Designing consensus catalytic geometries for *de novo* multistep enzymes

Juan Sebastian Alvarez

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
Ottawa, Ontario, Canada

Abstract

Natural enzymes that catalyze multistep reactions possess highly organized active-site geometries that require minimal atomic rearrangement during catalysis (<1.5 Å inter-state average RMSD). While *de novo* enzymes for multistep reactions have previously been designed, none have explicitly leveraged this principle - a limitation likely contributing to their poor catalytic performance. In this thesis, I introduce a computational framework for the design of multistep *de novo* enzymes based on pre-optimized consensus geometries, using a Michael addition as a model system. Using density functional theory, theozymes were constructed for all seven transition states of a proposed mechanism mediated by a Lys/Tyr/His catalytic triad. Optimizing for a shared consensus geometry yielded a theoretical pathway with an average inter-state RMSD of 1.1 Å, successfully mirroring the organization of natural multistep enzymes. To physically realize this theoretical model, the consensus theozymes were embedded into a structural ensemble derived from a TIM-barrel scaffold. However, experimental characterization revealed that the designed sequences failed to adopt the target fold, largely reflected by inclusion body formation. Consequently, the viability of the consensus theozyme hypothesis remains experimentally untested. Nevertheless, this negative result stress-tests the design hypothesis and isolates limitations in the protein backbone redesign workflow. By diagnosing these failure points, this work establishes a foundation for deriving multistep consensus theozymes and provides an actionable roadmap for their future implementation using *de novo* generative scaffolding models.

Acknowledgments

The work presented in this thesis would not have been possible without the following individuals, to whom I give my sincerest gratitude.

Firstly, I would like to thank Dr. Roberto Chica for his mentorship and guidance. Roberto gave me the opportunity to pursue an education in chemistry and computational research. His guidance, patience, understanding and empathy were the key factors that enabled me, an engineer with no formal biology education, to navigate the beauty of enzymatic chemistry. I will miss your tutelage as you pursue your new position at the University of Manchester, and I hope we will continue to work together in the future.

To my lab members - Dr. Hang Pham, Dr. Cindy Klaus, Dr. Niayesh Zarifi, Dr. Amir Bunyat-Zada, Ben Smith, Rui Guo, and Ahana Harsha - I am infinitely grateful for every conversation and moment spent together. Your companionship brought so much joy as I began a new life in Ottawa. I hope we continue to be friends for a long time.

To my parents, Octavio Alvarez and Nancy Amaya: your support and dedication to education have been the bedrock of my career. Your boundless love and indomitable will have been my guiding light. To my sister and best friend, Natalia Alvarez: I thank you for every laugh, every text, and every embrace. I am so lucky to be your brother.

To my wife, Dr. Anika Zaman: Thank you for your company every late night while I was in the lab, coding, or writing. Thank you for listening to every rant and ramble, and for being my source of strength.

Finally, to my grandfather, Carlos Amaya. This thesis is dedicated to you. As I wrote, I thought about you every day. Your passion for education has inspired hundreds, and within me lives the values you taught for over half a century. I am but a ripple from the stones you cast throughout your life.

Thank you to all, for everything and more.

Statement of Authorship

All research described in this thesis was conducted by the author, except for the preparation of ensemble-refinement calculations of the indole-3-glycerolphosphate synthase structures (PDB ID: 1A53) in Chapter 3. These calculations were conducted by Dr. Niayesh Zarifi.

Table of Contents

List of Figures

Chapter 1: Background and Overview

1.1 Introduction

Richard Feynman famously wrote on his blackboard, “What I cannot create, I do not understand” [25]. This dictum defines the underlying ethos of *de novo* enzyme design. While the natural protein universe encompasses a vast landscape of biological reactions, it rarely overlaps with the synthetic transformations demanded by modern industry [26]. To address this, *de novo* enzyme design seeks to create bespoke biocatalysts for new-to-nature reactions. Beyond industrial utility, building enzymes from scratch serves as the ultimate test of our fundamental understanding of enzymatic catalysis.

The foundational challenge of *de novo* enzyme design lies in the precise structural scaffolding of a catalytic residue arrangement - a spatial motif that represents a specific mechanistic hypothesis for the target reaction. Historically, enzymes designed in this manner exhibited modest catalytic activity. This underperformance is largely attributed to structural inaccuracies, specifically the failure of static scaffolds to maintain sub-angstrom organization of the catalytic geometry in solution. While recent advancements in artificial intelligence have revolutionized protein structure prediction and generation, these tools primarily reflect improvements in sequence-structure mapping rather than the implementation of new catalytic design principles [27]. As such, generating scaffolds that can accommodate catalytic function remains a frontier challenge.

This challenge is magnified in multistep reactions. Natural enzymes that catalyze multi-transition-state reactions possess highly organized active sites capable of accommodating shifting geometric requirements with minimal atomic rearrangement throughout the catalytic cycle [1,2]. Although *de novo* enzymes for multistep reactions have been reported, none have explicitly leveraged this

principle - a structural limitation that may contribute to poor catalytic performance. The focus of this thesis is to develop a computational framework that explicitly maps and scaffolds these shifting geometric requirements into a single minimal-RMSD pathway, a paradigm herein referred to as *geometric consensus.*

The content of the thesis is as follows. Chapter 1 focuses on the computational background of *de novo* enzyme design, a brief history of the field with an emphasis on relevant case studies of enzymes designed for multistep reactions, and the specific aims of this work. Chapter 2 focuses on the computational design of an enzymatic Michael addition mechanism, the model multistep reaction focused on in this work. Chapter 3 focuses on the computational design workflow and experimental evaluation to design Michaelase enzymes around the proposed mechanism in Chapter 2. Finally, Chapter 4 is a discussion and conclusion of the computational and experimental results throughout the work.

1.2 Computational Methods in *De Novo* Enzyme Design

The methodological principles underlying *de novo* enzyme design are a direct extension of computational protein design (CPD). CPD methods involve the identification of protein sequences that can adopt a specific three-dimensional fold and possess a desired property such as increased stability, binding affinity, or altered specificity [28]. In computational enzyme design, CPD methods are employed for the construction of a catalytic site into a selected protein scaffold that does not natively bind the target substrate or catalyze the reaction. As one of the key ways that enzymes navigate catalysis is through transition state stabilization, the design objective of computational enzyme design is the identification of sequences that can: Orient the designed catalytic residues towards the transition state complex in a catalytically competent arrangement, and stabilize the

transition state complex via active site packing. Historically, these protocols are calculated using combinations of physics-based potential energy functions and quantum mechanics (QM). While recent advancements in machine-learning models have supercharged the development of the field, the general philosophy of transition state complex orientation and organization remains at the heart of *de novo* design objectives. Here, the critical concepts underlying the methodologies are overviewed.

1.2.1 Computational Protein Design Through Stability Scoring

Most CPD algorithms consist of four steps. First, a protein fold is selected based on its relevance to the design objective. Second, amino acid sidechains are modelled using discrete rotamer libraries at selected positions on the template. Energy calculations - typically from empirically derived potential energy functions - evaluate the interactions between each designed rotamer and the template (one-body energies), as well as between pairs of designed rotamers (two-body energies), where the sum of the energies represent the compatibility of an amino acid identity at a given position relative to both the protein backbone and all other residues on the protein. The potential energy functions enable direct comparison of different sequences. Finally, a sequence optimization algorithm navigates the combined rotamer and sequence space, identifying sequences that yield the most stable *in silico* energy scores [28].

1.2.2 Multistate Sequence Design Principles

Traditionally, CPD relies on single-state design (SSD), optimizing an amino acid sequence to stabilize a single, fixed backbone template. While SSD is effective for rigid structural targets, it is ill-equipped for applications that rely on the natural conformational fluctuations of proteins. Multistate Sequence Design (MSD) accounts for conformational plasticity by evaluating a single amino-acid sequence across multiple chemical or structural states simultaneously.

Computationally, this is achieved by performing independent rotamer optimizations for each structural state within the ensemble, allowing the target sequence to adopt distinct optimal-side chains in response to variations in the protein backbone. The individual energetic scores from these discrete states are aggregated into a unified fitness function. By optimizing this aggregate fitness value, MSD drives the selection of sequences that are energetically compatible with multiple static substates. This provides the computational framework for design objectives that arise as a function of the proteins' conformational fluctuations. In enzyme design, multistate design has been utilized as the computational framework for substrate selectivity and enantioselectivity for single-state reactions, as well as for engineering the exchange of conformational states [5,29,30,44,58].

1.2.3 Transition State Theory and Calculation of the Theozyme

Transition state theory (TST) models the rates of elementary chemical reactions. According to TST, a reaction is defined by the quasi-equilibrium between reactants and an activated complex situated at the saddle point of the potential energy surface [31]. The activation energy required to reach this high-energy transition state (TS) is quantified by the Eyring equation[32]. A foundational assumption of computational enzyme design is the utilization of the theozyme - a direct structural extension of TST.

Transition states are frequently computed using density functional theory (DFT). Unlike wavefunction theory, which treats each electron individually, DFT utilizes a functional to map the electron density to the system's ground-state energy. This approach relies on the theorem that all ground-state properties of an electronic system are uniquely determined by its electron density. By treating electron density as a continuous functional, the ground state is defined by the energy minima of that functional [33]. These DFT models describe the relationship between a systems'

energy and its geometry. As such, the task of computing either a ground state or transition state manifests as a geometry optimization problem. Here, the DFT algorithm seeks out stationary points on the potential energy surface. Ground states are defined as minima, where the potential energy rises locally in all directions along the relevant geometric reaction coordinate. Conversely, the transition state is located at a saddle point of the energy surface, corresponding to an energy maxima in one direction - along the reaction coordinate - and a minimum across all others. Because these definitions apply to both local and global stationary points, the convergence of a DFT algorithm does not guarantee that the global minimum for the molecular system has been found. As the number of atoms in a system increases, the potential energy surface becomes exponentially more complex, generating a landscape with a vast number of stationary points [33].

TST is a design principle used in enzyme design, specifically through the theozyme. Houk first described the concept of the theozyme model as a transition state structure combined with the minimally required catalytic residue atoms arranged in an optimal orientation to stabilize the TS. Theozymes are typically calculated using QM methods, and most commonly via DFT [34]. Fundamentally, theozyme calculations are similar to computing a non-enzymatic transition state. The primary distinction is that the functional groups selected to facilitate catalytic motion correspond directly to amino-acid side chains or cofactors. The DFT-optimized geometries are then used as design inputs to protein design algorithms. This marks a distinct departure from the potential energy functions used during standard protein design, as molecular mechanics force fields do not explicitly model electronic terms. As such, the theozyme is typically computed prior to designing the active site.

To construct a theozyme, an initial ‘guess’ geometry is specified for both the transition state and the catalytic groups, based either on existing structural knowledge or chemical intuition. This

geometry is then optimized to a first-order saddle point, characterized by a single imaginary frequency. Practically, the QM software backend manages the optimization through iterative perturbation and calculation of the electron density, until convergence criteria corresponding to the stationary point are achieved. Practically, this process is managed internally by the QM software itself, through workflows such as the Berny algorithm [35]. To verify the reactants and product connected by this saddle point, an Intrinsic Reaction Coordinate (IRC) calculation is performed. The IRC traces the reaction pathway starting from the saddle point, displacing the system geometry downhill in both directions toward the closest stable intermediates [36]. The resulting theozyme represents the transition state complex, with the side chain reactive atoms optimally oriented for catalysis. This geometry defines the structural criteria of the design objective.

1.2.4 Ligand Placement and Matching Algorithms

The design objective of computational enzyme design is to recapitulate the catalytic contacts present in the DFT-derived theozymes, into the selected protein scaffold using catalytic residues. Given that catalysis cannot be explicitly designed in protein design software as molecular mechanics forcefields do not account for electronic structure, the selected active site is designed to stabilize the geometric poses defined in the theozyme. Each catalytic contact between the catalytic side chain and TS atoms are defined by 6 geometric degrees of freedom [34,37]**.** These contacts are defined as constraints to identify positions on the template backbone, where the necessary catalytic contacts can be made to the transition state. Matching algorithms are then utilized to generate these placements. Initially, a protein backbone template and theozyme ligand must be provided as inputs. Catalytic side chain rotamers and TS binding poses are sampled during

the calculation. Rotamers are generated using user-defined torsional degrees of freedom, while TS poses are built off of the catalytic side chains.

1.2.5 *Machine Learning Directed Design*

The advent of Machine Learning (ML) tools are a tremendous force in the evolution of enzyme design. Availability of ML models has enabled access to sequence-structure space that is unexplored by natural evolution, enabling the design of new protein and enzyme systems. Computational design workflows utilizing ML design tools nominally follow three steps. The generation of a novel protein backbone using diffusion models, with the goal of scaffolding catalytic motifs or completely unconditioned backbones [59,60]. Then, inverse-folding tools are used to design sequences that are predicted to adopt the target backbone fold [61–64]. The validation of the designed sequences is often conducted by using a structure prediction model, where the objective is identifying sequences that adopt the selected backbone structure at high confidence [65–67].

1.3 *De Novo* Enzyme Design for Multistep Reactions

1.3.1 A brief history of *de novo* enzyme design

The field of computational enzyme design has undergone rapid evolution over the past two decades. To contextualize this current state, it is valuable to briefly examine the progression of the discipline and highlight the foundational paradigms that have driven its success. One of the earliest proof-of-concept studies demonstrating that enzyme-like catalytic activity could be computationally introduced into an inert protein scaffold was reported by Bolon and Mayo [38]. Targeting the ester hydrolysis of *p*-nitrophenyl acetate, the authors modeled the multistate reaction using a single tetrahedral intermediate as a transition state proxy. Here, a ligand-placement algorithm was used to scan for optimal geometric positions of a nucleophilic histidine within the

active site, which successfully generated functional biocatalysts. While the resulting catalytic activity was modest, this study established the benchmark proving the viability of rational *de novo* enzyme design.

In subsequent years, the Kemp elimination emerged as a model system for the field. Given it proceeds via single-step proton abstraction and concerted bond cleavage, it serves as an ideal benchmark for design strategies focused on transition state stabilization. Most early studies used a variation of the standard ligand-matching approach, which yielded minimally active enzymes that required extensive directed evolution [39–41]. At the same time, features of highly active enzymes began to emerge from structural dynamics-type studies. The correlation between active site preorganization and catalytic proficiency was also observed in a number of studies [42,43]. Broom and Rakotoharisoa evaluated the fluctuations that arise from crystallographically restrained dynamics, and explored the concept of ensemble-based design [30,44]. These approaches used a library of scaffolds of the same protein template as input to the ligand-placement algorithm. This approach led to 250-fold order of magnitude improvement in catalytic efficiency relative to the starting template, approaching the performance of natural enzymes (median natural $k_{cat}/K_M \sim 10^5$ $M^{-1}s^{-1}$). As *de novo* enzymes are notorious for reduced activity compared to their natural counterparts, these results provided a valuable design technique. The work demonstrates the viability of designing the active site as dynamic states, capable of sampling productive conformational substates, and the subsequent improvement in the designed activity.

Bond-forming reactions have also been explored in *de novo* enzyme design. One of the first examples of such a system was demonstrated for the single-step Diels-Alder reaction [45]. The complexity of this system is non-trivial, as reactivity is dependent on the precise orientational control of two substrates. In this example, designs did not exceed bisubstrate catalytic efficiencies

of 10 $M^{-2}s^{-1}$. More recently, the field has explored utilizing deep-learning generative models to construct active sites and scaffolds from scratch [46]. In one example, machine-learning hallucination algorithms were used to design entirely novel scaffolds for luciferases. While this approach successfully generated tailored active-site pockets with high shape complementarity for the anionic substrates, the success rate of their design library was notably low (~0.04%). This metric underscores a recurring theme in the field observed over many examples: building a structurally complementary pocket is not synonymous with explicitly designing competent catalytic machinery.

The trajectory of *de novo* enzyme design over the past two decades represents immense progress in protein engineering. For unimolecular bond cleavages and single-state transformations, computational strategies - particularly those incorporating conformational ensembles - have successfully yielded biocatalysts with efficiencies approaching those of natural enzymes. However, extending these principles to multistep, bond-forming reactions presents a fundamentally different paradigm. These complex reactions impose stringent demands on the active site, requiring precise orientational control of multiple substrates, the stabilization of sequential and geometrically distinct transition states, and the accommodation of specific conformational dynamics along the reaction coordinate. Because the field has historically relied on single-state design algorithms, the explicit computational design of true multistep mechanisms remains in its infancy. To evaluate the limitations of modern design strategies when applied to these systems, three studies were selected and examined as relevant efforts in the design of multistep enzymes. These case studies highlight both the impressive structural capabilities of modern generative algorithms and the persistent limitations of current static design approaches.

1.3.2 Case study 1: Computational Design and Evolution of Retroaldolases

One of the most rigorously studied multistate mechanisms in *de novo* enzyme design is from the RA95 retro-aldolase family, pioneered by the Baker and Hilvert groups. The retro-aldol mechanism proceeds via enamine catalysis utilizing a catalytic lysine. The reaction initiates with a nucleophilic attack by the lysine ε-amine on the substrate ketone to form a tetrahedral carbinolamine intermediate. Subsequent water elimination yields a covalently bound Schiff base, which acts as an electron sink to facilitate the deprotonation of the β-hydroxyl group and the concomitant C-C bond cleavage. This releases the first product (6-methoxy-2-napthaldehyde) (6MNA) and generates an enamine intermediate, which then tautomerizes back to an imine before undergoing hydrolysis to release the final product and regenerate the apo-state [15].

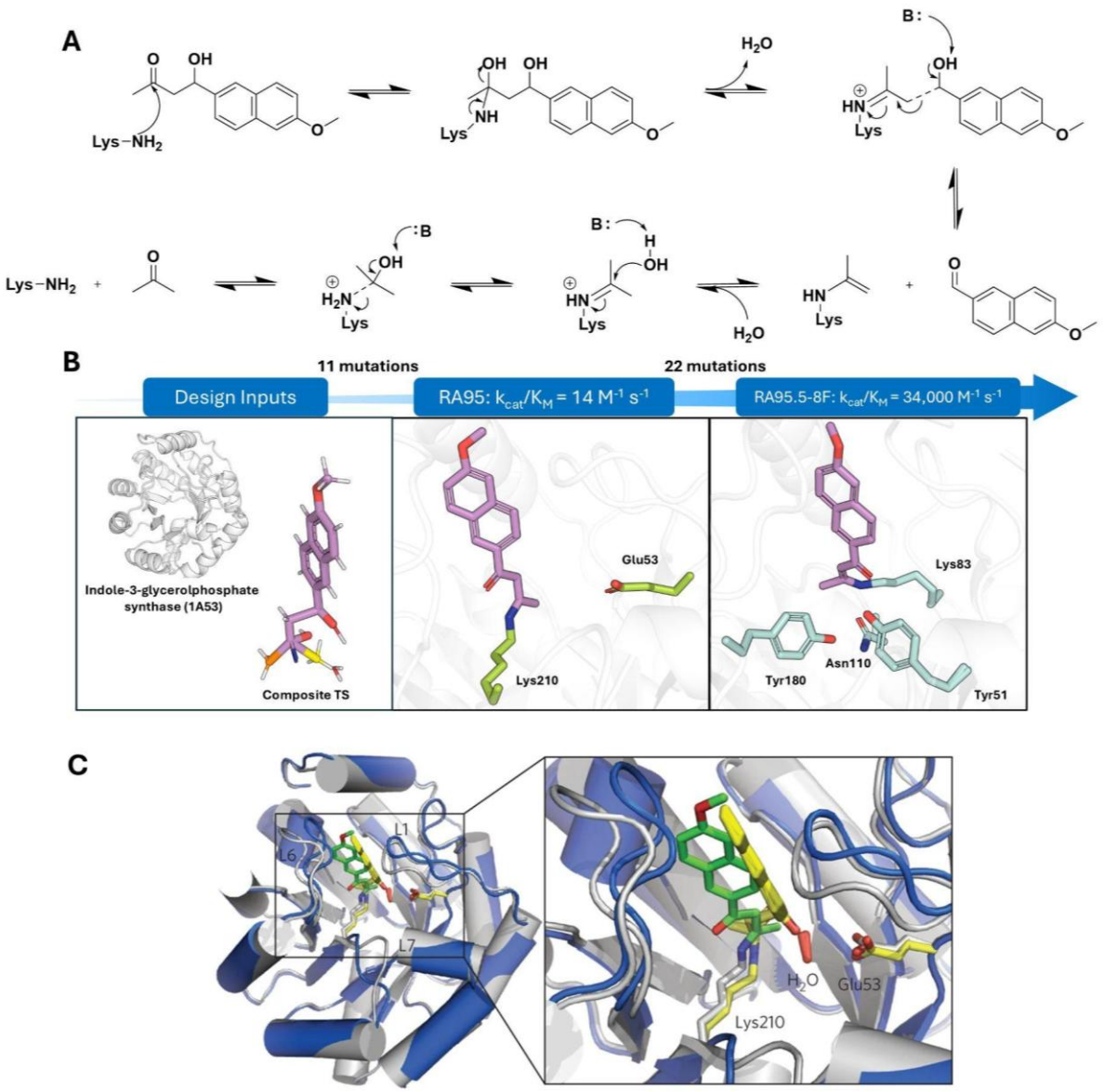

Figure 1.1: Design and evolution of the RA95 family of retroaldolases

A.) Proposed enamine-mediated reaction mechanism of the retro-aldol reaction on methodol, utilized to guide the theozyme and *de novo* enzyme design of RA95. **B.)** Overview of evolutionary trajectory of RA95 to RA95.5-8F. Composite theozyme model (sticks, light pink) and indole-3-glycerolphosphate synthase (cartoon, white) are used as design inputs to the initial *de novo* design. Catalytic residues in the active site (sticks, light green) are evolved from K210/E53 in RA95, to Y51/K83/N110/Y180 in RA95.5-8F (sticks, light blue), labelled with the corresponding catalytic efficiency of the variant. **C.)** Overlay of the *in silico* design model of RA95 (cartoon, blue) with the composite theozyme (sticks, yellow) and the X-ray crystal structure of RA95 (cartoon, grey) with a covalently-bound mechanistic inhibitor (sticks, green). Figure reproduced from [16].

Initial computational designs by Jiang and Althoff attempted to scaffold this mechanism using a 'composite transition state' strategy [15,47]. To account for all of the catalytic requirements of the complete mechanism, the authors geometrically superimposed the structures of the carbinolamine, the water-elimination transition state, the Schiff base intermediate, and the C-C bond-breaking transition state into a single model. However, forcing an amalgamation of temporally distinct transition states into a static model creates an unphysical geometric chimera. This approach inflates the steric footprint of the ligand, demanding a swollen active site that compromises precise packing. Consequently, while the design featured a catalytic water molecule explicitly oriented by the glutamate residue (Glu53), the resulting enzymes also exhibited poor catalytic efficiency. The flaw in this static composite approach was illuminated by Giger et al: crystallographic characterization of RA95 bound to a mechanism-based inhibitor revealed that while the naphthyl ring occupied the designed hydrophobic pocket, it was rotated 114 degrees around its' long axis, pointing the reactive carbonyl away from the designed Glu53-water proton shuttling network [16]. Limitations of the composite design were resolved through extensive directed evolution [17]. Over 19 rounds of optimization, RA95 was evolved into the highly proficient variant RA95.5-8F. Notably, this evolutionary trajectory completely remodeled the original catalytic motif. While the lysine-mediated iminium was conserved, its position within the active site was moved to the opposite face of the active site. Furthermore, the rigid glutamate-water network was abandoned in favor of a dynamically coupled catalytic tetrad composed of the repositioned lysine, two tyrosines, and an asparagine. Subsequent QM/MM studies confirmed that the rate-limiting steps in these evolved enzymes dictate rapidly shifting electrostatic and geometric demands throughout the catalytic cycle - features that a static structural composite fails to capture [48,49]. The RA95 trajectory

demonstrates that by optimizing a rigid scaffold around a static composite theozyme cannot capture the dynamic networks capable of navigating a sequential reaction coordinate.

1.3.3 Case study 2: Computational Design and Evolution of Morita-Baylis-Hillmanases

The Morita-Baylis-Hillman (MBH) reaction is a challenging carboligation of interest to *de novo* design due to its capacity for stereocenter generation and its complex multistep mechanism. The reaction proceeds via nucleophilic Michael addition at the β-carbon of a conjugated carbonyl, followed by C-C bond formation at the α-carbon and subsequent product release.

