limitations, and broader implications of the methodology for extending computational enzyme design to increasingly complex multistep C-C bond-forming reactions.

Chapter 2. Computational design and experimental validation of *de novo* enzymes for the Friedel-Crafts alkylation

**Jean Roch Kouassi,^{1,2} Helena Giramé Rizzo³, Marc Garcia-Borràs³ Anthony Green⁴
and Roberto A. Chica^{1,2}**

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

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

³Institute of Computational Chemistry and Catalysis (IQCC)
Faculty of Sciences, University of Girona

⁴Manchester Institute of Biotechnology, Department of Chemistry, The University of Manchester

2.1 Statement of Contribution

Theozyme calculation and mechanism proposition were performed by Helena Giramé Rizzo and Dr Marc Garcia-Borràs. Product standards were furnished by Prof Anthony Green. Computational design was done by Jean Roch Kouassi and Roberto A. Chica. All other experimental and computational experiments were completed by Jean Roch Kouassi. This chapter was written by Jean Roch Kouassi.

2.2 Acknowledgements

Jean Roch Kouassi was a member of the Advanced Protein Engineering Training, Internships, Courses, and Exhibition (APRENTICE) program. He was the recipient of the

Dr. Yu Scholarship and a Collaborative Research and Training Experience (CREATE) Program scholarship funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

2.3 Introduction

Computational enzyme design has enabled the creation of *de novo* enzymes for a diverse range of chemical transformations through methodologies including theozyme-based design, motif scaffolding, hybrid approaches such as CANVAS. Collectively, these approaches have demonstrated that efficient catalytic architectures can be generated computationally by enabling the design of enzymes for reactions including Kemp elimination,^{10,16} retro-aldol cleavage,^{8,17,53} ester hydrolysis,^{19,20,27} Diels–Alder cycloaddition¹⁸, Morita–Baylis–Hillman catalysis^{8,28}. Furthermore, previous work has demonstrated that explicitly considering multiple catalytic states during enzyme design can improve the design of enzymes for multistep reactions compared with conventional single-state approaches^{16,17}. However, extending these methodologies to complex multistep C-C bond-forming reactions remains a major challenge, as such reactions require catalytic architectures capable of accommodating evolving catalytic requirements throughout the reaction coordinate.

One such reaction is the Friedel-Crafts alkylation⁴³ of indoles with α,β -unsaturated aldehydes. This complex multistep C-C bond-forming reaction is widely employed in synthetic organic chemistry for the preparation of pharmaceutically relevant indole derivatives and has been extensively studied through organocatalysis using iminium activation.⁵⁴ Its multistep mechanism and evolving catalytic requirements make it an ideal model system for evaluating computational methodologies aimed at extending *de novo* enzyme design to this class of reactions.

Here, we develop Multistate-CANVAS (MsC), a computational framework for the *de novo* design of enzymes for the Friedel-Crafts alkylation of 2-methylindole with *trans*-

cinnamaldehyde. Starting from a proposed reaction mechanism, QM calculations were used to characterize the reaction transition states and derive the geometric and chemical constraints. These constraints were subsequently integrated into a unified catalytic motif and incorporated into protein scaffolds using CANVAS followed by multistate stabilization. Experimental characterization demonstrated that multiple MsC-designed variants could be successfully expressed, purified, and folded while displaying product formation superior to the no-enzyme reaction, with the most active variant achieving a 254-fold enhancement in conversion over the uncatalyzed reaction. Despite the low catalytic efficiency of the designed enzymes, the efficiencies achieved are comparable to those reported for first-round *de novo* designed enzymes targeting other C-C bond-forming reactions. Therefore, MsC may provide a framework for exploring the *de novo* design of enzymes for new-to-nature complex multistep reactions of industrial relevance.

2.4 Characterization of the catalytic mechanism

2.4.1 Mechanistic framework

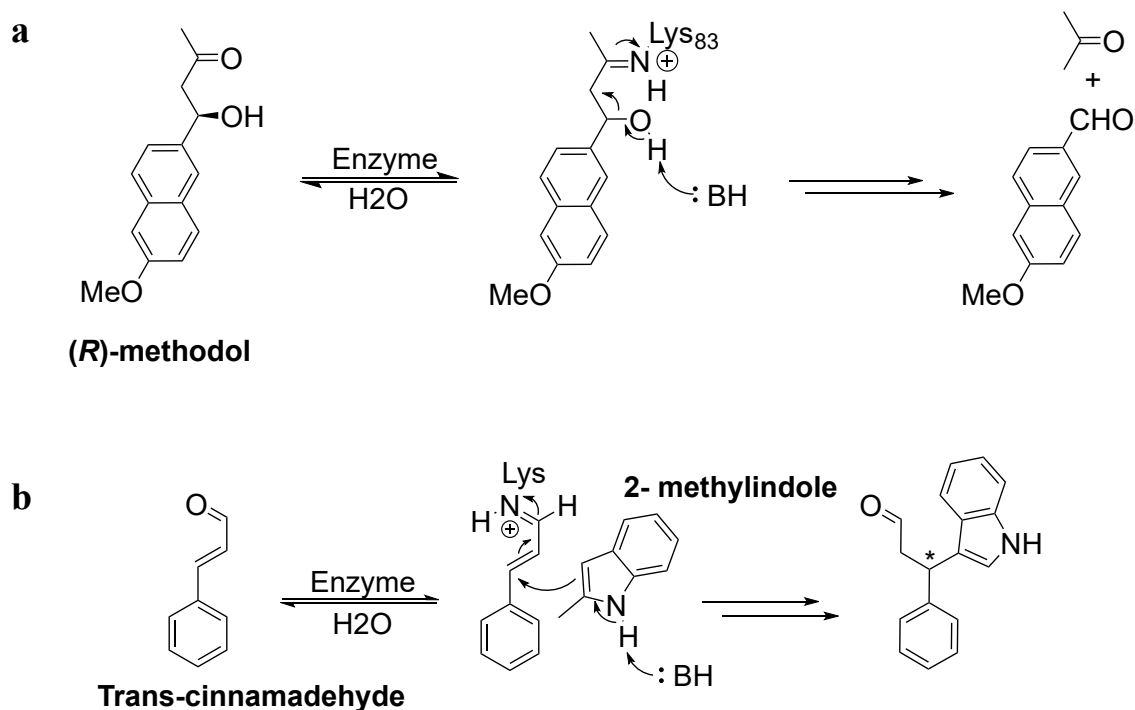


Figure 2.1. Iminium mediated catalysis of retro aldol (a) and Friedel-Crafts alkylation (b) reactions.

The Friedel-Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde to yield the C3-alkylated indole product (Figure 2.3b) was selected as the target reaction. The proposed catalytic mechanism, which serves as the working hypothesis underlying all subsequent design calculations, proceeds through nine transition states and is analogous to the aldol reaction, sharing the iminium-forming catalytic machinery established for the evolved retro-aldolase RA95.5-8. Iminium-based catalysis has also been demonstrated for Friedel-Crafts alkylation using LmrR_pAF⁵¹, an engineered enzyme incorporating a non-canonical catalytic amine, providing experimental support for this catalytic strategy. Here, we sought to extend this chemistry using only canonical amino acids.

On this basis, a catalytic pentad comprising Lys, Tyr1, Tyr2, His and Glu was proposed initially. Lysine was selected as the catalytic nucleophile responsible for Schiff base formation because it is the only canonical amino acid capable of forming a stable covalent iminium intermediate with an aldehyde under physiological conditions, particularly within the hydrophobic environment of a protein active site. Two tyrosine residues were introduced to mediate the proton transfers associated with carbinolamine formation and hydrolysis. This choice was motivated by directed evolution of RA95, where introduction of Tyr51 adjacent to the catalytic lysine replaced the water-mediated proton transfer network of the original design, demonstrating that precisely positioned tyrosine residues can efficiently mediate acid–base chemistry during iminium catalysis. A His residue was initially assigned to catalyze the dehydration of the carbinolamine intermediate; however, because this step was hypothesized to proceed via general acid catalysis or through solvent molecules present within the active site, it was not considered for further design. Finally, glutamate was selected as the general base responsible for deprotonating the indole nitrogen, thereby promoting nucleophilic attack on the β -carbon of the activated iminium intermediate.

2.4.2 Theozyme construction and QM validation

DFT calculations were carried out using the Gaussian 16C.01 software package to characterize each step of the proposed mechanism. All geometry optimizations and single-point energy calculations employed the B3LYP-D3(BJ) functional with the 6-31G(d) basis set. Solvent effects were modeled using the Polarizable Continuum Model (PCM) with a dielectric constant of 4.33 to approximate the low-polarity environment of an enzyme

active site. Transition states were located following a procedure consistent with transition state theory:

1. **Initial transition state guess:** An initial transition state geometry was proposed for each step of the reaction, drawing on the established aldol/retro-aldol mechanism as a guide for the iminium-forming and decomposition steps.
2. **Transition state optimization:** Each proposed geometry was refined by traversing the potential energy surface through iterative calculation of energy gradients and the Hessian matrix. A structure was accepted as a true transition state when the energy gradient converged to zero and the Hessian matrix displayed exactly one imaginary frequency, confirming the presence of a first-order saddle point.
3. **Intrinsic Reaction Coordinate (IRC) calculation:** IRC calculations were performed from each optimized transition state to verify connectivity with the intended reactant and product intermediates on either side of the energy barrier.
4. **Intermediate geometry optimization:** The geometries of all reactant and product intermediates flanking each transition state were independently optimized to their minimum energy configurations, enabling calculation of activation energies and reaction free energies for each step.

For each transition state, a truncated active site model was constructed comprising only the catalytic residues directly intervening at that step, in addition to Lys, which was included in all models as it remains covalently bound to the substrate throughout the catalytic cycle. The truncated residue models were capped with hydrogen atoms at the sites of truncation, and their protonation states were assigned according to the proposed mechanism at each

stage of the catalytic cycle. All QM clusters were constructed with an overall charge of 0, while local charge distributions varied according to the protonation states of the catalytic groups. Accordingly, TS1 and TS2 were modeled with Lys and Tyr1; TS3 was modeled with Lys and His but TS3 ultimately excluded from the design; TS4 and TS5 were modeled with Lys and Glu; TS6 and TS7 were modeled with Lys and Tyr2 and TS8 and TS9 with Lys and Tyr1.

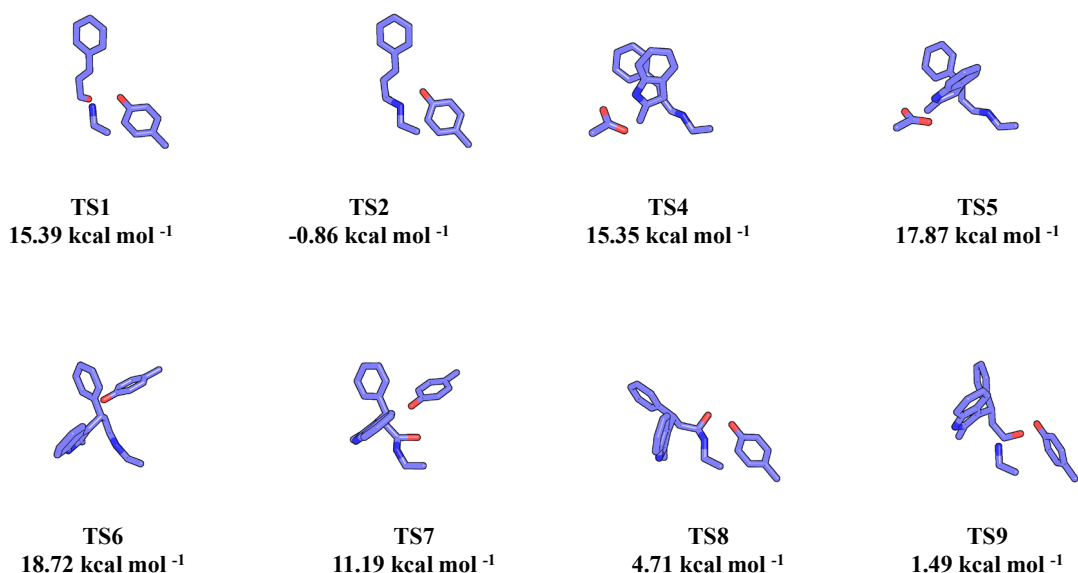


Figure 2.2. QM calculated theozymes for the Friedel-Crafts alkylation between 2-methylindole and *trans*-cinnamaldehyde.

The mechanistic proposal outlined above should be regarded as a working hypothesis rather than an established fact, and all design calculations rest on the assumption that catalysis proceeds through this pathway. As with any computationally derived mechanism, a number of caveats apply. The quantum mechanical calculations were conducted on simplified

representations of the active site, limited to the catalytic residues directly participating in each step, and therefore cannot fully reproduce the complexity of a real enzyme active site including backbone flexibility, explicit solvent, and long-range electrostatic contributions from the surrounding protein. These simplifications may introduce systematic errors in the computed energy barriers, particularly for steps involving charged or polar intermediates where accurate treatment of solvation is critical. It is also possible that the true mechanism differs from the one proposed, either through alternative pathways or through the involvement of catalytic factors not captured by the model. Notwithstanding these limitations, the proposed mechanism provides the necessary framework for theozyme construction and multistate transition state stabilization, and each step is characterized in detail in the following sections.

2.4.3 Stepwise mechanism

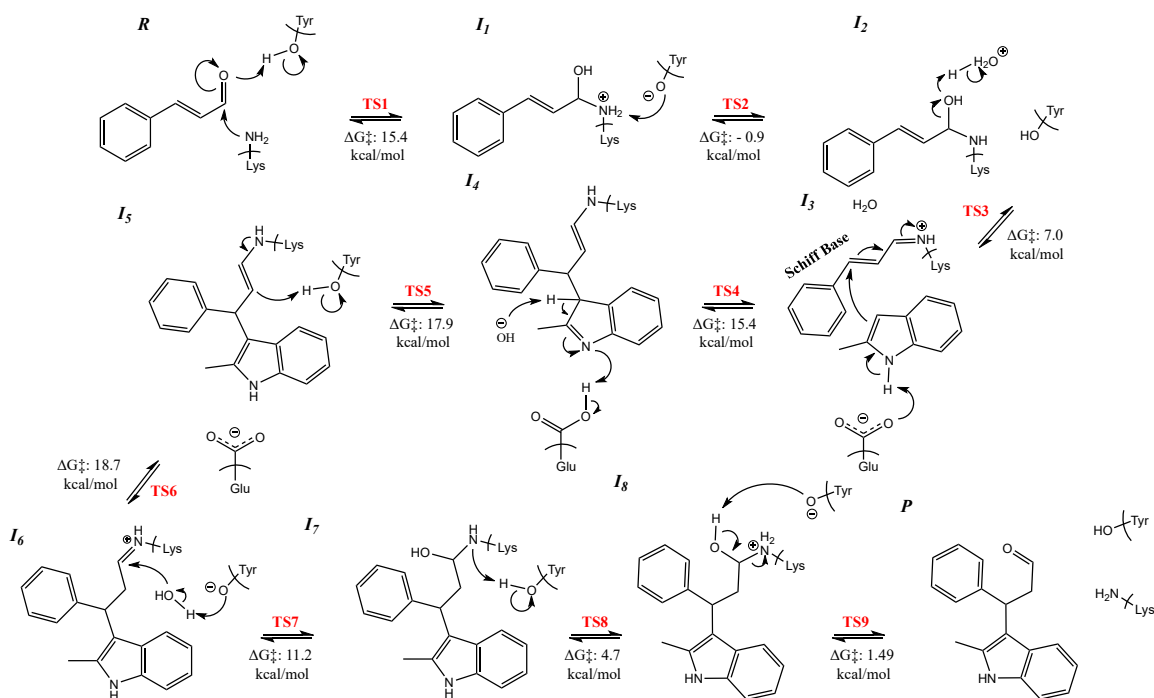


Figure 2.3. Proposed reaction mechanism for the Friedel-Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde.

The first step of the mechanism ($R \rightarrow TS1 \rightarrow I1$) involves the nucleophilic attack of the catalytic lysine on the aldehyde carbonyl of *trans*-cinnamaldehyde. Concurrently, the first tyrosine acts as a general acid, donating its hydroxyl proton to the developing alkoxide, yielding a protonated carbinolamine intermediate ($I1$) in which the first tyrosine is left in its deprotonated tyrosinate form. In the second step ($I1 \rightarrow TS2 \rightarrow I2$), the tyrosinate acts as a general base, deprotonating the positively charged nitrogen of the carbinolamine, regenerating the neutral tyrosine and yielding the neutral carbinolamine intermediate ($I2$). The carbinolamine-forming process represented by $TS1$ and $TS2$ was deconvoluted into two discrete steps to separately capture the catalytic interactions associated with nucleophilic attack and proton transfer. In related computationally characterized mechanisms,

including the evolved retro-aldolase^{55,56} and Michaelase⁵⁷ systems, the corresponding carbinolamine-forming chemistry is represented as a single overall step. Consequently, treating the two transformations independently results in TS2 lying lower in calculated free energy than I1, giving rise to the formally negative activation free energy reported for this step. This value should therefore not be interpreted as a physical negative activation barrier, but rather as a consequence of the deconvolution of the overall process. The third step (I2 $\rightarrow$ TS3 $\rightarrow$ I3) involves dehydration of the carbinolamine, in which the hydroxyl group departs as water to form the Schiff base, which tautomerizes to the reactive iminium ion (I3), activating the β -carbon toward nucleophilic attack. In the fourth step (I3 $\rightarrow$ TS4 $\rightarrow$ I4), the glutamate acts as a general base, deprotonating the indole nitrogen, which increases the electron density of the five-membered ring and drives nucleophilic attack at C3 onto the iminium β -carbon, forming the new C-C bond. This step constitutes the key Friedel-Crafts bond-forming event, upon which the C3 carbon of the indole transitions from sp² to sp³ hybridization, disrupting aromaticity and yielding a nonaromatic intermediate (I4). The fifth step (I4 $\rightarrow$ TS5 $\rightarrow$ I5) involves rearomatization of the indole ring. The π system of the newly formed C-C bond abstracts a proton from the protonated glutamate, returning the C3 carbon to sp² hybridization and restoring the aromatic character of the indole ring (I5). In the sixth step (I5 $\rightarrow$ TS6 $\rightarrow$ I6), the second tyrosine acts as a general acid and protonates the carbon of the enamine double bond. The enamine π -electrons shift toward the lysine-bound nitrogen, converting the enamine into the corresponding iminium intermediate I6 and leaving the second tyrosine in its deprotonated tyrosinate form. The seventh step (I6 $\rightarrow$ TS7 $\rightarrow$ I7), involves a

proton abstraction by the second tyrosinate from a water molecule to generate an hydroxyle. The hydroxyle acts as a nucleophile and simultaneously attacks the electrophilic iminium carbon, converting the planar iminium into the tetrahedral product-side carbinolamine intermediate I7. The second tyrosine is reprotonated during this step. This transformation corresponds to the reverse of the carbinolamine-forming chemistry represented by TS1. In the eighth step (I7 $\rightarrow$ TS8 $\rightarrow$ I8), the neutral first tyrosine acts as a general acid, donating its phenolic proton to the carbinolamine nitrogen. This proton transfer converts the tyrosine into its deprotonated tyrosinate form and generates the protonated carbinolamine intermediate (I8). In the ninth and final step (I8 $\rightarrow$ TS9 $\rightarrow$ P), the first tyrosinate abstracts the proton from the hydroxyl group of the protonated carbinolamine. Deprotonation of the hydroxyl promotes reformation of the aldehyde carbonyl, which in turn drives C–N bond cleavage, releasing the aldehyde product and the free catalytic lysine, thereby completing the catalytic cycle. As the QM calculations were performed targeting the (*R*) enantiomer of the product, and the design targeted the (*S*) enantiomer, the theozymes corresponding to TS4 through TS9 were inverted in PyMOL⁵⁸ prior to subsequent design calculations. The catalytic motif assembly therefore proceeded using the theozymes corresponding to TS1 through 9, excluding TS3.

