Backbone and loop remodelling is essential for design of efficient *de novo* enzymes

Serena Hunt

Thesis submitted in partial fulfillment of the requirements for the

degree of Master of Science in Chemistry

Department of Chemistry and Biomolecular Sciences

Faculty of Science

University of Ottawa

Abstract

The creation of artificial enzymes to catalyze desired reactions is a major goal of computational protein design. However, *de novo* enzymes display low catalytic efficiencies, requiring the introduction of activity-enhancing active site and distal mutations through directed evolution. A better understanding of how mutations introduced by directed evolution contribute to increased enzymatic activity will guide the development of design methods such that efficient enzymes can be designed *de novo*. Here, we evaluate the structural, functional, and dynamical impacts of active site and distal mutations introduced by directed evolution of the *de novo* retro-aldolase RA95, an enzyme that presents an important case study in enzyme design due to the significant structural remodelling that was observed during evolution. We observe that the variant RA95-Core, containing only active site mutations introduced by directed evolution, displays activity within one order of magnitude of the fully evolved variant. This suggests that computational enzyme design methods can be improved to create much more efficient enzymes than what was previously achieved in RA95. However, structural changes induced by distal mutations prevent computational recapitulation of the evolved active site on the original design template, indicating that the optimized active site identified through directed evolution could not have been designed *de novo* using current design methodologies. We suggest strategies for the incorporation of backbone remodelling into design procedures that would allow recapitulation of the evolved retro-aldolase active site, as well as the *de novo* design of highly efficient enzymes without the need for optimization by directed evolution.

Acknowledgements

There are many people I would like to thank for their contributions during my graduate studies. Without them, I would not have had such a successful experience over the past few years.

Firstly, I would like to thank my supervisor, Dr. Roberto A. Chica. Thank you for your supervision and continuous support during my years as an honours and masters student in your lab, and for always encouraging me to reach my goals. I appreciate the opportunity to not only improve as a scientist, but also to develop valuable skills in communication, networking, and leadership that will help me in my future career. I would also like to thank Dr. Michael C. Thompson and all the members of the Thompson lab for hosting me during my internship at UC Merced, which allowed me to obtain crystal structures for my research project.

Thank you to past and present members of the Chica lab, with a special mention to Rojo Rakotoharisoa, Niayesh Zarifi, and Dr. Hang Pham, for your friendship and support both in and out of the lab. Without our lab dinners and late nights at the office, I would not have had this much fun as a graduate student. A special thank you also goes to Dr. Adam Damry, Dr. Matthew Eason, and Marc Mayer, who guided and trained me during my time as an honour's student and as an early graduate student. I appreciate the opportunity to contribute to some of your research projects and I am grateful for the time you spent teaching me in the lab.

Finally, I would like to thank my parents, Lorna Woodrow and Dave Hunt, as well as all of my other family and friends who supported me both scientifically and personally during my studies. My experience as a graduate student would not have been as enjoyable without your guidance and advice.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

CD – Circular dichroism

CPD – Computational protein design

FPLC – Fast protein liquid chromatography

IPTG – Isopropyl β-D-1-thiogalactopyranoside

LB – Lysogeny broth

KE – Kemp eliminase

KIC – Kinematic closure

MBH – Morita-Baylis-Hillman

MD – molecular dynamics

PC – Principal component

PCA – Principal component analysis

PDB – Protein Data Bank

QM – Quantum mechanics

RA – Retro-aldolase

RMSD – Root-mean-square deviation

SPM – Shortest path map

TIM – Triose phosphate isomerase

Tm – Melting temperature

TS – Transition state

WT – Wild-type

6-MNA – 6-methoxy-2-naphthaldehyde

Chapter 1. Introduction

Enzymes are remarkable biocatalysts, able to facilitate chemical reactions through transition state (TS) stabilization and allowing rate accelerations of up to 26 orders of magnitude.[1–3] Their high efficiencies, substrate selectivities, and environmentally-friendly nature make them attractive catalysts for a wide range of industrial applications. However, our ability to apply enzymes in industrial processes relies on their ability to catalyze industrially relevant reactions. Unfortunately, many desired reactions do not have naturally occurring enzymes able to catalyze them. While protein engineering techniques have been successfully used to modify the functions of natural enzymes for desired applications,[4] these methods often rely on random mutagenesis strategies, which are limited by the ability to screen the vast protein sequence space. It is therefore a major goal of protein engineering to develop robust computational methods for the design of efficient *de novo* enzymes for any desired reaction.