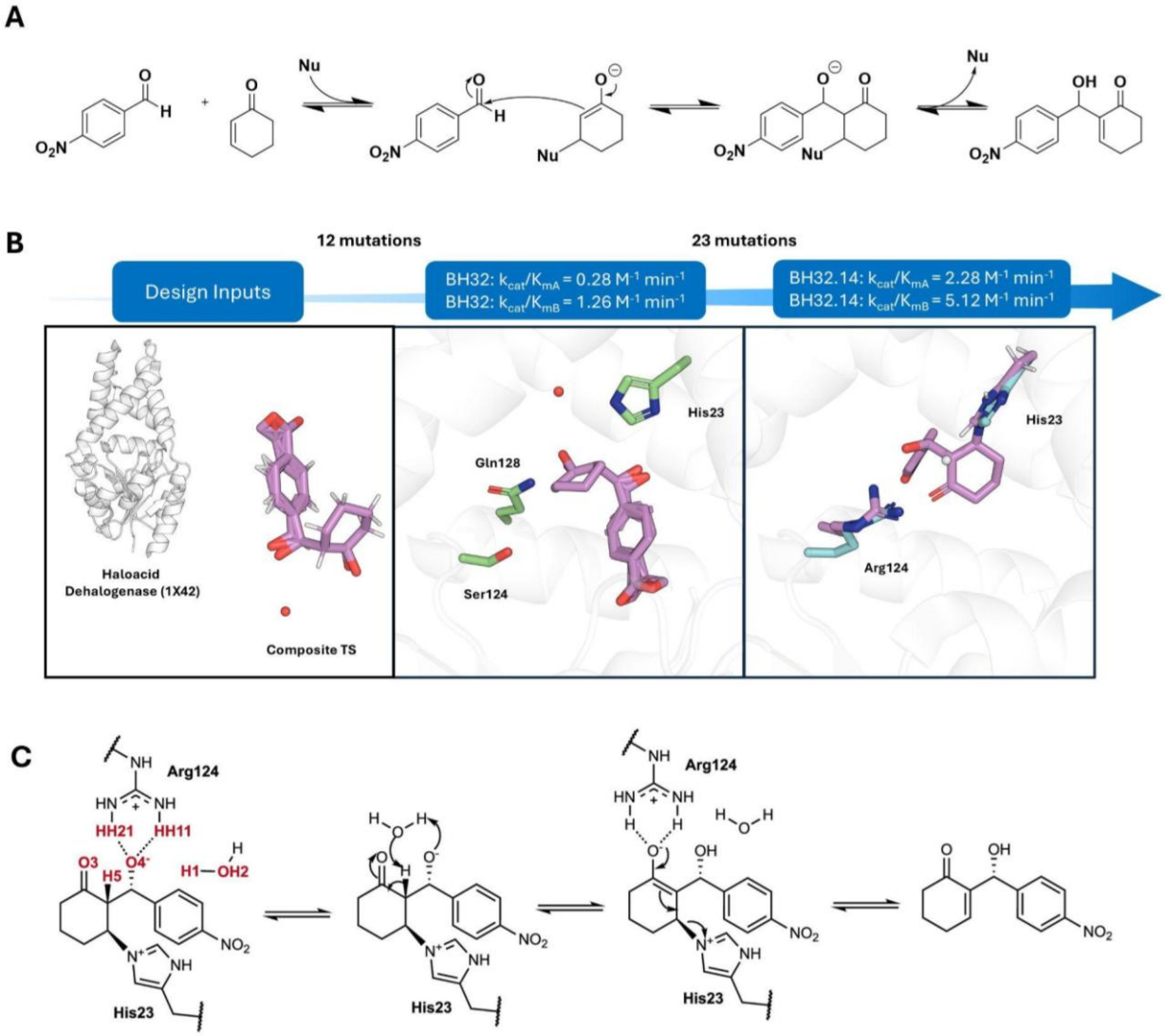

Figure 1.2: Design and evolution of the BH32 family of Morita-Baylis-Hillmanases

A.) Proposed reaction mechanism of the MBH reaction on 4-nitrobenzaldehyde and 2-cyclohexenone, utilizing histidine-mediated nucleophile activation of the ketone. The mechanism is utilized to guide the theozyme and *de novo* enzyme design of BH32. **B.)** Overview of evolutionary trajectory of BH32 to BH32.14. Composite theozyme model (sticks, light pink) and haloacid dehalogenase (cartoon, white) are used as design inputs to the initial *de novo* design. Catalytic residues in the active site (sticks, light green) are evolved from H23/E128/S124 in BH32, to H23/R124 in BH32.14 (sticks, light blue), labelled with the corresponding catalytic efficiency of the variant. **C.)** Mechanistic overview of the evolved BH32.12 enzyme as determined by QM/MM path sampling studies, demonstrating positioning of the R124 acting as an oxyanion hole for the developing negative charges on O3/O4. Figure reproduced from [52].

Initial efforts by Bjelic et al. attempted to scaffold this mechanism using a static, minimal theozyme [50]. To account for the multiple chemical states, the authors utilized the composite theozyme approach - superimposing the geometries of the C-C bond formation transition state and subsequent enolate intermediate. One of the resulting designs, BH32, utilized a histidine for nucleophilic attack, alongside a serine glutamine network intended to rigidly hydrogen-bond to the developing oxyanions. While the enzyme demonstrated trace activity, it also suffered from limited catalytic turnover.

Limitations of the static composite strategy were illuminated through directed evolution. Over 14 rounds of mutagenesis, Crawshaw et al. evolved the initial BH32 template into the benchmark variant BH32.14 [51]. This variant demonstrated a 710-fold improvement in relative conversion, as well as improved apparent catalytic efficiency on both substrates. Notably, retrospective structural and QM/MM analysis revealed that evolution overhauled the originally designed hydrogen-bond network, replacing it with a single arginine residue. Furthermore, evolution optimized the active site to stabilize an alternative transition state geometry as well: the two developing oxyanion groups, which were originally designed to point away from each other, rotated to align in the same direction. This alternative TS geometry is catalytically superior because it enables the flexible

arginine side chain to dynamically coordinate both oxyanions as the reaction progresses. QMMM transition path sampling studies corroborate this, demonstrating that catalytic proficiency of the evolved MBHases is driven by the coupling of protein dynamics to the reaction coordinate [52]. This evolutionary trajectory highlights a critical flaw in traditional single-state design: optimizing a scaffold around a static minimal DFT active site does not guarantee catalytic competence. Instead, efficient turnover relies on alternative TS geometries that are coordinated by the functional flexibility of the surrounding active site.

1.3.4. Case Study 3: Design of Serine Hydrolases

A recent benchmark study by Lauko et al. represents a state-of-the-art approach to the *de novo* design of enzymes for multistep reactions, using serine hydrolases as a model system [54].

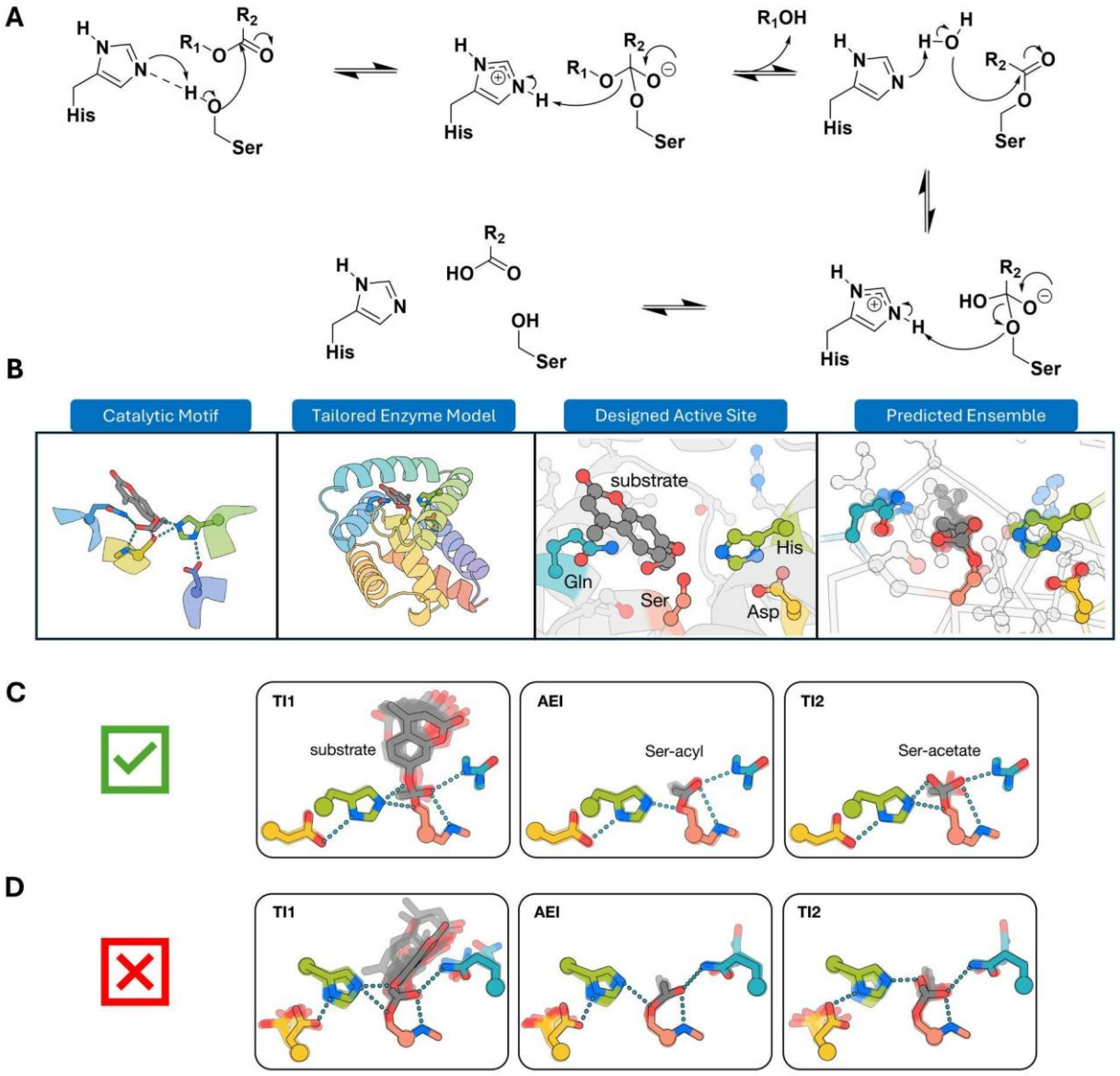

Figure 1.3 : Machine-learning guided design of serine esterases

A.) Reaction mechanism of serine hydrolases, used to guide theozyme and *de novo* enzyme design. **B.)** Computational pipeline overview for the design of serine hydrolases utilizing catalytic motif scaffolding. Catalytic motifs are scaffolded utilized RFdiffusion3 and sequence-optimized with iterations of LigandMPNN and Rosetta FastRelax. Active site rigidity is scored utilizing a predicted conformational ensemble with PLACER. **C.)** Example of a high-scoring active site design, where rigidification of the active site residues is prioritized. **D.)** Example of a low-scoring design, where large flexibility in active site residues are penalized. Figures adapted from [54]

Rather than searching for geometric matches within existing protein scaffolds, the authors developed a generative computational pipeline to construct novel backbones tailor-made for a desired active site. The methodology begins with the generation of target active site motifs, utilizing a DFT-optimized transition state geometry combined with explicit conformers or rotamers of active site residues derived from natural serine hydrolases. To realize these sites, RFdiffusion was employed to generate novel protein backbones, while LigandMPNN and Rosetta FastRelax were utilized iteratively for sequence design and structural optimization.

The use of PLACER - a deep learning ensemble generation network - is a defining feature of this study. PLACER is used to filter sequences by assessing the structural 'preorganization' of the active site across the entire reaction coordinate, including the apo state, the acyl-enzyme intermediate (AEI), and the flanking tetrahedral intermediates. By penalizing structural deviation across these states, the pipeline enforces a highly rigid active site geometry. This feature is approached as a structural compromise to accommodate the entire mechanism with minimal side-chain interstate movement. While the resulting enzymes exhibit catalytic efficiencies (k_{cat}/K_M) that impressively approach those of natural serine hydrolases, this efficiency is largely driven by highly optimized, sub-micromolar substrate affinities (K_M). Conversely, the catalytic turnover rates (k_{cat}) remain relatively poor, with the most active variants struggling to surpass one turnover per second. This work highlights that the most current design strategies for multistep enzymes prioritize rigid preorganization, yielding excellent binding. On the other hand, these design strategies are not as effective at improving the catalytic turnover, which may be explained by how active site rigidification could restrict subtle interstate conformational movement necessary for catalytic turnover and product release.

1.3.5. Takeaways on Current Design Methods for Multistep Enzymes

The evolution of *de novo* enzyme design is fundamentally intertwined with the advancement of computational methodologies, enabling designers to target increasingly complex chemical transformations. The field now widely recognizes that incorporating backbone conformational dynamics is important for achieving high catalytic proficiency. Despite these advances, contemporary design strategies remain tethered to paradigms that enforce strict active-site rigidity. As demonstrated, historical attempts to accommodate multistep mechanisms by forcing composite transition states into rigid scaffolds generate unphysical geometries requiring extensive directed evolution to rescue catalytic function. Similarly, modern ML-driven methods operate as a rigidification filter; bypassing the sequential demands of a multistep reaction coordinate.

Consequently, the explicit design principles required to choreograph the dynamic geometries of multistep catalysis remain underexplored. To address this limitation, design methodologies should pivot from enforcing static rigidity towards explicitly accommodating the intrinsic structural dynamics of the reaction mechanism.

1.4. Active Site Rearrangement via The Consensus Geometry

Natural enzymes that catalyze multistep reactions possess a nuanced structural feature - they must sequentially stabilize distinct chemical states that possess inherently different geometries. Hannes described this succinctly: “For catalysis, flexible but not too flexible, as well as rigid but not too rigid, is essential” [55]. In a more discrete definition, natural enzymes utilize highly organized active-site architecture governed by a **consensus geometry** - a structural feature that allows the active site to accommodate multiple intermediate states with the absolute minimum required atomic motion.

This ‘least-motion’ phenomenon was first quantitatively defined by Smith and Houk [1], utilizing the multistep catalytic mechanism of serine esterases as a model. The authors categorize these necessary atomic movements into two regimes: intrastep reorganizations (reorientation of functional groups required for specific bond cleavage or formation) and interstep reorganizations (repositioning of catalytic units between distinct chemical steps). Utilizing conformational ensembles of QM-derived theozymes, the authors demonstrate that the positioning of catalytically relevant residues in natural enzymes aligns excellently to a lowest-RMSD consensus geometry. By evaluating these transition-state ensembles across strict RMSD thresholds, they revealed that natural esterase architectures cluster tightly around geometries that mathematically minimize interstep motions without compromising the ideal geometry required for any single intrastep transition state, corresponding to a global RMSD average of ~1.3Å. Furthermore, this structural property is highly conserved across multiple fold families, confirming nature convergently evolved active sites that require the least motion theoretically possible to navigate the reaction coordinate.

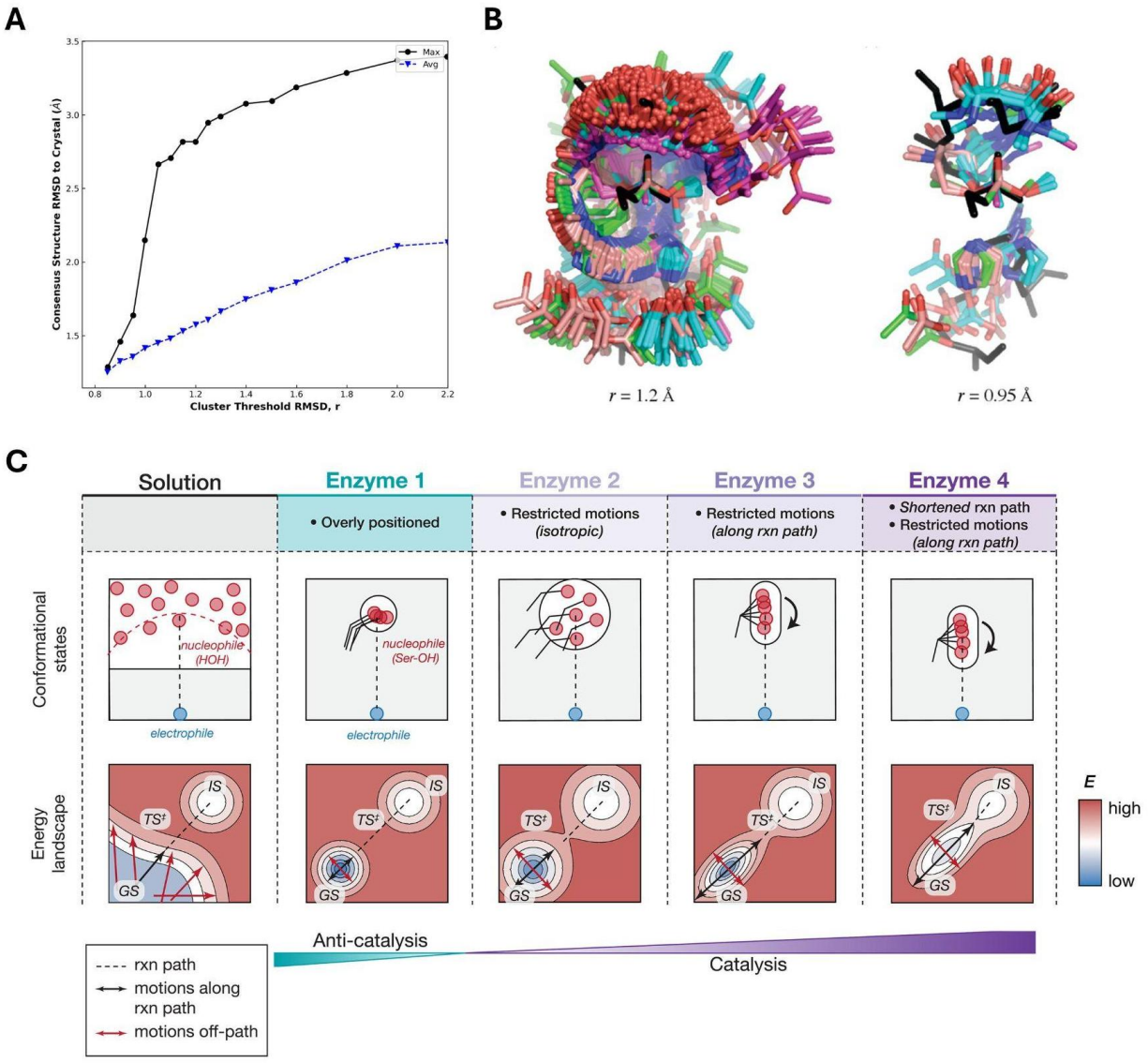

Figure 1.4: Two examples of minimal catalytic mechanism rearrangement involving serine hydrolases

A.) Plot of average and maximum catalytic atom RMSD of consensus structures from the serine esterase crystal structure (PDB ID: 1XLV) at different cluster threshold RMSDs (*r)*, for defining consensus structures. Blue diamonds give the average catalytic atom RMSD from the crystal structure over all consensus geometries at each threshold *r* value. Black circles show the associated largest catalytic atom RMSD from the crystal found for a single theozyme structure at each threshold *r* value. Figure adapted from [1] **B.)** Representation of theozyme conformer ensembles aligned to serine esterase crystal structure (sticks, black) utilizing clustering thresholds (*r*) at 1.2 Å and 0.95 Å. Figure adapted from [1] **C.)** Abstraction of on-path mechanistic motions utilized in the nucleophilic attack of serine proteases. Top row: demonstration of the conformational states comparing ineffective and effective enzymes, where the nucleophilic serine (red) and the electrophilic substrate (blue). Bottom row: demonstration of a model 1D reaction coordinate

(dotted line) energy landscape corresponding to each positioning strategy. The energy contours represent the relative energy level. Arrows in black represent the "on-path" reaction coordinate, arrows in red represent the "off-path" reaction coordinate. Hypothetical enzyme 1 would hinder catalysis, as indicated by the green color bar at the bottom of the figure ("Anti-catalysis"), while reactions on hypothetical enzymes 2, 3, and 4 would be increasingly faster, as indicated by the purple color bar
("Catalysis") Figure labels are GS = ground state; TS‡ = transition state; IS = intermediate state. Figure adapted from [2].

The physical manifestations and thermodynamic consequences of this consensus geometry were recently elucidated by Du[2]. By generating structural "pseudo-ensembles" from static crystallographic snapshots of serine proteases in apo, ground-state, and transition-state analog conformations, the authors demonstrated how minimizing interstate reorganization actively drives catalysis. Enforcing a minimal-motion consensus physically results in "path-shortening." By preorganizing the active site to favor the transition state, the enzyme forces the ground-state reactants into highly-restricted, slightly destabilized conformers. This ground-state destabilization - driven by reduced conformational entropy and unfavorable steric interactions absent in solvent - establishes a downhill free energy gradient that actively propels the reaction forward.

A striking physical example of this path-shortening is the conserved torsional rotation in the χ_1 angle in the nucleophilic serine. To realize the necessary atomic contacts preceding the transition state, the active site leverages precise rotameric shifts rather than large backbone reorganizations. Compared to an equivalent reaction in solution, the overall enzymatic nucleophilic attack path is shortened by ~1.1Å . This large distance reduction is achieved by replacing the disorganized translational motion of a solvated attack with highly targeted, minimal rotameric shifts and localized hybridization changes.

Together, these studies reveal that natural enzymes navigate multistep mechanisms through two coupled strategies: mathematically minimizing interstep reorganizational motions around a consensus geometry, and leveraging highly restricted rotameric shifts (path-shortening) to physically execute those motions.

These characteristics represent underexploited principles for the *de novo* design of artificial multistep enzymes. The persistent absence of these biomimetic principles in artificial catalysis is likely primarily a consequence of the field's historical reliance on "Single-State Design." By definition, single-state methodologies focus exclusively on optimizing an active site for an isolated, high-energy transition state. This approach is fundamentally incapable of accounting for consensus geometry, because the minimization of interstep motion is a dynamic property that only emerges when multiple sequential states are evaluated simultaneously.

Conversely, "Multistate" enzyme design methodology - which integrates multiple conformational or chemical states as simultaneous design inputs - are inherently suited to capture these dynamic principles [3–5]. By utilizing a conformational backbone ensemble, designers can explicitly accommodate a consensus geometry mechanism. Rather than attempting to trap multiple transition states within a rigidly compromised static backbone, the continuously sampled atomic coordinates of a structural ensemble can serve as dynamic design states. This approach allows for the engineering of a self-consistent mechanism, mirroring the "least-motion" proficiency of natural enzymes by ensuring the active site adapts across the entire reaction coordinate.

1.5. Thesis Overview and Objectives

As established throughout chapter 1, the field of computational *de novo* enzyme design remains highly reliant on static, single-state design paradigms. Whether achieved through unphysical composite transition states or the strict enforcement of rigid active-site preorganization,

contemporary methods fail to account for the intrinsic structural dynamics required to navigate a multistate reaction coordinate. Consequently, while modern generative algorithms excel at building highly complementary binding pockets, catalytic turnover for multistep bond-forming reactions remains severely limited, invariably requiring directed evolution to rescue activity. Therefore, the overarching objective of this thesis is to move beyond static structural constraints and explicitly engineer a *de novo* enzyme based on a **consensus geometry**. The central hypothesis of this thesis is that explicitly designing active sites to stabilize a minimal-RMSD consensus geometry - dynamically accommodated by protein backbone ensembles - can enable the design of efficient multistep *de novo* enzymes.

To evaluate this consensus-driven approach, the Michael addition was selected as the model transformation. Representing a complex multistep carbon-carbon bond-forming reaction, the Michael addition is a valuable synthetic transformation for which no *de novo* enzyme has been successfully designed to date. By utilizing this reaction as a model system, this thesis pursues the following specific aims:

Specific Aim 1: To computationally derive and define a minimal-RMSD consensus mechanism for a multistep Michael addition.

Before a dynamic scaffold can be engineered, the geometric requirements of the chemical mechanism need to be mapped. This aim focuses on the quantum mechanical (QM) evaluation of a 7-step Michael addition pathway. By analyzing the energetics and spatial trajectories of the required transition states and intermediates, we will derive a series of consensus theozymes, demonstrating how the designed catalytic motifs minimize interstep atomic deviation and satisfy the fundamental definition of a consensus geometry.

Specific Aim 2: To structurally scaffold the consensus mechanism utilizing multistate conformational ensembles.

To physically realize the consensus theozyme, the active site must be capable of subtle dynamic reorganization. This aim details the development and implementation of a multistate computational design protocol. By utilizing backbone conformational ensembles rather than static templates, we explicitly design protein scaffolds capable of dynamically accommodating the shifting geometric requirements of the consensus mechanism throughout the catalytic cycle.

Specific Aim 3: To experimentally validate the designed multistate enzymes.

This aim encompasses the recombinant expression, purification, and characterization of the computationally designed enzymes. By evaluating the limitations of the generated variants, we will identify immediate avenues for improvement, refining the consensus geometry protocol not only for the selected Michael addition, but in the broader context of multistep computational enzyme design.

Chapter 2: Consensus Theozymes for the Michael Addition Reaction

The central hypothesis of this thesis is that the use of minimal-RMSD consensus geometry can enable the design of efficient *de novo* multistep enzymes. To evaluate the utility of this design strategy, the Michael addition was selected as a model system. There are three reasons the Michael addition was selected. First, the Michael addition is a multistep reaction involving sequential substrate binding, carbon-carbon bond formation, and requisite proton transfers. These steps dictate constant shifts in atomic hybridization and charge distribution, making it ineffective to accurately represent the physical demands of the reaction coordinate with a single geometric snapshot. Second, carbon-carbon bond-forming reactions like the Michael addition possess synthetic utility; demonstrating *de novo* design in this space provides a useful framework for designing industrially relevant biocatalysts. Third, while highly efficient Michaelases have been discovered via promiscuous natural hydrolases or engineered through directed evolution, no enzyme has yet been explicitly designed *de novo* for this reaction from first principles. Consequently, the Michael addition serves as an ideal benchmark to validate the utility of consensus geometry in computational multistep catalysis.

2.1.1. Background on Enzymatic Michael additions

The discovery of enzymatic Michael additions represented an expansion of the biocatalytic repertoire, pushing enzymes into the realm of synthetic cofactor-free carbon-carbon bond formation. In overview, the field has evolved from leveraging the catalytic promiscuity of natural hydrolases to the engineering of mechanism-based biocatalysts that mimic small-molecule organocatalysis. Prior research on these enzymes are grouped into three main themes.