2.5 Multistate CANVAS (MsC)

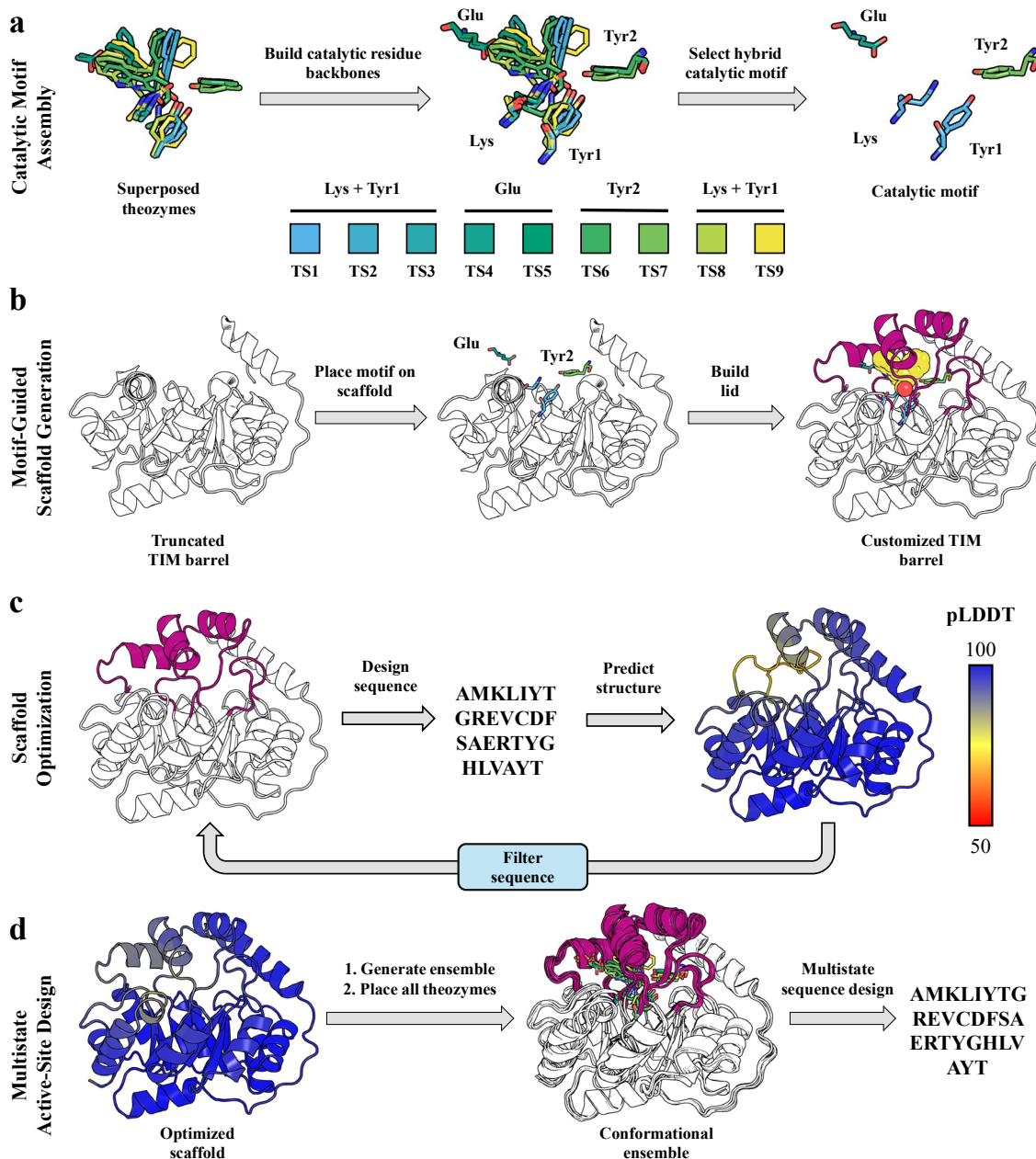


Figure 2.4. Multistate CANVAS workflow for *de novo* enzyme design. (a) Multiple transition-state theozymes are structurally aligned to construct a unified catalytic motif. (b) The catalytic motif is incorporated into a truncated TIM-barrel scaffold preceding *de novo* scaffold generation. (c) The resulting scaffold is optimized through sequence design and structure prediction to obtain a stable protein architecture. (d) A backbone ensemble is

generated from the optimized scaffold, enabling explicit multistate active-site sequence design by simultaneously optimizing the active site across all catalytic states.

2.5.1 Catalytic motif assembly

To translate the theozyme set into a realizable active-site architecture, a consensus catalytic motif was derived by integrating the different transition state geometries into a unified spatial framework. The theozymes corresponding to TS1, TS2, and TS4 through TS9 were superimposed in PyMOL⁵⁹ onto a common reference frame, (Figure 2.4.a). This alignment enabled identification of a shared geometric arrangement of the four catalytic residues across all mechanistic states. As defined by the proposed mechanism, Tyr1 participates in TS1, TS2, TS8, and TS9; Glu in TS4 and TS5; Tyr2 in TS6 and TS7; and Lys in all steps of the catalytic cycle. For each catalytic residue, backbone positions compatible with the required catalytic geometries were sampled using Invrotzyme⁶⁰, a method that generates inverted rotamer solutions by sampling backbone conformations around a fixed ligand while preserving catalytic geometry constraints.⁶¹ Because the structural alignment confirmed that a single backbone solution could, in principle, satisfy all transition states involving a given residue, Invrotzyme was applied once per catalytic residue using a representative transition state, yielding a residue-specific ensemble of backbone solutions. From these ensembles, incompatible solutions were filtered based on steric feasibility when all four catalytic residues were considered simultaneously. Specifically, backbone configurations that introduced main-chain steric clashes upon assembly of the full catalytic motif were discarded. Only solutions that remained mutually compatible across all four residues were retained. From the surviving set, a single hybrid catalytic motif was selected in PyMOL as the configuration in which all four catalytic backbone segments could be simultaneously accommodated without steric clashes or geometric distortion of the

required catalytic contacts. Following the initial geometric filtering, more than 10 catalytic assemblies remained. The catalytic motif was assembled by selecting Lys and Tyr1 from TS1, Glu from TS4, and Tyr2 from TS6. The selection was guided by the position of each residue along the proposed reaction pathway. Specifically, the earliest occurrence of each catalytic residue was prioritized when multiple states contained the same residue type. This was intended to ensure that the resulting hybrid motif retained the catalytic functionality required from the beginning of the reaction sequence. These surviving assemblies were subsequently evaluated for compatibility with the selected TIM-barrel scaffold. Rather than selecting an assembly solely on the basis of the catalytic geometry, the final assembly was chosen based on its ability to be connected to the existing scaffold backbone with minimal backbone remodeling. Preference was therefore given to the assembly that could be incorporated while preserving the existing backbone geometry and fold, without requiring substantial backbone restructuring. This approach was intended to minimize perturbation of the pre-existing scaffold while maintaining the required catalytic architecture. The selected motif was then used as a fixed geometric constraint for subsequent scaffold generation, as described in the following section.

2.5.2 Scaffold generation

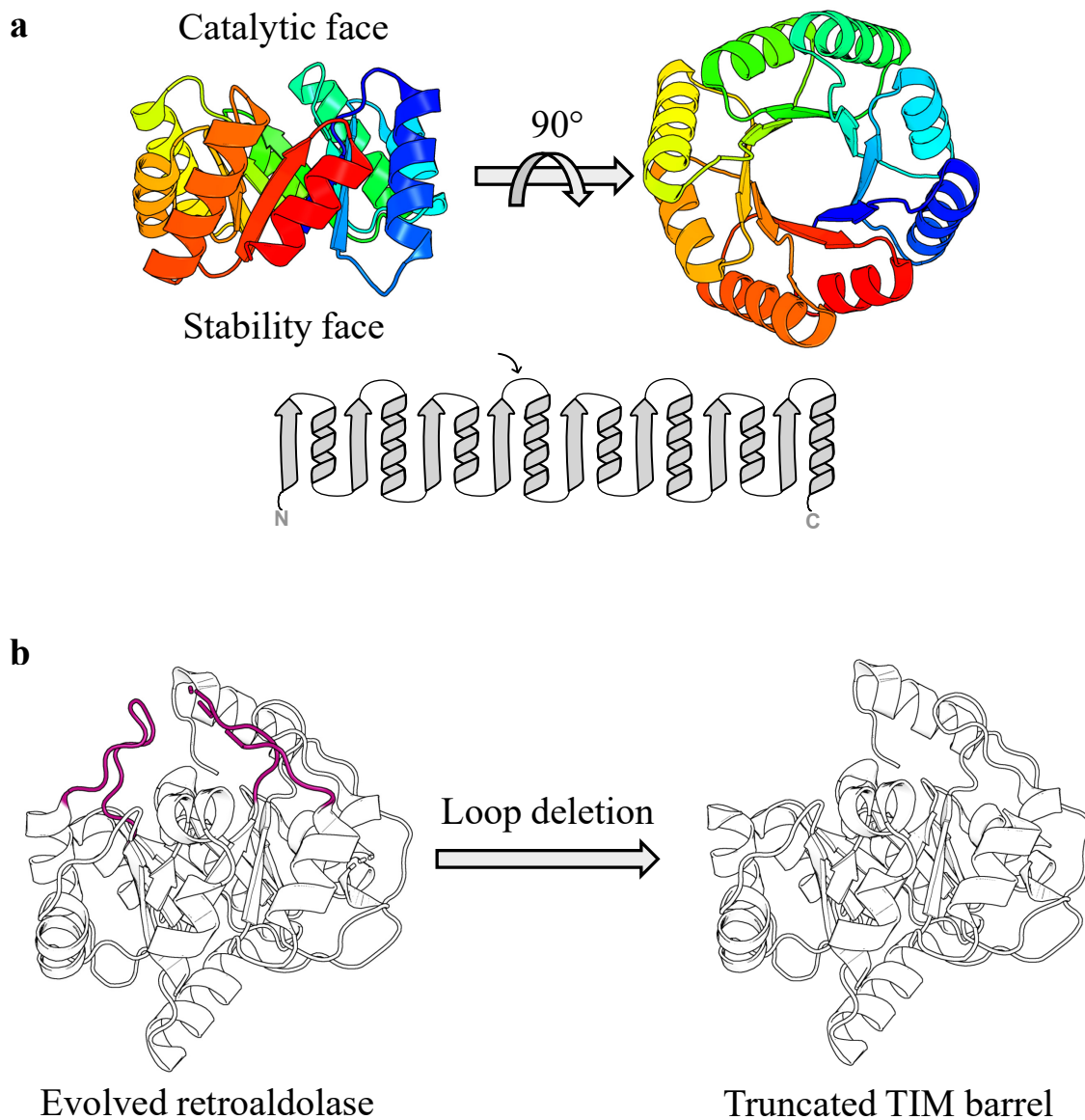


Figure 2.5. Preparation of a TIM-barrel scaffold for catalytic motif incorporation. (a) The TIM barrel architecture comprises a stability face and eight converging β -strands forming catalytic face suitable for active-site construction. The dark arrow indicates the position where residues can be inserted **(b)** Truncation of the evolved retro aldolase RA95.5-8 (PDB:5AN7), a TIM barrel, for scaffold design.

TIM barrel scaffolds are particularly well suited for enzyme design and have served as scaffolds for numerous successful protein engineering and computational design efforts.

As one of the most abundant enzyme folds in nature, the TIM barrel supports a remarkably diverse range of catalytic functions while maintaining a conserved structural architecture.^{61–63} Its eight C-terminal loops converge to form a well-defined catalytic face that facilitates precise positioning of active-site residues, while the surrounding hydrophobic core provides a stable environment capable of accommodating catalytic residues and modulating their pKa values (Figure 2.5.a). Notably, the evolved retro-aldolase, RA95.5-8 which employs lysine-mediated iminium catalysis analogous to that proposed here, is itself based on a TIM barrel scaffold. For these reasons, a TIM barrel scaffold was selected for *de novo* scaffold generation, with RA95.5-8 used as a representative example of this fold. As both the active site and surface were extensively redesigned in the subsequent steps, the initial scaffold primarily provided the TIM-barrel framework upon which the catalytic architecture was constructed. Prior to scaffold generation, loops 1 and 6 were deleted, creating a truncated TIM barrel, to eliminate steric clashes with the catalytic motif (Figure 2.5.b) . To identify a suitable position for the catalytic lysine within the scaffold, Lys was placed at each position on the catalytic face of the TIM barrel, and for each placement, the ability of the remaining catalytic residues (Tyr1, Tyr2, and Glu) to be accommodated by nearby backbone positions was evaluated. A position was selected where the catalytic lysine could be placed and the surrounding backbone positions were sufficiently close to accommodate the remaining catalytic residues of the motif. With the lysine position established, Fold conditioning, a method within RFDiffusion⁶⁴ that enables local backbone remodeling while preserving the overall protein fold, was applied to remodel the existing TIM barrel backbone to connect to the backbone fragments of the selected catalytic residues. For backbone fragments that could

not be connected through fold conditioning alone, RFDiffusion was used to generate *de novo* connecting secondary structure elements. Finally, a lid was built *de novo* on top of the scaffold using RFDiffusion, as surface loops and lids closing over the active site have been shown to play important roles in catalysis by shielding the active site from bulk solvent and improving substrate positioning.⁶⁵ The resulting scaffold was then iteratively optimized through cycles of sequence design and structure prediction. In each cycle, sequences were designed using LigandMPNN, a deep learning-based inverse folding model conditioned on both the backbone geometry and the coordinates of the transition state and catalytic residues to ensure compatibility with the active site environment. The predicted structures were then assessed using ESMFold, and the backbone RMSD between the predicted structure and the model scaffold, as well as the per-residue backbone pLDDT confidence score, were used as filtering criteria. This iterative process was repeated until convergence between the predicted structure and the model scaffold was achieved. Convergence was defined as an average pLDDT score >90 and a backbone RMSD <1 Å between the predicted structure and the design model. The resulting structure was used as the final optimized scaffold for multistate sequence design (Figure 2.4.c).