1.1. Methods in computational enzyme design

Computational enzyme design is the creation of *de novo* enzymes using computational protein design (CPD) methodologies.[5,6] CPD algorithms, which have undergone continuous improvement over the past 25 years,[7] are used to search for amino acid sequences that adopt a desired protein fold. These methods have been used to develop novel proteins with desired properties, from improved thermal stability[8] to altered substrate specificity.[9] CPD algorithms (Figure 1.1) generally begin with the selection of a single fixed protein backbone template. A sidechain placement step follows, whereby discrete sidechain rotamers[10] are placed onto user-specified positions of the template. Next, an energy calculation is completed, where interaction

energies between each rotamer and the backbone (one-body energies) and between each pair of rotamers (two-body energies) are computed using an empirical potential energy function.[7,11–13] Finally, a sequence optimization step is performed, where a search algorithm[14–17] explores both rotamer and sequence space to identify sequences that yield lowest *in silico* energy scores, which represent the predicted stability of the sequence on the backbone template. Following these four steps, a list of sequences ranked by their energy values is generated along with their corresponding structures (Figure 1.1b). This procedure is useful in protein engineering as it allows researchers to search through potential amino acid sequences on a scale that would not be possible through experimental screening methods.

According to TS theory, what makes enzymes such effective catalysts of chemical reactions is their ability to stabilize high energy TS structures along the reaction coordinate, allowing the potential energy barrier of the reaction to be overcome. Thus, active sites used in computational enzyme design comprise a high-energy reaction intermediate such as a TS structure and one or more catalytic sidechains optimally oriented around the TS for catalysis of the desired reaction. Together, the catalytic sidechains and TS are referred to as a theozyme or theoretical enzyme[18], whose structure can be derived from quantum mechanics (QM) calculations. However, general CPD algorithms do not contain steps for the incorporation of theozymes, as these algorithms are designed to optimize sequence stability and not catalytic activity. Thus, an additional step is required in computational enzyme design, where a user-defined theozyme is placed onto the protein backbone template according to a set of geometric constraints believed to be necessary for the desired reaction. Through specialized theozyme placement algorithms, ideal active sites of desired chemical reactions can be incorporated into the CPD procedure. In this way, CPD can be

used to identify protein sequences that not only adopt the desired fold but are also predicted to catalyze the reaction of interest.

Theozyme placement algorithms are used to identify positions in the protein backbone template where the modelled TS and necessary catalytic residues can fit, which can be done using either a ligand placement[19] or matching[20] algorithm (Figure 1.2) In both ligand placement and matching, an inert protein backbone template and a theozyme are required as inputs (Figure 1.2a). Catalytic sidechain rotamers and TS binding poses are sampled during calculation, where rotamers are generated using user-defined torsional degrees of freedom. TS poses are then generated from the catalytic sidechains using torsional, rotational, and translational degrees of freedom to maintain user-defined catalytic contacts. In ligand placement (Figure 1.2b), the TS pose is generated from each rotamer of the first catalytic sidechain according to the first catalytic contact, followed by the testing of each rotamer of additional catalytic sidechains to identify combinations of rotamers that satisfy all catalytic contacts. In matching (Figure 1.2c), the TS pose is generated from each rotamer of every catalytic sidechain according to each associated catalytic contact individually, and overlapping TS poses that were generated from each sidechain are identified as a match.