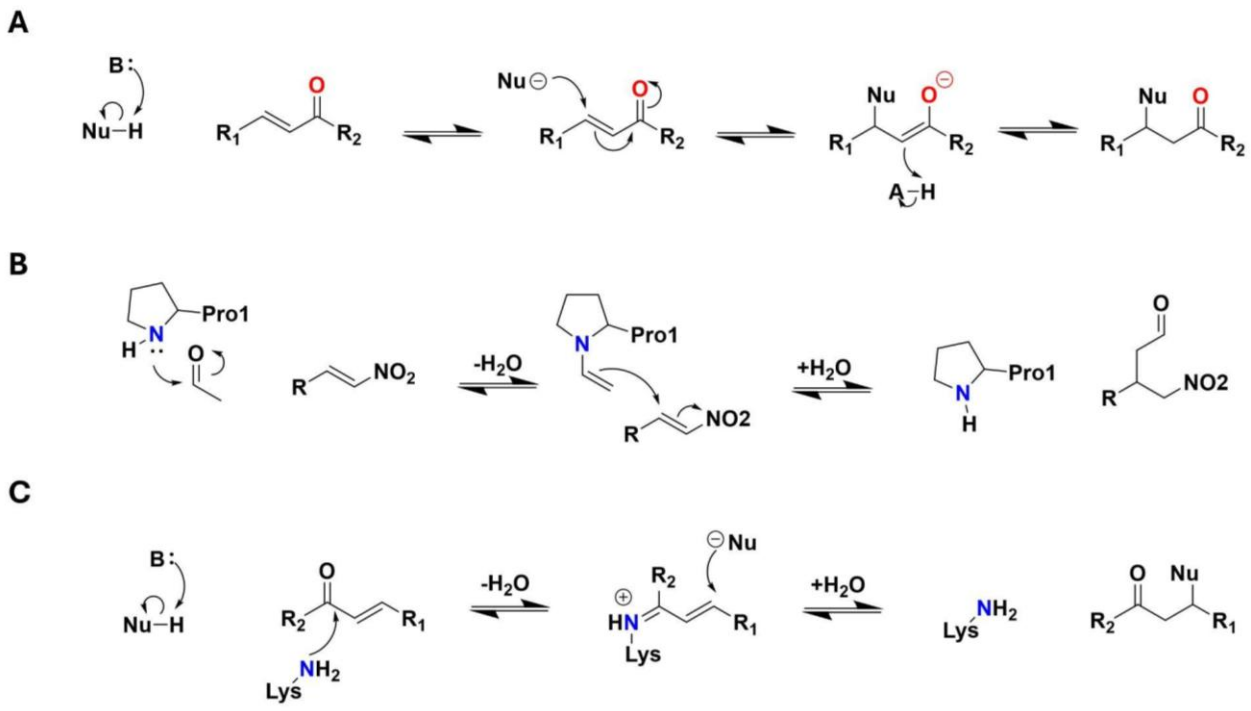

Figure 2.1: High-level mechanistic overview of enzymatic Michael additions

A.) General mechanism for a Michael addition catalyzed by stabilization on the developing oxyanion, typically catalyzed by promiscuous hydrolases. **B.)** Michael-type addition proceeds through enamine activation of the Michael donor (typically acetaldehyde) to a nitroalkene, catalyzed by an N-terminal proline. **C.)** Example michael addition proceeding through inamine activation of an unsaturated ketone/aldehyde, catalyzed by a nucleophilic lysine.

2.1.1.1 Promiscuous Michaelase Activity via Oxyanion Stabilization

Initial explorations into enzymatic Michael additions relied heavily on the catalytic promiscuity of natural hydrolases, primarily of lipases and proteases. Early work demonstrated that enzymes such as *Candida antarctica* lipase B (CALB) could catalyze additions between 1,3-dicarbonyls and α,β-unsaturated carbonyls [6]. It is hypothesized that these reactions proceed through a canonical 1,4-conjugate addition mechanism. The reaction is driven by the stabilization of a charged transition state, where the activated substrate carbonyl is stabilized via hydrogen bonding within the enzyme's preexisting oxyanion hole, while a general base (typically an Asp-His dyad) facilitates proton transfers.

Further studies on lipases from *C. antarctica, Mucor miehei, Streptomyces griseus,* and *Thermomyces lanuginosus* highlighted a broad substrate scope, though they often required highly nucleophilic CH-acidic donors and strongly electron-withdrawing substituents on the alkene to proceed [7,10]. To overcome the inherent limitations of these promiscuous active sites, co-catalyst additives have been explored. For example, acetamide has been used to accelerate the addition of poorly activated ketones to nitroolefins [8,9]. Kinetic isotope effects suggest the amide group assists in the activation and subsequent enolate stabilization of the donor ketone, increasing the acidity of the α-carbon proton, promoting the C-C bond formation. In all these cases, the active site relies on non-covalent stabilization and general acid/base chemistry to navigate the reaction.

2.1.1.2 Covalent Catalysis via the Enamine Mechanism

A notable shift in the field was the utilization of covalently bound intermediates, directly inspired by the enamine mechanisms used in organocatalysis. In this pathway, a reactive amine residue within the active site attacks the carbonyl carbon of the Michael donor substrate to form an iminium ion. Subsequent deprotonation yields a highly nucleophilic enamine species that is poised to attack the electrophilic double bond of the Michael acceptor [68]. The Poelarends group has extensively demonstrated this using 4-oxalocrotonate tautomerase (4OT), which relies on a catalytic N-terminal proline (Pro1) [11,12]. Acting initially as a nucleophile, Pro1 forms the covalent adduct, activating the donor substrate to drive C-C bond formation. Proof of this enamine mechanism was captured via hydrogen-deuterium exchange assays and X-ray crystallography, which revealed continuous electron density extending from the Pro1 pyrrolidine ring to the covalently bound substrate [13,56].

2.1.1.3 Covalent Catalysis via the Iminium Mechanism

Conversely, while the enamine mechanism activates the nucleophilic donor, the iminium functions by activating the electrophilic Michael acceptor. Here, a reactive amine in the active site forms a reversible, covalent Schiff base (an iminium ion) with the α,β-unsaturated carbonyl compound. This covalent NC bond significantly lowers the lowest unoccupied molecular orbital (LUMO) of the Michael acceptor, enhancing its electrophilicity and directing the subsequent nucleophilic attack from the donor [68,69].

A recent success in navigating this mechanism was achieved through the directed evolution of the class I aldolase, 2-deoxy-D-ribose-5-phosphate aldolase (DERA) [14]. While natural class I aldolases typically proceed via enamine-like aldol additions, this study evolved DERA to repurpose the wild-type Lys167 residue to form the critical iminium intermediate with the Michael acceptor, resulting in highly enantioselective additions.

Within the realm of computational enzyme design, the iminium mechanism has a long history. The RA95 retro-aldolase lineage, which was originally designed to catalyze the aldol cleavage of methodol via enamine formation with a nucleophilic lysine, has also been shown to be highly adaptable. The Hilvert group demonstrated that the evolved nucleophilic Lys83 in RA95.5 is promiscuously active for the iminium-mediated Michael addition for a wide range of α,β-unsaturated ketones and a diverse set of carbon donors [18]. Further studies demonstrated that the RA95.5 scaffold could be further evolved to catalyze all the stereoproducts of the model addition between ethylcyanoacetate and 4-4(-Methoxyphenyl)-3-buten-2-one [19].

2.1.1.4. Takeaways on Enzymatic Michael Additions

These three previously explored methods of enzymatic Michael additions - promiscuous oxyanion stabilization, enamine nucleophile activation, and iminium electrophile activation - demonstrate

that biological scaffolds can successfully orchestrate this carboligation cofactor free with canonical amino acids. While mechanistically distinct, these approaches do share a critical commonality: they must all navigate a cascade of chemically and geometrically distinct states. Whether accommodating the shifting hybridization of a covalent Schiff base or managing the charge distribution of a 1,4-conjugate transition state, these enzymes navigate a sequence of highly reactive intermediates.

While directed evolution and catalytic promiscuity have yielded functional Michaelases, no enzyme has yet been explicitly designed *de novo* for this reaction from first principles. This gap likely persists because of two factors. First, the experimental and evolutionary methods can afford to treat the underlying reaction coordinate as a black box. Second, the field of computational *de novo* design has only recently begun to seek methods to incorporate the inherent dynamics of enzymatic catalysis into the design process. As a rigidly designed active site must coordinate multiple shifting intermediates without over stabilizing a single covalent intermediate, designing a *de novo* Michaelase requires high geometric precision. Therefore, before a bespoke protein scaffold can be generated, the complete contiguous reaction pathway must first be formally mapped to extract the spatial trajectories of each discrete step. Guided by the discussed biological precedents, the following section outlines the explicit mechanistic proposal for this *de novo* design, focusing on the derivation of an iminium-based pathway.

2.1.2. Known Reactive Elements & Mechanistic Proposal

Guided by the biological precedents of electrophile activation via the iminium pathway, we select the Michael addition between 4-4(-Methoxyphenyl)-3-buten-2-one (Michael acceptor) and ethylcyanoacetate (Michael donor) as our target reaction. The viability of mitigating this specific carboligation via iminium catalysis was previously described by the Hilvert group through the

directed evolution of the RA95.5 retro-aldolase [18]. The electronic properties of these substrates are well suited for this mechanism: the α,β-unsaturated ketone (Michael acceptor) readily forms the covalent Schiff base required to drive electrophilic activation, while the electron-withdrawing nitrile and ethyl ester substituents of ethyl cyanoacetate (Michael donor) sufficiently enhance the C-H acidity of the donor carbon. Consequently, we sought to explicitly design a *de novo* Michaelase built around iminium catalysis.

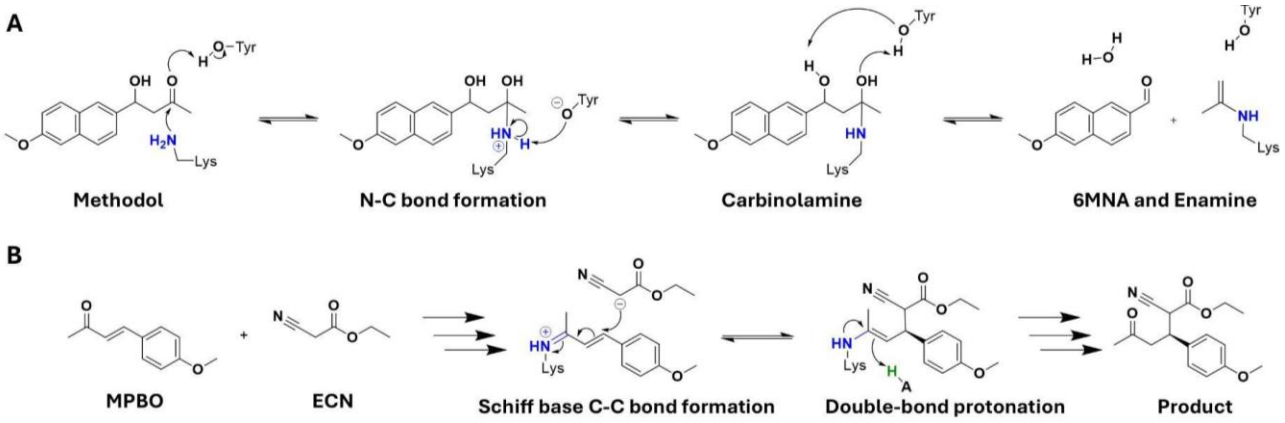

Figure 2.2: Mechanistic schematic of known reactive elements for designing a Michaelase **A.)** Retroaldol mechanism proposed for RA95.5-8F utilizes a nucleophilic lysine and tyrosine dyad to catalyze the formation of a carbinolamine intermediate and subsequent C-C bond cleavage of Methodol into 6-Methoxy-2-napthaldehyde (6MNA) and an enamine adduct. **B.)** Proposed mechanism of an evolved Michaelase, RA95.5-8 T53L/K210H, which utilizes a nucleophilic lysine to generate a Schiff base intermediate that catalyzes the C-C bond formation between 4-4(-Methoxyphenyl)-3-buten-2-one (MPBO) and ethyl cyanoacetate (ECN) to form ethyl 2-cyano-3-(4-methoxyphenyl)-5-oxohexanoate.

Our mechanism proposal begins with a focus on the iminium mechanism, generated by a lysine-tyrosine dyad using the reactive elements of the RA95 family as inspiration. In these evolved enzymes, the lys/tyr dyad is responsible for the formation of and collapse of a carbinolamine intermediate [17]. Prior QM/MM studies on RA95.5-8F demonstrates synergetic cooperation in this dyad; the tyrosine acts as an acid/base to generate the carbinolamine, which precedes the formation of the enamine and release of the covalently bound product [49]. Building on this precedent, this

dyad was selected as the foundation to which to generate the Schiff base intermediate for this mechanism.

In the evolved Michaelases from the RA95 lineage, the Michaelase catalytic efficiency is quite low for an evolved enzyme (153 $M^{-1}s^{-1}$ for the ketone, 0.58 $M^{-1}s^{-1}$ for the donor) - partially due the high K_M for the ethyl cyanoacetate (16.5 mM) [18]. This was expected to stem from the lack of explicitly designed binding pocket for the ECN donor, potentially also suggesting that the carbanion Michael donor is not efficiently generated by the enzymes' active site. Furthermore, while the RA95.5-8F mechanism utilizes the tyrosine of the dyad to protonate the enamine and regenerate the Schiff base, QM/MM studies identify this Schiff base formation as a rate-limiting step [49]. Conversely, early studies on promiscuous Michaelase activity frequently highlight the utilization of histidine as a general acid/base to facilitate these proton transfers [57].

Here, a 7-step atom-economical mechanism is proposed. The proposed mechanism is divided into three mechanistic themes: (1) the formation and collapse of the carbinolamine intermediates preceding the Schiff base; (2) the generation and hydrolysis of the Schiff base via water elimination and hydration; and (3) the core Michael addition, encompassing donor nucleophile generation, C-C bond formation, and enamine protonation.

The mechanism design is approached by first identifying the general acid/base network that will be responsible for navigating the proton transfer, by comparing the energetic and geometric consequences of utilizing a histidine versus a tyrosine to mediate the expected rate-limiting protonation step. Then, the remainder of the theozymes are constructed to utilize the same catalytic residue for acid/base catalysis by searching for valid transition states that demonstrate minimal reorganization of all the catalytic atoms. Finally, reorganizational geometries and energies are evaluated for the complete proposal.

2.2 Theozyme Construction Results

2.2.1 Derivation of the Catalytic Triad

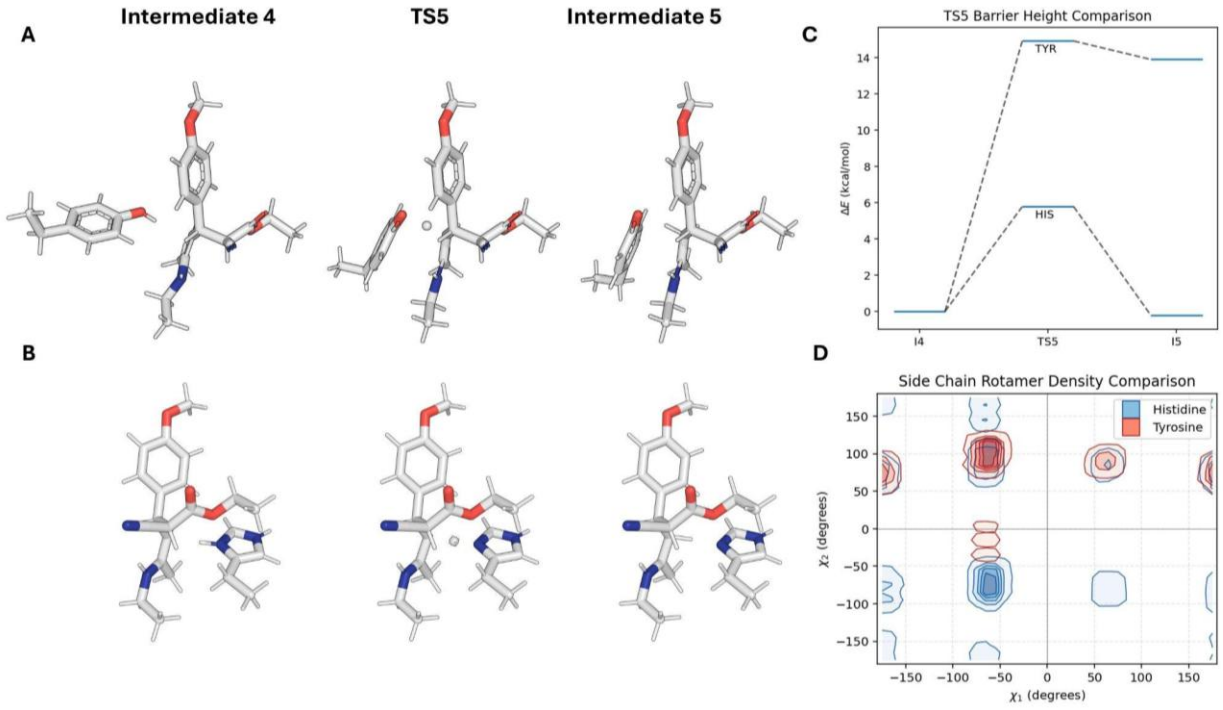

Figure 2.3: Comparison of utilizing Tyrosine and Histidine as a general acid for the protonation of the covalently bound Michael adduct

A.) Intermediate and transition state geometry of the double-bond protonation step utilizing neutral tyrosine. **B.)** Intermediate and transition state geometry of the double-bond protonation step utilizing the imidazolium cation of histidine. **C.)** Comparison of the barrier heights computed for tyrosine and histidine mediated protonation at the B3LYP 6-31G(d,p) D3BJ level of theory. **D.)** Janin plot comparison of tyrosine and histidine components of the theozyme geometry measured to generate catalytic motifs in downstream enzyme design.

To resolve the expected rate-limiting protonation step inherent to the RA95.5 catalytic tetrad, we first evaluated the viability of incorporating either a tyrosine or a histidine to mediate the requisite acid/base chemistry. The transition states were computed with density functional theory at the B3LYP/6-31G(d,p) level with D3BJ empirical dispersion.

Initial geometric analysis revealed a steric limitation in the tyrosine-mediated pathway. In its reactant pose, the tyrosine hydroxyl proton is roughly perpendicular to the sp2 protonation site. To transfer this proton in the transition state, the side chain must undergo a large conformational shift, a reorganization that would require a large backbone fluctuation. In contrast, the smaller, unbranched imidazole ring of histidine fits snugly beneath the ECN substrate, accommodating the proton transfer without requiring large spatial repositioning.

Thermodynamically, the tyrosine-mediated reaction exhibits a free energy barrier of ~14.5 kcal/mol. Furthermore, the resulting product state is nearly isoenergetic with the transition state, highlighting the destabilizing effect of the charged tyrosine enolate intermediate. Conversely, the histidine-mediated step proceeds with a smaller barrier of ~6.0 kcal/mol, with nearly isoenergetic reactant and product states, consistent with the stability of the imidazole protonation states. Finally, mapping the side-chain dihedral angles on a Janin plot utilizing the Dunbrack rotamer library that will be used for enzyme design in Chapter 3, confirms a notably larger rotameric space than that of tyrosine. This inherent flexibility suggests histidine is better suited to accommodate the geometric shifts required across the full reaction coordinate with minimal backbone deviation. Based on these energetic and geometric advantages, the Lys/Tyr/His triad was selected as the catalytic triad, and all subsequent reaction steps were mapped utilizing this architecture.

2.2.2 TS1 & TS7 - Substrate Binding and Release via Carbinolamine Intermediate

Having established the Lys/Tyr/His catalytic triad, the complete reaction coordinate was mapped by grouping symmetric mechanistic themes. The pathway initiates (TS1) with the formation of the carbinolamine intermediate. To avoid falling into a high-energy enolate state, a concerted mechanism is proposed wherein the initial N-C bond formation between the nucleophilic lysine and the ketone substrate occurs synchronously with proton transfer. Specifically, the tyrosine

hydroxyl donates a proton to the forming ketone oxyanion while concertedly abstracting a proton from the attacking lysine, generating a neutral carbinolamine. This concerted proton slide is observed by the displacement vectors of both the amino and hydroxyl protons in the computed hessian, and the downstream intermediates computed by the intrinsic reaction coordinate (IRC). TS1 was found to be endergonic with a barrier height of ~12.8 kcal/mol.

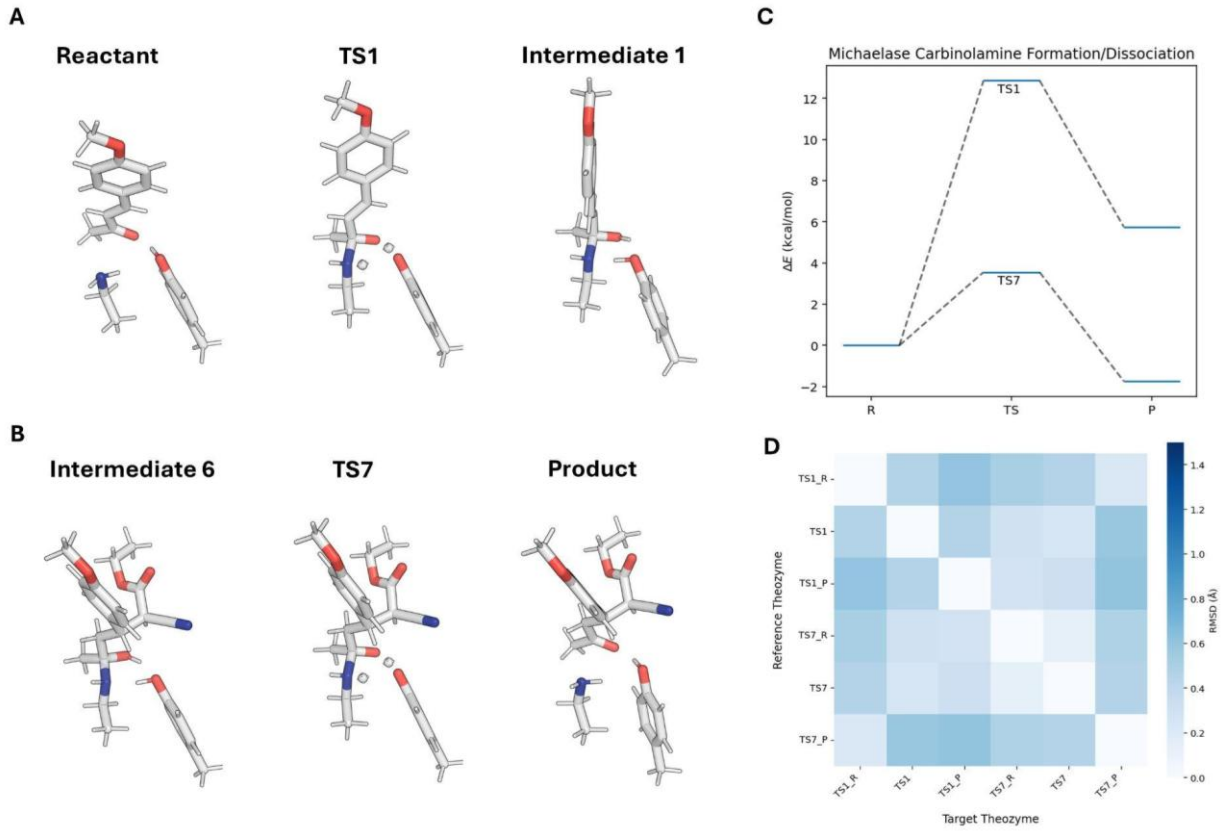

Figure 2.4: Comparison of the formation and release of the covalently bound carbinolamine intermediate

A.) Intermediate and transition state geometry of the carbinolamine formation in TS1 via concerted proton slide between lysine and tyrosine residues. **B.)** Intermediate and transition state geometry of the carbinolamine collapse in TS7 via concerted proton slide between lysine and tyrosine residues **C.)** Comparison of the free energy barrier heights computed for the formation and release steps at the B3LYP 6-31G(d,p) D3BJ level of theory. **D.)** Pairwise catalytic atom alignment over all computed states for the carbinolamine formation and collapse. 'R' refers to the backwards intermediate relative (reactant of step) to the transition state. 'P' refers to the forward intermediate relative to the transition state (product of step). Darker shades on the matrix correspond to a larger catalytic atom pairwise RMSD.

The terminal release step (TS7) proceeds via a mechanistic mirror image. The step is initiated by the protonation of the carbinolamine nitrogen by the tyrosine, concerted with N-C bond cleavage and proton abstraction from the substrate hydroxyl group. This restores the initial protonation states of the Lys/Tyr dyad and releases the final ketone product. Notably, while mechanistically symmetric to TS1, TS7 is more exergonic, possessing a reduced barrier height of ~3.6 kcal/mol. The expected hybridization shifts (TS1: $sp^2 \rightarrow sp^3$, TS7: $sp^3 \rightarrow sp^2$) associated with carbinolamine formation and release were captured in the IRC trajectories for both steps. Theozymes for TS1 and TS7 are also well aligned, as the pairwise alignment demonstrates minimal rearrangement across all the computed steps.

2.2.3 TS2 & TS6 - Schiff Base Formation and Collapse

Following carbinolamine formation, the pathway proceeds through the generation (TS2) and subsequent collapse (TS6) of the covalent Schiff base. In TS2, the carbinolamine undergoes dehydration to form the iminium ion. This water elimination is facilitated by the derived histidine residue acting as a general acid to protonate the leaving hydroxyl group. Both proton transfer and C-O bond cleavage occur concurrently, proceeding through a barrier of ~6.5 kcal/mol.

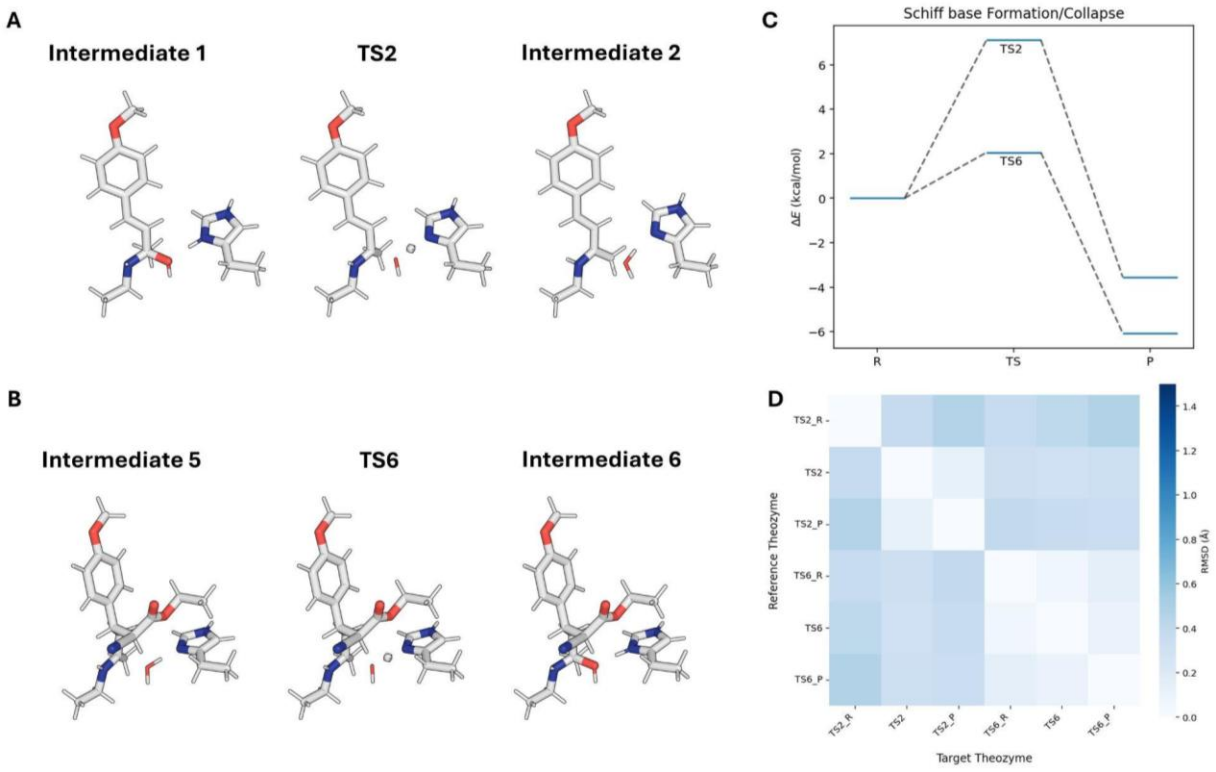

Figure 2.5: Comparison of water elimination and rehydration steps for the Schiff base intermediate

A.) Intermediate and transition state geometry of the generation of the Schiff base intermediate in Step 2 via water elimination through the histidine. **B.)** Intermediate and transition state geometry collapsing the Schiff base intermediate in Step 6 via rehydration through the neutral histidine tautomer. **C.)** Comparison of the free energy barrier heights computed for the formation and collapse steps at the B3LYP 6-31G(d,p) D3BJ level of theory. **D.)** Pairwise catalytic atom alignment over all computed states for the two compared states. 'R' refers to the backwards intermediate relative (reactant of step) to the transition state. 'P' refers to the forward intermediate relative to the transition state (product of step). Darker shades on the matrix correspond to a larger catalytic atom pairwise RMSD.