2.5.3 Theozyme placement

With the optimized scaffold in hand, theozyme placement was performed to position the catalytic residues within the active site in geometries compatible with all eight transition states. To capture the inherent backbone flexibility of enzyme active sites, a backbone ensemble of 120 members was generated by applying Backrub^{66,67} to the optimized scaffold. Prior to theozyme placement, each ensemble member was energy-minimized using TRIAD⁶⁸ to correct geometric imprecision introduced by structure prediction models,

as small errors in backbone coordinates that could prevent successful catalytic geometries during placement. Theozyme placement was then performed using TRIAD, placing all eight theozymes simultaneously onto each member of the backbone ensemble under the required catalytic geometric constraints. Only solutions satisfying all catalytic constraints and displaying negative interaction energies were retained. To ensure that the transition state placements were spatially proximal within the active site, solutions were further filtered by computing pairwise distances between the geometric centers of the phenyl ring across all theozyme placements. The phenyl ring was used as the reference moiety as it is the structural element that remains geometrically invariant across all transition states. Solutions exhibiting the lowest pairwise distances were prioritized and carried forward as input for multistate sequence design, as described in the following section

2.5.4 Multistate sequence design

Sequence design encompassed two complementary steps, beginning with multistate design of the active-site residues. MenD optimizes a single sequence across multiple transition states simultaneously by evaluating a combined energy function that assesses overall sequence fitness across all chemical states. Because the geometric and chemical requirements of the active site vary between states, optimizing the sequence toward a single transition state could favor interactions specific to that state at the expense of those required by other states. For each of the eight transition states, two structural representations were included in the multistate optimization: the transition-state-containing structure and a corresponding structure from which the transition-state model was removed, resulting in a total of 16 states. The transition-state-free structures were included to favor sequences that remained energetically favorable independently of interactions with the transition-state

models. A sum fitness function was used as the multistate objective, whereby the energies of all 16 states were summed to obtain a combined fitness score for each candidate sequence that was optimized using a Monte Carlo algorithm. Following multistate sequence design, a series of filters were applied to identify the most promising designs for experimental characterization. The metrics were selected to evaluate specific properties considered desirable for the designed enzymes, including favorable accommodation of each transition state, energetic balance across states, ligand enclosure, catalytic lysine nucleophilicity, and recovery of the designed structure and ligand geometries. First, only sequences displaying negative energies across all eight transition states individually were retained, ensuring that each transition state was favorably accommodated according to the design energy function. Second, the energy range across transition states (ΔE) was computed as the difference between the highest and lowest per-state energies, and only sequences displaying a ΔE (< 400) were retained, ensuring that no individual transition state was disproportionately destabilized relative to the others. Third, the degree of ligand enclosure within the active site was assessed by computing the solvent-accessible surface area (SASA) of each transition state ligand. Following its previous use as a design-selection metric⁵⁷, a minimum ligand burial of 75% was used to favor designs in which the ligand was substantially enclosed within the active site. Fourth, sequences were filtered based on the predicted pKa of the catalytic lysine. As a neutral lysine is required for nucleophilic attack and subsequent iminium formation, designs with lower predicted pKa values were favored. A cutoff of pKa < 9 was empirically selected based on the distribution of predicted values to enrich the candidate pool for designs with comparatively lower catalytic lysine pKa values. Application of these filters reduced the candidate pool from 500 to 35 designs,

which were subsequently carried forward for shell-residue redesign using LigandMPNN while keeping the active-site residues fixed. This shell sequence redesign strategy has been shown to reliably produce soluble proteins ⁸, and 100 sequences were generated per design using this approach, resulting in a total of 3,500 sequences. Final filtering was performed using Boltz-2⁶⁹, a structure prediction model capable of co-predicting protein structure in complex with a ligand. Each design was independently predicted in complex with both the substrate and the product, and the resulting structures were evaluated based on backbone RMSD, substrate RMSD, and cat RMSD with respect to the designed model, as well as average pLDDT confidence score. Backbone RMSD was used to assess recovery of the designed protein architecture, while substrate and product RMSDs were used to assess recovery of the corresponding designed ligand poses. Average pLDDT was used to assess confidence in the predicted structures. Designs displaying lower RMSDs and higher pLDDT scores were therefore favored during the final selection (Supplementary table 6). The numerical cutoffs applied throughout the filtering process were empirically selected to prioritize designs exhibiting the desired properties and progressively reduce the candidate pool. These values were not considered strict physical thresholds known to determine catalytic activity, nor established optimal cutoffs for the corresponding properties. This process narrowed the 3500-candidate pool down to 11 designs, which were selected for experimental characterization.

Table 1. Multistate active-site design filtering metrics.

Metric	Maximum
Active site energy ^a	0
ΔE ^b	500
Predicted Lysine pKa	9
SASA ^c	25 %

a Energies are calculated by subtracting the energy of an all-Gly structure.

b ΔE is the difference between the highest and lowest transition-state stabilization energies for a given sequence.

c Solvent accessible surface area of the transition states in the Triad models.

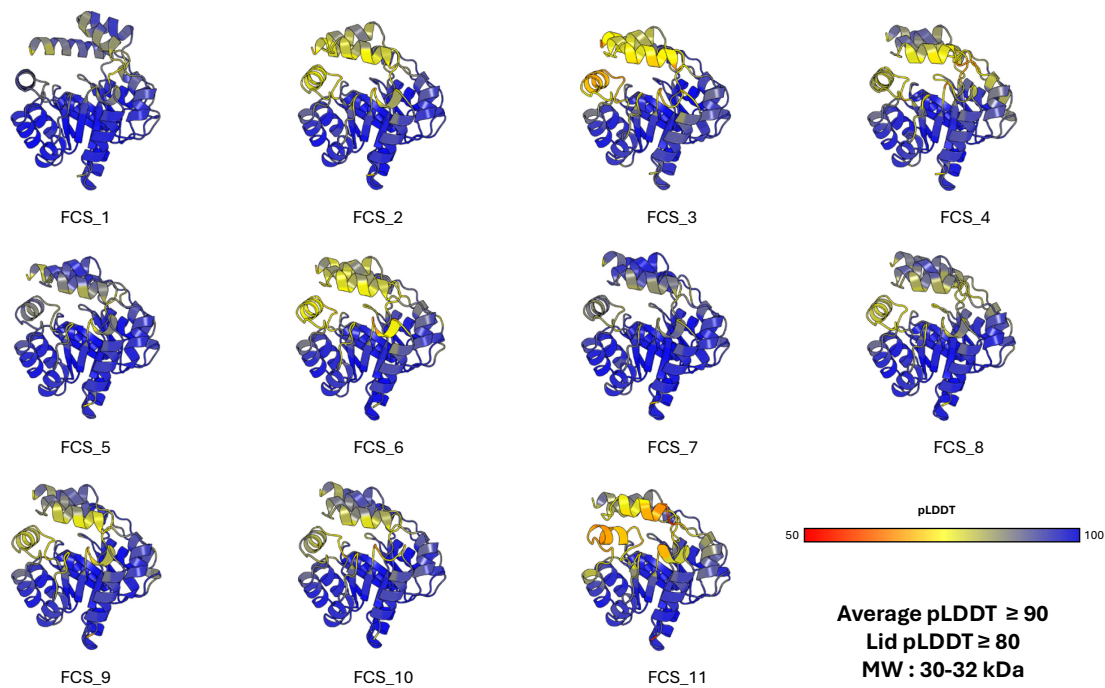


Figure 2.6. Boltz-2 models of selected variants. Structures of variants are coloured by pLDDT.

2.6 Results

2.6.1 Structural characterization of designed enzymes

2.6.1.1 SDS-PAGE

Of the eleven designs selected for experimental characterization, nine were successfully expressed in *E. coli* and purified by immobilized metal affinity chromatography (IMAC). SDS-PAGE analysis on a stain-free gel imaged with ImageLab confirmed successful purification of nine variants, with primary bands migrating at approximately 25–30 kDa, consistently below the expected molecular weight of at least 30 kDa for the designed enzymes (Figure 2.7). This anomalous migration behavior is consistent with either partial unfolding under denaturing SDS-PAGE conditions or a high density of charged surface residues, both of which are known to alter the SDS-binding stoichiometry and affect migration behaviour.^{70,71} While densitometric analysis of band intensities revealed a heterogeneity in purity across the variant set, the majority of variants were obtained at high purity. Four variants: FCS_1, FCS_5, FCS_6, and FCS_7 displayed single dominant bands with no discernible contaminant species, yielding estimated purities exceeding 95% and rendering them immediately suitable for downstream biophysical characterization. FCS_2 and FCS_4 were estimated at approximately 90% purity, each exhibiting a faint contaminant band at ~ 65–70 kDa most likely corresponding to co-purifying *E. coli* proteins, present at levels unlikely to significantly interfere with downstream characterization. FCS_9 yielded an estimated purity of approximately 85%, with a moderately reduced main band intensity and faint higher-molecular-weight smearing, but no discrete secondary bands. FCS_10 exhibited more substantial contamination at an estimated purity of approximately 75%, with two visible contaminant bands at approximately 40 and 65 kDa alongside a reduced main band intensity indicative of lower

overall yield. Notably, FCS_8 displayed an anomalous migration pattern distinct from the rest of the set, with its primary band running at an apparent molecular weight exceeding that of the other variants and accompanied by substantial lane smearing, yielding the lowest estimated purity of approximately 70%. Given these observations, all nine purified variants were carried forward for circular dichroism spectroscopy and thermal denaturation. Seven variants: FCS_1, FCS_2, FCS_4, FCS_5, FCS_6, FCS_7, and FCS_9, were characterized directly, with the acknowledgment that minor contaminant contributions in variants estimated below 95% purity may marginally affect spectral baselines but are unlikely to preclude meaningful secondary structure or thermostability assessment. FCS_8 and FCS_10 were subjected to an additional purification step by fast protein liquid chromatography (FPLC) prior to biophysical characterization, given the presence of discrete contaminant bands and, in the case of FCS_8, substantial lane smearing suggestive of aggregation or degradation. With all nine variants in hand at sufficient purity, far-UV circular dichroism (CD) spectroscopy was performed to assess their secondary-structure characteristics and determine whether the purified proteins adopted folded conformations.

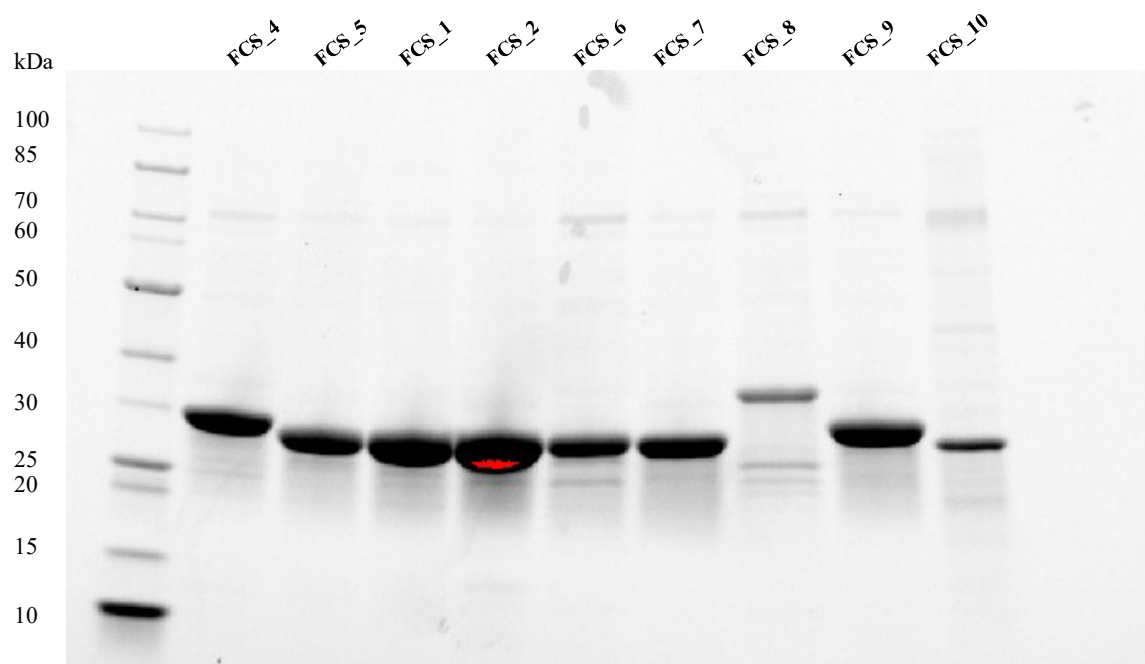


Figure 2.7. SDS-PAGE analysis of purified enzyme variants. Purified enzymes were analyzed by SDS-PAGE following Ni-NTA affinity chromatography. Samples were resolved on 4–15% precast polyacrylamide gels (BioRad) alongside an unstained broad-range protein standard (10–200 kDa, ThermoFisher), with 8 μ L of each purified protein loaded per lane. The average molecular weight of variants is between 30 and 32 kDa.

2.6.1.2 Circular dichroism and thermal denaturation assays

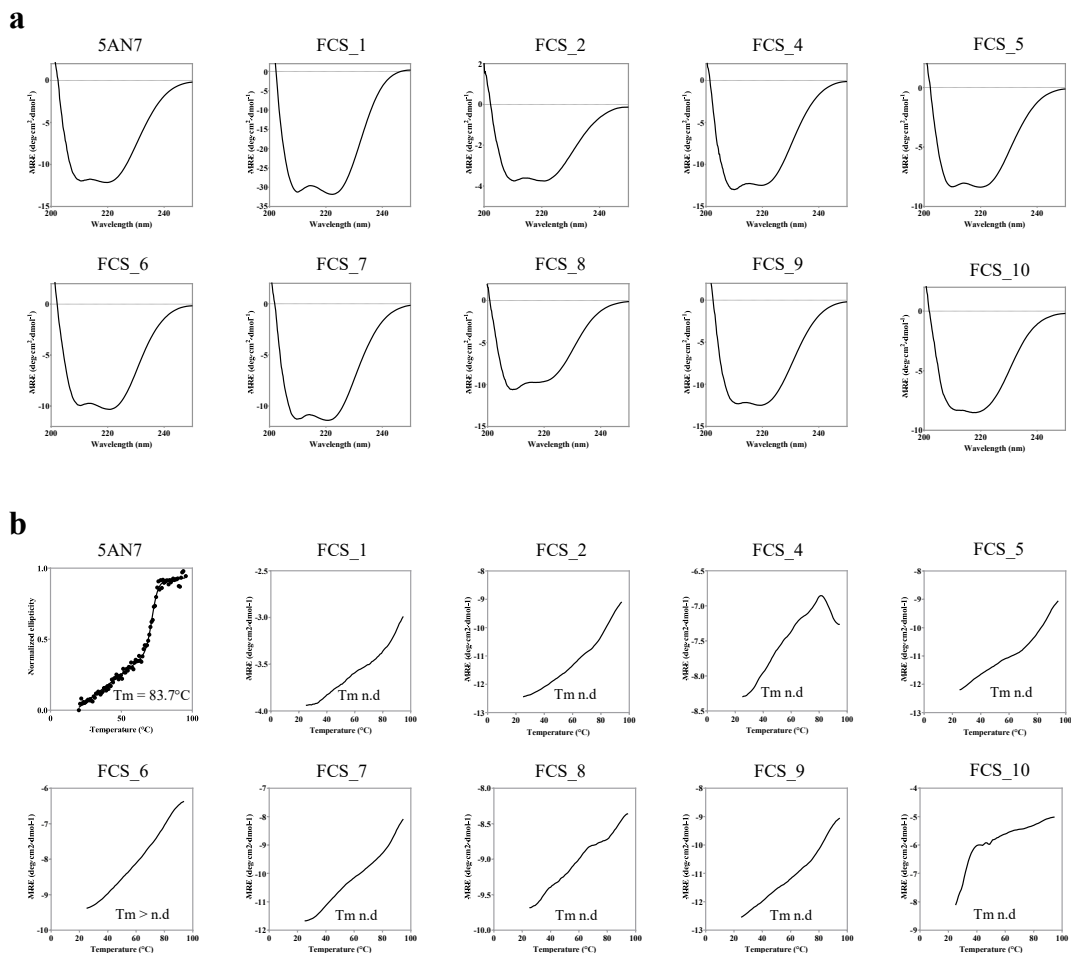


Figure 2.8. Structural characterization of designed FCS variants. (a) CD spectra (b) Thermal denaturation curves.

Far-UV CD spectra were recorded for all nine purified variants and compared with the parental TIM barrel scaffold 5AN7 as a structural reference (Figure 2.8). The variants displayed spectra consistent with the presence of substantial secondary structure, although differences in both spectral shape and intensity were observed across the set and relative to 5AN7. Several variants displayed positive ellipticity around 190 nm and a double minimum in the regions of 208 and 222 nm, consistent with the mixed α/β secondary-structure composition expected for a TIM-barrel protein. In contrast, FCS_8 exhibited a markedly

different spectral profile, displaying a broad, shallow minimum rather than the characteristic double minima indicating differences in its secondary-structure characteristics relative to the other variants and 5AN7. Collectively, the CD spectra indicate that the majority of the purified variants adopted folded, secondary-structure-rich conformations. Collectively, the CD spectra indicate that the majority of the purified variants adopted folded, secondary-structure-rich conformations with spectral characteristics consistent with the α/β composition expected for the intended TIM-barrel fold. Despite the differences in spectral shape and intensity observed across the set, most variants retained the prominent secondary-structure features expected from the designed proteins.

Thermal stability was assessed by monitoring the CD signal at 222 nm as a function of temperature between 25 and 95 °C, again using 5AN7 as a reference. As expected, 5AN7 displayed a sharp cooperative unfolding transition with a melting temperature (T_m) of 83.7 °C, consistent with its well-documented thermostability. In contrast, the designed variants exhibited gradual changes in ellipticity that did not reach a well-defined unfolding plateau within the experimental temperature range, precluding reliable determination of melting temperatures. The thermal denaturation profiles were therefore evaluated using the non-normalized CD signal, and no T_m values were assigned to the designed variants. For most variants, the change in ellipticity between 25 and 95 °C remained relatively small, and substantial ellipticity at 222 nm was retained at the highest temperature examined rather than approaching the low-magnitude signal expected upon extensive loss of ordered secondary structure. This behavior suggests that substantial secondary structure persisted throughout the temperature range examined and could be consistent with several possible

thermal behaviors, including high thermal stability with unfolding extending beyond the accessible temperature range, an interpretation consistent with the high thermostability previously reported for proteins generated using ProteinMPNN-based sequence design approaches, non-cooperative unfolding, or structural transitions that retain substantial secondary structure at elevated temperatures. Collectively, the thermal CD profiles indicate thermal behavior distinct from the cooperative unfolding transition observed for 5AN7, with considerable secondary structure retained up to 95 °C.

Taken together, the CD analyses suggest that the majority of the MsC-designed variants are folded and display substantial secondary structure consistent with the expected α/β composition of the designed proteins. Although their thermal denaturation profiles differed from that of the parental scaffold, substantial secondary structure was retained across the temperature range examined. These results provide a biophysical characterization of the structural and thermal properties of the designed variants prior to subsequent functional characterization.

2.6.2 Functional characterization of the designed enzymes

2.6.2.1 Activity screening

Catalytic activity of the nine purified variants was assessed by incubating each purified enzyme with 2-methylindole and *trans*-cinnamaldehyde, and product formation was quantified by comparison to an authentic standard. In the absence of enzyme, background conversion was negligible at 0.11%. All nine variants displayed detectable conversion well above this background, with conversions ranging from 4% to 28%, representing a 36 to 254-fold enhancement over the uncatalyzed reaction. A clear hierarchy of activity was observed across the variant set, with FCS_1 emerging as the most active variant at

approximately 28% conversion, followed by FCS_9 at approximately 12%, and the remaining variants in the 4–9% range. While the observed enhancement establishes that the designed proteins promote product formation relative to the no-enzyme condition, additional controls would be required to determine the extent to which the observed activity originates specifically from the designed active sites. Given its superior conversion, FCS_1 was selected as the lead variant for further characterization, and the pH dependence of its activity was subsequently investigated to probe consistency with the proposed iminium-based mechanism.