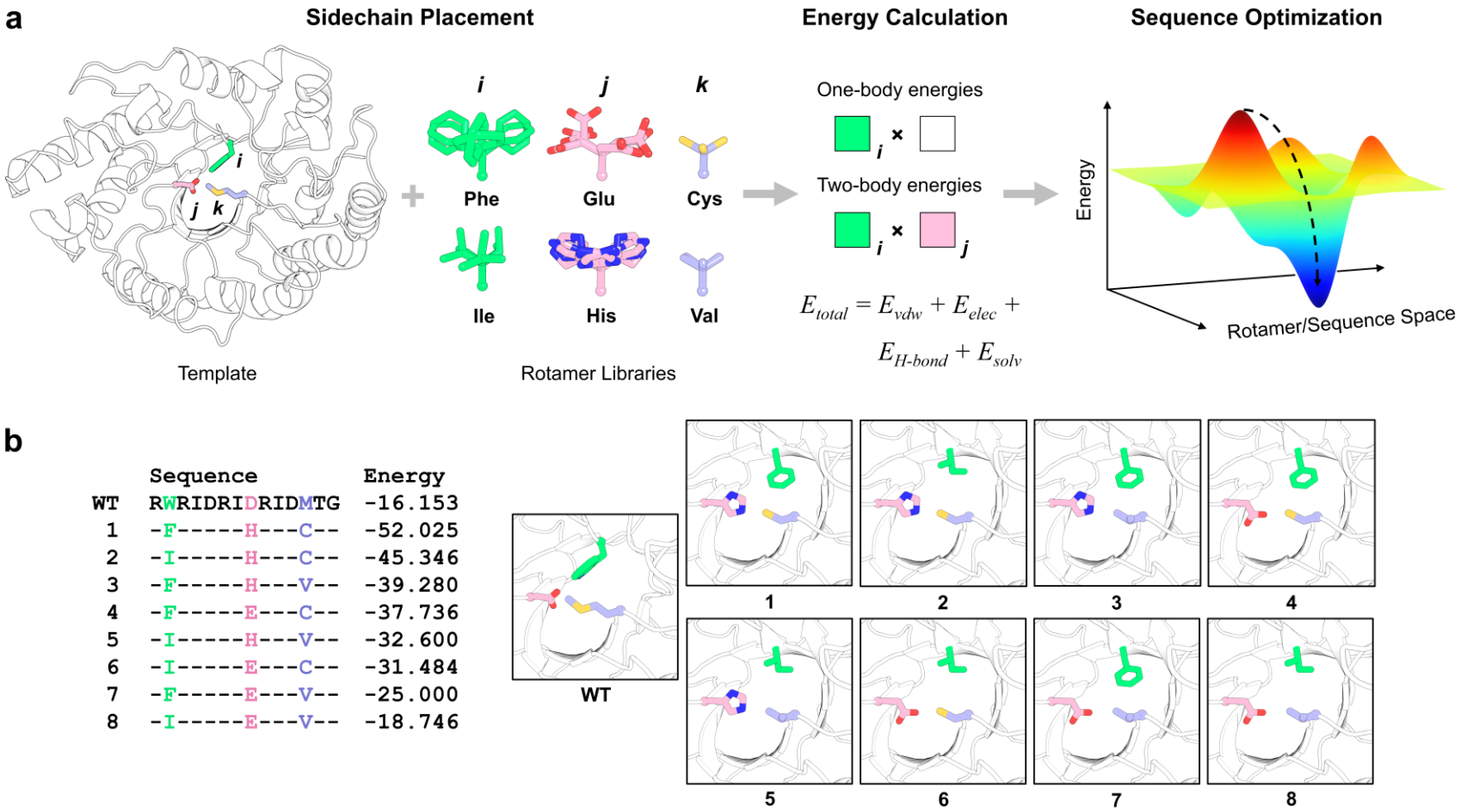

Figure 1.1 Computational protein design (CPD) algorithm. (a) The general CPD algorithm consists of four steps: (1) selection of a protein backbone template; (2) placement of sidechain rotamer libraries for each allowed amino acid type at user-specified positions in the backbone template (Phe and Ile at position i; Glu and His at position j; and His and Val at position k); (3) an energy calculation where interaction energies between the backbone template and each rotamer (one-body energies) and between pairs of rotamers (two-body energies) are computed using a potential energy function including van der Waals (vdw), electrostatic (elec), hydrogen bonding (H-bond), and solvation (solv) terms; and (4) sequence optimization where a search algorithm explores rotamer and sequence space to identify optimal sequences that yield lowest energy values. (b) A list of ranked sequences is returned along with their corresponding structures. Sequences are ranked according to an energy value that represents its predicted stability on the backbone template. Sidechains of designed positions are shown. WT indicates the wildtype sequence