The reverse hydration step (TS6) occurs via nucleophilic attack of a water molecule on the sp^2 carbon of the Schiff base. Here, the histidine acts as a general base, abstracting a proton from the attacking water to regenerate the carbinolamine. The barrier height is lower than that of TS2, at about ~2 kcal/mol. TS2 and TS6 are both exergonic, maintaining the continuous downward thermodynamic trajectory required for efficient turnover. The two transition states are also

extremely well aligned, with the histidine adopting an almost identical rotamer. Furthermore, the leaving/entering hydroxyl group is in a nearly identical pose.

2.2.4 TS3 & TS4 & TS5 - The Core Michael Addition Steps

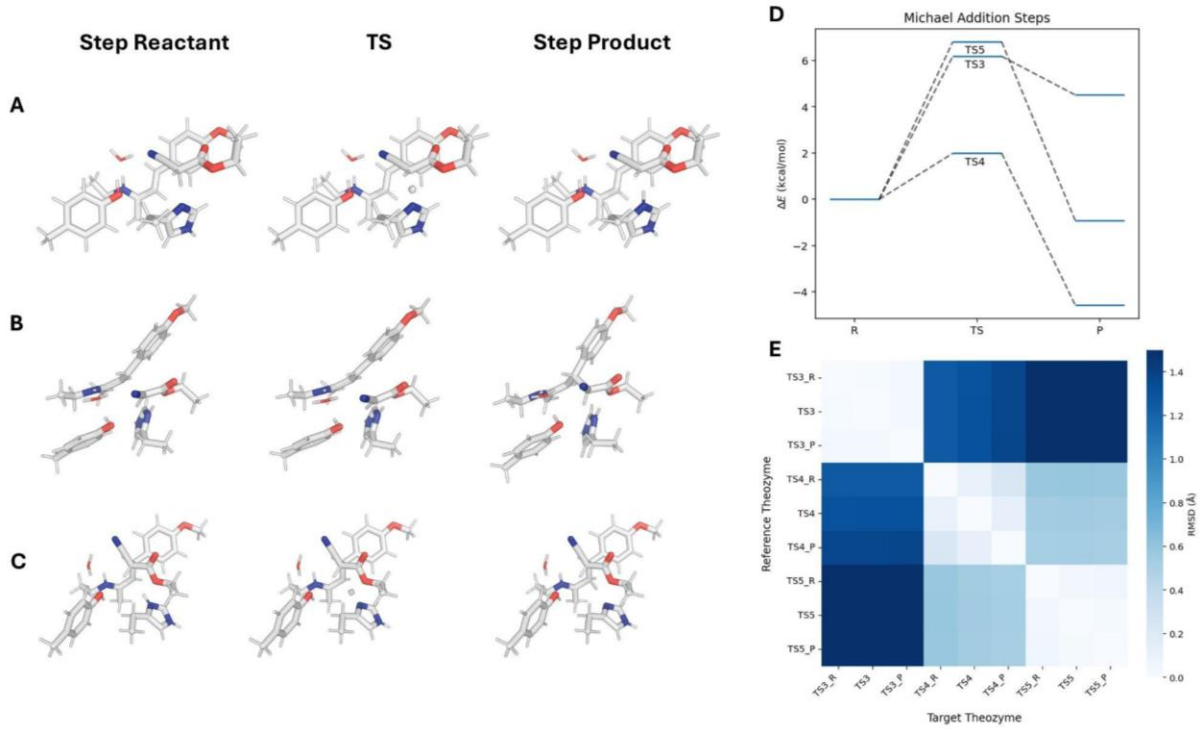

Figure 2.6: Comparison of core michael addition steps

A.) Intermediate and transition state geometry corresponding to TS3, the generation of carbanion Michael donor by deprotonation through histidine. **B.)** Intermediate and transition state geometry corresponding to TS4, the C-C bond formation between the Michael donor and acceptor. **C.)** Intermediate and transition state geometry corresponding to TS5, the enamine protonation by the histidine cation. **D.)** Comparison of the free energy barrier heights computed for the core Michael addition steps at the B3LYP 6-31G(d,p) D3BJ level of theory. **E.)** Pairwise catalytic atom alignment over all computed states for the Michael addition steps. 'R' refers to the backwards intermediate relative (reactant of step) to the transition state. 'P' refers to the forward intermediate relative to the transition state (product of step). Darker shades on the matrix correspond to a larger catalytic atom pairwise RMSD.

The central carboligation phase of the mechanism is composed of three transition states. In TS3, histidine acts as a general base to deprotonate the ECN donor, generating the necessary carbanion.

This step is endergonic ($\Delta G_{rxn} \approx 5.0$ kcal/mol) with a barrier of ~6.0 kcal/mol due to the instability of the high-energy carbanion. Notably, the histidine undergoes a distinct χ_2 rotameric shift to properly align the imidazole nitrogen for proton abstraction.

An interesting structural observation emerged when modelling this step. By explicitly retaining the spectator water generated in TS2, alongside the complete atomic coordinates of the catalytic triad, an emergent hydrogen bond network was observed. This network bridges the nitrile group of the ECN, the tyrosine hydroxyl, the Schiff base amine, and the spectator water. The extensive H-bonding helps to rigidify the transition state complex, enabling tighter structural alignment across TS1, TS2, TS6 and TS7.

In TS4, the hydrogen-bond network is maintained while the C-C bond is formed via nucleophilic attack of the ECN carbanion onto the electrophilic iminium acceptor. This bond-forming step is exergonic ($\Delta G_{rxn} \approx -4.6$ kcal/mol) with a barrier height of ~2 kcal/mol. Furthermore, an additional stabilizing contact emerges, with the ND1 proton forming a transient hydrogen bond with the ECN ester oxygen, further anchoring the substrate. Finally, in TS5, the resulting enamine is protonated by the acidic histidine. During this step, the histidine rotates back around its χ_2 dihedral to its initial pose, repositioning the triad for the carbinolamine collapse in TS6. Within the core C-C bond formation steps, TS4 and TS5 are notably more similar to each other than to TS3. The deviation mainly arises from the positioning of the ECN, as its approach conformation requires repositioning of the remainder of the reactive complex to accommodate it. The step is unique in the sense that the covalently-bound Michael donor is isolated from the reactive part of the complex, aside from the hydrogen bond network established by the explicit water molecule.

2.2.5 Geometric Rearrangement Evaluation of Designed Mechanism

Evaluating the free energy over the complete reaction surface shows that it is thermodynamically favorable towards the product. With the exception of the endergonic initial carbinolamine formation (TS1) and carbanion generation (TS3), the overall pathway maintains a continuous exergonic trajectory. The energy landscape ensures a strong thermodynamic driving force for catalytic turnover, preventing accumulation of intermediates.

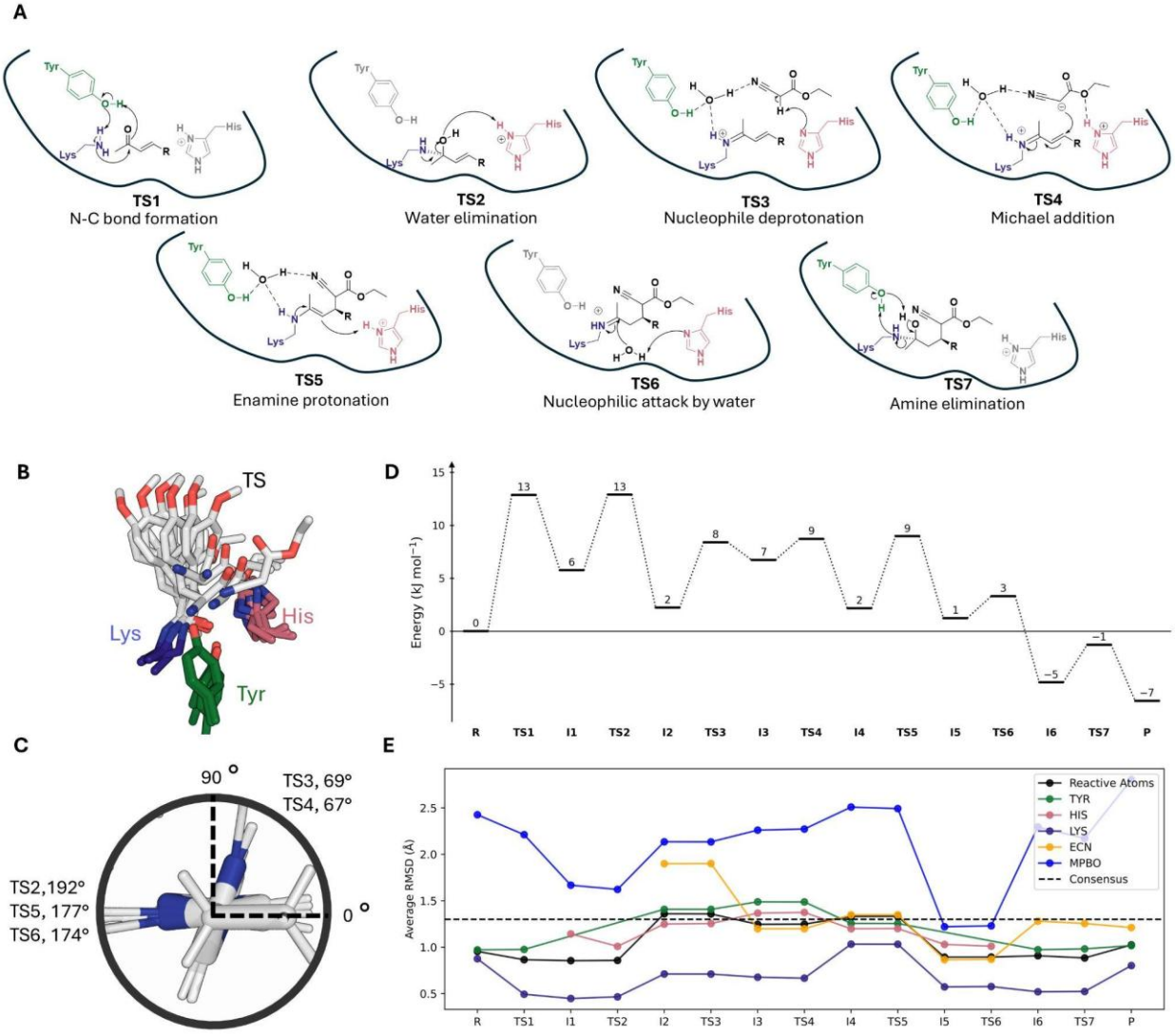

Figure 2.7: Energetic and geometric evaluation of the designed Michael addition mechanism

A.) 2D representation of each theozyme corresponding to each transition state. Atoms included in the theozyme calculations are colored, and groups not-modelled are in light grey. **B.)** Superimposition of the computed theozyme geometries. **C.)** His χ_2 rotation across theozymes after alignment to histidine CA-CB-CG carbons. Labelled angles represent χ_2 dihedral based on ND1-CG-CB-CA. **D.)** Free energy landscape of the complete mechanism. **E.)** Averaged per heavy-atom RMSD of each state relative to the complete reaction coordinate. Alignment atoms represent: lysine terminal N; tyrosine terminal O, CB, CG; carbonyl C, O, CA, CB ; Michael donor carbanionic C and ester O. TYR/HIS/LYS/ECN/MPBO represents all heavy atoms in the group, including both catalytic and non-catalytic atoms. The dashed line represents the average mechanism-wide RMSD observed by Smith and Houk, for the serine esterase active site (1.3 Å).

Beyond thermochemistry, structural alignment of the discrete theozymes confirms the viability of a consensus geometry. When superimposed across all transition states and intermediates, the catalytic atoms (Lys/Tyr/His and substrates) exhibit a low mechanism-wide atomic deviation of ~1.1 Å RMSD. Critically, this variance aligns well with the observed deviation of multistep mechanisms derived by Smith and Houk [1]. Analyzing the step-wise RMSD fluctuations isolates the core Michael addition steps as the main source of reorganization. The largest geometric deviation among the catalytic atoms occurs during TS3 (Figure 2.7, black line), a structural accommodation for the entry and deprotonation of the ECN. Furthermore, the largest positional variance within the complex is primarily within the bulky aromatic ring of the Michael acceptor. This motion stems from a lever-arm effect, an artifact from aligning the complexes by their catalytic atoms. It demonstrates that while the distal regions of the substrate deviates, the catalytic core remains well-conserved.

Finally, isolating the geometric trajectory of the histidine further highlights its utility in the mechanism. To navigate the shifting catalytic demands of the multistep mechanism, the histidine side chain undergoes a reversible rotameric toggle. The χ_2 dihedral rotates ~110° from a locally planar orientation (relative to the CA-CB-CG plane) into a perpendicular conformation to

facilitate carbanion generation and C-C bond formation (TS3 and TS4). It subsequently rotates back to its original planar geometry to execute the enamine protonation (TS5) and subsequent hydration step. This choreographed χ_2 sweep confirms how a single, anchored histidine can accommodate the entire acid/base requirements, minimizing the interstep reorganizational motions required to progress the catalytic mechanism.

2.3 Chapter Summary

The computational derivation of this 7-step pathway demonstrates that full-mechanism mapping by theozyme design is feasible and produces a self-consistent geometry. By evaluating the rate-limiting protonation requirements of the retro-aldol iminium pathway, a catalytic Lys/Tyr/His triad capable of overcoming thermodynamic bottlenecks in retro-aldolases was derived. Importantly, tracking the continuous trajectories of these atoms revealed non-obvious structural dependencies that a single-state design methodology would have missed. Specifically, the retention of the eliminated spectator water molecule following Schiff base formation. Rather than diffusing away, this water serves as a structural determinant, participating in an emergent extended hydrogen-bond network that bridges the substrate, the catalytic tyrosine, and the iminium intermediate. This network explicitly rigidifies the transition state complex during the core carbon-carbon bond forming steps. Consequently, this rigidification helps to anchor the catalytic atoms, enabling a uniform, low-RMSD alignment across all seven transition states.

These results confirm that it is possible to derive a set of geometrically self-consistent theozymes. By modelling the continuous pathway, a consensus structural geometry that satisfies the spatial and electronic demands of the Michael addition could be derived. To translate this chemical blueprint into a functional enzyme, the 3D coordinates must be scaffolded within a stable protein

environment. Having defined the geometric requirements of the active site, the next phase of this work (Chapter 3) focuses on matching this consensus theozyme into protein scaffolds to design Michael addition enzymes.

Chapter 3: Designing Enzymes around Consensus Theozymes

3.1 Introduction

The *de novo* design of enzymes capable of catalyzing complex, multistep reactions remains a challenge in protein engineering. While multistate design (MSD) has been successfully employed to engineer sequences compatible with discrete chemical or conformational states, its application to contiguous multistep reaction coordinates is still highly limited. The primary precedent for applying MSD to a multistep mechanism within this lineage of work was established by Zarifi, who focused on the stabilization of two distinct transition states for the Michael addition utilizing a Lys/Tyr/Tyr catalytic triad [3]. While these initial designs displayed improved apparent K_M for the Michael acceptor relative to the starting scaffold, they demonstrated minimal catalytic turnover. This low catalytic efficiency was hypothesized to arise from the over-stabilization of targeted transition states, shifting the rate limiting step to non-targeted transition states and trapping the turnover in a thermodynamic pit.

This chapter details the extension of multistate design to encompass a complete, contiguous reaction coordinate, explicitly designed as a consensus mechanism. The central hypothesis of this aim is that applying MSD across the entirety of a reaction pathway will allow for the explicit design of the conformational substates necessary to enable continuous catalytic turnover. This work represents the first application of multistate design to structurally realize a consensus mechanism.

3.1.1 Computational Pipeline Proposal

Computational enzyme design is driven by the objective of anchoring an idealized theozyme within a protein scaffold and subsequently stabilizing that geometry through active-site repacking. For multistep reactions, the scaffold must possess the inherent backbone flexibility required to accommodate the subtle conformational shifts of the reaction coordinate. To achieve this, a scaffold ensemble is leveraged, which allows a wider sampling of potential side-chain positions to effectively accommodate various sequential intermediates and transition states. The overview of the computational design workflow is as follows:

1. **Scaffold Selection and Motif Anchoring**: The crystal structure of indole-3-glycerolphosphate synthase (PDB: 1A53) was selected as the scaffold. As the evolutionary precursor to the highly adaptable RA95 retro-aldolase lineage, 1A53 is well suited to accommodate the Lys/Tyr dyad responsible for carbinolamine formation in both the retro-aldol and Michael addition mechanisms. Utilizing an ensemble of 1A53 conformations generated via ensemble refinement, the Lys/Tyr dyad was placed to accommodate the initial and terminal steps (TS1/TS7) of the pathway.
2. **Targeted Scaffold Reconstruction via a CANVAS-Inspired Workflow:** To integrate the catalytic histidine, a workflow inspired by CANVAS (Customizing Amino Acid Networks for Virtual Active-Site Scaffolding) was developed [70]. CANVAS is an enzyme design computational workflow that introduces a structural lid onto a minimal TIM barrel to anchor catalytic residues. In the standard CANVAS protocol, one catalytic side chain is built into the scaffold, while the Cα atoms of the remaining residues are positioned in the empty space above the TIM barrel's catalytic face. This serves as an anchor for constructing a custom active-site lid using RFdiffusion[60], followed by sequence design with ProteinMPNN[64] and structural prediction with AlphaFold2[65]. In this work, the core

philosophy of CANVAS is adapted: rather than constructing a novel lid, the native scaffold was modified. The catalytic histidine was positioned relative to the Lys/Tyr dyad utilizing the consensus theozymes, and an adjacent α-helix was reconstructed outward from the histidine backbone to anchor the new triad.

3. **Iterative Scaffold Refinement and Ensemble Generation:** Since the newly reconstructed α-helix occupies a geometry not co-evolved with the native 1A53 scaffold, a refinement protocol adapted from the Baker lab was developed to relieve structural strain. The scaffold was iteratively subjected to sequence redesign (LigandMPNN[61]), structural prediction (Boltz-2[66]), and energetic minimization (Triad) to correct geometric errors until sequence-structure convergence was achieved. Following refinement, Rosetta Backrub was utilized to generate an artificial conformational ensemble of the stable scaffold, providing the backbone sampling to accommodate the shifting geometries of the 7-step consensus mechanism.
4. **Multistate Sequence Design:** With the artificial ensemble generated, the full sequence of consensus theozymes was placed, and the active-site sequences were optimized via multistate design to simultaneously stabilize the entire pathway. Following active-site filtering based on criteria expected to confer catalysis, the remainder of the sequence was optimized via Shell Redesign (using LigandMPNN and Boltz-2) to ensure the active site and ligand-binding poses were structurally maintained.

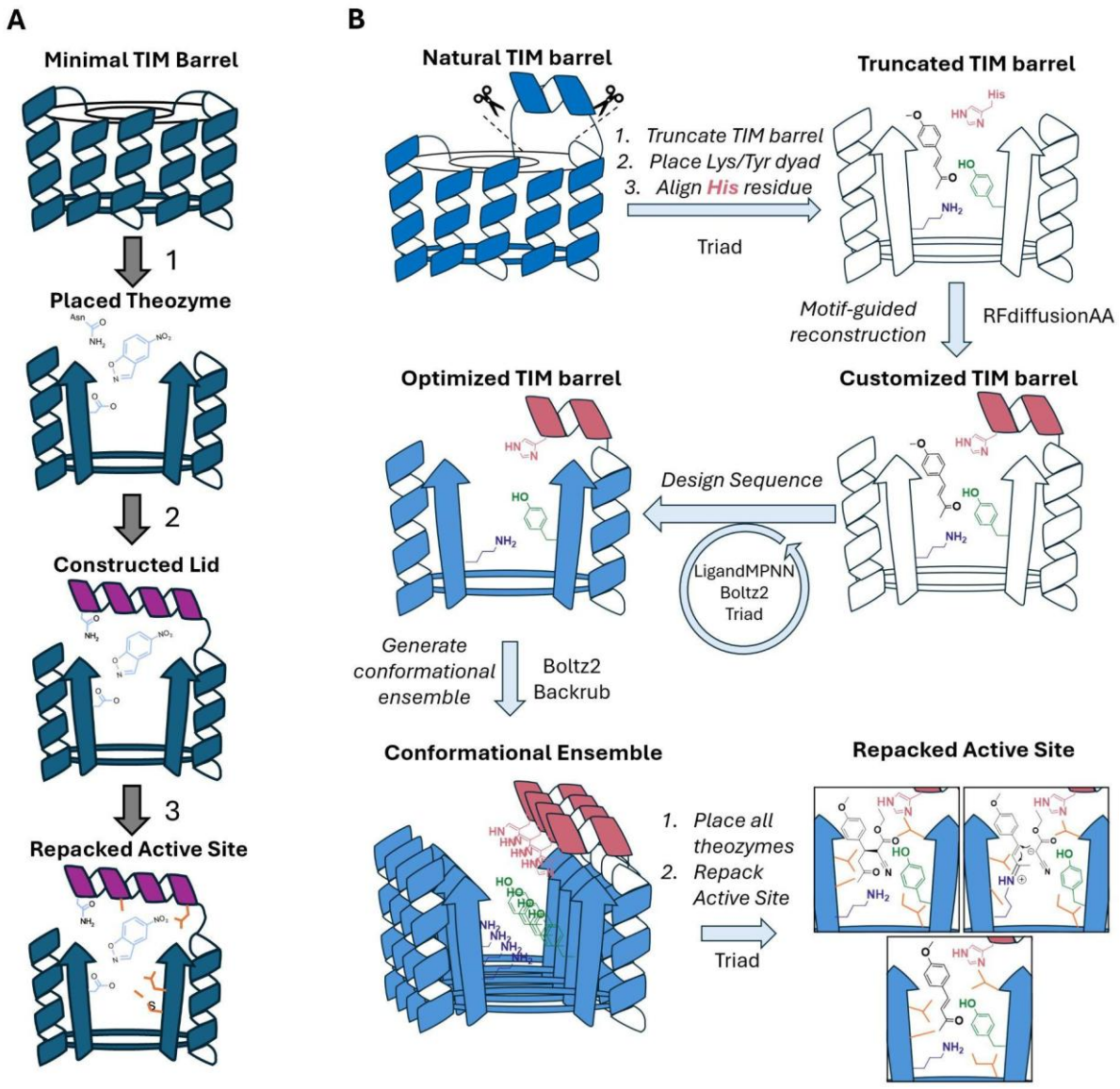

Figure 3.1: Computational pipeline proposed for the design of a *de novo* Michaelase

A.) Simplification of the CANVAS workflow, adapted from [71]. **B.)** Design workflow for scaffolding 7-step theozymes. Workflow begins with the truncation of the secondary structure near the activity face of the TIM barrel. Following, the Lys/Tyr dyad is reconstructed using both TS1 and TS7. Once a high-quality placement is achieved, a helix motif and corresponding connecting loops are reconstructed to scaffold the catalytic histidine. Once the customized TIM barrel is designed, the structure and sequence are redesigned to a higher-quality scaffold by utilization of a scaffold refinement workflow and by optimizing for the predicted CA-CB positions of all three catalytic residues relative to the RFdiffusion design model at the prior step. Then, a conformational ensemble is generated with Boltz2 and Rosetta Backrub. The sampled backbone conformers are used to position all 7 consensus theozymes, and repack the active site with multistate sequence design.

The described workflow was utilized to design protein sequences that were expected to fold into stable structures to accommodate the consensus mechanism. To explicitly evaluate the consensus geometry hypothesis, two sets of sequences were designed. A set of "Positive" (called consensus) sequences were explicitly designed to minimize the predicted RMSD of the catalytic histidine across the reaction coordinate, as an approximation to the consensus geometry. Another set of "Negative" (called anti-consensus) sequences were designed to maximize the predicted histidine RMSD while maintaining all other catalytic contacts and structure fold. Given the expectation that the positive consensus sequences should manifest the designed catalytic geometry, any improvement of their activity relative to the negative anti-consensus sequences could be correlated to activity conferred by the consensus design strategy.

3.2 Computational Design Results

3.2.1 Motif generation for the Lys/Tyr Dyad

The initial phase of the computational workflow required scaffolding the foundational Lys/Tyr dyad capable of catalyzing the carbinolamine formation critical to both the initial and terminal (TS1/TS7) steps of the pathway. To achieve this, a full combinatorial scan of all tip β-strands facing the interior of the 1A53 β-barrel was evaluated against the geometric constraints of the consensus theozymes.