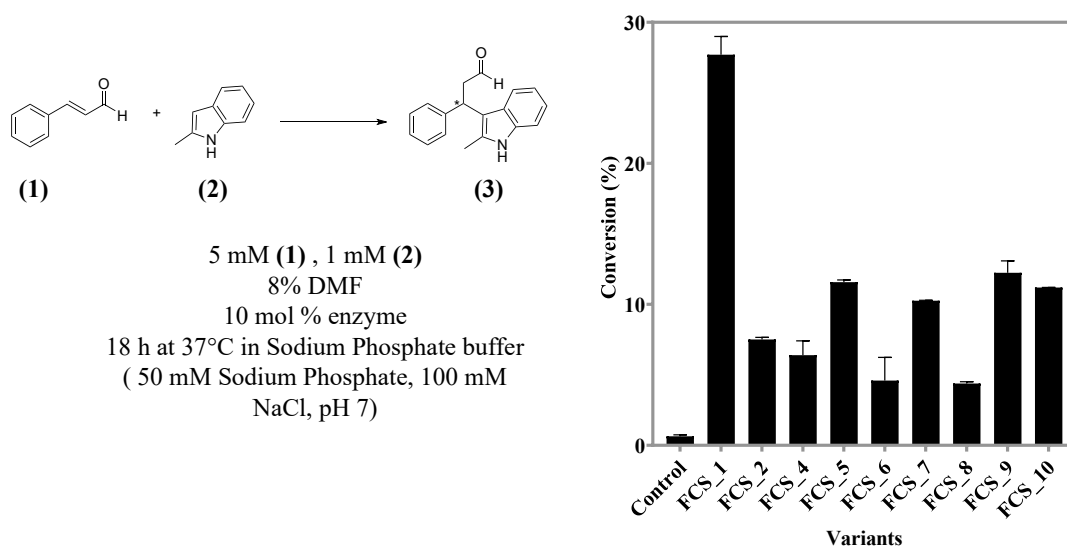


Figure 2.9. Conversion assay. Conversion was monitored for all soluble variants. Reactions were performed with 5 mM *trans*-cinnamaldehyde and 2 mM 2-methylindole and 10 mol % catalyst in sodium phosphate buffer (100 mM sodium phosphate, 50 mM NaCl) with 8 % DMF at 37 °C for 18 h. Data represents the mean $\pm$ SD of two technical replicates.

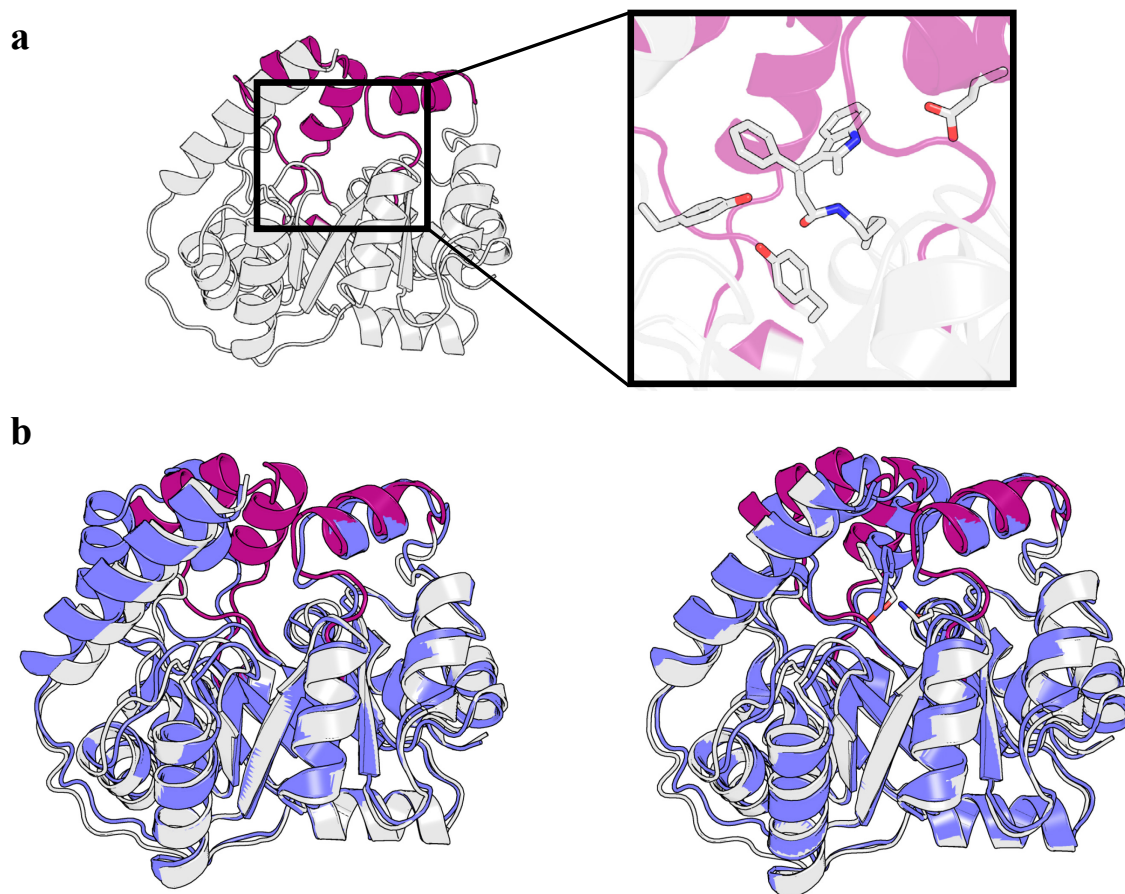


Figure 2.10. FCS_1 structure. (a) Designed model of FCS_1 in a cartoon representation with a zoomed-in view of the active site showing the catalytic residues and TS9 in sticks. The lid is colored in magenta. (b) Boltz-2-predicted structures (blue) of FCS_1 with (left) and without (right) substrate bound are overlaid on the Triad design model (grey). The predicted substrate binding pose shows good agreement with the design model.

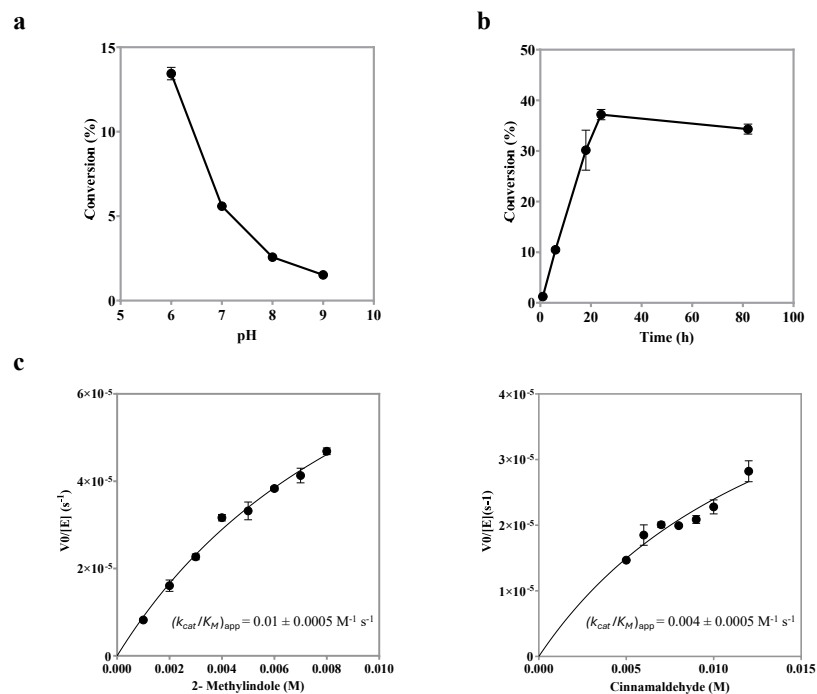


Figure 2.11. Functional characterization of FCS_1. (a) Effect of pH on FCS_1 activity. Reactions were performed with 5 mM *trans*-cinnamaldehyde and 2 mM 2-methylindole 8 % DMF (v/v) and 10 mol % catalyst in Britton-Robinson buffer at pH 6, 7, 8 and 9 at 37 °C for 18 h. The reaction is quenched with equal volume acetonitrile and analyzed by reverse phase HPLC. (b) Time course of FCS_1. Reactions were performed with 5 mM *trans*-cinnamaldehyde and 2 mM 2-methylindole 8 % DMF (v/v) and 10 mol % catalyst in sodium phosphate buffer, pH 7 at 37 °C for 18 h. (c) Michaelis-Menten kinetics of FCS_1 for 2-methylindole (left) and *trans*-cinnamaldehyde (right).

2.6.2.2 pH profile

The pH dependence of catalytic activity of FCS_1 was assessed across a pH range using Britton-Robinson buffer (Figure 2.11.a). The enzyme displayed a clear preference for acidic conditions, with maximal conversion of approximately 10% observed at pH 6, declining progressively to approximately 1.5% at pH 9. The enzyme was not soluble below pH 6, precluding assessment of activity under more acidic conditions. As the pH profile was determined from endpoint conversion rather than initial rates, differences in conversion across the pH range may arise from both the pH dependence of catalysis and

differences in enzyme stability over the course of the 18-h reaction. The different buffer and pH conditions used relative to the initial activity screen may likewise contribute to the differences in absolute conversion observed between the two experiments. This pH dependence was initially unexpected, as iminium catalysis relies on the condensation of the catalytic lysine with the aldehyde substrate to form an iminium ion, a step that is generally favored at basic pH where the lysine amine is in its neutral, nucleophilic form. However, the observed preference for acidic conditions can be rationalized by considering the cumulative acid-base requirements of the downstream catalytic steps in the proposed mechanism. The two active site tyrosines, which act as general acids at key steps of the mechanism, require their protonated forms to be sufficiently populated to effectively donate protons to the appropriate species. Similarly, although glutamate is predominantly deprotonated at pH 6 and can therefore effectively perform its role as a general base, the equilibrium at this pH is also sufficiently favorable for the protonated form to persist long enough to subsequently perform its role as a general acid. At higher pH, this balance is disrupted; the protonated form of glutamate becomes too transient to contribute productively, and the tyrosines lose their capacity to act as effective general acids. Therefore, the observed pH optimum might reflect a sweet spot where the acid-base requirements of multiple catalytic residues are simultaneously satisfied, rather than the optimization of any single step in isolation.

Alternatively, the acidic pH preference raises the possibility that the observed product formation may not be predominantly dependent on K259-mediated Schiff-base formation. The extent of this dependence was further examined through the catalytic-residue knockout experiments described below. Having established the optimal pH conditions for FCS_1,

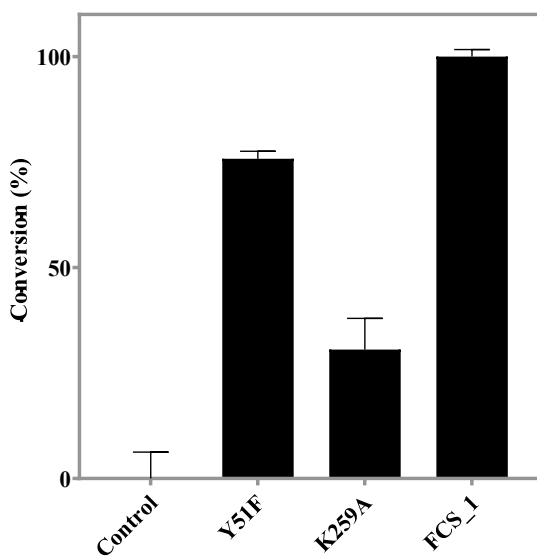
the time dependence of substrate conversion was subsequently monitored to further characterize its catalytic behavior.

2.6.2.3 Conversion time course

The time dependence of substrate conversion by FCS_1 was monitored over a period of approximately 82 hours under the optimal assay conditions established by the pH profile (Figure 2.11.b). Conversion increased linearly during the first 24 hours, reaching a maximum of approximately 38 %, corresponding to 3.8 total turnover numbers (TTN), before plateauing. Beyond 82 hours, a decrease in detected product was observed, the origin of which remains to be established but could reflect product degradation under the prolonged assay conditions, enzyme inactivation, or product inhibition. Based on these observations, 24 hours was selected as the optimal time window for subsequent kinetic assays, capturing the linear phase of conversion during which initial rate assumptions hold, and steady-state kinetic parameters can be reliably determined. To confirm that the observed catalytic activity is mechanistically dependent on the designed active site rather than reflecting a non-specific effect, single point mutants of each catalytic tetrad residue were generated and characterized.

2.6.2.4 Catalytic knockouts

a



b

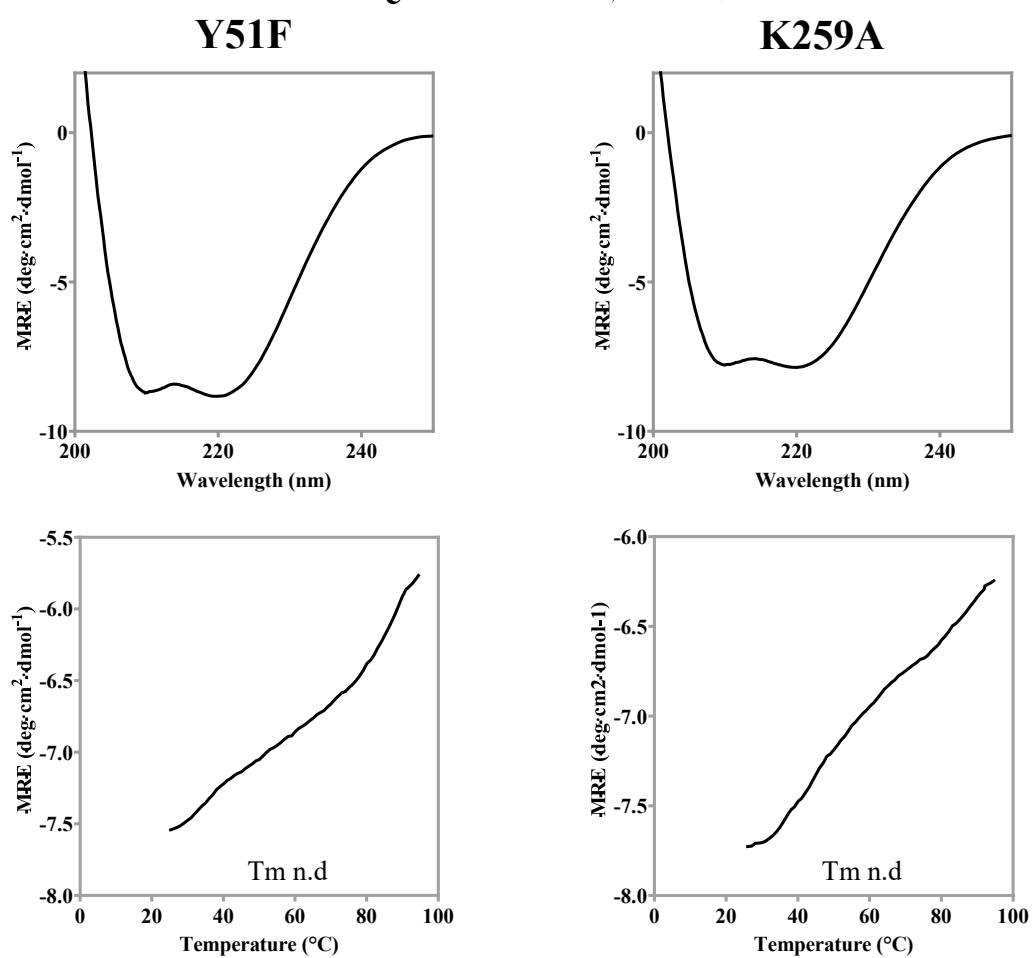


Figure 2.12. Characterization of FCS_1 knockouts. (a) SDS-PAGE of knockout mutants. (b) Conversion of knockouts. Reactions were performed with 5 mM *trans*-cinnamaldehyde and 2 mM 2-methylindole, 8 % DMF (v/v) and 2 mol % in sodium phosphate buffer at pH 7 at 37 °C for 18 h. (c) Structural characterization of knockouts. Top panel: CD spectra of Y51F and K259A. Bottom panel: Melting curves of Y51F and K259A.

To investigate the contribution of the designed catalytic residues to activity, four single-point mutants were generated by substituting the two tyrosines with phenylalanine and the glutamate and lysine with alanine. Of the four mutants, only Y51F and K259A could be successfully expressed and purified and were subsequently characterized structurally and functionally. SDS–PAGE analysis indicated that Y51F was obtained at an estimated purity exceeding 90%, whereas K259A was estimated to be approximately 80% pure. CD spectroscopy indicated that both mutants retained substantial secondary structure with spectral characteristics similar to those of FCS_1, suggesting that neither mutation substantially altered their secondary-structure composition. In contrast, the Y106F and E208A variants could not be obtained in sufficient soluble quantities for characterization. Although the origin of this insolubility remains unclear, it raises the possibility that these residues contribute not only to catalysis but also to the structural stability or folding of the designed enzyme. Additional experiments would be required to distinguish between effects on protein folding, stability, expression, or solubility. Thermal denaturation experiments revealed profiles similar to those observed for FCS_1, suggesting that the mutations did not substantially alter the thermal behaviour of the proteins under the conditions examined. While Y51F was obtained at high purity as confirmed by SDS–PAGE, K259A contained substantial contaminant bands, which should be considered when interpreting its functional characterization.