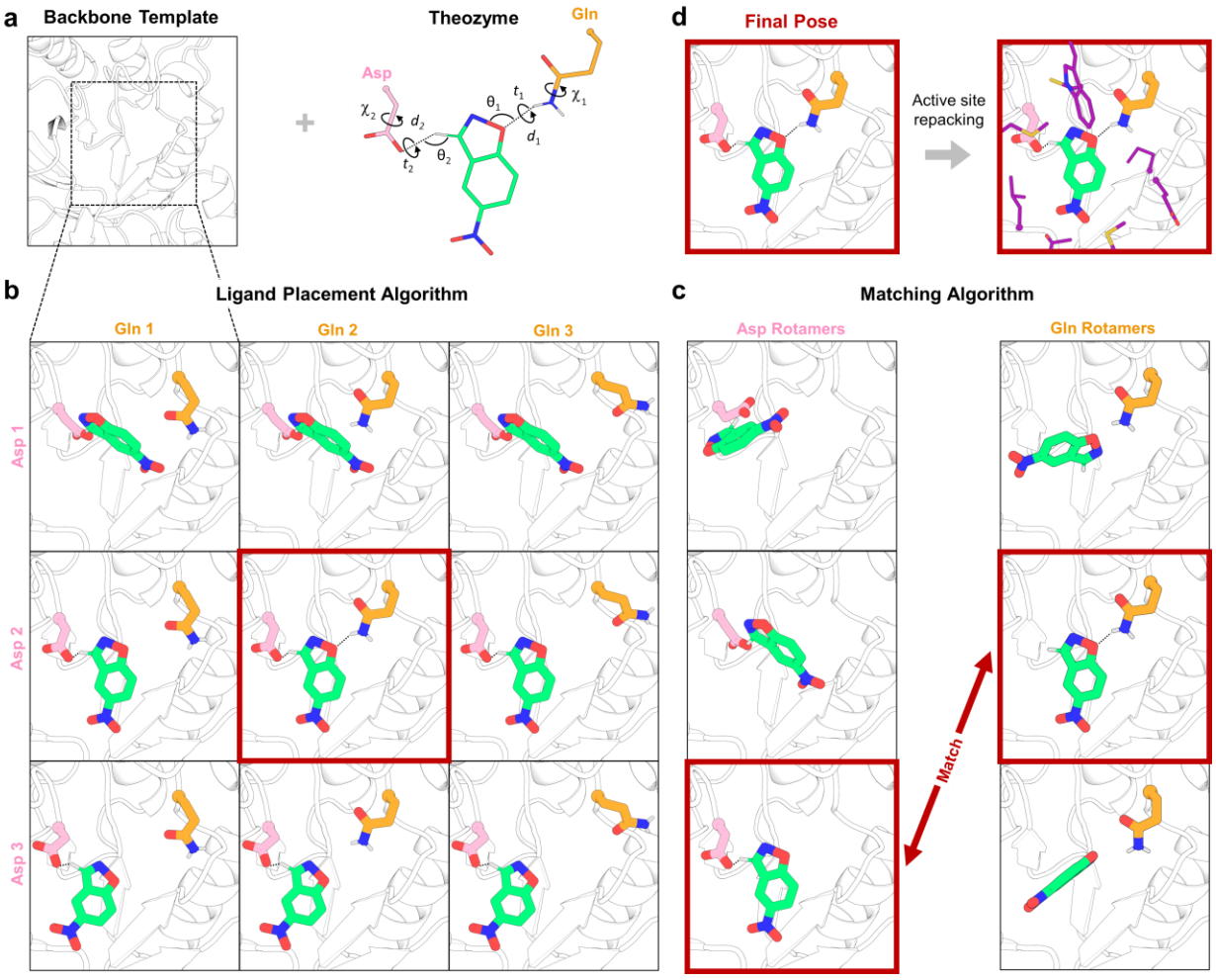

Figure 1.2. Ligand placement and matching algorithms used for the incorporation of theozymes in computational enzyme design. (a) A protein backbone template and a theozyme are required as inputs for each algorithm. Catalytic sidechain rotamers and TS binding poses are sampled during calculation to compensate for the use of a fixed backbone template. Rotamers are generated using user-defined torsional degrees of freedom (χ), while TS poses are generated from the catalytic sidechains using torsional (t), rotational (θ), and translational (d) degrees of freedom to maintain user-defined catalytic contacts (dashed lines). The theozymes are placed into the backbone template using a (b) ligand placement or (c) matching algorithm. For each algorithm, three rotamers for each of the theozyme sidechains (Asp and Gln) and one TS binding pose are shown. (b) In ligand placement, the TS pose is generated from each Asp rotamer according to the first catalytic contact (columns). Each Gln rotamer is then tested (rows) to identify combinations of Asp and Gln rotamers that satisfy both catalytic contacts (dashed lines). Active sites containing both catalytic contacts (red box) are chosen. (c) In matching, the TS pose is generated from each Asp rotamer (left column) and Gln rotamer (right column) according to each associated catalytic contact. Overlapping TS poses that were generated by each of the contacts are identified as a match (red arrow and boxes). (d) Chosen active sites (red box) are repacked and ranked according to their predicted stability on the backbone template.

Once the positions of catalytic residues have been determined by ligand placement or matching, a step known as active site repacking is performed (Figure 1.2d). In repacking, amino acids surrounding the previously placed theozyme are designed to optimize substrate binding and stabilization of the TS. While placement of a theozyme ensures that essential catalytic and ligand binding interactions are present, further stabilization of the TS via non-bonding interactions introduced during the active site repacking step is needed to confer catalytic activity to the designed enzyme. Resulting sequences are finally ranked