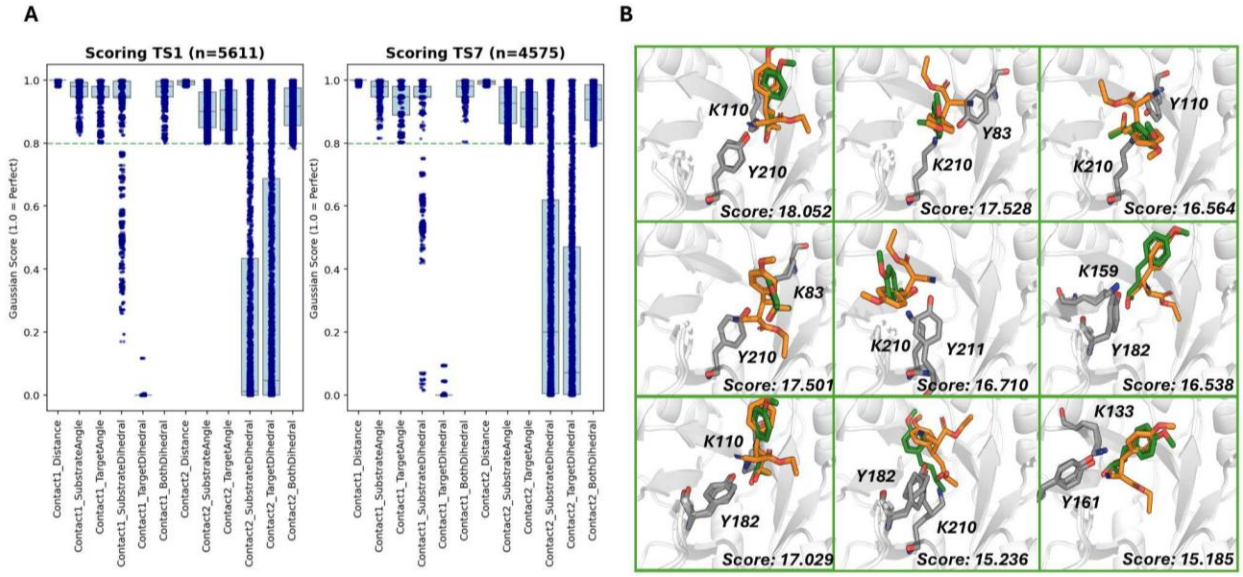

Figure 3.2: Lys/Tyr dyad placements on 1A53 ensemble members and scoring of best-quality motifs

A.) All generated Lys/Tyr motifs on all 1A53 ensemble members were scored by their accuracy to all 6 geometric degrees of freedom for each relevant residue. Lys contact (Contact_1) and Tyr_contact (Contact_2) are both present in TS1/TS7, each with 6 geometric degrees (perfect score = 24). Contact geometries are scored based on a Gaussian profile normalization allowing direct comparison of all different motifs. **B.)** Comparison of the passing generated motifs, with TS1 (sticks, green), TS7 (sticks, orange), and the catalytic residues (sticks, grey) shown.

Filtering for theozyme contact quality yielded 9 unique motifs capable of accommodating the dyad. The highest-scoring design positioned the dyad at K110 and Y210. The top tier of these motifs utilized positions 83, 110, and 210. Interestingly, these residues are catalytically relevant positions that arise throughout the RA95 evolutionary trajectory. These residues are repeatedly known to act as either direct catalytic participants or active-site hydrogen-bond donors. Given these findings, the K110/Y210 dyad motif was selected as the foundation upon which the remainder of the active site would be constructed.

3.2.2 Scaffold Refinement and Conformational Ensemble Generation

Following the successful anchoring of the Lys/Tyr dyad, the remaining histidine-mediated theozymes of the consensus mechanism were attempted to be placed. However, the native 1A53 scaffold lacked the necessary coordinates to support the catalytic histidine with high geometric precision. To overcome this, a CANVAS-inspired structural modification was used to rebuild the active site.

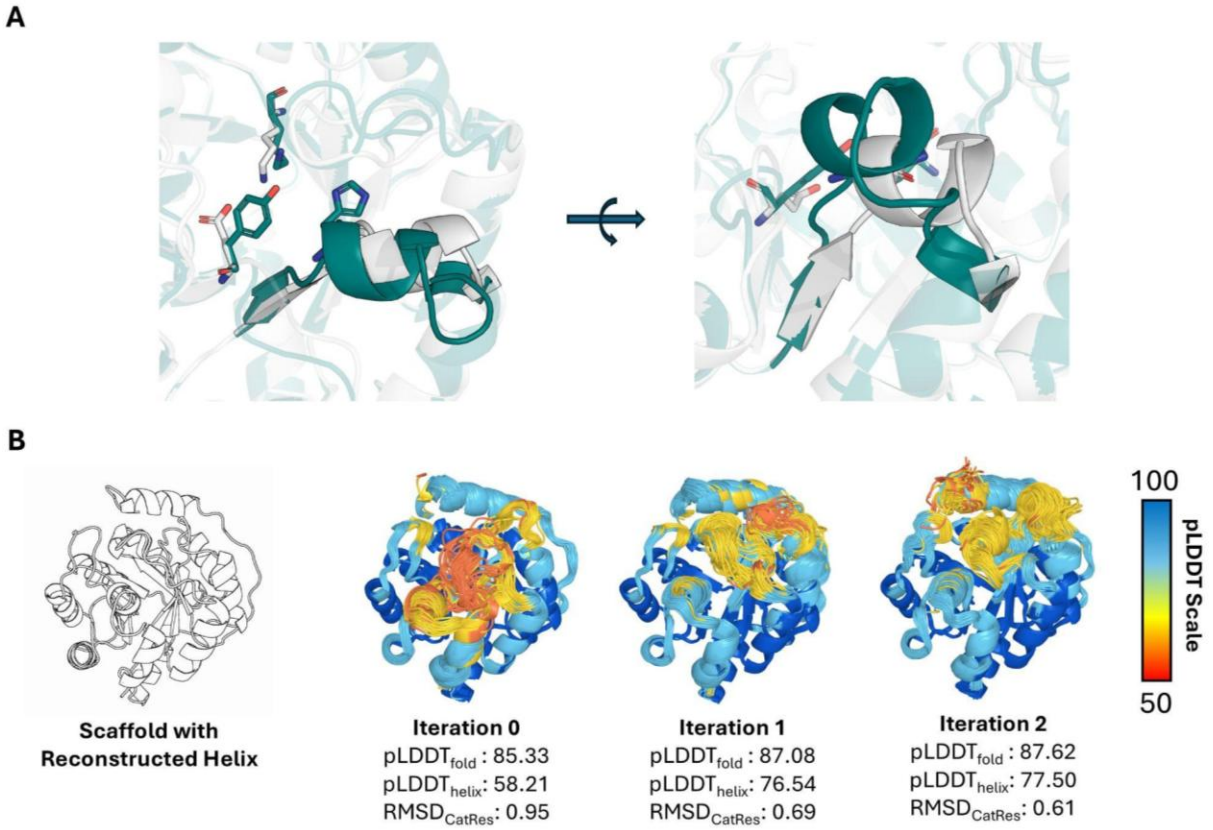

Figure 3.3: Histidine-accommodating helix reconstruction and scaffold refinement results

A.) RFDiffusionAA-generated backbone (dark green) aligned to 1A53 ensemble-member accommodating the lys/tyr dyad (white). Designed helix is opaque shading, with positions 110, 210 and 233 shown in sticks in both the parent scaffold and the designed template. **B.)** Results of the iterative scaffold refinement process after 2 iterations. The scaffold with the reconstructed helix is the customized TIM barrel backbone (cartoon, white). Boltz-2 ensembles are pictured for each top-scoring design at each iteration, with key design metrics pictured. Ensembles are colored from red (worst)-to-blue (best), indicating the CA pLDDT at the given position.

An optimal α-helical fragment was generated to provide the ideal backbone vectors for the histidine contact. This structural element was connected to the modified 1A53 scaffold utilizing RFdiffusionAllAtom. Introducing this non-native secondary structure generated local distortion in the backbone geometry.

To relieve the geometric distortion, the modified scaffold was subjected to two cycles of iterative backbone refinement. The refinement cycles demonstrated a clear and rapid convergence toward structural designability. While the overall confidence of the global fold marginally improved (average pLDDT increasing from ~85.3 to 87.6), the local structural metrics demonstrated validation of the refinement process. The pLDDT of the newly reconstructed α-helix increased from 58.2 to 77.5. As the backbone minimized into a more confident state, the active site geometry was preserved: the positioning of the catalytic triad Cα-C$^{\beta}$ vectors settled to a 0.61Å RMSD relative to the designed scaffold.

With the artificially reconstructed helix minimized and geometry stabilized after iteration 2, the static template was converted into an ensemble for multistate design. A synthetic ensemble was generated utilizing Rosetta Backrub, producing the backbone sampling necessary to use as templates for the remainder of the consensus mechanism theozymes.

3.2.3 Motif Generation of the Lys/Tyr/His Triad and Multistate Sequence Design

With the artificial helix-containing scaffold generated, the remainder of the consensus mechanism theozymes were placed using the K110/Y210 dyad and the newly constructed H233 position. Since each placement calculation generates a library of poses, the lowest RMSD alignment of the ligand heavy atoms was utilized to maintain the continuous geometry characteristic of the consensus mechanism. The placements captured the requisite rotameric shifts of the catalytic histidine across

the reaction coordinate, achieving high-precision geometries with distance deviations of less than 0.1 Å and angle deviations below 5° relative to the DFT theozyme structures.

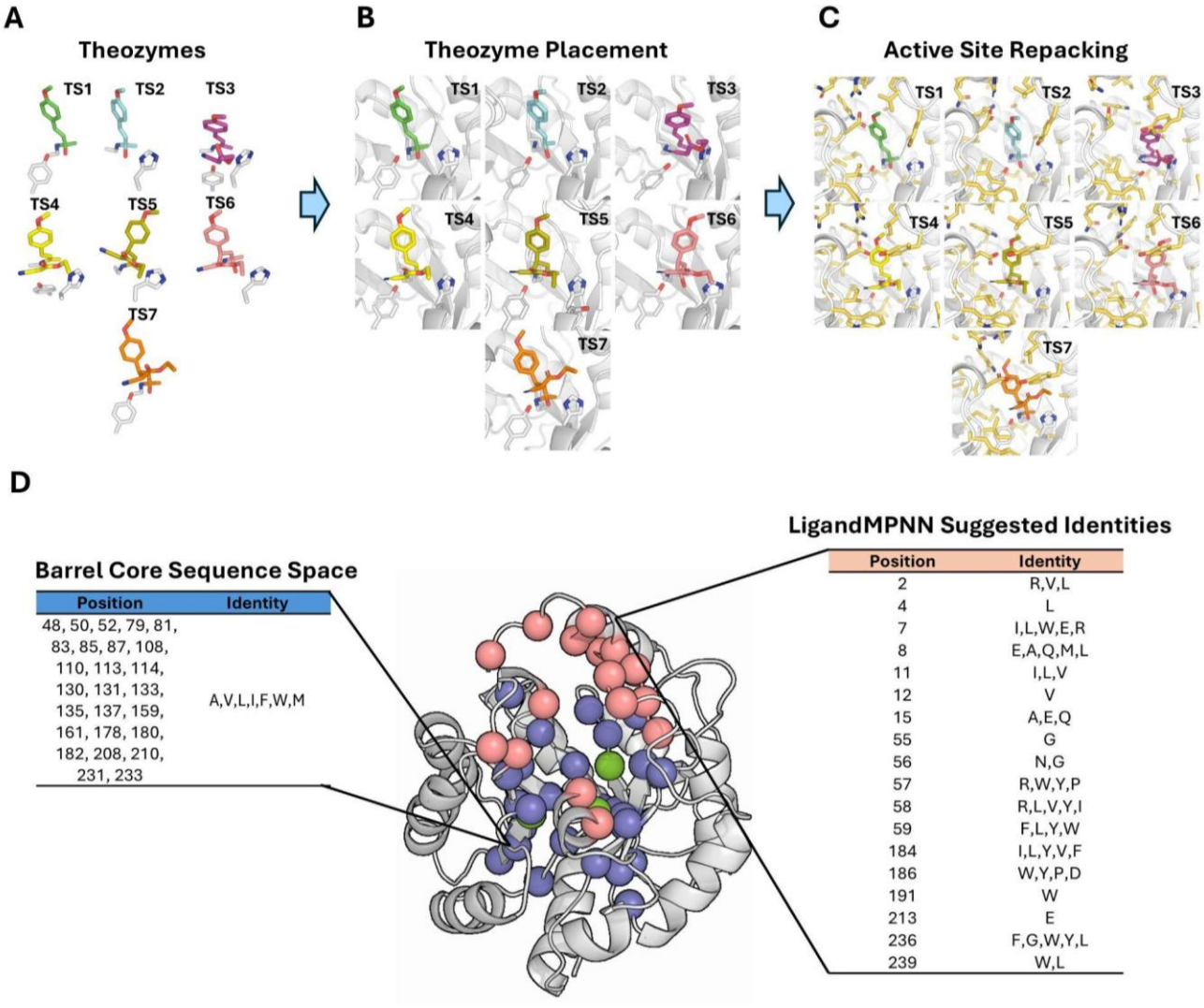

Figure 3.4: Lys-Tyr-His motif generation through complete mechanism theozyme placement and multistate active site design

A.) All seven theozymes are input into the placement algorithm across the generated backbone ensemble (colored, sticks). **B.)** Placed theozymes across the backbone ensemble utilizing the three designed catalytic residues (white, sticks). **C.)** Multistate sequence design (sticks, yellow) is utilized to repack the active site of all 7 theozyme placements, simultaneously. **D.)** An example of one of the backbone members input to the MSD algorithm. The evaluated sequence space is partitioned into a first shell barrel core (spheres, blue), and auxiliary set (spheres, pink). Green residues are the catalytic triad and remain fixed throughout the design.

Following integration of the catalytic triad, the surrounding active-site sequence space was defined for multistate design. A 44-residue design shell was constructed, comprising 26 first-shell and 18 auxiliary positions. To desolvate the active site, perturb the catalytic lysine pKa, and accommodate the Michael acceptor, the 26 first-shell positions were restricted to hydrophobic amino acids. Conversely, to pre-pack the active site and restrict the degrees of freedom of flexible loop regions during the MSD calculations, the sequence space of the 18 auxiliary positions were selected with LigandMPNN. This sequence space was subsequently injected to the MSD algorithm. Utilizing a sum combination function - which optimizes for the lowest cumulative energetic state across the entire evaluated ensemble - 4000 unique active-site sequences were generated to stabilize the complete reaction coordinate.

3.2.4 Consensus Filtering and Active Site Selection

In order to filter down the designed sequence list to a number that is experimentally manageable, a series of biophysical and structural filters were applied. For all filtering metrics, thresholds were defined explicitly with the objective of filtering the design list down to 15 sequences:

- Active site energy: Sustained scores below -300 kcal/mol across all transition states to ensure no active site clashes were introduced in the design, with a maximum energetic variance of 100 kcal/mol across the reaction coordinate ensuring no thermodynamic traps were introduced.
- Desolvation: A maximum relative solvent accessible surface area (rSASA) of 15% for all of the transition states to maintain active site hydrophobicity.
- Nucleophile activation: A maximum predicted pKa of 9.0 for the catalytic lysine in TS1 and TS7, ensuring lysine nucleophilicity for carbinolamine formation.

Finally, to isolate active site sequences that align with the consensus hypothesis, an interstate reorganization metric was incorporated. An active site design was classified as structurally "organized" if at least 60% of the designed first-shell residues maintained a maximum side-chain RMSD of 2.0 Å and a maximum interstate dihedral shift ($\Delta\chi$) of 30° between any adjacent states. This penalized sequences that would require large and energetically prohibitive side-chain rearrangements to accommodate sequential transition states.

	Active site energy score							LYS pKa		Structural scores		
Design	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS6**	**TS7**	**TS1**	**TS7**	**rSASA**	**Organization**	**$RMSD_{interstate}$**
1	-386	-390	-384	-432	-431	-399	-389	8.1	7.7	0.15	63.6	2.0
2	-348	-416	-373	-421	-434	-382	-348	8.1	7.6	0.14	68.2	1.7
3	-344	-406	-355	-412	-427	-371	-351	7.8	7.3	0.12	65.9	1.6
4	-342	-414	-366	-398	-414	-385	-316	8.1	7.9	0.14	65.9	1.7
5	-342	-414	-363	-400	-416	-379	-318	8.1	7.9	0.13	61.4	1.7
6	-342	-414	-363	-398	-413	-381	-317	8.1	7.9	0.13	63.6	1.7
7	-343	-415	-359	-419	-430	-395	-355	8.2	7.8	0.15	65.9	1.9
8	-351	-410	-357	-418	-429	-394	-351	8.2	7.9	0.14	63.6	1.7
9	-375	-392	-373	-411	-426	-367	-367	8.1	7.8	0.14	61.4	1.8
10	-349	-411	-357	-418	-429	-394	-350	8.1	7.9	0.14	63.6	1.8
11	-344	-416	-361	-414	-427	-391	-350	8.1	7.8	0.14	61.4	1.7
12	-352	-406	-357	-407	-421	-390	-343	8.1	7.8	0.15	65.9	1.6
13	-353	-401	-349	-408	-424	-393	-344	7.9	7.5	0.14	70.5	1.7
14	-381	-387	-391	-430	-435	-390	-375	8.1	7.6	0.14	65.9	1.8
15	-361	-389	-368	-429	-433	-404	-389	7.8	7.5	0.13	63.6	1.9

Table 1. Filtering parameters for resulting 15 designed active sites

Following the application of the catalytic and consensus-driven metrics, the 4000 generated designs from the MSD were filtered down to 15 active sites. These 15 active sites served as the templates for the subsequent shell redesign phase, enabling the design of the "positive" (consensus) and "negative" (anti-consensus) sequences.

3.2.5. Outer Shell Redesign and Sequence Selection

With the 15 core active-site sequences optimized and filtered, the remainder of the protein scaffold required sequence optimization to support the positioning of the catalytic cores. Designed active-site models were subjected to Shell Redesign utilizing LigandMPNN. To evaluate the stability of these sequences across the reaction coordinate, the generated sequences were predicted using Boltz-2, co-folded explicitly with both the reactant substrates and the final Michael addition product.

The objective of this phase is to ensure global and local structural stability independent of the ligand's reaction state. Sequences were filtered based on structural criteria: the ensemble deviation and overall fold RMSD required to remain below 1.5 Å relative to the Triad active site models. Furthermore, to ensure the active site did not collapse between chemical states, the catalytic side chains across triplicate predictions were restricted to a maximum RMSD of 2.0 Å, with a maximum $C\alpha$-C^{β} interstate RMSD between the reactant and product-bound states of 1.0 Å.

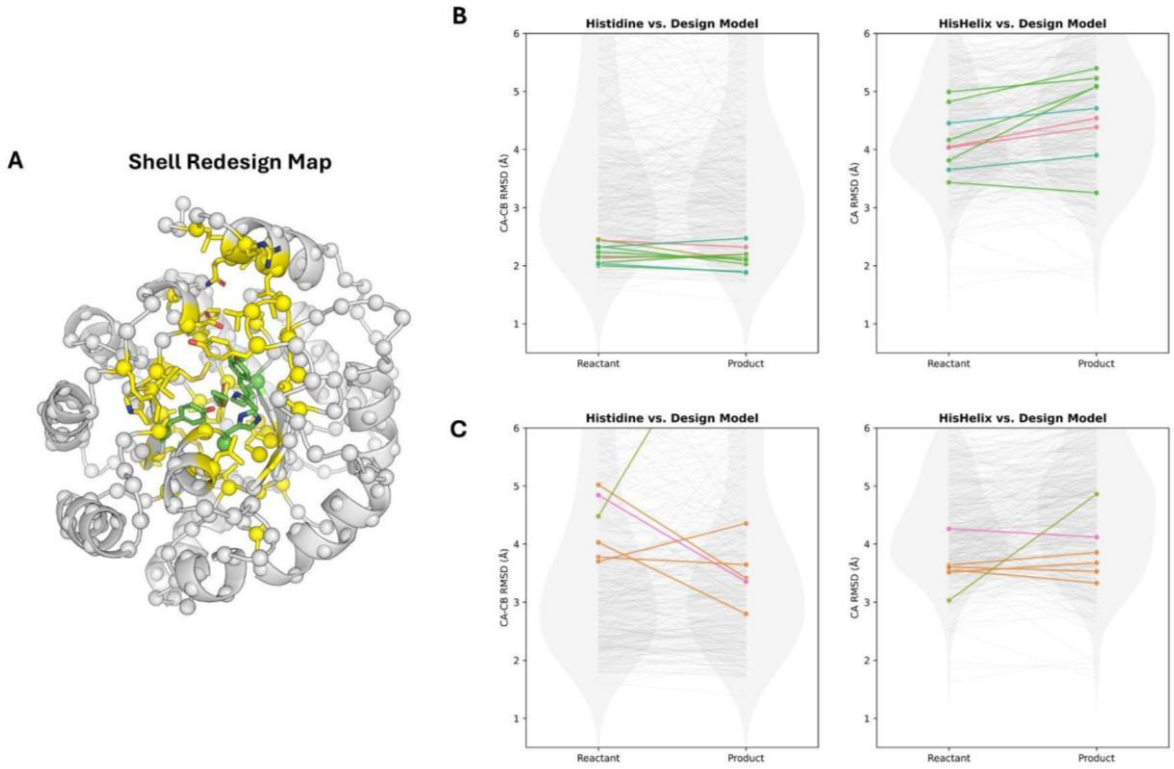

Figure 3.5: Sequence redesign of the outer shell residues on the modified backbone

A.) All 47 designed active site residue identities and rotamers were frozen (spheres, yellow and green), while shell sequence (spheres, white) was redesigned with LigandMPNN. The TS3 ligand is used as ligand context for the design model (sticks, green). In the two design graphs, the grey lines are individual sequence designs that did not pass the structural filter cut offs, while the colored lines are sequences that do pass the structural filter cut offs. **B.)** Positive consensus sequences are filtered by minimizing the predicted RMSD of the reactant and product states for the histidine. Histidine vs design model depicts the prediction-average CA-CB RMSD of the histidine side chain atoms relative to the active site design models. HisHelix vs. Design Model refers to the CA RMSD of the entire designed helix fragment that scaffolds the histidine, relative to the design model. **C.)** Negative anti-consensus sequences were selected for maximizing the histidine RMSD prediction-average CA-CB RMSD of the histidine side chain atoms relative to the active-site design models.

With this, the final control strategy designed to test the consensus hypothesis was implemented.

For each of the 15 active sites, two sequences were selected:

1. Consensus designs: sequences where the catalytic histidine maintained a maximum predicted RMSD of < 2.5 Å relative to the active site models, while maintaining all other structural filters.
2. Anti-consensus designs: Sequences where the catalytic histidine was predicted to deviate significantly, defined by a minimum RMSD of > 2.5 Å, while maintaining all other structural filters.

This paired selection yielded 30 final designs (15 positive, 15 negative) targeted for experimental evaluation.

3.2.5.1 Computational Observations of Shell Redesign

During the structural analysis of the Boltz-2 calculations a limitation was observed. While the global fold metrics and catalytic side-chain geometries satisfied the filtering criteria, the specific α-helix reconstructed to support the catalytic histidine exhibited high backbone RMSD relative to the design models. Despite the catalytic side chains remaining correctly positioned in the Boltz-2 predictions, the localized misfolding of the structural backbone was noted preceding experimental characterization.

3.3 Experimental Evaluation Results

3.3.1 Protein Expression and Biophysical Characterization

To evaluate the physical viability of the computationally generated enzyme designs, the 30 selected sequences (15 consensus “positive” and 15 anti-consensus “negative” designs) were recombinantly expressed in *E.coli* BL21(DE3)-Gold. The native 1A53 scaffold and the highly

active, laboratory-evolved Michaelase RA95.5-8 T53L/K210H (referred to as RA95_E for brevity) were expressed in parallel as controls.

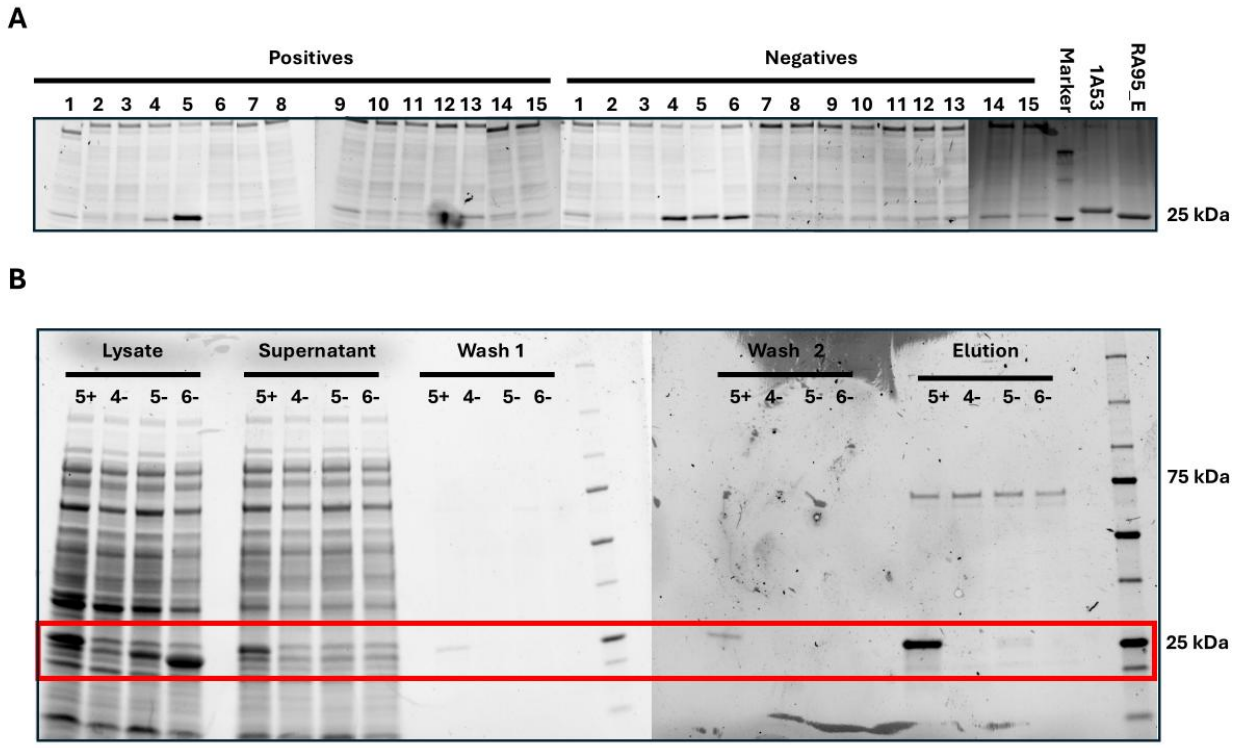

Figure 3.6: Expression screening of designed Michaelase candidates

A.) SDS-PAGE analysis (~25kDa-75 kDa) of small-scale expression & purification of designed Positive consensus sequences and Negative anti-consensus sequences. Parent template 1A53 and RA95.5-8 T53L/K210H variants (RA95_E) are loaded as positive controls. All 47 designed active site residue identities and rotamers were frozen (spheres, yellow and green), while shell sequence (spheres, white) was redesigned with LigandMPNN. The TS3 ligand is used as ligand context for the design model (sticks, green). **B.)** Scaled up expression and purification of best-expressing positive consensus sequences and their negative anti-consensus counterparts, aligned with the purification fractions.