Both mutants were assayed at 2 mol%, compared with 10 mol% FCS_1, because of their lower expression yields, and the resulting conversions were normalized relative to FCS_1 under the corresponding assay conditions, with FCS_1 defined as 100% activity (Figure 2.12.b). Under these conditions, Y51F retained approximately 85% of the FCS_1 activity, whereas K259A retained approximately 50%. The reduced activity of both variants is consistent with contributions from Tyr51 and Lys259 to catalysis. Notably, however, the substantial activity retained by K259A indicates that product formation persists upon removal of the designed catalytic lysine, raising the possibility that the observed catalysis is not strictly dependent on K259-mediated Schiff-base formation. The lower purity of the K259A preparation should also be considered when interpreting the magnitude of this residual activity. Together, the mutational results suggest contributions from the designed residues while also raising questions regarding the extent to which the observed activity depends on the originally proposed catalytic mechanism, which are considered further in the discussion. The kinetic properties of FCS_1 were therefore investigated next to quantify the catalytic efficiency of the designed enzyme.

2.6.3 Steady state kinetics

Steady-state kinetic parameters of FCS_1 were determined by measuring initial rates as a function of substrate concentration for both 2-methylindole and *trans*-cinnamaldehyde (Figure 2.11.c). In both cases, saturation could not be achieved within the solubility limits of the substrates under the assay conditions, precluding reliable independent determination of k_{cat} and K_M . Only the apparent second-order rate constants $(k_{cat}/K_M)_{app}$, which remain well-defined even under sub-saturating conditions could be determined and were $0.01 \pm$

0.0005 M⁻¹s⁻¹ and 0.004 ± 0.0005 M⁻¹s⁻¹ for 2-methylindole and *trans*-cinnamaldehyde respectively and represent the most reliable kinetic parameter. Notably, the uncatalyzed rate of the reaction was below the detection limit of the method employed, precluding formal calculation of the rate acceleration but indicating that the enzyme-mediated rate enhancement exceeds what can be quantified under the current experimental conditions. To our knowledge, FCS_1 displays among the highest levels of product formation reported for a *de novo* designed protein targeting a Friedel–Crafts alkylation using exclusively canonical amino acids. Although its catalytic efficiency is modest, this result demonstrates that the MsC pipeline is capable of generating functional enzymes for reactions with no natural or computational precedent. To further rationalize the activity differences observed across the variant set and assess whether the designed active site geometries are maintained in solution, molecular dynamics simulations were subsequently performed.

2.6.4 Molecular dynamics analysis

To gain structural insight into the activity differences observed across the variant set, two complementary molecular dynamics (MD) simulation strategies were employed. A single 1- μ s apo simulation was performed for each variant to assess maintenance of the designed protein structure over an extended trajectory, while ten independent 100-ns product-bound simulations were performed for each variant to sample the active-site conformational ensemble and evaluate the persistence of the designed catalytic geometry. The apo and product-bound simulations were therefore used to assess global structural stability and active-site catalytic geometry, respectively. The product-bound state was selected because it contained the molecular features required to simultaneously evaluate the relevant contacts between the ligand and catalytic residues within a common structural state. It also

provided a non-covalently bound state that could be readily treated using the classical MD framework, whereas several intermediates along the proposed iminium pathway are covalently linked to the catalytic lysine and would require additional parameterization of the corresponding covalent adducts. The product geometry most closely resembles that of TS9 and therefore provided a consistent structural framework for assessing the maintenance of the designed catalytic contacts throughout the simulations. Catalytic competency was defined as the percentage of snapshots from the product-bound simulations in which two key geometric constraints were simultaneously satisfied: a distance of less than 3.5 Å between the glutamate oxygen and the ligand nitrogen and a distance of less than 5 Å between the catalytic lysine nitrogen and the C10 carbon of the ligand. Snapshots satisfying both constraints simultaneously were classified as catalytically competent, and competency was calculated as their percentage across the product-bound trajectories for each variant. A positive correlation was observed between catalytic competency and log(conversion) across the variant set ($R = 0.77$), indicating that variants that more frequently sampled the defined active-site geometry tended to exhibit higher conversion. FCS_1 displayed the highest catalytic competency at approximately 7%, consistent with its superior conversion in the activity assays. To evaluate the contribution of FCS_1 to this relationship, the correlation was recalculated following its exclusion, with a positive correlation retained across the remaining variants ($R = 0.52$). Thus, although FCS_1 strengthens the overall relationship, the association between catalytic competency and conversion persists across the remaining variant set. Taken together, the product-bound MD simulations reveal a positive association between maintenance of the designed active-site geometry and experimental catalytic output across the designed variants.

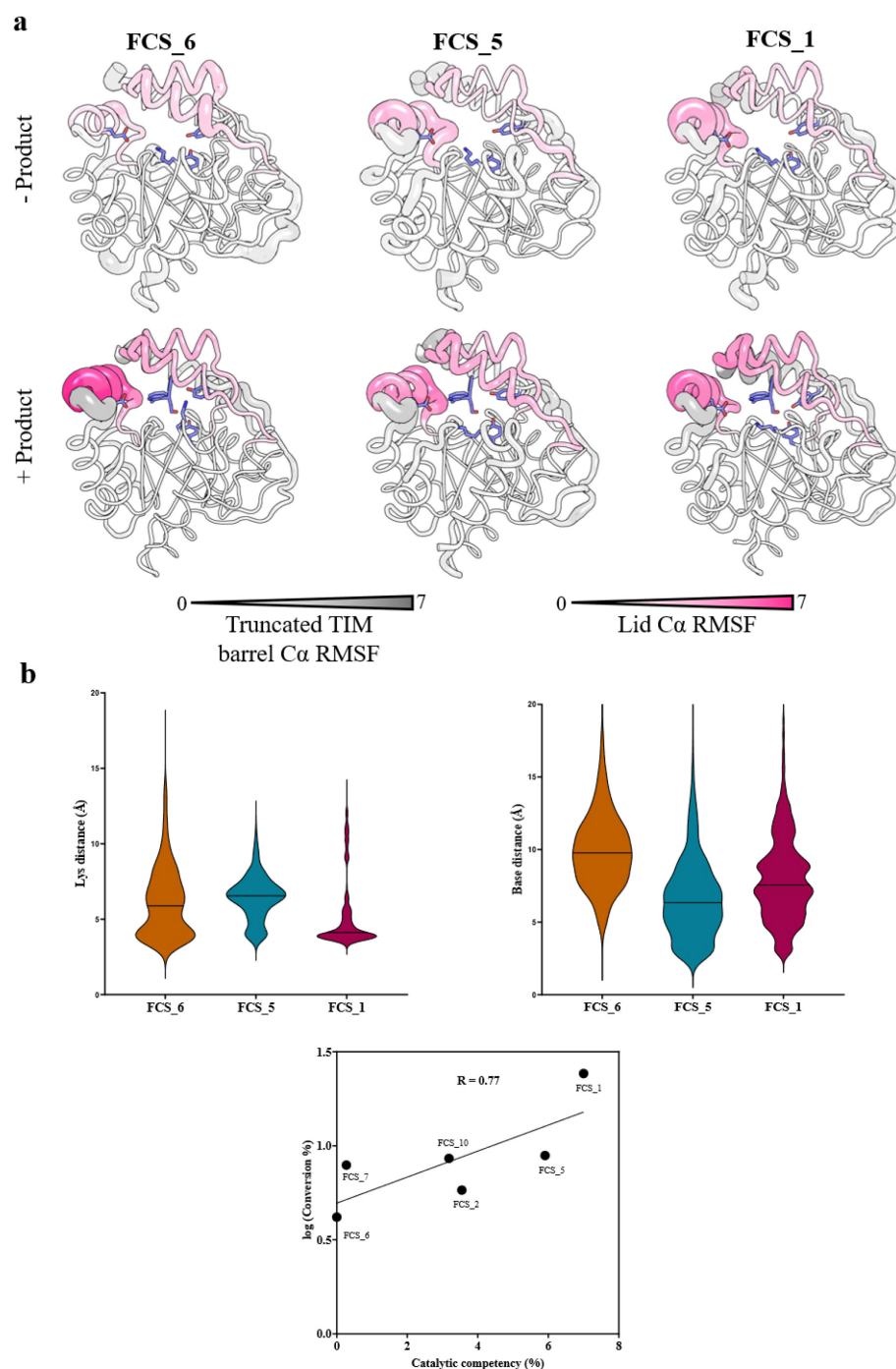


Figure 2.13. Molecular dynamics simulations of FCS variants in the absence and presence of the product. Simulations of the product-bound complexes were performed as 10 replicates of 100 ns each, while simulations of the free (apo) proteins consisted of a single 1- μ s replicate. **(a)** Root-mean-square fluctuations (RMSF) of the protein backbone, represented as cartoon putty, reveal different levels of flexibility in the lids of variants with

different activities: FCS_6, the least active; FCS_5, with higher activity; and FCS_1, the most active. Product and catalytic residues are depicted as sticks. **(b)** Distributions of catalytic contact distances between the product and the nucleophile or catalytic base, alongside the catalytic competency, presented as violin and scatter plots, respectively with the median value indicated by a black line (left and right panels).

2.7 Discussion

In this study, Multistate-CANVAS (MsC) was developed and applied to the *de novo* design of proteins targeting the Friedel–Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde. Most selected designs could be expressed and purified as substantially structured proteins, and all purified variants produced greater product conversion than the no-enzyme reaction, with FCS_1 emerging as the most active design. These results suggest that the MsC workflow generated proteins associated with enhancement of the target reaction. At the same time, the combined structural, functional, mutational, and computational characterization raises important questions regarding the origin and mechanism of this enhancement and the specific contribution of multistate optimization.

The biophysical characterization indicates that extensive sequence redesign was generally compatible with the formation of structured proteins. The substantial secondary structure retained at elevated temperatures further suggests considerable structural robustness, potentially reflecting high thermal stability, non-cooperative unfolding, or structural transitions that preserve secondary structure. High thermal stability is particularly plausible given the use of ProteinMPNN/LigandMPNN-derived sequences in the design workflow. Nevertheless, this behavior was not uniform across the variant set, as illustrated by the distinct biophysical properties of FCS_8. More broadly, the considerable variation in catalytic performance among predominantly structured variants suggests that differences

in global folding and thermostability are unlikely to represent the principal limitation for most designs, instead pointing toward local active-site organization and the chemical environment as important determinants of catalytic performance.

Although all nine purified variants displayed conversion above the no-enzyme background, the origin of this enhancement remains to be established. Activity across multiple independently redesigned sequences is encouraging and demonstrates that measurable product formation was not restricted to a single design. However, comparison exclusively with the uncatalyzed reaction does not distinguish activity arising from the designed catalytic architectures from potential scaffold-dependent or more general protein-mediated effects. This distinction is particularly relevant for FCS_8, which retained measurable conversion despite its different biophysical characteristics. Thus, the observed activity supports enhancement of the reaction in the presence of the designed proteins, while its specific attribution to the intended active sites requires further validation.

The magnitude of this activity is nevertheless notable in the context of first-generation computationally designed enzymes for C–C bond formation (Table 2). Although substantially below the efficiencies of natural enzymes, the activities obtained here are comparable to those of several initial *de novo* catalysts for related transformations, including Morita–Baylis–Hillman and Diels–Alder reactions. The best designs also fall within the range reported for engineered Friedel–Crafts catalysts incorporating noncanonical amino acids, whereas the present proteins were constructed exclusively from the 20 canonical amino acids. The significance of the observed activity therefore lies primarily in generating protein designs capable of enhancing a chemically demanding

multistep C–C bond-forming transformation rather than in their absolute catalytic efficiencies.

Further characterization of FCS_1 raises questions regarding the chemical mechanism underlying this enhancement. Its preference for acidic conditions could reflect the competing protonation requirements of the proposed catalytic cycle, in which K259 nucleophilicity must be balanced against the general acid–base functions of the other catalytic residues. However, because the pH profile was determined from 18 h endpoint conversions, differences in enzyme stability or retention of activity across the pH range may contribute to the observed trend. The pH dependence therefore demonstrates that the chemical environment and protonation state influence FCS_1 performance but does not uniquely establish the catalytic processes responsible for this behavior.

The mutational analysis further raises questions regarding the precise roles of the designed catalytic residues and the mechanism underlying product formation. K259 was designed as the catalytic nucleophile responsible for Schiff-base formation, yet K259A retained approximately 50% of the activity of FCS_1. Similarly, Y51F retained substantial activity despite removal of the phenolic hydroxyl group proposed to participate in proton transfer. These results indicate that both positions influence catalytic performance but that their proposed functions are not strictly required for product formation under the conditions examined. One possibility is that the loss of these catalytic functions is partially compensated by other residues within the active site. For K259A, another accessible amine, such as a nearby lysine or the protein N-terminus, could potentially substitute for K259 and preserve imine-mediated catalysis. For Y51F, the second catalytic tyrosine could potentially adopt or contribute to the proton-transfer function assigned to Y51 following

local active-site reorganization, with water-mediated proton transfer providing an additional possibility. Alternatively, the residual activity of K259A may reflect a pathway that does not require formation of the designed Lys-derived Schiff base. Such a pathway could involve reversible aldehyde hydration followed by acid-assisted proton transfer and dehydration to generate an electrophilic oxocarbenium-like species, consistent with established carbonyl hydration and Brønsted-acid-catalyzed dehydrative C–C bond-forming chemistry.^{72–74} Glu208 could potentially facilitate proton transfer or indole activation, enabling C–C bond formation (Figure 2.14). Thus, the mutational results support contributions from K259 and Y51 to catalytic performance but also suggest mechanistic plasticity or an alternative mechanistic pathway within the designed active site. Interpretation of the complete catalytic tetrad remains limited by the inability to characterize soluble Y106F and E208A variants, and further experiments would be required to establish the roles of the tetrad and distinguish between functional rescue and an alternative reaction pathway.

Alternatively, the residual activity of K259A may reflect a pathway that does not require formation of the designed Lys-derived Schiff base. A water-mediated, acid-assisted pathway could involve reversible aldehyde hydration followed by proton transfer and dehydration to generate an electrophilic oxocarbenium-like species, consistent with established carbonyl hydration and Brønsted-acid-catalyzed dehydrative C–C bond-forming chemistry (Buschmann et al., 1982; Estopiñá-Durán and Taylor, 2021). Glu208 could potentially facilitate proton transfer or indole activation, enabling C–C bond formation (Figure 2.14). Although this pathway remains speculative, the residual activity of K259A suggests that the designed K259-derived iminium intermediate may not be

strictly required for product formation. Further mechanistic studies will be required to distinguish between functional rescue of the intended pathway and an alternative reaction mechanism.

The MD simulations provide a complementary perspective on local active-site organization. Catalytic competency correlated positively with experimental conversion across the complete variant set ($R = 0.77$), and a weaker positive association remained following removal of FCS_1 ($R = 0.52$). Thus, variants that more frequently maintained the selected designed contacts tended to display greater conversion, although FCS_1 substantially strengthens this relationship. The choice of the product-bound state for this analysis should also be considered when interpreting catalytic competency. The product provided a practical representation because it contains the atoms required to monitor the selected catalytic contacts and avoids the additional complexity associated with modeling covalently bound intermediates. However, the product represents only one region of a multistep catalytic cycle and may not be the state that most effectively distinguishes catalytically competent from incompetent active-site arrangements. Intermediates associated with the most chemically demanding steps may provide more informative representations of catalytic competency, particularly those surrounding the highest activation barrier or the nucleophilic attack responsible for C–C bond formation. States such as I4, I5, or I6 may therefore provide alternative probes of whether the active site maintains the geometric organization required during the central catalytic events. Together, the biophysical characterization and MD simulations suggest that overall protein structure does not necessarily reflect active-site organization. While most designs displayed

substantial secondary structure, their ability to maintain the designed catalytic geometries varied considerably. These differences in active-site organization may therefore contribute to the observed variation in catalytic activity among the designs.

The current MsC objective also raises a broader consideration regarding the energetic consequences of multistate transition-state optimization. MsC simultaneously optimizes sequences against multiple transition-state representations, while the neighboring intermediates are not explicitly considered in the sequence objective. Because a transition state and its neighboring intermediate may share similar structural and interaction requirements, interactions introduced to stabilize the transition state could also stabilize the corresponding intermediate. If an intermediate is stabilized more strongly than the transition state leading out of it, the resulting activation barrier could increase, potentially creating an energetic trap that slows progression through the catalytic cycle. Thus, effective multistate optimization requires not only stabilization of individual transition states, but their preferential stabilization relative to neighboring intermediates.

This energetic perspective complements the conformational limitations suggested by the MD simulations. The current workflow does not explicitly account for the substrate-bound Michaelis complex, reaction intermediates, or product-bound states, and the intentionally minimal catalytic tetrad lacks the extensive second-shell networks characteristic of highly optimized natural and evolved enzymes. Such interactions can contribute to substrate positioning, electrostatic organization, active-site solvation, and modulation of catalytic residue pKa values. The measurable activity combined with the relatively limited population of the designed active-site geometries is therefore consistent with designs that

capture important features of the target chemistry but remain incompletely optimized across both their conformational and catalytic energy landscapes.

Finally, the present results distinguish the feasibility of the MsC workflow from the specific advantage of multistate optimization. The study demonstrates that proteins generated under a multistate objective can be experimentally produced as structured proteins associated with enhancement of the targeted reaction. However, because equivalent single-state CANVAS designs were not experimentally benchmarked alongside the MsC variants, the contribution of simultaneous optimization across multiple states cannot be isolated from the overall design strategy. The present work therefore establishes the feasibility of incorporating multistate optimization into CANVAS, while its advantage relative to an equivalent single-state approach remains to be established.

The unresolved questions identified by the present study provide a basis for further evaluation and development of the MsC approach. In particular, establishing the origin of the observed catalytic enhancement, resolving the contributions of the designed catalytic residues and the operative reaction mechanism, and determining the specific contribution of multistate optimization will be important for assessing the current designs. These questions, together with potential strategies for addressing the conformational and energetic limitations identified here, provide the foundation for the future directions discussed in Chapter 3.