Initial expression screening highlighted systemic limitations within the designed sequence space. While the 1A53 and RA95_E controls yielded robust soluble protein concentrations (~100 and ~77 mg/$L_{culture}$, respectively), the computationally designed variants exhibited compromised expression profiles. Total protein content for the designed variants rarely exceeded 23 mg/$L_{culture}$

following Ni-NTA purification and desalting, with SDS-PAGE analysis revealing that the majority of this yield consisted of expression chassis co-contaminants (notably a prominent ~70 kDa species) rather than the target proteins (~25 kDa). Comparison of the whole-cell lysate and soluble supernatant fractions confirm that most designed sequences partitioned into the insoluble fraction, indicating possible misfolding and inclusion body formation.

Of the 30 screened variants, only a single consensus design - variant "5+" - exhibited sufficient partitioning into the soluble phase in a scaled up expression to warrant detailed characterization. Variant "5+" was further purified by Size Exclusion Chromatography (SEC) alongside the 1A53 template. Chromatographic analysis revealed that the major elution peak for variant 5+ shifted approximately 1 mL earlier than that of the native 1A53 scaffold, potentially indicative of aggregation or other structural defects. SDS-PAGE of the SEC fractions also demonstrate a clear isolated band for both proteins of interest, confirming purification quality.

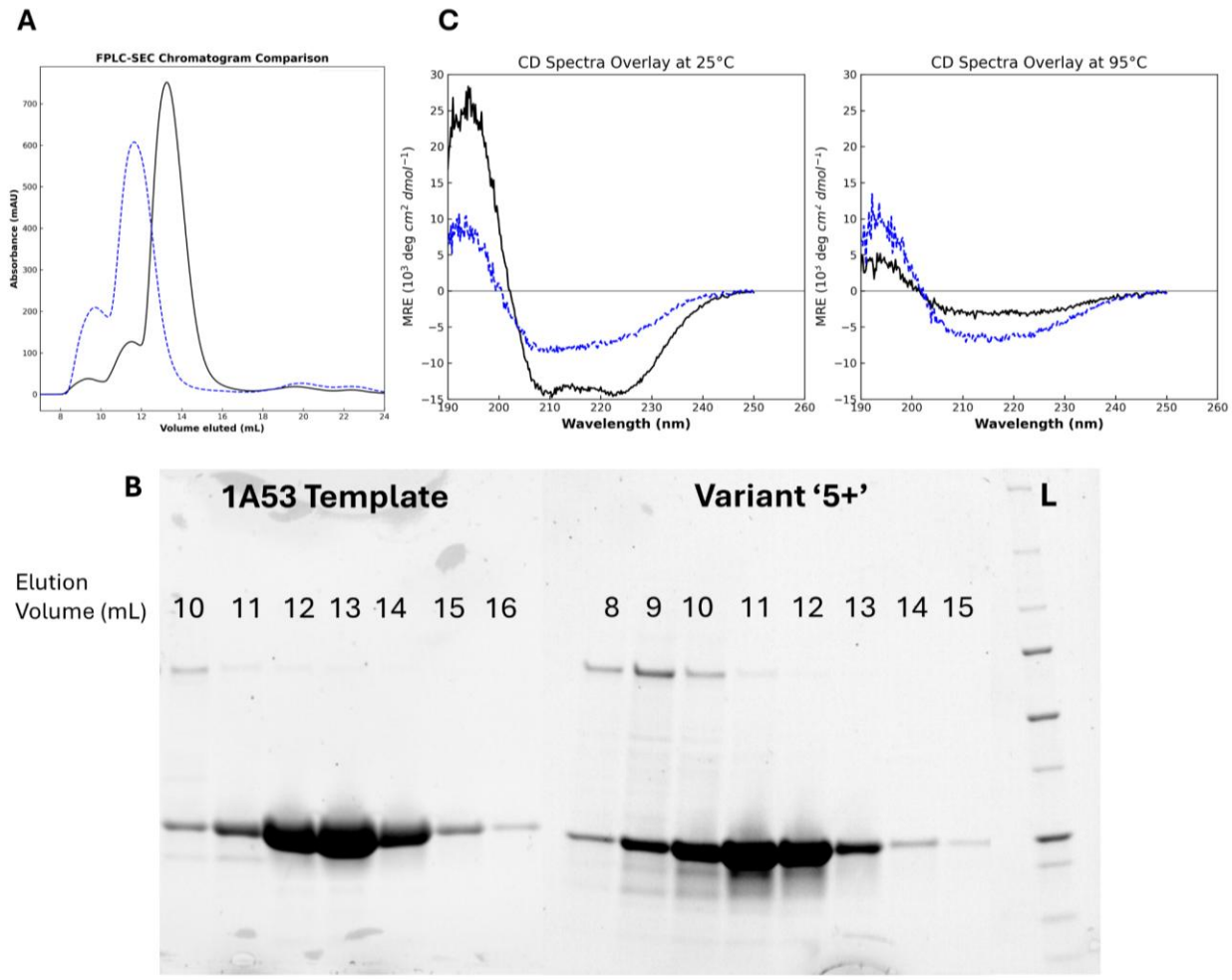

Figure 3.7: Size-exclusion chromatography and structural analysis of variant '5+' and parent template 1A53

A.) SEC chromatogram of the elution profile of both proteins (blue-dotted: variant '5+', black-solid: 1A53). **B.)** SDS-PAGE of selected volume range comparing the parent template and the Variant '5+'. **C.)** Comparison of CD spectra for both variants at 25 °C and 95 °C, scanned between 190 and 250 nm.

Table 2. Predicted secondary structure composition of SEC purified variants with BeSTSel Server [85]

Protein	Temperature	Alpha helix	Parallel Sheet	Anti Parallel Sheet	Turn	Other
1A53	25	40.8	7.5	2.8	13.7	35.2
	95	1.8	1.3	31.9	14.4	50.6
5+	25	18.1	7.7	18.9	14.4	41
	95	8.1	9.6	25.5	12.9	43.9

Subsequent structural evaluation of the SEC-purified protein via Circular Dichroism (CD) spectroscopy corroborated the chromatographic analysis. At 25° C, the 1A53 template exhibited the canonical α-helical signature expected of a native TIM-barrel fold. Conversely, the 5+ variant demonstrated an attenuated signal dominated by β-sheet characteristics. Furthermore, thermal denaturation obtained at 95° C revealed that while 1A53 underwent expected unfolding (a ~39% reduction in α-helical signal), the 5+ exhibited minimal structural shift (~10% reduction). Together, the SEC elution-peak shift, increased β-sheet CD profile, and minimal structural shift suggest that the 5+ variant did not adopt the target TIM-barrel fold and potentially solubilized as an aggregate.

3.3.2 Biochemical Evaluation of Designed Michaelases

Despite the biophysical indicators of structural misfolding, variant 5+ was evaluated for Michaelase activity to assess catalytic competence. Kinetic assays were conducted over a 6-hour time course across four Michael acceptor concentrations, comparing the purified 5+ variant against the RA95_E positive control, the 1A53 negative control, and a baseline no-enzyme control.

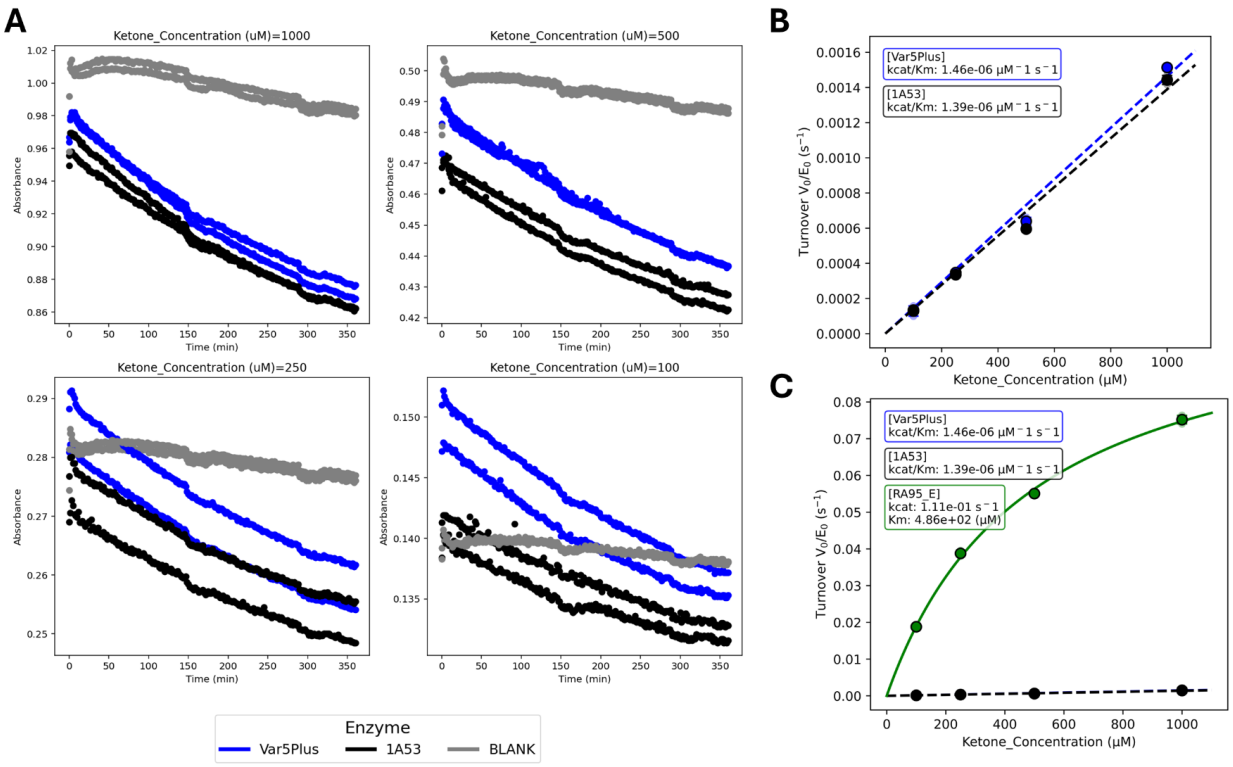

Figure 3.8: UV-vis Kinetic assay monitoring on variant 5+ and control enzymes

Assays were conducted at 29°C in the assay buffer (25 mM HEPES, 100 mM NaCl, pH 7.5) monitoring **370 nm.** Results here were conducted in technical duplicate (*n=2)*. Four concentration ketone (michael acceptor) gradient was used to evaluate apparent activity of the designed enzymes, with the ethylcyanoacetate (ECN/michael donor) in excess. **A.)** Kinetic traces over all evaluated substrate concentrations comparing variant 5+ with the parent template 1A53, and a no-enzyme negative control. **B.)** First order michaelis-menten fit of variant 5+ and template. **C)** Comparison of kinetic data and parameter fits for the evolved Michaelase RA95_E, variant 5+, and 1A53.

While variant 5+ exhibited a kinetic trace with an elevated activity above the no-enzyme control at all substrate concentrations, its activity profile was not significantly distinguishable from the native 1A53 template. In addition, over the substrate range evaluated the kinetic curve did not demonstrate Michaelis-Menten kinetics. The absence of observable catalytic rate enhancement relative to the starting scaffold confirms that the multistate design protocol failed to install the targeted consensus geometry into a functional scaffold. The minimal substrate consumption

observed is more likely due to the background promiscuity of the reaction with the protein rather than due to the computationally designed active site.

3.4 Chapter Summary

In summary, this chapter details the development and implementation of a multistate computational design pipeline aimed at realizing a contiguous, 7-step consensus mechanism for the Michael addition. By pairing a CANVAS-inspired localized structural reconstruction with a scaffold refinement protocol and Rosetta Backrub ensemble generation, a derivative of the native 1A53 scaffold was designed with a Lys/Tyr/His catalytic triad. Subsequent Multistate Sequence Design coupled with consensus-derived filtering generated an active-site sequence space optimized to explicitly accommodate the targeted reaction coordinate. However, experimental evaluation suggested a failure of these designed sequences to adopt the targeted TIM-barrel fold *in vitro*. This misfolding is indicated by inclusion body formation during expression of the majority of the designed sequences. The sole expressible variant, sequence 5+, demonstrated a non-native β-sheet-rich character relative to the design template. Crucially, the *in silico* Boltz-2 structural predictions generated during the outer-shell redesign phase foreshadowed this outcome, highlighting local design instability and degradation of the artificially reconstructed histidine-containing α-helix.

Chapter 4: Discussion and Future Perspectives

4.1 Introduction

This work derived and mapped a complex 7-step transition state coordinate for a Michael addition (Specific Aim 1). However, translating this theoretical consensus mechanism into a functional biocatalyst (Specific Aims 2 and 3) revealed limitations in current machine learning-guided multistate design protocols. While the sequence redesign successfully minimized energetic clashes *in silico*, experimental validation yielded inactive proteins with potential misfolding of the target structure. Consequently, it is highly unlikely that the computationally designed active-site geometries manifested *in vitro* - preventing direct evaluation of the consensus mechanism hypothesis. Therefore, the scope of this discussion pivots to evaluate the theoretical viability of the designed consensus mechanism, assess the methodological failures that led to misfolding, and propose modifications to the current work to overcome these limitations in future enzyme designs.

4.2 Theoretical Viability of the Consensus Mechanism

4.2.1 Energetic Comparisons Against Relative Enzymes

To evaluate the physical viability of the derived 7-step consensus mechanism, the activation barriers are compared against carboligase and lyase enzymes, specifically the Class I fructose-6-phosphate aldolase and the evolved RA95 lineage. Notably, the computed mechanism for the Michael addition does not exhibit any single barrier exceeding 13 kcal/mol. The rate-limiting steps were identified as TS1 (initial carbinolamine formation), followed by the dehydration and double-bond protonation events (TS2 and TS5). This compares well to two benchmark RA95 variants. The RA95.5-5 variant exhibited a rate-limiting C-C bond scission barrier of 29.1 kcal/mol. The

further evolved variant RA95.5-8F shifted the rate-limiting step to the enamine-to-Schiff base water elimination with a barrier height of 20.1 kcal/mol [49]. The activation energy for TS1 in the derived consensus mechanism is 12.9 kcal/mol, highly similar to the 11.9 kcal/mol barrier of RA95.5-8F [49].

The exact mechanistic steps involving the carbinolamine formation and collapse also varies between the derived Michael addition and prior studies. In the RA95 lineage, the carbinolamine formation/collapse proceeds though a two-step process, initiated by the concerted tyrosine proton abstraction by the ketone and N-C bond formation. The tyrosine enolate intermediate, while unstable (activation barrier of <1kcal/mol in RA95.5 and ~2.5 kcal/mol in RA95.5-8F), demonstrates a distinct stationary point on the QM/MM energy surface. This enolate subsequently deprotonates the carbinolamine nitrogen. A similar mechanism is observed in fructose-6-phosphate aldolase - the carbinolamine proceeds through three transition states [72]. First, the nucleophilic lysine forms a zwitterionic carbinolamine, which deprotonates the adjacent tyrosine to form the tyrosine enolate, which then abstracts the proton from the carbinolamine lysine. The computed barrier heights are similar to those of the RA95 variants. While the exact mechanistic pathway and barrier heights vary, potentially due to how this sort of cluster-DFT model lacks the heterogenous dielectric environment of a fully solvated protein, the relative energetic parity with full-enzyme QM/MM studies suggests the theoretical mechanism is chemically viable.

4.2.2 Rotameric Interstate Shifts as a Path Shortening Mechanism

A theoretical advantage of consensus mechanisms was the observation of finding side-chain repositionings that would align to rotameric shifts. The consensus mechanism revealed that these theozymes navigate the mechanism through efficient side chain dynamics, rather than a large backbone reorganization. Specifically, the catalytic histidine demonstrates χ_2 rotamer

interconversion between states. Operating first as an acid to generate the Schiff base, the histidine then acts as a base to generate the ethylcyanoacetate carbanion following a distinct χ_2 rotation. The resulting imidazolium hydrogen-bonds to the ECN ester oxygen, positioning it for C-C bond formation, before undergoing another χ_2 rotation back to the original pose to protonate the double bond and regenerate the imidazole. There are examples of rotameric interconversions in other catalytic mechanisms, such as human carbonic anhydrase II [73]. Here, histidine rotates between an unprotonated "inward" state and a protonated (solvent-facing) "outward" state. In serine proteases, the rotameric shifts of the $\chi 1$ serine is critical for generating the first tetrahedral intermediate. The rotameric interconversion in the derived consensus mechanism acts as a path-shortening mechanism, demonstrating how a minimal-RMSD consensus geometry can dynamically satisfy multiple transition states provided the active site allows these rotamer interconversions.

4.2.3 Steric Context Enables Consensus Design via H-bonding Network

The localized rotameric flipping of the catalytic histidine is strictly dependent on the steric and electrostatic context provided by the active-site spectator network, specifically the explicitly defined water molecule and adjacent tyrosine. Generated during the water elimination step (TS2), this water molecule was found to participate in an extended hydrogen-bonding network bridging the nitrile group of the ethyl cyanoacetate donor, the tyrosine hydroxyl, and the Schiff base amine in steps 3-5. Incorporating this bridging network was mandatory for achieving low-RMSD alignments. However, relying on an explicit water molecule to maintain active site organization would require a large entropic penalty through restriction of the translational and rotational degrees of freedom. While the cluster-model suggests that this hydrogen-bonding lattice is necessary to rigidify the transition state complex, it may also not be a requirement that this network is mediated exclusively by a free water molecule. Instead, this water molecule can act as a structural blueprint

for the geometric requirements of the active site. Further design protocol modifications could replace this solvent molecule with a covalently tethered amino-acid side chain acting as a hydrogen-bonding partner to structurally recapitulate the bridging network. As such, a tethered residue could preserve the critical low-RMSD steric constraints of the consensus mechanism while eliminating the entropic cost of ordering an explicit solvent molecule in the catalytic pocket.

4.3 Structural Limitations of Machine Learning Guided Scaffolding

4.3.1 Scaffold Designability and Structural Incompatibility of Helix Repositioning

From a purely computational perspective, the multistate design protocol functioned as intended: the theozyme placements demonstrated low RMSD relative to the DFT models, and the final sequence space did not demonstrate energetic clashes within the active site while maintaining features relevant to catalytic turnover. This demonstrates that the MSD algorithm can stabilize a consensus mechanism *in silico*. However, the folding of the designed sequences does not enable direct testing of the consensus hypothesis. The structural deviation observed in the CD spectra and the increase in expected β-sheet character in sequence 5+, alongside the systemic inclusion body formation across the designed sequences point toward a limitation in the scaffold design strategy. By forcing the native TIM-barrel α-helix into a geometry dictated by the catalytic residue positioning, the backbone may have been pushed into an undesignable sequence-structure space. For example, the loop following the β-strand was reconstructed to accommodate the histidine. However, this loop conformation may not be compatible with the highly conserved β-barrel geometry. While the CANVAS workflow demonstrates that functional helix motifs can be introduced on TIM-barrels, the design strategy might be enabled directly by the use of minimal *de*

novo TIM-barrels as the initial scaffold. These types of idealized topologies might tolerate these types of structural modifications better compared to evolved scaffolds.

4.3.2 Minimal Scaffold Refinement May Limit Designability

The widespread misfolding observed across the designed sequences could be physically correlated to the short structure refinement phase. The protocol utilized only two iterations of alternative sequence design and structural prediction. While short refinement cycles (e.g. 3 iterations) have been shown to be sufficient for simple *de novo* binders or minimal helical bundles, designing a deep active site for a ~250 residue α/β-topology like a TIM barrel might be more difficult to achieve than a minimal structure. Recent literature from the Polizzi lab demonstrates a pipeline to address this structural self-consistency challenge through up to 32 refinement cycles utilizing dual neural-network workflows (LazerMPNN, RosettaFold-AllAtom) [74]. In addition, they show that utilizing Rosetta minimization to relax the structure over iterations does not converge the design metrics at all. The widespread aggregation and inclusion body formation observed in the expression screening here might be due to this "incomplete" optimization. At cycle 2, the algorithm may have satisfied the local geometry of the Lys/Tyr/His triad, but non-active site structure might've been trapped in a minima containing unresolved geometric errors, reducing the designability of the scaffold. This is specifically observed by the lower quality design metrics of the histidine helix after the Shell Redesign. If the initial scaffold was designable, then redesigning the shell sequence should not have resulted in low quality designs.

4.3.3 Unstructured Active Site Loops are Challenging to Predict with High Confidence

The use of the folding model Boltz-2 in the Shell Redesign also revealed large structural deviance within the active-site loops. These active site loops consistently demonstrate high RMSD and

lower pLDDT, which may be concerning as parts of the loops constitute the active site packing interactions designed by the MSD algorithm. A recurrent challenge in ML-guided design is decoupling sequence designability from intrinsic loop elasticity [75,76]. Because these models predict the final folded state rather than the folding path, it remains ambiguous whether a low loop pLDDT indicates a fundamentally undesignable structure or if the model is simply (and accurately) reflecting high conformational elasticity inherent to that region. However, given that the consensus mechanism requires organized positioning of the active site, building a complex 7-step mechanism against loops prone to high structural variance introduced a vulnerability to the design workflow.

4.4 Improvements and Future Directions

4.4.1 Development of an Automated Consensus Theozyme Protocol

While this work successfully derived a minimal-RMSD consensus mechanism for the Michael addition, the current methodology relies on a highly time- and labor-intensive framework. Deriving the 7-step coordinated required manual intervention, including iterative structural correction, constrained geometric scans, and manual alignment to identify the consensus-qualifying theozymes. As the enzyme design field scales toward increasingly complex, non-natural multistep cascades, the natural evolution of this aim is to develop an automated/generalized computational workflow capable of deriving consensus theozymes for any user-defined reaction coordinate.

A proposed algorithmic solution entails a two-cycle, “core-to-shell” expansion pipeline building upon established theozyme frameworks. In the initial cycle, each step of the reaction mechanism would be evaluated *in vacuo*, isolating only the reactant substrates and primary catalytic atoms to generate DFT-validated transition states (TS) devoid of spectator atoms. In the secondary cycle, the necessary spectator groups of the mechanism (explicit waters of hydrogen bond donors) are

introduced into the computed geometries. An automated conformer search algorithm, driven by rapid semi-empirical methods (GFN2-xTB) or machine-learning potentials (MLPs), would then sample conformers while constraining the DFT-validated TS core distances and angles [77,78].

Finally, an automated algorithm would evaluate the multi-step conformational ensembles, minimizing the heavy-atom RMSD across the reaction coordinate to identify overlapping geometric states. This automated approach provides two distinct advantages. First, it yields a highly generalizable, reaction agnostic software pipeline. Second, rather than producing a single consensus mechanism, the algorithm would generate a diverse library of consensus mechanisms. These distinct candidate theozymes could then be computationally ranked by their overall activation barriers relative to the reactant ground states, ensuring that downstream sequence-structure scaffolding is only performed on the most thermodynamically accessible catalytic geometries.

4.4.2 Methodological Improvements to Sequence-Structure Scaffolding

The structural collapse of the 1A53-derived sequences indicates that imposing large structural modifications onto naturally evolved backbones to scaffold complex consensus mechanisms creates fatal topological incompatibilities. Therefore, an immediate methodological improvement should involve transitioning towards highly designable and structurally idealized templates, such as minimal *de novo* TIM barrels. Early iterations of these synthetic folds were heavily rigidified; to maintain stability in the absence of evolved core packing, their β-α loops were co-opted for purely structural hydrogen-bonding. Consequently, introducing polar or charged mutations deep into the central β-barrel to install a catalytic network was commonly destabilizing. While approaches like CANVAS partially alleviate this spatial restriction by appending a catalytic lid-motif, a more fundamental structural solution lies in the utilization of *de novo* ovoid TIM barrels,

such as OT1 (PDB ID: 7EUK) [79]. By adopting an ovoid curvature, OT1 naturally breaks core symmetry to generate a defined central cleft, decoupling the β-α loops from their stabilizing role. Importantly, the thermodynamic stability of this ovoid scaffold is not primarily driven by deep-core hydrophobic packing, allowing for the introduction of buried polar or charged residues that would destabilize a circular fold. For the specific consensus mechanism derived in this thesis, this plasticity unlocks the utilization of deep-barrel positions to anchor the Lys/Tyr dyad, dramatically expanding the designable motif space. Scaffolds like OT1 may provide a clean, evolutionarily unburdened scaffold ideally suited to absorb the geometric shock of the complete Lys/Tyr/His triad and its accompanying explicitly water network.

Another approach would be to move beyond the utilization of existing templates, and towards fully *generative,* catalytic motif-driven scaffolding. However, generative diffusion models often need to generate large libraries of designed outputs to yield viable biocatalyst scaffolds. Drawing inspiration from recent *de novo* design studies, such as the RiffDiff pipeline which utilizes predefined metrics for localized ligand burial (e.g. targeting > 7.1 Cα atoms within 8Å of the substrate), future protocols should define rigorous structural boundary conditions prior to diffusion [80]. By establishing formalized heuristic criteria for multistep carboligases - quantifying ideal secondary structure content, active-site radius of gyration, and localized solvent accessibility - computational models could be explicitly conditioned to generate backbones that inherently possess the requisite desolvation environments and packing geometries. This ensures the generative output is topologically primed to support the multistep consensus mechanism from the very first step of sequence-structure generation.

4.4.3 Algorithmic Improvements to Sequence-Structure Refinement

A systemic limitation in the current design workflow was the breakdown in sequence-structure self-consistency following the Shell Redesign phase. In current machine learning-guided workflows, candidate scaffolds are iteratively optimized via some sequence-design and folding prediction loop, until the predicted structure meets the selected convergence criteria. The truncation of this refinement phase to two cycles fundamentally handicapped this refinement. As demonstrated by the degradation of the histidine-anchoring α-helix in the shell redesign phase, two cycles provided insufficient optimization depth for the sequence and folding models to converge on a highly designable solution.

Resolving this self-consistency bottleneck requires transitioning to high-cycle iteration. While much *de novo* literature relies on shallow refinement - such as the 3-cycle iterations used in the serine hydrolase, retroaldolase, and MBHase design pipelines - these protocols are used in application to relatively simple folds like 4-helix bundles or NTF2 folds. Increasingly complex structures like TIM-barrels may require deeper optimization. The NISE pipeline demonstrates that achieving true self-consistency requires prolonged cross-talk between the sequence generator and folding model. Future work should adopt these high-iteration pipelines, enforcing stricter convergence filters (e.g. active-site pLDDT>95 and active-site RMSD < 1.0Å) prior to experimental evaluation.

However, if in the case that an initial scaffold resides in a completely undesignable region of sequence-structure space, high-number of refinement cycles may not guarantee a designable fold. Therefore, a critical algorithmic improvement is the parallel maintenance of diverse structural pools throughout the refinement loop. Rather than bottleneck the entire computational pipeline through a single target fold (e.g., the specific 1A53 derivative utilized here), future workflows should evaluate multiple (e.g. 3-5) distinct candidates simultaneously. By maintaining a parallel

trajectory approach, the design algorithm minimizes the risk of single-point scaffold failure. If one topological variant is unable to converge on a high-pLDDT active site during sequence refinement, it can be culled, reallocating computational resources to the more designable scaffolds within the pool.

4.5 General Conclusions

The *de novo* design of multistep enzymes remains one of the most challenging frontiers in computational protein design. Historically, the field has largely restricted its focus to single step mechanisms optimized around a single, rate-determining transition state. This thesis addressed a previously unexplored aspect of *de novo* biocatalysis by deriving a contiguous, minimal-RMSD consensus mechanism for a complex, 7-step Michael addition. The quantum mechanical evaluation conducted in this work demonstrates that the shifting geometric demands of highly complex multistep mechanisms can be theoretically satisfied within a designed active site. Crucially, this theoretical viability is not achieved through unphysical static compromises, but by leveraging local dynamic plasticity - specifically the path shortening χ_2 toggling of the catalytic histidine, supported by a reorienting explicit spectator H-bond network. Energetic comparisons against mechanistically relevant enzymes suggest that this derived consensus mechanism is chemically viable, fulfilling the objectives of Specific Aim 1 and providing a novel theoretical blueprint for multistep enzyme design.

However, the subsequent attempt to physically realize this theoretical mechanism underscored a critical bottleneck in the scaffolding phase of the design pipeline. The observed misfolding of the designed 1A53 derivatives demonstrated that a rigorously derived consensus mechanism cannot

be experimentally evaluated if the chosen protein chassis lacks the topological designability to support it. The experimental outcomes of Specific Aims 2 and 3 revealed that forcing complex, non-natural geometric constraints of a multistep theozyme onto a naturally evolved TIM-barrel pushes the backbone into an undesignable sequence-structure space. Furthermore, the reliance on shallow, low-cycle machine learning refinement protocols proved inadequate for resolving these designability barriers, manifesting as misfolded structures and systemic inclusion body formation.

While the engineered sequences failed to yield functional biocatalysts, this negative result identified specific methodological limitations of the current design workflow. By diagnosing the exact failure points of native template redesign, this work establishes an actionable roadmap for future multistate enzyme design protocols. Transitioning to unburdened *de novo* templates or explicitly conditioning generative diffusion models to hallucinate bespoke topologies, coupled with deep, parallel-trajectory refinement loops, provides the necessary computational framework to translate complex theoretical theozymes into physically active biocatalysts.

Methodology

Theozyme Design

Quantum mechanical calculations were conducted using semi-empirical tight binding theory and density functional theory. For each given reaction step, Grimme's GFN2-xTB was utilized to scan the reaction coordinate of interest[81]. The highest energy image was subsequently optimized to the saddle point at the B3LYP-6-31G(d,p) level of theory with the D3BJ dispersion correction. Transition state optimization and energy calculations were all computed with implicit solvent correction utilizing the Conductor-like Polarizable Continuum Model (CPCM) using diethyl ether as the solvent with a dielectric constant of 4.33 to mimic the enzymatic active site hydrophobicity. Transition states were confirmed by vibrational frequency analysis utilizing the FREQ keyword. Individual TS calculations were computed utilizing the catalytically active functional groups relevant to that reaction step. The lysine residue is modelled as an ethyl amine, tyrosine as 4-methylphenol, and histidine as 4-ethylimidazole. Consensus transition states were computed by aligning catalytic atoms to adjacent or reversal transition states in PyMOL and then undergoing a three-step workflow. 1.) Scanning the reaction coordinate of interest (GFN2-xTB). 2.) freezing the catalytic atoms in the highest-energy image, optimizing non-frozen atoms, and computing the analytical hessian (keyword: OPT FREQ). 3.) re-optimizing the prior steps' converged system (keyword: OptTS) in the absence of constraints utilizing the prior hessian input, until a validated saddle point was identified. Free energy values computed for the theozymes were utilized for plotting the energy diagram of the computed mechanism. Intermediate calculations were computed by using the intrinsic reaction coordinate (keyword: IRC) method, optimizing reactant and product endpoints by freezing the position of the terminal methyl-carbon of each intermediate. All calculations were carried out in the ORCA 6.1.0 package [82,83].

Ensemble Refinement

Time-averaged ensembles were generated with phenix.ensemble_refinement in the PHENIX v.1.15.2-3472 package. In the ensemble refinement method, local atomic fluctuations are sampled using molecular dynamics simulations accelerated and restrained by electron density to produce ensemble models fitted to diffraction data. The crystal structure input from PDB code 1A53 was edited to remove low occupancy conformers and assigned an occupancy of 1.0 to the remaining conformer. Following hydrogen addition, parallel ensemble refinement simulations were performed using a combination of p_TLS (0.6,0.8,0.9,1.0), tao_x (0.5,1.5,2.0) and w_(ray) (2.5,5.0,10.0). p_TLS describes the percentage of atoms included in a translation-libration-screw (TLS) model used to remove the effects of global disorder, tao_x is the simulation time-step, and w_x_ray is the coupled tbath offset, which controls the extent to which electron density contributes to the simulation force field such that the simulations runs at a target temperature of 300 K. The ensemble that generated the lowest R_free value was selected for design and analysis.

Ligand Attribute Generation and Ligand Placement

All ligand placement calculations were performed with the Triad protein design software (Protabit, Pasadena, CA, USA). Transition state ligands were extracted from QM design calculations, converted to PDB files from XYZ files, and atom naming was standardized across all 7 transition state ligands. Molecular attribute files were defined and generated utilizing the `makeLigandAttributes.py` and `makeHdefTemplate.py` apps in the Triad Software Suite.

The 2002 Dunbrack backbone-independent rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 were utilized to sample side chain conformations (keyword:

(bbind02.May.e2) [84]. Transition state poses were generated utilizing the targeted ligand placement pose generator utility in PheonixMatch.py called, utilizing all 6 contact geometries extracted from the appropriate theozyme model. Poses were generated with a ± 0.1 A translational distance, ± 5° angular range over two-degree intervals for the two angular contacts, and a ± 5° torsional range over 1 degree intervals for the torsional contacts. An exception was made on the torsional contact range for the lysine contact torsional range, allowing a 228-300° range over 10 intervals to allow rotamer sampling around the χ_4 dihedral whilst avoiding high-energy clashes with the transition-state complex.

TS pose energies were calculated using a modified version of the Phoenix energy function utilizing a Lennard-Jones 12-6 van der Waals term from the Dreiding II force field with atomic raddi scaled by 0.85, a direction-dependent hydrogen bond term with a well-depth of 8.0 kcal mol^(-1), an equilibrium donor-acceptor distance of 2.8 A, and an electrostatic energy term modelled using Coulomb's law with a distance-dependent dielectric of 10. An energy benefit of -100 kcal mol ^(-1) was applied to satisfy theozyme contacts geometries utilizing the ContactGeometryBiasFactory.py class in Triad, and an energy penalty of 500 kcal/mol was applied to poses that did not satisfy the theozyme contact geometries.

Two tiers of ligand placement calculations were conducted to manage the combinatorial space of the design calculations. Tier 1 included design calculations placing TS1 and TS7 were conducted first, where the catalytic lysine and tyrosine pair were allowed at all beta strand tip-positions of the beta-barrel core of 1A53, as well as all *N_tip+2* and *N_tip-2* positions. Tier 1 calculations utilized poses generated off of the catalytic lysine position tested. Following Tier 1 calculations, a matching calculation was conducted to determine combinations of Lys/Tyr positions that enable optimal alignment of TS1/TS7. Following matching, Tier 2 calculations included TS2-TS6, and

were conducted using fixed Lys/Tyr positions identified from Tier 1 calculations and tested positions for the catalytic histidine on all residue positions with the Cα-Cβ positions pointing into the beta-barrel. In all ligand placement calculations, all non-catalytic residue positions were mutated to Alanine to avoid steric clashes with designed catalytic residues.

Ligand Placement Insterstate Matching Algorithm

To identify catalytic motifs/sub-ensembles capable of stabilizing the complete reaction coordinate, a multistate consensus algorithm was implemented using a graph-theroretic clique detection approach, inspired by the consensus coverage algorithm implemented by [1]. Initially, an ensemble of scaffold structures with placed ligands (*S_i)* is iteratively evaluated against the theozyme geometries of each targeted transition state (*TS_j)*. Catalytic motifs lacking valid structural placements for the complete set of *N* transition states are discarded, ensuring catalytic motifs with full transition state coverage proceed to consensus mapping.

For each retained catalytic motif, the valid transition state placements are abstracted into an unidirected graph where each placement represents a node. Edges are constructed to represent the geometric consistency between differing transition states. An edge is drawn between a node representing state *TS_i* and a node representing *TS_j* according to two criteria 1.) the root-mean-square deviation of the aligned anchor atoms is below a defined threshold of ϵ =1.5, and 2.) the nodes represent a distinct transition state (*k_i =/= k_j).* This spatial relationship is formalized through an adjacency matrix *A*, defined mathematically as: A_ij = {1 if RMSD_ij < epsilon and k_j =/= k_j, and 0 otherwise}.

The extraction of a consistent multistate design is formulated as finding a *K-clique* within this graph. A valid multistate consensus is defined as a subgraph *Q* that strictly fulfills the two

topological criteria. First, it must exhibit geometric consistency, requiring that all nodes within the subgraph be mutually compatible within the RMSD threshold ($\forall u, v \in Q, \quad (u, v) \in E$). Second, it must satisfy the state coverage condition, ensuring that the unique transition state labels associated with the vertices in *Q* perfectly map to the full set of required transition states *T*. By isolating *K*-cliques where all component nodes possess distinct transition state labels, the algorithm extracts the exact combinations of generated motifs capable of accommodating the complete reaction mechanism.

C-terminal Helix Repositioning via Scaffold Refinement

To accommodate the catalytic histidine onto the 1A53 scaffold, the TS3 theozyme geometry was aligned to the TS1/TS7 matched TS ligand placements. The 4-ethylimidazole group in TS3 was utilized as an alignment anchor for an histidine inverse-rotamer library generated with the Invrotzyme application and then selected for structural similarity to the wild-type C-terminal helix [60].. The selected Histidine-containing helix was reconnected to the scaffold through reconstructing the appropriate N and C terminal peptide bonds with RFDiffusion-AllAtom (RFAA) [59]. Specifically, the 2 top β8 strand residues were reconstructed to connect to the N-term of the new motif, and 6 residues were remodelled to connect to the C-terminal. In addition, the loop connecting the β7 strand and α10 helix were simultaneously remodelled to sample a backbone conformation to accommodate the repositioned catalytic histidine motif. The three catalytic-face side loops on the WT 1A53 TIM-barrel were remodelled to better accommodate the placed TS1/TS7 TS ligands. For all diffusion calculations, the standard model weights were selected ("RFDiffusionAA_paper_weights.pt") with 100 diffusion steps. 50 designs were generated and

selected for maintaining α-helical secondary structure in the remodelled segment, as well as maintaining the preceding β-strand secondary structure.

Following scaffold generation, a refinement-style workflow was implemented. LigandMPNN was utilized to predict 1000 sequences on the designed scaffold, while keeping the catalytic residue positions of the Lys-Tyr-His triad frozen [61]. The co-folding model, Boltz-2, was then utilized to predict the fold of each sequence in the presence of the Michael addition product [66]. Each sequence was scored based on the RMSD of the Cα-Cβ over 3 predictions, relative to the RFAA design model. The best scoring sequence was then threaded onto the original scaffold, following which was refined in Triad. First, the protein was corrected for hydrogen addition with the addH.py application, and then refined with the StandardizeStructure.py application, allowing iterations of rotamer repacking and backbone minimization, altering the scaffold to accommodate the new sequence. The catalytic motif was kept frozen utilizing the -pointRestraints' flag, with the maximum force constant setting of 1000. All refinement calculations were conducted utilizing the 'cphoenix' forcefield.

Following scaffold refinement, a new set of non-catalytic sequences were then predicted with LigandMPNN, scored with Boltz-2, and then re-threaded onto the scaffold. This refinement process was conducted until no new sequences were identified as better-scoring than the parent at iteration. LigandMPNN sequence predictions were computed with the 'ligandmpnn_v_32_010_25.pt' weights, the '--ligand_mpnn_use_side_chain_context' flag, and a batch size of 10. The sidechain predictions for sequence threading were computed with the 'ligandmpnn_sc_v_32_002_16.pt' weights as well as the "--_pack_side_chains" and "--pack_with_ligand_context" flags. Boltz-2 calculations were conducted in single-sequence mode utilizing all default model settings. Scaffold alignment and RMSD calculations were conducted

with the Biotite package, utilizing the 'biotite.structure.superimpose.py and biotite.structure.rmsd.py' methods respectively

Backrub Ensemble Generation

Following selection of a final scaffold, a backrub ensemble was generated utilizing PyRosetta version 2024.39. A standard Backrub_Mover class was utilized with MonteCarlo sampling at a temperature of 1.0 for metropolis criterion. The backrub simulation was set to 10000 moves, and set to generate a 100-member ensemble. Then, each ensemble member was refined prior to ligand placement calculations in Triad. First, each ensemble member was corrected for hydrogen addition with the addH.py application, and then refined with the StandardizeStructure.py application. The catalytic motif was kept frozen utilizing the '-pointRestraints' flag, with the maximum force constant setting of 500. All refinement calculations were conducted utilizing the 'cphoenix' forcefield.

Multistate Active Site Design

Multistate design calculations were performed with consensus placements of the complete reaction coordinate, utilizing the multi_design.py application in the Triad software suite. Transition state ligands were modelled as non-canonical amino acids, with geometries built out of the consensus geometry poses selected. Covalent bonds were added between the $/eta$-amino group of the nucleophilic lysine and the carbonyl-carbon moiety of each TS, defined with the interconnect Triad keywords and a 'bbtype' definition of 'natural'. The 'sum' type combination fitness function was utilized to calculate the energy score across all transition states. Energies were calculated utilizing the 'phoenix' energy function with atomic radii scaled by 0.9. Monte Carlo was utilized for the

sequence optimizer, with a high and low temperature of 4000 and 150, respectively, and 5000 steps.

Multistate Design Analysis and Filtering

The stabilization of each transition state for each designed sequence was evaluated by an estimation of the designed active site energy, minimized active site energy, relative solvent accessible surface area of the TS ligand over each designed state, pKa of the initial and final transition states, and interstate reorganization of the designed residues.

Active site energy was defined by the energy difference of the packed structure with designed residues at the designed positions, and the hollow state. Hollow states are defined as ligand-containing scaffolds, with glycine mutations at designed residue (except for the Lys-Tyr-His catalytic triad). Minimized Active energies were estimated utilizing the same procedure as the active site energies, except that a brief backbone minimization was applied to the designed state with point restraints applied to the catalytic residues. To compute the states necessary for the active site energies, we applied the following protocol: For a given designed state, the hollow state were generated by mutating designed residues to glycines utilizing the triad.py application, scored with the '-score' flag and the 'phoenix' energy function. For the minimized active site energies, the states were generated by utilizing the 'cartesianMinimize' flag and the 'cphoenix energy function, applied glycine mutations for the hollow state, and then scored with the 'phoenix' energy function. Catalytic lysine pKa for all states were computed using propka v3.5.1 on the minimized packed structure, after removing the transition state ligand in the structure file.

rSASA was computed using the ratio of the SASA of the unbound ligand to the SASA bound-ligand. SASA values were computed with the Shrake-Rupley algorithm available with the Biopython v1.83 package with a default probe radius of 1.4 and other default parameter settings. Interstate reorganization was computed by evaluating the χ_1 or χ_2 dihedrals over each relevant residue, as well as the RMSD of the side-chain heavy atoms of each residue utilizing Biotite v0.40.0. A designed residue was classified as "organized" if across all designed states, the maximum side-chain RMSD was 3.0, maximum average side-chain RMSD was 2.0, and the maximum χ_1 or χ_2 difference across any adjacent state was less than 30°.

The following filtering parameters were applied. A maximum value of any active site energy score of -300 kcal/mol, with a max ΔE of 100 kcal/mol across all designed states. rSASA was kept to a maximum of 15%, and designs were considered to demonstrate a minimized inter-state reorganization if at least >60% of the designed residues were organized.

Shell Redesign and Filtering

All non-active site residues were redesigned utilizing LigandMPNN in multistate mode with the "--homo_oligomer" flag. Input ensembles were prepared by copying each designed state for a given selected active-site sequence into a single PDB file with each state separated by 200 A in the x-axis coordinate. The ligand for each state was cleaned for LigandMPNN by removing the covalent bond between the η-amino group and the carbonyl-carbon moiety, and changing the atom type of the lysine residue to the standard ATOM code. LigandMPNN sequence predictions were computed with the 'ligandmpnn_v_32_010_25.pt' weights, the '--ligand_mpnn_use_side_chain_context' flag, and a batch size of 10. Boltz-2 calculations were

conducted in single-sequence mode for apo, reactant, and products states utilizing all default model settings. Ligands were input as SMILES strings.

LigandMPNN-designed sequences were filtered based on metrics computed from Boltz-2 predictions. For all states, a minimum pLDDT cutoff of 80 was applied. A maximum state-average Cα RMSD relative to the designed fold of 1.5 was applied for the reactant and product state predictions. An ensemble average side-chain RMSD of 2.0 was applied to the reactants and product states for the Lys-Tyr residues, and 2.5 for the histidine. A Cα-Cβ interstate RMSD of 1 was applied to all catalytic residues. Negative designs were selected by removing either the interstate or ensemble average RMSD filter of the histidine residue. Scaffold alignment and RMSD calculations were conducted with the Biotite package, utilizing the 'biotite.structure.superimpose.py and biotite.structure.rmsd.py' methods respectively.

Protein Expression and Purification for Variant Screening

Michaelase designs were synthesized and cloned into the pET-29b(+) vector via NdeI and XhoI restriction sites, by Twist Biosciences. Plasmids were transformed into *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth, and plated on LB-agar plates with supplemented 100 ug/mL kanamycin. Individual colonies were picked for 5-mL overnight cultures in TB media supplemented with 100 ug/mL kanamycin, and then inoculated into 100 mL expression cultures supplemented with 100 ug/mL kanamycin. Cultures were grown at 37 °C at 250 rpm to an OD600 of 0.6-0.8, and expressed with 1 mM isopropyl -β-D−1-thiogalactopyranoside (IPTG). Following addition of IPTG, cultures were incubated at 16°C for 24 hrs at 250 rpm. Cells were harvested by centrifugation at maximum speed, resuspended in 8 mL of lysis buffer (5 mM imidazole, 100 mM potassium phosphate, pH 8.0), supplemented with lysozyme, benzonase nuclease (Merck

Millipore), and protease II inhibitor cocktail (ThermoFisher Scientific). Resuspended cells were lysed with an EmulsiFlex-B15 cell disruptor (Avestin), and then centrifuged at max speed for 1hr to isolate supernatant. The supernatant was filtered through a 0.2 uM PES filter, and then nutated for 1 hr with 0.5 mL resin beds of Ni-NTA agarose**.** Resin was pre-equilibrated with the lysis buffer in Econo-Pac gravity-flow columns (Bio-Rad), prior to nutating. Resin bed was washed with 20 column volumes of wash 1 buffer (10 mM imidazole, 100 mM potassium phosphate, pH 8.0), and then 20 column volumes of wash 2 buffer (20 mM imidazole, 100 mM potassium phosphate, pH 8.0). Bound protein was eluted with 5 column volumes of elution buffer (250 mM imidazole, 100 mM potassium phosphate, pH 8.0). Eluted protein was then buffer-exchanged into the assay buffer (25 mM HEPES, 100 mM NaCl, pH 7.5). Buffer-exchanged protein concentration was calculated using the Bradford Assay**,** using commercial BSA standards.

Michael Addition Activity Screening

Assays were carried out at 29°C in the assay buffer (25 mM HEPES, 100 mM NaCl, pH 7.5) over a 6 hr interval. Michael addition reaction between (E) -4-(4-methoxyphenyl)but-3-en-2-one (A2B Chem) and ethyl cyanoacetate (ECN) (A2B Chem) was monitored spectroscopically at 370 nm ($\Delta\varepsilon$ = 1858.5 M^{-1} cm^{-1}) using a BioTek () plate reader. Reactions were conducted at a volume 200 uL in UV-STAR plates, 100 uM of the ketone, 50 mM of ECN, and a final volume of 3% v/v acetonitrile. Reactions were initiated by addition of a 150 uL of a ECN+Ketone mastermix to a 50 uL aliquot of the enzyme variant. Enzymes were loaded to a final reaction concentration of 4 uM, determined via Bradford assay and computed molecular weight of the variant. Activities were measured by using the most linear and steepest region of the kinetic trace, measured over intervals of 50 timesteps.

Circular Dichroism and Thermal Denaturation

Far-UV Circular Dichroism measurements were carried out using 300-uL aliquots of selected enzymes in a 1-mm path-length quartz cuvette (Jasco) at a concentration of 5 uM in 20 mM sodium phosphate buffer (pH 7), using a Jasco J-815 spectrometer. Secondary structure scans were completed by circular dichroism spectrum from 180 to 250 nm, sampled every 0.2 nm at a rate of 10 nm min ^-1, with 3 scans were acquired and averaged for each sample. Thermal denaturation curves were obtained by heating samples at a rate of 0.5 °C per minute, and ellipticity at 208 nm and 222 nm.

References

1. Smith, A. J. T. *et al.* Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J. Am. Chem. Soc.* **130**, 15361–15373 (2008).
2. Du, S. *et al.* Conformational ensembles reveal the origins of serine protease catalysis. *Science* **387**, eado5068 (2025).
3. Zarifi, N. Next-generation enzyme design: incorporation of multiple conformational and chemical states. (University of Ottawa, 2025).
4. Davey, J. A. Multistate Computational Protein Design: Theories, Methods, and Applications. (Université d'Ottawa / University of Ottawa, 2016).
5. Davey, J. A. & Chica, R. A. Multistate computational protein design with backbone ensembles. *Methods Mol. Biol.* **1529**, 161–179 (2017).
6. Svedendahl, M., Hult, K. & Berglund, P. Fast carbon-carbon bond formation by a promiscuous lipase. *J. Am. Chem. Soc.* **127**, 17988–17989 (2005).
7. Strohmeier, G. A. *et al.* Investigation of lipase-catalyzed Michael-type carbon–carbon bond formations. *Tetrahedron* **65**, 5663–5668 (2009).
8. Cai, J.-F., Guan, Z. & He, Y.-H. The lipase-catalyzed asymmetric C–C Michael addition. *J. Mol. Catal. B Enzym.* **68**, 240–244 (2011).
9. Chen, X.-Y., Chen, G.-J., Wang, J.-L., Wu, Q. & Lin, X.-F. Lipase/acetamide-catalyzed carbon-carbon bond formations: A mechanistic view. *Adv. Synth. Catal.* **355**, 864–868 (2013).
10. Wu, L.-L., Li, L.-P., Xiang, Y., Guan, Z. & He, Y.-H. Enzyme-promoted direct asymmetric Michael reaction by using protease from Streptomyces griseus. *Catal. Letters* **147**, 2209–2214 (2017).

11. Miao, Y., Geertsema, E. M., Tepper, P. G., Zandvoort, E. & Poelarends, G. J. Promiscuous catalysis of asymmetric Michael-type additions of linear aldehydes to β-nitrostyrene by the proline-based enzyme 4-oxalocrotonate tautomerase. *Chembiochem* **14**, 191–194 (2013).
12. Zandvoort, E., Geertsema, E. M., Baas, B.-J., Quax, W. J. & Poelarends, G. J. Bridging between organocatalysis and biocatalysis: asymmetric addition of acetaldehyde to β-nitrostyrenes catalyzed by a promiscuous proline-based tautomerase. *Angew. Chem. Int. Ed Engl.* **51**, 1240–1243 (2012).
13. Poddar, H., Rahimi, M., Geertsema, E. M., Thunnissen, A.-M. W. H. & Poelarends, G. J. Evidence for the formation of an enamine species during aldol and Michael-type addition reactions promiscuously catalyzed by 4-oxalocrotonate tautomerase. *Chembiochem* **16**, 738–741 (2015).
14. Kunzendorf, A. *et al.* Unlocking asymmetric Michael additions in an archetypical class I aldolase by directed evolution. *ACS Catal.* **11**, 13236–13243 (2021).
15. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
16. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
17. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
18. Garrabou, X., Beck, T. & Hilvert, D. A promiscuous DE Novo retro-aldolase catalyzes asymmetric Michael additions via Schiff base intermediates. *Angew. Chem. Int. Ed Engl.* **54**, 5609–5612 (2015).
19. Garrabou, X., Macdonald, D. S., Wicky, B. I. M. & Hilvert, D. Stereodivergent evolution of artificial enzymes for the Michael reaction. *Angew. Chem. Weinheim Bergstr. Ger.* **130**, 5386–5389 (2018).
20. Xu, J.-M., Zhang, F., Liu, B.-K., Wu, Q. & Lin, X.-F. Promiscuous zinc-dependent acylase-mediated carbon-carbon bond formation in organic media. *Chem. Commun. (Camb.)* 2078–2080 (2007).
21. Jiang, R., Casilli, F., Thunnissen, A.-M. W. H. & Roelfes, G. An artificial copper-Michaelase featuring a genetically encoded bipyridine ligand for asymmetric additions to nitroalkenes. *Angew. Chem. Int. Ed Engl.* e202423182 (2025).