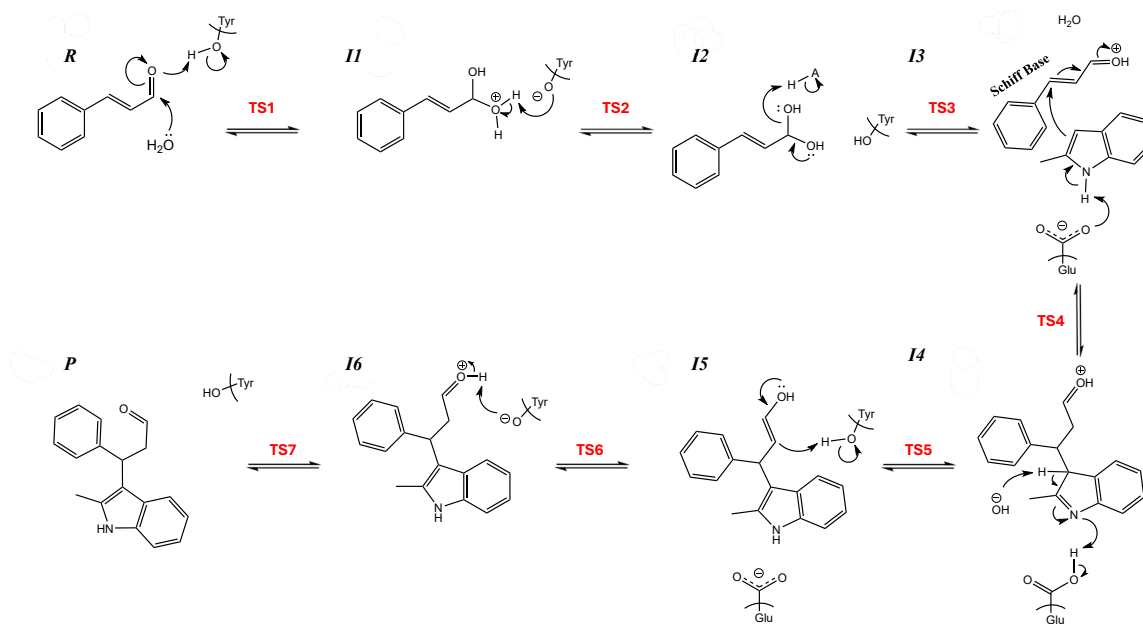


Figure 2.14. Alternate proposed mechanism for the Friedel-Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde.

Table 2. Catalytic properties of *de novo* designed C-C bond-forming enzymes.

Enzyme	Reaction	Conversion (%)	TTN	k_{cat} (min ⁻¹) ^a	K_{M} (mM) ^a	$k_{\text{cat}}/K_{\text{M}}$ (M ⁻¹ s ⁻¹) ^b	Reaction conditions	Reference
<i>Canonical motifs</i>								
BH32	MBH	0.54	0.2	0.002	8.00 ^{cyclo} 1.77 ^{benzal}	0.0046 0.021	3 mol%, 31 h	Crawshaw , 2022 Braun, 2025
MBH18	MBH	3.01	1.5	0.008	8.33 ^{cyclo} 8.55 ^{benzal}	0.017 0.015	2 mol%, 8 h	Braun, 2025
MBH48	MBH	16	8	0.025	25.2 ^{cyclo} 1.42 ^{benzal}	0.014 0.30	2 mol%, 8 h	Braun, 2025
DA_20_00	DA	~78	3.9	0.0017	146.0 ^{diene} 72.8 ^{dieno}	0.0002 0.0004	20 mol%, 60 h	Siegel, 2010
FCS_1	FCA	28	2.8	—	—	0.015 ^{indole} 0.0039 ^{cinna}	10 mol%, 18 h	Kouassi, 2026
<i>Noncanonical motifs</i>								
LmrR_pAF_RGN	FCA	4	2	0.64	5.2	1.25	2 mol%, 16 h	Leveson-Gower, 2021
V15aY	FCA	14	2.3	—	—	3.3	6 mol%, 16 h	Brouwer, 2025
<i>Natural enzymes</i>								
PfTrpB	FCA			18.6	1.2 ^{serine} 0.077 ^{indole}	258 ^{serine} 4428 ^{indole}		Buller, 2015

2.8 Methods

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

2.8.1 Computational Enzyme Design

All calculations were performed with the Triad protein design software (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. To provide sidechain conformations, a backbone-independent conformer library (bda-bbind_1.0)¹³ was used for theozyme placement, and a backbone-independent rotamer library (bbind02.May.e2)⁷⁵ 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 function⁷⁶ consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field⁷⁷ with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal mol⁻¹ and an equilibrium donor–acceptor distance of 2.8 Å,⁷⁸ and an electrostatic energy 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 (Supplementary Table 3), as described by Lassila *et al.*¹³

2.8.2 Catalytic motif assembly

The theozymes corresponding to TS1, TS2, and TS4–TS9 were aligned in PyMOL by superimposing the stereocenter carbon atom of the reacting intermediate to generate a common reference frame (Figure 2.4a). Based on the proposed catalytic mechanism, Tyr1

was assigned to TS1, TS2, TS8, and TS9, Glu to TS4 and TS5, Tyr2 to TS6 and TS7, and Lys to all catalytic states. For each catalytic residue, Invrotzyme was used to generate backbone conformations compatible with the required catalytic geometry using a representative transition state. The resulting backbone ensembles were combined to assemble complete catalytic motifs, and assemblies containing backbone steric clashes or incompatible residue geometries were discarded.

2.8.3 Lid generation

Following catalytic motif assembly and incorporation into the TIM barrel scaffold, the glutamate-containing fragment of the catalytic motif could not be accommodated within the barrel architecture while preserving the required catalytic geometry. This fragment was therefore scaffolded onto a short α -helical lid using Invrotzyme. The deleted L1 and L6 $\beta\alpha$ -loops of the 5AN7 scaffold were subsequently reconstructed using RFdiffusion to connect the helical lid to the TIM barrel. Fragment lengths of 20–30 residues were sampled for the primary lid connection, while 7–10 residues were sampled for the glutamate anchor. For each design, 100–200 RFdiffusion trajectories were generated. Designs lacking regular secondary structure, introducing steric clashes with the catalytic motif or transition states, or displacing the catalytic residues from their target geometries were discarded. The remaining backbone models were retained for subsequent sequence design and structural validation.

2.8.4 Ensemble generation

Local conformational sampling of the optimized TIM-barrel scaffold was performed using the Rosetta Backrub protocol. Backbone and side-chain torsional degrees of freedom were sampled using a BackrubMover with Monte Carlo acceptance based on the Rosetta score

function. Each trajectory consisted of 10,000 Backrub moves, after which the lowest-energy conformation sampled during the trajectory was recovered. A total of 120 independent Backrub decoys were generated for and used for subsequent theozyme placement.

2.8.5 Theozyme placement

A 120-Backrub ensemble from the optimized scaffold, was used as backbone template for theozyme placement. Following Cartesian minimization of the scaffold in Triad, neighbouring residues were mutated to alanine to avoid steric clashes with the TS. Solutions were further filtered by computing pairwise distances between the geometric centers of the phenyl ring across all theozyme placements. Solutions exhibiting the lowest pairwise distances were prioritized and carried forward as input for multistate active-site sequence design.

2.8.6 Active-site repacking

Active-site repacking calculations were performed on the structures with this theozyme placement. In these calculations, residues near the catalytic amino acids and TS (10 Å from the TSs phenyl ring) 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 (Supplementary Table 4).

2.8.7 Multistate design

Calculations were performed for the apo structures corresponding to each TS (without any TS) to evaluate the physical stability of the overall generated sequence and its ability to

exist by itself. Multistate design calculations were performed for TS1-TS9 motifs excluding TS3 using an energy combination function to determine the overall fitness score for a sequence across all transition states. The combination function minimizes the energy difference between all TSs. These calculations were performed using the multi_design.py application with the Triad software suite. To compare energies, the energy difference between each structure and a corresponding all-Gly structure in the absence of each corresponding TS was calculated.

2.8.8 Analysis and filtering

For the each state multistate designs 500 sequences were generated. To reduce the number of sequences to test experimentally, we discarded those who presented positive overall energies as well as active site energies. Also , the sequence energy difference between the highest energy state and lowest energy state was computed and only sequences having a $\Delta E < 400$ were selected. Additionally, the catalytic lysine pKa for all TSs was determined using propka3⁷⁹ and sequences with a predicted lysine pKa greater than 9 were eliminated. Finally, the solvent-accessible surface area (SASA) of TS4 and TS6 was calculated using MDTraj (Python), and only sequences where $\geq 75\%$ of the TS surface is buried were retained.

2.8.9 Surface redesign

Sequence optimization of selected structures was performed using LigandMPNN²³. Amino-acid identities of all residues designed during active-site repacking were retained (Supplementary Table 6). Hundred sequences were generated per structure, omitting cysteine. The generated sequences were used for structure prediction with Boltz-2 in presence of the substrate and product.⁶⁹ Predictions were filtered based on their backbone

RMSD, substrate and product centroid RMSD, lid RMSD as well as backbone pLDDT and lid pLDDT compared to their corresponding input structures from active-site repacking. (Supplementary Table 6).

2.8.10 Protein expression and purification

Genes constructs for all proteins were ordered as codon-optimized fragments flanked by restriction sites for *Nde*I and *Xho*I from Twist Bioscience (South San Francisco, CA, USA). *E. coli* BL21 (DE3) competent cells (Novagen) were transformed with the corresponding plasmid DNA and plated on LB agar supplemented with kanamycin (50 µg/mL). Plates were incubated overnight at 37 °C. Single colonies were used to inoculate overnight precultures under the same conditions. The following day, 5 mL of preculture was used to inoculate 500 mL of LB medium supplemented with kanamycin (50 µg/mL), and cultures were grown at 37 °C with shaking until an OD600 between 0.4 and 0.6 was reached. Protein expression was induced by addition of isopropyl-β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM, after which cultures were transferred to 16°C and shaken at 225 rpm overnight. Cells were collected by centrifugation, and pellets were either stored at -20°C or processed immediately for purification.

For purification, cell pellets were resuspended in 8 mL of buffer D (5 mM imidazole, 100 mM potassium phosphate, pH 8.0) containing lysozyme (1 mg/mL, MP Biomedicals) and benzonase nuclease (2 U/mL, Merck Millipore). Cell lysis was achieved by homogenization using an Avestin EmulsiFlex-B15 cell disruptor. Clarified lysates were subjected to immobilized metal affinity chromatography using Ni-NTA agarose resin (Qiagen) packed in Econo-Pac gravity-flow columns (Bio-Rad) pre-equilibrated with buffer D. Non-specifically bound proteins were removed by sequential washes with buffer

E (10 mM imidazole, 100 mM potassium phosphate, pH 8.0) followed by a wash with 20 mM imidazole in the same buffer. Target proteins were eluted with 5 mL of buffer F (250 mM imidazole, 100 mM potassium phosphate, pH 8.0). Purification quality was assessed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Protein concentration was determined spectrophotometrically by measuring absorbance at 280 nm and applying the Beer-Lambert law using theoretical extinction coefficients calculated with the ExPASy ProtParam tool. (<https://web.expasy.org/protparam/>)

2.8.11 Circular dichroism spectroscopy and thermal stability measurements

CD spectra were collected with a Jasco J-815 spectrometer. Spectra were recorded using a 1-mm quartz cuvette at 25 °C with scanning speed 10–100 nm min⁻¹, bandwidth 1 nm, response time 1 s. Experiments were performed in 20 mM sodium phosphate buffer pH 7.0, using a protein concentration of approximately 12 µM. For each protein, 3 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 25 to 95 °C at a rate of 1 °C per minute and ellipticity at 222 nm was measured every 1 °C.

2.8.12 Conversion assays

Conversion assays were carried out in kinetic buffer (50 mM sodium phosphate buffer, 100 mM NaCl, pH 7) with 8% DMF as a cosolvent. Duplicate 400 µL reactions with varying concentrations of freshly prepared *trans*-cinnamaldehyde (Achemica) or 2-methylindole (Achemica) dissolved in DMF (8 % final concentration) were initiated by the addition of

the different purified enzymes (10 mol% final concentration). Product formation was quantified by reverse- phase HPLC (Agilent) after 18 h of reaction at 37 °C followed by quenching with equal volume of acetonitrile. The different areas under the curves for the product, were converted into product quantity by fitting them into a standard curve generated with different varying concentrations of the standard (Supplementary figure 2).

2.8.13 pH profile determination

pH profile determination were carried out in a Britton-Robinson buffer (0.04 M boric acid, 0.04 M phosphoric acid, 0.04 M acetic acid) at varying pH values. Duplicate 400 µL reactions with varying concentrations of freshly-prepared *trans*-cinnamaldehyde (Achemica) or 2-methylindole (Achemica) dissolved in DMF (8 % final concentration) were initiated by the addition of the different purified FCS_1 enzyme (10 mol % final concentration). Product formation was quantified by reverse- phase HPLC (Agilent) after 18 h of reaction followed by quenching with equal volume of acetonitrile. The different areas under the curves for the product, were converted into product quantity by fitting them into a standard curve generated with different varying concentrations of the standard.

2.8.14 Conversion time course

Conversion time course was carried out in kinetic buffer (50 mM sodium phosphate buffer, 100 mM NaCl , pH 7) with 8% DMF as a cosolvent. Duplicate 400 µL reactions with varying concentrations of freshly prepared *trans*-cinnamaldehyde (Achemica) or 2-methylindole (Achemica) dissolved in DMF (8 % final concentration) were initiated by the addition of the different purified enzymes (10 mol % final concentration). Product formation was quantified by reverse- phase HPLC (Agilent) 1, 6, 18, 24 and 82 h of reaction at 37 °C followed by quenching with equal volume of acetonitrile. The different areas under

the curves for the product were converted into product quantity by fitting them into a standard curve generated with different varying concentrations of the standard (Supplementary figure 2).

2.8.15 Knockouts conversion assays

Conversion time course was carried out in kinetic buffer (50 mM sodium phosphate buffer, 100 mM NaCl , pH 7) with 8% DMF as a cosolvent. Duplicate 400 μ L reactions with varying concentrations of freshly prepared *trans*-cinnamaldehyde (Achemica) or 2-methylindole (Achemica) dissolved in DMF (8 % final concentration) were initiated by the addition of the different purified enzymes (2 mol % final concentration). Product formation was quantified by reverse- phase HPLC (Agilent) 1, 6, 18, 24 and 82 h of reaction at 37 °C followed by quenching with equal volume of acetonitrile. The different areas under the curves for the product were converted into product quantity by fitting them into a standard curve generated with different varying concentrations of the standard (Supplementary figure 2).

2.8.16 Steady-state kinetics

Steady-state kinetic assays were carried out at 25°C in kinetic buffer (50 mM sodium phosphate buffer, 100 mM NaCl , pH 7) with 8% DMF as a cosolvent. Duplicate 400 μ L reactions with varying concentrations of freshly-prepared *trans*-cinnamaldehyde (Achemica) or 2-methylindole (Achemica) dissolved in DMF (8 % final concentration) were initiated by the addition of the purified FCS_1. Product formation was quantified by reverse- phase HPLC (Agilent) after 1, 2 and 3 h of reaction followed by quenching with equal volume of acetonitrile. The different areas under the curves for the product, were converted into product quantity by fitting them into a standard curve generated with

different varying concentrations of the standard. Linear phases of the kinetic traces were used to measure initial reaction rates. k_{cat} and K_{M} were determined by fitting the data to the Michaelis Menten model.

2.8.17 Molecular dynamics (MD) simulations

The Amber 2020 software (<http://ambermd.org/>) with the ff19SB protein force field⁸⁰, gaff2 ligand force field⁸¹ and OPC water force field⁸² was used for all simulations. A cutoff of 10 Å was applied to electrostatics modeled using the particle mesh Ewald method⁸³. All MD trajectories were run with a time step of 2 fs. Models of each design generated using Triad as described above were prepared for MD. MD trajectories were run for each enzyme in the presence and absence of the product. Parameters for the product were generated using the Antechamber package⁸⁴. Hydrogen atoms were added using Reduce⁸⁵, 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 addions2 algorithm in Amber was used to place counterions to neutralize the system. The system was first minimized through the method of steepest descent using a force constant of 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 thermostat⁸⁶ and Berendsen barostat⁸⁷, respectively. Following equilibration, one replicate 1-μs and 10 replicates of 100-ns production simulations were run. To analyze catalytic interactions and RMSF, 5×10^4 snapshots separated by 20 ps each

were extracted from the production trajectory. Catalytic interactions and RMSF values were extracted using CPPTRAJ⁸⁸.

2.9 Summary

The overarching objective of this thesis was to extend the scope of computational enzyme design to complex multistep C–C bond-forming reactions. Although substantial progress has been made in the development of computational enzyme design methodologies, most successful examples have focused on reactions that can be adequately represented by a single catalytic state or rely on catalytic motifs derived from natural or previously engineered enzymes. Consequently, extending *de novo* enzyme design to complex multistep reactions with evolving catalytic requirements remains a significant challenge.

To address this challenge, this work developed Multistate-CANVAS (MsC), a computational framework integrating multistate active-site optimization with hybrid scaffold generation for the *de novo* design of proteins targeting complex multistep reactions. As a proof of concept, MsC was applied to the Friedel–Crafts alkylation of 2-methylindole with *trans*-cinnamaldehyde, a C–C bond-forming reaction lacking a naturally evolved biocatalyst for the targeted transformation. A catalytic mechanism was first characterized using quantum mechanical calculations to generate transition-state models representing different stages of the reaction. These models were subsequently integrated into a unified catalytic motif, incorporated into a TIM-barrel scaffold through the CANVAS framework, and further optimized using multistate computational enzyme design.

Experimental characterization showed that multiple designed variants could be successfully expressed and purified, with the majority displaying secondary-structure

signatures consistent with a TIM-barrel fold. All purified variants produced greater product conversion than the no-enzyme reaction, with FCS_1 displaying the highest activity. Although the observed catalytic efficiencies remain modest relative to those of natural enzymes, they are comparable to those reported for several first-generation *de novo* catalysts for related C–C bond-forming reactions. Mutational characterization of FCS_1 showed reduced but substantial residual activity following substitution of K259 and Y51, indicating that these residues influence catalytic performance while leaving their precise roles and the operative reaction mechanism unresolved. Molecular dynamics simulations further revealed a positive relationship between maintenance of the defined catalytic geometries and experimental conversion, suggesting that active-site organization contributes to catalytic performance.