22. Witayakran, S. & Ragauskas, A. J. Cocatalytic enzyme system for the Michael addition reaction of in-situ-generated *ortho*-Quinones. *European J. Org. Chem.* **2009**, 358–363 (2009).
23. Xie, B.-H., Guan, Z. & He, Y.-H. Promiscuous enzyme-catalyzed Michael addition: synthesis of warfarin and derivatives. *J. Chem. Technol. Biotechnol.* **87**, 1709–1714 (2012).
24. Zhu, Z., Hu, Q., Fu, Y., Tong, Y. & Zhou, Z. Design and evolution of an enzyme for the asymmetric Michael addition of cyclic ketones to nitroolefins by enamine catalysis. *Angew. Chem. Int. Ed Engl.* **63**, e202404312 (2024).
25. Way, M. What I cannot create, I do not understand. *J. Cell Sci.* **130**, 2941–2942 (2017).
26. Buller, R. *et al.* From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**, eadh8615 (2023).
27. Ai, H.-W. *et al.* Protein engineering: Status report. *Protein Eng. Des. Sel.* gzag020 (2026).
28. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Chapter 4. Computational enzyme design: Successes, challenges, and future directions. in *Catalysis Series* 88–116 (Royal Society of Chemistry, Cambridge, 2018).
29. St-Jacques, A. D. *et al.* Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. *Nat. Commun.* **14**, 6058 (2023).
30. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
31. Appendix B: Transition state theory and enzymology: Enzyme catalytic power and inhibitor design. in *Kinetics of Enzyme Action* 263–273 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011).
32. Truhlar, D. G., Garrett, B. C. & Klippenstein, S. J. Current status of transition-state theory. *J. Phys. Chem.* **100**, 12771–12800 (1996).
33. Sholl, D. & Steckel, J. A. *Density Functional Theory: A Practical Introduction*. (Wiley-Blackwell, Hoboken, NJ, 2009).
34. Tantillo, D. J., Chen, J. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).

35. Shajan, A., Manathunga, M., Götz, A. W. & Merz, K. M., Jr. Geometry optimization: A comparison of different open-source geometry optimizers. *J. Chem. Theory Comput.* **19**, 7533–7541 (2023).

36. Maeda, S., Harabuchi, Y., Ono, Y., Taketsugu, T. & Morokuma, K. Intrinsic reaction coordinate: Calculation, bifurcation, and automated search. *Int. J. Quantum Chem.* **115**, 258–269 (2015).

37. Zhang, X. *et al.* Quantum mechanical design of enzyme active sites. *J. Org. Chem.* **73**, 889–899 (2008).

38. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14274–14279 (2001).

39. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

40. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).

41. Khersonsky, O. *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10358–10363 (2012).

42. Doustmohammadi, H., Sanchez, J., Ram Mahato, D. & Osuna, S. Evolution enhances Kemp eliminase activity by optimizing oxyanion stabilization and conformational flexibility. *Chem. Eur. J.* **31**, e202403747 (2025).

43. Peccati, F., Noey, E. L., Houk, K. N., Osuna, S. & Jiménez-Osés, G. Interplay between substrate polarity and protein dynamics in evolved Kemp eliminases. *ChemCatChem* **16**, e202400444 (2024).

44. Rakotoharisoa, R. V. *et al.* Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J. Am. Chem. Soc.* **146**, 10001–10013 (2024).

45. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309–313 (2010).

46. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).

47. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes: Robust Design and Optimization of Retroaldol Enzymes. *Protein Sci.* **21**, 717–726 (2012).
48. De Raffele, D., Martí, S. & Moliner, V. Understanding the directed evolution of DE Novo retro-aldolases from QM/MM studies. *ACS Catal.* **10**, 7871–7883 (2020).
49. De Raffele, D., Martí, S. & Moliner, V. QM/MM theoretical studies of a de Novo retro-aldolase design. *ACS Catal.* **9**, 2482–2492 (2019).
50. 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).
51. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita-Baylis-Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).
52. Balasubramani, S. G., Korchagina, K. & Schwartz, S. Transition path sampling study of engineered enzymes that catalyze the Morita-Baylis-Hillman reaction: Why is enzyme design so difficult? *J. Chem. Inf. Model.* **64**, 2101–2111 (2024).
53. Shen, Z. *et al.* Computational design of generalist cyclopropanases with stereodivergent selectivity. *Nat. Commun.* **17**, 1620 (2026).
54. Lauko, A. *et al.* Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).
55. Hammes, G. G., Benkovic, S. J. & Hammes-Schiffer, S. Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry* **50**, 10422–10430 (2011).
56. van der Meer, J.-Y. *et al.* Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective Michaelases. *Nat. Commun.* **7**, 10911 (2016).
57. Zhao, H. Recent advances in enzymatic carbon-carbon bond formation. *RSC Adv.* **14**, 25932–25974 (2024).
58. Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A. Rational design of proteins that exchange on functional timescales. *Nat. Chem. Biol.* **13**, 1280–1285 (2017).
59. Ahern, W. *et al.* Atom level enzyme active site scaffolding using RFdiffusion2. *Biochemistry* (2025).

60. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

61. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods* **22**, 717–723 (2025).

62. Shuai, R. W., Lu, T., Bhatti, S., Kouba, P. & Huang, P.-S. Ensemble-conditioned protein sequence design with Caliby. *Bioengineering* (2025).

63. Shuai, R. W., Widatalla, T., Huang, P.-S. & Hie, B. L. Sidechain conditioning and modeling for full-atom protein sequence design with FAMPNN. *Bioengineering* (2025).

64. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

65. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

66. Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxivorg* 2025.06.14.659707 (2025) doi:10.1101/2025.06.14.659707.

67. Listgarten, J. & Jiang, H. How artificial intelligence is reengineering protein engineering. *Science* **392**, 159–166 (2026).

68. Mukherjee, S., Yang, J. W., Hoffmann, S. & List, B. Asymmetric enamine catalysis. *Chem. Rev.* **107**, 5471–5569 (2007).

69. Xu, G. & Poelarends, G. J. Unlocking new reactivities in enzymes by iminium catalysis. *Angew. Chem. Int. Ed Engl.* **61**, e202203613 (2022).

70. Beck, J. *et al.* Customizing the structure of minimal TIM barrels to craft efficient de novo enzymes. *Nat. Chem. Biol.* 1–8 (2026).

71. Beck, J. *et al.* Customizing the structure of a minimal TIM barrel to craft a DE Novo enzyme. *bioRxiv* 2025.01.28.635154 (2025) doi:10.1101/2025.01.28.635154.

72. Wang, D. *et al.* Engineering of fructose-6-phosphate aldolase for one-carbon conversion to mannitol in a designed biotransformation system. *Synth. Syst. Biotechnol.* **14**, 447–458 (2026).

73. Paul, S., Paul, T. K. & Taraphder, S. Reaction coordinate, free energy, and rate of intramolecular proton transfer in human carbonic anhydrase II. *J. Phys. Chem. B* **122**, 2851–2866 (2018).

74. Fry, B., Slaw, K. & Polizzi, N. F. Zero-shot design of drug-binding proteins via neural iterative selection-expansion. *Nature* 1–13 (2026).

75. Stevens, A. O. & He, Y. Benchmarking the accuracy of AlphaFold 2 in loop Structure Prediction. *Biomolecules* **12**, 985 (2022).

76. Bertoline, L. M. F., Lima, A. N., Krieger, J. E. & Teixeira, S. K. Before and after AlphaFold2: An overview of protein structure prediction. *Front. Bioinform.* **3**, 1120370 (2023).

77. Finta, S., Kalikadien, A. V. & Pidko, E. A. Data-driven virtual screening of conformational ensembles of transition-metal complexes. *J. Chem. Theory Comput.* **21**, 5334–5345 (2025).

78. Froitzheim, T., Müller, M., Hansen, A. & Grimme, S. G-xTB: A general-purpose extended tight-binding electronic structure method for the elements H to Lr (Z=1–103). *ChemRxiv* (2025) doi:10.26434/chemrxiv-2025-bjxvt.

79. Chu, A. E., Fernandez, D., Liu, J., Eguchi, R. R. & Huang, P.-S. De Novo design of a highly stable ovoid TIM barrel: Unlocking pocket shape towards functional design. *Biodes. Res.* **2022**, 9842315 (2022).

80. Braun, M. *et al.* Computational enzyme design by catalytic motif scaffolding. *Nature* **649**, 237–245 (2026).

81. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).

82. Neese, F. The ORCA program system: The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).

83. Tetenberg, T. *et al.* ORCA meets Python─The ORCA Python Interface OPI. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.5c02141 (2026) doi:10.1021/acs.jctc.5c02141.

84. Dunbrack, R. L., Jr. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).

85. Micsonai, A. *et al.* BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res.* **50**, W90–W98 (2022).

Supplementary Information

Table SI 1: Amino acid sequences of designed enzymes

Enzyme	Amino acid sequence
1+	PVPLTGRQRELVELALNLPPIRLPRTRPIVPLPEAIREAVAAGRTPVLVAYAPYGGPIYPGLAPPEDPVEYAKRASRLAV ALAIAGAPEPGQGRLEVMRRIAEAVDIAIVWKGWVVDERHIDTALALGAAAVLLLLFVLDRERARRLYERVQSYGMT PAILIRNEEELELALEIGAKFIWVAALDDFTGEWDRERALELIKKIPKEILKLAYAAFLGLPEIEELKKLGVDAFAPHGSV LLRGGGLLEQLPELTA
2+	PVPLTGRQRELVELALNLPPIDIPRTRPIVPLLEAIREAVKAGRTPVLVAYAPYGGPIYPGLADPRDPVEYAKEAAKYAV ALAIAGAPEPGTGRDEVLRRIAEAVDIAIVYKGAVVNEKQIDRAIALGAAAVMLFTAILDEERLRRLYEYVKSYGMTPI IAISNFEQLELALKIGAEMIAVAAYEDFTNAWNRELCYELIKKIPKNIAKLAYICFTGLPDIEELRKLGVDMFAPHGSVL LSGGGLLEQLPELTA
3+	PRPLTGRQRELVERALNLPPIDLPRTRPIVPLLEAIKRAVAEGRRPVIVALAPFGGPVYPGLADPRDPVEYAKEAAKHA VALAIAGAYEPGTGRLEVLRRIAEAVDIAIVYKGATVNEKQIDTALALGAAAVFLWTAILDREELRRLYEYVQSYGMT PIIMISNEEQLKLALEIGAKMIAVAAIDDPSGAWNKELAYKLIKKIPKNIAKIAYAAFTGLPEIDELEKLGVDAFAPHGSV LLSGGGLLEQLPELTA
4+	PVPLTGRQREIVERARNIPPIDIPRTRPIVPLLERIKEAVAAGRRPVIVAYAPYGGPVYPGLADPRDPVEYAKEAAKRAV ALAIAGWPEPGTGREEVLRRIAKAVDIAIVWKGAVVDRRHVDRAVALGAAAVFLLLAILDEERARELYEYVKSYGME PIIMISNEEQLELALKIGAKMIAVAAIDDFTGAWNKELCYELIKKIPKDIAKIAYACFDGLPDIDKLVELGVDAFAPHGS VLRAGGGLLEQLPELTA
5+	PVPLTGRQRELVERAENLPPIDLPRTRPIVPLLERIREAVAAGRRPVLVALAPYGGPIYPGFADPRDPVEYAKEAAKHA VALAIAGAPEPGTGREEVVRRIAEAIDIAIVWKGAVVNRKHVDRAVALGAAAVMLFTAILDEERLRELMEYVRSFGM EPIVAISNEEQLELALRLGAKMIAVAAYEDFTGAWNKERCYELIKKIPADVAKLAYICFDGLPDIDELWKLGVDMFAP HGSVWLRGGGLLEQLPELTA
6+	PVPLTGRQRALVELAKNLPPINEPRTRPIVPLLERIREAVAAGRRPVLVAYAPYGGPIYPGFADPRDPVEYAKEASKLA VALAIALAPEPGTGRLEVLKRIAEAVDIAIVYKGAVVNKKQIDTALALGAAAVMLFLAILNEEQARELYEYVQSYGMT PIIAISNEEQLKLALEIGAKMIAVAAFDDFTGSWDKEKAYELIKKIPKEIAKLAYIAFTGLPDIEELEKLGVDMFAPHGSV WLRGGGLLEQLPELTA
7+	PVPLTGRQRELVERALNLPPIDIPRTRPIVPLLEVIRRAVEEGRTPVIVAYAPYGGPVYPGLADPRDPVEYAKEAAKYAV ALAIAMFPEPGTGREEVLERIAKEIDIAILYKGATVNEKHIDRAVALGAAAVALLLFILDEEQARRLYEYVQSYGMTPA LMISNFEQLKLALEIGAKFIAVAAIRDFTGEWNKELVRELIKKIPKDIAKLCYAGFTGLPDIEELKKLGVDAFAPHGSVL LRGGGLLEQLPELTK
8+	PIPLTGRQRELVELAENLPPIHIPRTRPIVPLLEAIRRARAEGRRPVIVAYAPFGGPIYPGLAPPDDPVAYAKRAAERAVA LAIAMFPEPGTGREEVLRRIAEAVDIAILYKGAVVNPKHIDRAVALGAAAVALLLFILDRERARELYEYVKSYGMTPAI MIYNEEQLKLALEIGAEFIVVAAIQDFLGEWRRDLAYELIKKIPKNIAKLAYAAFLGLPDIDRLIELGVDAFAPHGSVLL SGGGLLEQLPELTA
9+	PRPLTGRQREIVELARNLPPIDIPRTRPIVPLPEVIREAVAAGRTPVIVAYAPFGGPVYPGLADPRDPVEYAKEAAKHAV ALAIAAAPRPGEGREEVIKRIAEAIDIAILYKGWTVDEKHIDRALALGAAAVLLLLFILDEEEARRLYEYVKSYGMTPAI MISNEEQLELALKIGAEFIAVAAIDDPTGSWNKELVRELIKKIPKNIAKLAYAGFTGLPDIEELKKLGVDAFAPHGSTLL RGGGVLEQLPELTA
10+	PVPLTGRQRELVERARALPPIDLPRTRPIKPLLEVIKEAVEKGRTPVIVAYAPYGGPVYPGLADPEDPVEYAKRAAEHA VALAIALWPEPGTGREEVMRRIAEAVDIAILAKGATVDRSQVDRALALGAAAVALILFILDEERARELYEYVKSYGMT PAIFISNEEQLKLALEIGAKFIAVAAIRDFTNAWNKELAYELIKKIPKNIAKLAYAAFTGKEEIDELKKLGVDAFAPHGS VLLSGGGLLEQLPELTA
11+	PVPLTGRQRELVELAENLPPIDEPRTRPIVPLLERIREAVAAGRRPVIVAFAPYGGPIYPGLADPRDPVEFAKEAAKHAV ALAIAGFPEPGTGREEVMRRIAEAVDIAILYKGAVVNKKQIDRAVALGAAAVALLLFILNEEQARELYEYVKSYGMTP AIMISNFEELELALKIGAKFIAVAAIRDFTNSWDKDLVYELIKKIPKNIAKLAYAAFTGLPEIDELVKLGVDAFAPHGSV LLSGRGLLEQLPELTA
12+	PVPLTGRQRELVELALNLPPIDIPRTRPIVPLPEVIREAVAAGRTPVIVAYAPYGGPVYPGLAPKEDPVEYAKRAAKHAV ALAIAGLPRPGEGREEVLRRIAEAVDIAILYKGAVVNKKQIDRAIALGAAAVALLLFILNREQARELYEYVKSYGMTPA IFISNEEELELALEIGAEFIAVAAIRDPSGSWDREKVRELIKKIPKEIAKLAYAGFLGLPEIEELKKLGVDAFAPHGSVLLS

GGGLLERLPELTA

13+ PVPLTGRQRELVERERNLPPIDLPRTRPIVPLLDVIREAVAAGRTPIIVAYAPYGGPVYPGLADPRDPVEYAKEASKYAV
ALAIAGVPEPGTGREEVLRRIAEAIDIAILYKGAVVNKKHIDRARALGAAAVLLLLFILNEEQARELYEYVESFGMTPAI
MISNEEQLELALKIGAKFIVVAAIDDPTGAWNRDLAYELIKKIPKDIAKLAYAAFTGLPDIEELKKLGVDAFAPHGSVW
LSGGGLLERLPELTK

14+ PVPLTGRQRELVELAKNLPPIDEPRTRPIVPLLDAIRAAVAAGRTPVLVAFAPYGGPIYPGLADPRDPVEYAKEAAKHA
TALAVAGWPEPGEGREEVLRRIAEAVDIAILYKGAVVNEKQIDRAVALGAAAVLLLLFILDRERARRLYEYVRRYGM
TPAVMIRNEEELELALEIGAEFIAVAAFDDFTGEWNEERAIELIKKIPKNIAKLAYAAYRGLPDIEKLKKLGVDAFAPHG
PVLLRGGGKLEQLPELTA

15+ PVPLTGRQRELVEAELNRPPIDIPRTRPIVPLLERIREAVAAGRTPVIVAYAPYGGPVYPGLADPRDPVEYAKEAAEHAV
ALAIAGAYEPGTGREEVLERIAKAVDIAILWKGWTVSEKQIDRAVALGAAAVLLLTFVLDEERLRRLYDYVRSFGMTP
AIMIRNEEQLKLALEVGAEFIWVAAFDDFTGEWDEERCIRLIKKIPKHIAKLAYACFRGLPDIERLKKLGVDAFAPHGP
VLLTGRGLLEQLPELTA

1- PRPLTGRQRELVELARNLPPIDLPRTRPIRPLPEAIREAVAAGRTPVLVAYAPYGGPIYPGLAPREDPVEYARAAAERAV
ALAIAGAPEPGTGREEVLRRVAEAVDLAIVWKGWTVNEKQVDRALALGAAAVLLLTFILDREELRRLYEYVKSYGM
TPAILIRNHEQLELALEIGAEFIWVAALDDATGEWNEELCRELIEKIPAHVLKLAYACFRGLEEIEALKKLGVDAFAPHG
SVLLRGGGLLEQLPELTA

2- PVPLTGRQRELVELALNLPPIDIPRTRPIVPLLEAIRRAVAAGRTPVLVAYAPYGGPIYPGLADPRDPVEYAKEAAKHAV
ALAIAGAPEPGTGRDEVLRRIAEAVDIAIVYKGAVVNEKQVDRAIALGAAAVMLFTAILDEEQLRRLYEYVKSYGMTP
IIAISNFEQLELALKVGAEMIAVAAYEDFTGAWNKELCYELIKKIPKNIAKLAYICFDGLPDIEELRKLGVDMFAPHGSV
LLSGRGLLEQLPELTA

3- PLPLTGRQRELVELARNRPPIDIPRTRPIVPLLEAIREAVAAGRRPVIVAYAPFGGPVYPGLADPRDPVEFAKEAAKRAV
ALAIAAAPEPGQGREEVLRRIAEAVDIAIVYKGAVVDEKQVDRAVALGAAAVFLWLAILDREQARRLYEYVRSYGMT
PILMISNEEQLKLALEIGAEMIAVAAIRDPEGRWDKELCYELIKKIPKNIAKIAYACFTGLPDIDELVKLGVDAFAPHGS
VLLSGGGLLERLPELTA

4- PVPLTGRQREIVERAKNLPPIDIPRTRPIVPLLERIKEAVAAGRRPVIVAYAPYGGPVYPGLADPRDPVEYAKEAAKRAV
ALAIAGWPEPGTGREEVLRRIAEAVDIAIVWKGATVDKKHIDRAIALGAAAVFLLLAILDEKQARELYEYVKSYGMEP
IIMISNEEELELALKIGAKMIAVAAIRDFEGAWDKELCYRLIKKIPKDIAKIAYACFDGLPDIDRLVELGVDAFAPHGEV
LRAGGGLLEQLPALTA

5- PVPLTGRQRELVERARNLPPIDLPRTRPIVPLLERIREAVAAGRRPVLVALAPYGGPIYPGFADPRDPVEYAKEAAKHA
VALAIAGAPEPGTGREEVLRRIAEAVDIAIVYKGAVVDRKHVDRAVALGAAAVMLFTAILDEEQLRELMEYVRSYGM
TPIVAISNEEELELALRLGAKMIAVAAYRDFEGAWDRERCYELIKRIPADVAKLAYICFDGLPDIDELWKLGVDMFAPH
GSVWLRGGGLLEQLPELTA

6- PLPLTGRQRELVELARNLPPIDIPRTRPIVPLLEAIREAVAAGRTPVLVAYAPFGGPIYPGLAPPDDPVEYAKRAAELAV
ALAIAGAPEPGTGREEVLRRVAEAVDIAIVWKGATVDRKHIDRAVALGAAAVMLFTAILDEERLRELYEYVQSYGMT
PIIAISNEEELELALKIGAKMIAVAAFRDFTGAWDKEKVYELIKKIPKEIAKLAYIGFTGLPEIEELKKLGVDMFAPHGSV
WLSGGGLLEQLPALTA

7- PRPLTGRQRELVELALNLPPIDIPRTRPIVPLPEVIREAVEKGRRPIIVAYAPYGGPVYPGLADKRDPVEYAKEASKYAV
ALAIAGFPEPGTGREEVLRRIAEAVDIAILYKGAVVNKKHIDRAIALGAAAVALLTFILNEEELKELYEYVKSYGMTPAI
MISNFEELELALKIGAEFIAVAAIDDPSNSWNKELVRELIKKIPKNIAKLAYAGFTGLPEIEELVKLGVDAFAPHGSVLLS
GGGLLEQLPELTK

8- PVPLTGRQRELVELAENLPPIDVPRTRPIVPLLKAIEEARKAGRRPVIVAYAPYGGPIYPGLAPPEDPVAYAKEAAKHAV
ALAIAMFPLPGYGLEEVLRRIAEAVDIAILYKGAVVSPKHIDRALALGAAAVALLLFILDRERARELYERVKSYGMTPA
LMISNEEQLKLALEIGAEFIVVAAIDDPTGSWDRERAYELIKKIPKNIAKLAYAAFTGLPDIDKLIELGVDAFAPHGSVLL
SGGGLLEQLPELTA

9- PRPLTGRQREIVERALNLPPIDIPRTRPIVPLPEVIREAVAAGRTPVIVAYAPFGGPVYPGLADPEDPVEYAKRAAERAV
ALAIAAAPEPGQGREEVLRRIAEAIDIAILYKGWVVNEKHIDRAIALGAAAVLLLLFILNEEEARRLYEYVKSYGMTPAI
MISNEEQLELALKIGAEFIAVAAIDDPSGRWDKELVYELIKKIPKNIAKLAYAAFDGLPDIEELKKLGVDAFAPHGSTLL
RGGGILEQLPELTA

10- PVPLTGRQRELVELAENLPPIDIPRTRPIRPLLEAIREAVAAGRTPVIVAYAPYGGPVYPGLAPKEDPVEYAKRAAEHAV
ALAIAGWPRPGEGREEVLREIAEAIDIAILYKGAVVNKKHIDRALALGAAAVALILFILNEEQARELYEYVKSYGMTPA
VFISNEEELELALKLGAEFIAVAAIRDFEGAWDKERAYELIKKIPKDIAKLAYAAFDGLPDIEKLKKLGVDAFAPHGSVL
LAGGGLLERLPELTA

11- PRPLTGRQKELVELAKNLPPIDLPRTRPIRPLLEAIREAVAAGRRPVIVAYAPFGGPIYPGLADPRDPVEFAKEAAKHAV
ALAIAGFPRPGEGREEVLRRIAEAVDIAILWKGAVVDKKHVDRAIALGAAAVALLLFILDEKEAEELYNYVKSYGMTP
ALMISNEEQLELALKIGAEFIAVAAIDDPTGAWNKELCYELIKKIPKNIAKLAYACFDGLPDIDELWKLGVDAFAPHGS
VLRSGGGILERLPELTA

12- PVPLTGRQRELVELALNLPPINLPRTRPIRPLLEAIREAVAAGRTPVIVAYAPYGGPVYPGLAPKEDPVEYAKRAAEHA
VALAIAMLPEPGKGREEVLKRIAEAIDIAILYKGAVVNEKHIDRAIALGAAAVALLLFILDREQARRLYEYVKSYGMTP
AIFISNEEQLELALEIGAEFIAVAAIRDPDNAWDKERVRELIKKIPKDIAKLAYAGFTGLPEIEELRKLGVDAFAPHGSVL
LAGGGLLEQLPELTA

13- PRPLTGRQRALVEAARNLPPIDIPRTRPIKPLLEAIRKAVKEGRIPIIVAYAPYGGPVYPGFADKRDPVEYAKEASKYAV
ALAIAMVPRPGEGREEVMRRIAKAVDIAILAKGATVNEKQIDRAIALGAAAVLLLLFILNREEARRLYEYVKSYGMTP
AIMISNFEELELALEIGAEFIVVAAIRDPDNAWNKELCYELIKKIPKNIAKLAYACFDGLPEIEELKKLGVDAFAPHGSV
WLRGGGLLEQLPELTK

14- PVPLTGRQRELVELEKNLPPIDLPRTRPIVPLLEAIRRAVAAGRTPVLVAYAPYGGPIYPGLADPRDPVEYAKEAAKHA
VALAIAAWPEPGTGREEVLRRIAEAVDIAILYKGAVVNEKQIDRAVALGAAAVLLLLFILNEEQARRLYEYVRRYGMT
PALMISNEEQLELALKIGAEFIAVAAFDDFTGEWDEERCRRLIEKIPADIAKLAYACFRGLPEIEELRKMGVDAFAPHGP
VLLRGGGLLEQLPELTA

15- PRPLTGRQRELVELELNRPPIDIPRTRPIVDLLEAIREAVAAGRTPVIVAYAPYGGPVYPGLAPKYDPVEYAKEAAKHA
VALAIAMAPEPGEGREEVLRRIAKEVDIAILAKGWTVSKKQVDRALALGAAAVLLLLFVLDEKQARELYEYVRSYGM
TPAVMIRNEEELELALRLGAKFIWVAAFDDFTGEWDRERCLRLIRAIPADIAKLAYACFLGRPDIDRLRALGVDAFAPH
GPVLLSGGGLLEQLPELTA