Collectively, this work establishes the feasibility of applying multistate computational design to proteins targeting a complex multistep C–C bond-forming reaction. MsC provides a framework for integrating catalytic requirements from multiple stages of a reaction into a common active-site architecture and realizing these designs in structured proteins associated with enhanced product formation. Further experimental validation is required to establish the origin of this enhancement, the contribution of the designed catalytic motif, and the specific advantage of multistate optimization. These results provide a foundation for the continued development and experimental validation of multistate approaches for designing enzymes targeting increasingly complex chemical transformations.

Chapter 3. Conclusions and perspectives

The results presented in this thesis demonstrate the feasibility of applying Multistate-CANVAS (MsC) to the design of proteins targeting a reaction lacking natural precedent. While the designed proteins showed enhanced product formation, further experimental validation is required to establish the origin of this activity and the contribution of the designed catalytic architecture. These questions, together with opportunities to improve the methodology, provide the basis for future development of the MsC framework.

3.1 Validation of Designed Activity

A central limitation of the current experimental characterization is that the observed enhancement in product formation cannot yet be conclusively attributed to the designed catalytic motif. Addressing this limitation will require controls that progressively distinguish the intended active-site contribution from other possible sources of activity. Free lysine will establish the extent to which general amine chemistry can promote the reaction, while the parental scaffold will determine whether the original TIM-barrel possesses intrinsic activity toward the transformation. To further separate effects arising from the protein itself from those introduced through design, a denatured-enzyme control will assess whether an organized protein structure is required for activity, whereas a scaffold-only redesign lacking the catalytic motif will determine whether the redesigned protein environment alone can account for the observed enhancement. These controls will be complemented by a complete catalytic-tetrad knockout, providing a more direct test of the contribution of the designed motif as a whole. Kinetic characterization of all knockout variants that can be successfully expressed and purified will then establish how removal of individual catalytic residues, or of the complete tetrad, affects catalytic efficiency relative

to FCS_1. Together, these experiments will help determine whether the observed enhancement originates from the catalytic architecture intentionally introduced during design and clarify the contribution of its individual components.

3.2 Structural validation and conformational sampling



Figure 3.1. Crystal structures for enhanced conformational sampling.

Although circular dichroism spectroscopy showed that most designs display secondary-structure signatures consistent with a TIM-barrel fold, their unusual thermal behavior limited further assessment of their stability. Most variants retained substantial secondary structure throughout the temperature range examined without displaying clear cooperative unfolding transitions, preventing determination and comparison of their melting temperatures. Chemical denaturation monitored by CD could provide a complementary means of differentiating the stability of the variants by assessing their resistance to chemically induced unfolding. Importantly, chemical stability and thermostability represent distinct properties and should therefore be considered independently. Beyond these global biophysical properties, high-resolution structural characterization remains essential to evaluate the accuracy of the computational models. Determination of X-ray crystal structures of the most active variants would enable direct comparison between the designed and experimentally realized structures, particularly the positioning of catalytic

residues and the geometry of the substrate-binding pocket. These structures could also provide the basis for generating experimentally informed conformational ensembles, either through ensemble refinement or in combination with molecular dynamics simulations. Such crystallographically informed ensembles may better represent the conformational landscape accessible to the proteins than the Backrub-generated ensembles currently employed in MsC. Given the observed relationship between catalytic activity and maintenance of the designed active-site geometries, incorporating experimentally informed conformational ensembles into future iterations of MsC could improve the representation of catalytically relevant conformations during multistate optimization.

3.3 Experimental validation of the catalytic mechanism

The catalytic mechanism proposed in this work served as the basis for constructing the catalytic motif and therefore remains an important aspect to validate experimentally. Future studies should focus on determining whether the reaction proceeds through the intended iminium-mediated pathway or through an alternative mechanism. Formation of the proposed Schiff-base intermediate could be directly investigated through reductive trapping followed by mass spectrometric analysis, with peptide-level LC–MS/MS used to identify the residue involved in covalent adduct formation. Comparing FCS_1 with K259A would further determine whether Schiff-base formation is specifically dependent on the designed catalytic lysine or can be mediated by another nucleophile within the protein. Complementary isotope-labeling experiments using H₂O¹⁸ could probe the involvement of water in the alternative pathway, while solvent kinetic isotope effects could assess the contribution of proton-transfer events to catalysis. Time-resolved LC–MS or NMR analyses could additionally be used to detect and characterize reaction intermediates

associated with either pathway. Together, these experiments would help distinguish the intended iminium-mediated mechanism from alternative routes to product formation and establish whether the chemistry implemented by the designed active site corresponds to the mechanism used to construct the catalytic motif.

3.4 Evaluation of enantioselectivity

An important unresolved question concerns the stereochemical outcome of the reaction. The design calculations were performed using transition-state models corresponding to the target (*S*) enantiomer of the product, and the active site was therefore explicitly optimized to favor this stereochemical pathway. However, the enantiomeric excess of the reaction products was not determined experimentally. Measurement of product enantioselectivity therefore represents an important next step, as it would provide a stringent test of whether the designed active sites successfully encode the intended stereochemical preferences. Demonstration of selective formation of the target enantiomer would constitute strong evidence that the geometric information contained within the transition-state models is faithfully translated into catalytic function.

3.5 Alternative assessment of catalytic competency

An additional direction would be to evaluate catalytic competency using a structure-prediction-based sampling approach independent of molecular dynamics simulations. Mechanistically relevant covalently bound intermediates could be modeled with each design, and multiple structures generated using Boltz to sample alternative conformations of the corresponding enzyme–intermediate complexes. The same geometric criteria used to define catalytic competency could then be evaluated across these predicted structural

ensembles, providing an independent estimate of the population of catalytically competent conformations. Using a common intermediate across all designs would preserve direct comparability of the resulting competency values, while selection of an intermediate associated with a key catalytic step may provide a more informative representation than the product-bound state currently used for MD analysis. Comparison of Boltz-derived competency with both experimental activity and the MD-derived metric would help determine whether the observed relationship between catalytic geometry and activity is robust to the sampling method. Agreement between the two independent approaches would provide additional support for catalytic competency as a useful metric for evaluating and prioritizing MsC designs.

This thesis provides a proof of concept for the application of Multistate-CANVAS (MsC) to the *de novo* design of proteins targeting a complex multistep reaction lacking natural precedent. The enhanced product formation observed for the designed proteins supports the potential of the framework, while further experimental validation is required to establish the origin of this activity and the contribution of the intended catalytic architecture. Addressing these questions, together with further refinement of the computational methodology, provides a foundation for the continued development of MsC as a framework for the *de novo* design of enzymes for complex reactions beyond those catalyzed in nature.

References

1. Mak, W. S. & Siegel, J. B. Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
2. Edwards, D. R., Lohman, D. C. & Wolfenden, R. Catalytic Proficiency: The Extreme Case of S–O Cleaving Sulfatases. *J. Am. Chem. Soc.* **134**, 525–531 (2011).
3. Wolfenden, R. & Snider, M. J. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).
4. Bornscheuer, U. T. The fourth wave of biocatalysis is approaching. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**, (2018).
5. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nature Catalysis* 2020 3:3 **3**, 203–213 (2020).
6. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).
7. Zanghellini, A. de novo computational enzyme design. *Curr. Opin. Biotechnol.* **29**, 132–138 (2014).
8. Braun, M. *et al.* Computational enzyme design by catalytic motif scaffolding. *Nature* 2025 649:8095 **649**, 237–245 (2025).
9. Ahern, W. *et al.* Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods* 2025 23:1 **23**, 96–105 (2025).
10. Beck, J. *et al.* Customizing the Structure of Minimal TIM Barrels to Craft Efficient De Novo Enzymes. *bioRxiv* 2025.01.28.635154 (2026) doi:10.1101/2025.01.28.635154.
11. Garcia-Viloca, M., Gao, J., Karplus, M. & Truhlar, D. G. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. <https://www.science.org>.
12. Tantillo, D. J., Chen, J. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).
13. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16710–16715 (2006).

14. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* **15**, 2785–2794 (2006).
15. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De Novo Enzyme Design Using Rosetta3. *PLoS One* **6**, 19230 (2011).
16. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
17. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science (1979)*. **319**, 1387–1391 (2008).
18. Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science (1979)*. **329**, 309–313 (2010).
19. Richter, F. *et al.* Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).
20. Lauko, A. *et al.* Computational design of serine hydrolases. *Science (1979)*. **388**, (2025).
21. Watson, J. L. *et al.* De novo design of protein structure and function with RFDiffusion. *Nature 2023 620:7976* **620**, 1089–1100 (2023).
22. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science (1979)*. **378**, 49–56 (2022).
23. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods 2025 22:4* **22**, 717–723 (2025).
24. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).
25. Discovery, C. *et al.* Chai-1: Decoding the molecular interactions of life. *bioRxiv* 2024.10.10.615955 (2024) doi:10.1101/2024.10.10.615955.
26. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)*. **379**, 1123–1130 (2023).
27. Bolon, D. N. & Mayo, S. L. *Enzyme-like Proteins by Computational Design*. *PNAS December* vol. 4 www.pnas.org/cgi/doi/10.1073/pnas.251555398 (2001).
28. 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).

29. Dahiyat, B. I. & Mayo, S. L. De novo protein design: Fully automated sequence selection. *Science* (1979). **278**, 82–87 (1997).
30. Bar-Even, A. *et al.* The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).
31. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790–3795 (2012).
32. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nature Chemical Biology* 2013 9:8 **9**, 494–498 (2013).
33. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).
34. Listov, D. & Fleishman, S. J. De novo enzyme design: Controlling structure to design function. *Curr. Opin. Struct. Biol.* **98**, 103252 (2026).
35. Vaissier Welborn, V. & Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **119**, 6613–6630 (2019).
36. Zetsche, L. E. & Narayan, A. R. H. Broadening the scope of biocatalytic C–C bond formation. *Nat. Rev. Chem.* **4**, 334–346 (2020).
37. Sangster, J. J., Marshall, J. R., Turner, N. J. & Mangas-Sanchez, J. New Trends and Future Opportunities in the Enzymatic Formation of C–C, C–N, and C–O bonds. *ChemBioChem* **23**, e202100464 (2022).
38. Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Science* **21**, 1241–1252 (2012).
39. Ambroggio, X. I. & Kuhlman, B. Computational Design of a Single Amino Acid Sequence that Can Switch between Two Distinct Protein Folds. *J. Am. Chem. Soc.* **128**, 1154–1161 (2006).
40. Ashworth, J. *et al.* Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38**, 5601–5608 (2010).
41. Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins* **82**, 771–784 (2014).
42. St-Jacques, A. D., Eyahpaise, M. È. C. & Chica, R. A. Computational Design of Multisubstrate Enzyme Specificity. *ACS Catal.* **9**, 5480–5485 (2019).

43. C., F. & J.M., C. SUR UNE METHODE GENERALE NOUVELLE DE SYNTHBSE D'HYDROCARBURES, D'ACETONES, ETC. **84**, 1450–1450 (1877).
44. Gribble, G. W. . Indole ring synthesis : from natural products to drug discovery. (2016).
45. Welsch, M. E., Snyder, S. A. & Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **14**, 347–361 (2010).
46. Kaushik, N. K. *et al.* Biomedical Importance of Indoles. *Molecules* 2013, Vol. 18, Pages 6620-6662 **18**, 6620–6662 (2013).
47. Dunn, M. F. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Arch. Biochem. Biophys.* **519**, 154–166 (2012).
48. Para, N. A. & MacMillan, D. W. C. New Strategies in Organic Catalysis: The First Enantioselective Organocatalytic Friedel–Crafts Alkylation. *J. Am. Chem. Soc.* **123**, 4370–4371 (2001).
49. MacMillan, D. W. C. The advent and development of organocatalysis. *Nature* 2008 455:7211 **455**, 304–308 (2008).
50. MacMillan Imidazolidinone OrganoCatalysts Kit I | Sigma-Aldrich. <https://www.sigmaaldrich.com/CA/en/product/aldrich/674575?msocid=2d10feab e7f86ed21f39ee41e6526fd4>.
51. Leveson-Gower, R. B., Zhou, Z., Drienovská, I. & Roelfes, G. Unlocking Iminium Catalysis in Artificial Enzymes to Create a Friedel–Crafts Alkylase. *ACS Catal.* **11**, 6763–6770 (2021).
52. Brouwer, B., Della-Felice, F., Thunnissen, A.-M. W. H. & Roelfes, G. Genetically encoded 3-aminotyrosine as catalytic residue in a designer Friedel-Crafts alkylase †. <https://doi.org/10.1039/d5sc01055a> (2025) doi:10.1039/d5sc01055a.
53. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Science* **21**, 717–726 (2012).
54. Kang, Q. & You, S. L. Asymmetric Friedel–Crafts Alkylation Reactions. *RSC Catalysis Series* **2015-January**, 214–282 (2015).
55. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *NatCh* **9**, 50–56 (2017).

56. De Raffe, D., Martí, S. & Moliner, V. Understanding the Directed Evolution of De Novo Retro-Aldolases from QM/MM Studies. *ACS Catal.* **10**, 7871–7883 (2020).
57. Zarifi, N. Next-Generation Enzyme Design: Incorporation of Multiple Conformational and Chemical States. <https://doi.org/10.20381/RUOR-30941> (2025) doi:10.20381/RUOR-30941.
58. Title), L. S.-(No & 2015, undefined. The PyMOL Molecular Graphics System, Version 1.8. *cir.nii.ac.jp* <https://cir.nii.ac.jp/crid/1370576118684734988>.
59. PyMOL | pymol.org. <https://pymol.org/>.
60. GitHub - ikalvet/invrotzyme: Utility for building inverse rotamer assemblies out of a Rosetta matcher/enzdes constraint file. · GitHub. <https://github.com/ikalvet/invrotzyme>.
61. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
62. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the ($\beta\alpha$)8-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).
63. Huang, P. S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chemical Biology* 2015 12:1 **12**, 29–34 (2015).
64. GitHub - RosettaCommons/RFdiffusion: Code for running RFdiffusion · GitHub. <https://github.com/RosettaCommons/RFdiffusion>.
65. Wiese, J. G., Shanmugaratnam, S. & Höcker, B. Extension of a de novo TIM barrel with a rationally designed secondary structure element. *Protein Science* **30**, 982–989 (2021).
66. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **380**, 742 (2008).
67. 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).
68. Protabit, LLC. (2012). Triad [protein design software]. Pasadena, CA. <https://triad.protabit.com/>.
69. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* 2025.06.14.659707 (2025) doi:10.1101/2025.06.14.659707.

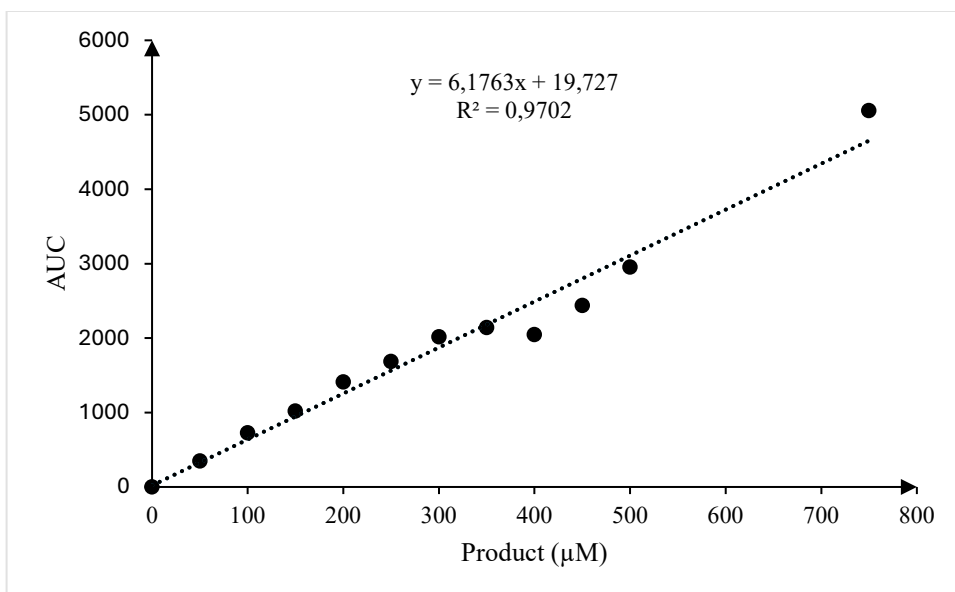
70. Shi, Y. *et al.* Abnormal SDS-PAGE migration of cytosolic proteins can identify domains and mechanisms that control surfactant binding. *Protein Science* **21**, 1197–1209 (2012).
71. Rath, A., Glibowicka, M., Nadeau, V. G., Chen, G. & Deber, C. M. Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1760–1765 (2009).
72. Crespi, A. F. *et al.* Reactivity of Imidazole- and Pyridine-Carboxaldehydes for Gem-Diol and Hemiacetal Generation: Theoretical and Experimental Insights. *ChemistryOpen* **14**, e202400411 (2025).
73. Tilgner, A. *et al.* Acidity and the multiphase chemistry of atmospheric aqueous particles and clouds. *Atmos. Chem. Phys.* **21**, 13483–13536 (2021).
74. Parmar, D., Sugiono, E., Raja, S. & Rueping, M. Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates. *Chem. Rev.* **114**, 9047–9153 (2014).
75. Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
76. Lee, F. S., Anderson, A. G. & Olafson, B. D. Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Engineering, Design and Selection* **36**, (2023).
77. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (2002).
78. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10172–10177 (1997).
79. Olsson, M. H. M., SØndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
80. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics* **78**, 1950–1958 (2010).
81. 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. *Journal of Chemical Physics* **153**, (2020).

82. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: a different approach. *ACS Publications* S Izadi, R Anandakrishnan, AV Onufriev *The journal of physical chemistry letters*, 2014 • *ACS Publications* **5**, 3863–3871 (2014).
83. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *Citation: J. Chem. Phys* **98**, 10089 (1993).
84. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, **25**, 247–260 (2006).
85. 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. *Journal of molecular biology*, **285**, 1735–1747 (1999).
86. Lzaguire, J. A., Catarello, D. P., Wozniak, J. M. & Skeel, R. D. Langevin stabilization of molecular dynamics. *pubs.aip.org* **114**, 2090–2098 (2001).
87. 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. *pubs.aip.org* **81**, 3684–3690 (1984).
88. Roe, D., and, T. C. I.-J. of chemical theory & 2013, undefined. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, **9**, 3084–3095 (2013).

Supplementary information

<u>Variants</u>	<u>Designed active sites</u>
FCS_1 FCS_2	
FCS_3	
FCS_4 FCS_5	
FCS_6 FCS_7 FCS_8	
FCS_9	
FCS_10	
FCS_11	

Supplementary figure 1. Designed active sites. Active sites of the designed FCS variants in complex with TS9. Residues optimized during active-site repacking are shown as orange sticks, and catalytic residues as blue sticks. The protein fold is shown in cartoon representation, with the lid colored in magenta.



Supplementary figure 2. Standard curve for quantification of the Friedel–Crafts alkylation product. AUC (area under the curve) obtained by reverse-phase HPLC is plotted as a function of product concentration (μM). The dotted line represents the linear regression fit ($y = 6.1763x + 19.727$, $R^2 = 0.9702$), used to convert measured AUC values to product concentrations in conversion assays.

Supplementary table 1. Amino acid sequences of the FCS variants

Enzyme	# residues	MW ^a (kDa)	Sequence ^b
FCS_1	286	31.1	GPEALAEELNAKLIEAHKALPDITIPRTRPLRDPVAAIREAAAEGRLLALLGYPPPTKITEEGVEAALRLGAAEDTI QRAIDSQTSVEEIVRVAQYVVTAILIAVDPIYGGSSFERIREIADKVDVGGVIAFALFVDEKQIDNAYNVGADAI AMSGSLPDELLRRLAEYARSIGLTPVMVATNLEQIERAARLGFPPIIIIFAGGSPGFAAEAAATPDGIPWDRVRE MVAAPDDQVKIVATDRLPPGALARLREMGVQAVLVKLEMFRLDELAKRLWEESHLEHHHHHHH
FCS_2	286	30.9	DPATLAALNAKLAEAMKALPEITIERTRPLRSPVEAIRRAAAEGRLLALLGYPPTEITEEGVEAALALGAAEGTI ETAKKNLTVERIVEVAPYVVTAILIAVDPIYGGSSFERIREIANRVDVGGVIAFGLFVDEKQIDNAYRVGADAI ALSHSLPDELLRELADYARSIGLTPVMVATNLEAIRRAAELGFPAILFAGGSPGFAAEAAAYPDGIPWDRVRE MVAAPDDQVRIVATPRLPPGALARLREMGVQAVLVKLELWRDLAHARRLWEESAALHHHHHHH
FCS_3	286	30.9	DPATLAELNAKRIAAMNALPEITVERTRPLRDPVAAIREERARGRLALLGYPPPTKITAEGVALALALGAAAGT IATAIRNQTTVEQIVEVAQYVSAILIAVDPIYGGSSFERIREIADKVDVGGVIAFALFNNKKQIDNAYRVGADAI ALSGSLPEEKLLLEAEYARSIGLTPIMVAFDLQIRLAELGFPAILFAGGSPGFAAEAAAYPDGIPWDRVREL VAAIPDDQVRIVATSRPPGALARLREMGVDAVLVLELWRDLALAKRLHEEAARLEHHHHHHH
FCS_4	286	31.3	GPEELARLNAKRIAAMKALPEITIERTRPLRSPVEAIRRAAEGRLLALLGYPPTEITEEGVEWALALGAAEGTI ETARRNLTTVEQIVEVAQYVVTAILIAVDPIYGGSSFERIREIADRVVGGVIAFALFTDEKQIDNAYRVGADAIA LSGSLPEELLRLRLAEYARSIGLTPIMVAFDLQIRLAELGFPAILFAGGSPGFAAEAAATPDGIPWDRVRE AAIPDDQVRIVSTGRPPGALARLREMGVDAVLVLELFRDLALAKRLWEESHRLHHHHHHH
FCS_5	286	30.9	DPATLAALNAKRAAMKALPEITVERTRPLRSPVEAIRRAAEGRLLALLGYPPTEITAEGVAAALALGAAAG TIATAIRSQASVEQIVEVAQYVDAILIAVDPIYGGSSFERIREIADKVDVGGVIAFALFVNEKQIDNAYRVGAD AIALSGSLPEELLRLRLAEYARSIGLTPIMVAFNLEQIRLAELGFPPIIIIFAGGSPGFAAEAAAYPDGIPWDRVRE MVAAPDDQVKIVSTSRPPGTLARLRAMGVQAVLVKLEMFRLDELAKRLWEESHLEHHHHHHH
FCS_6	286	31.3	DPEALAEELNAKRIAAMKALPVINTPRTRPLRDPVAAIREDRARGRLSLLGYPPTEITEEGVELALRLGAAEDFI ETAIRNQATVDEIVAVAPYVVTAILIAVDPIYGGSSFERIAEIASRVVGGVIAFALFVDRKQIDNAYAVGADAIA LSSSLPEELLRLRLAEYARSIGLTPIMVATDLEAVKRAAELGFPPIIIIFAGGSPGFAAEAAAYPDGIPWDRVRELVA AIPDDQVRIVATDRLPPGTLARLRAMGVDAVLVLELWRDLALARRLWEESRRLEHHHHHHH
FCS_7	286	31.7	GPEKLKELDAKRIEAMKALPEITTERTRPLRSPVEAIRERARGRLALLGYPPPTKITEEGVEWALS LGAAEDFI NEAIRNQATVEQIVEVAQYVVTAILIAVDPIYGGSSFERIREIADKVDVGGVIAFALFVDEKQIDNAYRVGADAI ALSGSLPDDLRLRLARYARSIGLTPIMVATNLEQIRRAAELGFPPIIIIFAGGSPGFAAEAAAYPDGIPWDRVREM VAAVPDEQVKIVATDRLPPGALARLREMGVDAVLVLELWRDLALAKRLWEESRRLEHHHHHHH
FCS_8	286	31.3	PPEVLAEVNAKRAEAMKALPEITVERTRPLRSPVEAIRERRARGELALLGYPPTEITEEGVALAKALGAAEDF IQEAIENLATVDRIVEVAPYVDAILIAVDPIYGGSSFERIREIADRVVGGVIAFALFVNNKKQIDNAYRVGADAI ALSDSLPDELLKELADYARSIGLTPIMVATNLEAIKRASELGFPPIIIIFAGGSPGFAAEAAAYPDGIPWDRVREM VAAIPDDQVRIVATDRLPPGALARLRAMGVDAVLVLELWRDLALAKRLWEESQQLHHHHHHH
FCS_9	286	31.2	DPAKLAELNAKRIAAMKALPEITVERTRPLRDPVAAIREARARGELALLGYPPTEITEEGVKWALS LGAAEDT INTAIESQATVEQIVEVAKYVVTAILIAVDPIYGGSSFERIREIRDKVDVGGVIAFALFVDEKQIDNAYRVGADAI ALADSLPRELLKRLADYARSIGLTPIMVATNLEAIRLAAELGFPAILFAGGSPGFAAEAAAYPDGIPWDRVRE MVAAPDDQVKIVSTSSLPGLTLARLRAMGVDAVLVLELFRDLALAKRLWEESQALEHHHHHHH
FCS_10	286	30.9	GPEELARLNAKRAEAMKALPEITVERTRPLRSPVEAIREARAGDLALLGYPPPTKITAEGVALALALGAAAG TIATAIKNQTTVEQIVAVAPYVVTAILIAVDPIYGGSSFERIREIADRVVGGVIAFGLFVDEKQIDNAYNVGADA IALSSSLPEDLLRLRLAEYARSIGLTPIMVAWDLDAIRLAARLGFPPIIIIFAGGSPGFAAEAAATPDGIPWDRVRE MVAAPDDQVRIVSTSLPPGALARLREMGVDAVLVLELWRDLALAKRLWEESRRLEHHHHHHH
FCS_11	286	31.0	GPEALAAALRLRLAEAMRALPEITTPRTRPLRDPVAAIRRAAEGRLLALLGYPPTEITAEGVAWALS LGAAAGT IATAIKNLATVEIVKVAPYVVTAILIAVDPIYGGSEKIREIAYKVDVGGVIAFGLFVNEKQIDNAYRVGADAI ALSHWLPRELLRLRLAEYARSIGLTPIMVVTDLGKVLASELGFPPIIIIFAGGSPGFAAEAAAYPDGIPWDRVRE LVAAPDDQVKIVSTSRPPGALARLREMGVQAVLVKLEMWDRDLALAKRLWEESQALEHHHHHHH

^a Molecular weight

^b Catalytic residues are underlined and bolded.

Supplementary table 2. Purification yields.

Enzyme	Yield (mg L⁻¹) ^a
FCS_1	15 ± 2
FCS_2	10
FCS_3	—
FCS_4	8 ± 1
FCS_5	14 ± 1
FCS_6	9 ± 1
FCS_7	20 ± 4
FCS_8	16
FCS_9	17 ± 9
FCS_10	9
FCS_11	—
Y51F	17
Y106F	—
E208A	—
K259A	15 ± 8

^a Data represent the average measurements from 1 to 2 independent protein batches, with standard deviation. FCS_3, FCS_11, Y106F and E208A, could not be expressed solubly.

Supplementary table 3. 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	Glu	Distance	OE1 or OE2	H70			1.0	2.5
		Angle	OE1 or OE2	H70	N44		145	180
		Angle	CD	OE1 or OE2	H70		105	125
		Torsion	CB	CD	OE1 or OE2	H70	150	190
H-bond donor	Tyr_1	Distance	HH	O24			1.0	3.0
		Angle	OH	HH	O24		150	180
		Angle	HH	O24	C10		140	160
	Tyr_2	Distance	HH	C10			1.0	3.0
		Angle	OH	HH	C10		160	180
		Angle	HH	C10	C12		120	140

* The Lysine was covalently bound to all the transition states and modeled as a non canonical amino acid. Different angles and torsion angles were allowed in the residue attribute file.

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

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

Supplementary table 4. Amino-acid and positions designed during active-site repacking.

Designed positions	Allowed amino acids
5	ILE, LEU, ALA, VAL, TRP, PHE
49	ILE, LEU, ALA, VAL, TRP, PHE
52	ILE, LEU, ALA, VAL, TRP, PHE, PRO
56	ILE, LEU, ALA, VAL, TRP, PHE
61	ILE, LEU, ALA, VAL, TRP, PHE
64	ILE, LEU, ALA, VAL, TRP, PHE
67	ILE, LEU, ALA, VAL, TRP, PHE
68	ILE, LEU, ALA, VAL, TRP, PHE, GLY, ARG
69	ILE, LEU, ALA, VAL, TRP, PHE
70	ILE, LEU, ALA, VAL, TRP, PHE
73	ILE, LEU, ALA, VAL, TRP, PHE, THR, GLN
77	ILE, LEU, ALA, VAL, TRP, PHE, GLY
98	ILE, LEU, ALA, VAL, TRP, PHE
100	ILE, LEU, ALA, VAL, TRP, PHE, THR
105	ILE, LEU, ALA, VAL, TRP, PHE
126	ILE, LEU, ALA, VAL, TRP, PHE
128	ILE, LEU, ALA, VAL, TRP, PHE
130	ILE, LEU, ALA, VAL, TRP, PHE
149	ILE, LEU, ALA, VAL, TRP, PHE
151	ILE, LEU, ALA, VAL, TRP, PHE, SER
153	ILE, LEU, ALA, VAL, TRP, PHE, SER
174	ILE, LEU, ALA, VAL, TRP, PHE
176	ILE, LEU, ALA, VAL, TRP, PHE
178	ILE, LEU, ALA, VAL, TRP, PHE
195	ILE, LEU, ALA, VAL, TRP, PHE
197	ILE, LEU, ALA, VAL, TRP, PHE, ARG
199	ILE, LEU, ALA, VAL, TRP, PHE, GLY
200	ILE, LEU, ALA, VAL, TRP, PHE, GLY
201	ILE, LEU, ALA, VAL, TRP, PHE, SER
205	ILE, LEU, ALA, VAL, TRP, PHE, LEU
209	ILE, LEU, ALA, VAL, TRP, PHE
210	ILE, LEU, ALA, VAL, TRP, PHE
236	ILE, LEU, ALA, VAL, TRP, PHE, SER

257	ILE, LEU, ALA, VAL, TRP, PHE
260	ILE, LEU, ALA, VAL, TRP, PHE

Supplementary table 5. Active-site energies of the FCS variants across transition states.

Active-site sequence	TS1 energy	TS2 energy	TS4 energy	TS5 energy	TS6 energy	TS7 energy	TS8 energy	TS9 energy
FCS_1, 2	-394.612	-319.288	-364.017	-765.221	-367.502	-344.159	-270.741	-362.226
FCS_3	-393.209	-324.151	-358.374	-756.755	-365.371	-355.399	-267.724	-368.951
FCS_4, 5	-393.960	-313.032	-366.056	-760.210	-358.438	-341.052	-261.413	-348.093
FCS_6,7,8	-384.210	-308.676	-364.961	-753.512	-362.623	-328.686	-266.514	-361.365
FCS_9	-386.937	-304.169	-352.850	-756.971	-363.319	-330.180	-261.753	-347.354
FCS_10	-385.315	-308.867	-353.424	-756.729	-351.837	-339.277	-257.778	-342.711
FCS_11	-385.900	-310.060	-355.666	-749.982	-347.661	-326.069	-251.439	-350.279

Supplementary table 6. Triad filtering parameters for the FCS variants.

Active-site sequence	(ΔE)	Predicted pKa range	Relative SASA range (%)
FCS_1, 2	494.480	7.58–8.91	3.63–12.26
FCS_3	489.030	7.54–8.96	5.96–14.59
FCS_4,5	498.797	7.52–8.84	4.28–11.25
FCS_6, 7, 8	486.998	7.36–8.90	3.63–10.74
FCS_9	495.218	7.56–8.87	4.28–12.11
FCS_10	498.951	7.54–8.84	5.15–14.63
FCS_11	498.543	7.53–8.87	5.13–14.16

Supplementary table 7. Boltz-2 filtering parameters for the FCS variants using the substrate-bound complexes.

Sequence	Average backbone RMSD	Substrate RMSD	Average backbone pLDDT	Big lid pLDDT	Small lid pLDDT	Big lid RMSD	Small lid RMSD
FCS_1	1.1	2.3	0.9	0.9	0.9	2.0	1.2
FCS_2	1.0	1.6	0.9	0.9	0.8	1.9	1.9
FCS_3	1.2	2.2	0.9	0.9	0.8	2.1	1.4
FCS_4	1.5	1.8	0.9	0.9	0.9	2.9	1.2
FCS_5	1.4	2.3	0.9	0.9	0.9	2.1	2.0
FCS_6	1.1	0.8	0.9	0.9	0.9	2.1	1.2
FCS_7	1.3	2.9	1.0	1.0	0.9	1.9	1.0
FCS_8	1.1	1.6	0.9	0.9	0.9	2.4	1.1
FCS_9	1.5	1.7	0.9	0.9	0.9	2.7	0.8
FCS_10	1.5	2.2	0.9	0.9	0.8	2.6	1.0
FCS_11	1.5	3.4	0.9	0.9	0.9	2.3	1.0

Supplementary table 8. Boltz-2 filtering parameters for the FCS variants using the product-bound complexes.

Sequence	Average backbone RMSD	Product RMSD	Average backbone pLDDT	Big lid pLDDT	Small lid pLDDT	Big lid RMSD	Small lid RMSD
FCS_1	1.3	2.6	0.9	0.9	0.9	2.9	1.0
FCS_2	1.2	3.4	0.9	0.9	0.8	2.5	0.9
FCS_3	1.4	2.6	0.9	0.9	0.9	2.9	0.7
FCS_4	1.3	1.8	0.9	0.8	0.8	2.8	0.6
FCS_5	1.5	2.8	0.9	0.9	0.8	2.5	0.9
FCS_6	1.4	2.7	0.9	0.9	0.8	2.6	1.1
FCS_7	1.3	2.2	0.9	0.9	0.9	2.2	0.7
FCS_8	1.5	2.9	0.9	0.9	0.9	2.8	0.6
FCS_9	1.5	2.5	0.9	0.8	0.9	2.9	0.6
FCS_10	1.4	2.0	0.9	0.8	0.8	2.9	1.0
FCS_11	1.6	2.3	0.9	0.8	0.8	2.6	1.0

Supplementary table 9. Boltz-2 filtering parameters and cutoff values.

Metric	Type	Min	Max
Ligand RMSD ^{a, x}	RMSD	–	5.0
Triad vs Boltz backbone RMSD ^{b, x}	RMSD	–	2.0
Lid RMSD ^{c, z}	^f Big lid RMSD	–	3.0
	^g Small lid RMSD	–	2.0
Backbone pLDDT ^{d, z}	pLDDT	90	–
Lid pLDDT ^{e, z}	^f Big lid pLDDT	80	–
	^g Small lid pLDDT	80	–

^a RMSD between the centroid of the benzene rings of the substrate and product in the Boltz2 model and transition-state 1 and 9 in the Triad model.

^b RMSD between the backbones of Triad models and the Boltz-2 predictions with substrate and product residues.

^c RMSD between designed lids of Triad models with transition states and the Boltz-2 predictions with substrate and product (51-86 and 196-211).

^d pLDDT values Boltz-2 predictions, represented by values in the b-factors column.

^e pLDDT values of designed lids in Boltz-2 predictions.

^f The big lid is the top part of the designed lid with the region spanning residues 52 to 86.

^g The small lid the lower part of the lid that anchors the Glutamate, with the region spanning residues 196 to 211.

^x RMSD measurements are given in Å.

^z pLDDT values are arbitrary and range from 0 to 100.