according to their predicted stability scores on the backbone template, and the highest ranked sequences are experimentally validated. In the following sections, current successes in computational enzyme design will be presented, followed by a discussion of the challenges that must be overcome in the future. Here, applications of computational enzyme design refer to the design of *de novo* enzymes for a predefined model organic reaction using CPD methodologies and placement of an ideal active site on the protein scaffold.

1.2. The first computationally designed *de novo* enzyme

The first example of the use of CPD to design a biocatalyst for a predefined model organic reaction was reported in 2001 by Bolon and Mayo.[21] Using the hydrolysis of *p*-nitrophenyl acetate to *p*-nitrophenol and acetate as a model reaction (Figure 1.3a), catalyzed by a nucleophilic histidine residue, esterase activity was introduced into the catalytically inert *Escherichia coli* thioredoxin scaffold. However, as theozyme placement methodologies had not yet been developed, the authors instead used a simplified approach implementing a "superrotamer" library to confer catalytic activity to the protein scaffold. Instead of stabilizing a TS for the reaction, rotamers of the tetrahedral reaction intermediate were introduced into each position of the backbone template to

identify sites where the catalytic histidine sidechain and intermediate could be positioned according to ideal catalytic geometries. In this way, a high-energy intermediate was used as a proxy for the TS during active site placement. Following active site repacking to further stabilize the TS proxy, two of the resulting designs were chosen for experimental characterization, the more active variant being PZD2 (Table 1.1). While successful, the minimal rate enhancement of PZD2 highlighted the limitations of the design approach used in this study, indicating that approximation of the TS with a high-energy intermediate is likely not sufficient for design of efficient *de novo* enzymes.

1.3. Continued success in computational enzyme design

Since the first implementation of CPD methods for the design of a *de novo* enzyme by Bolon and Mayo, the field of computational enzyme design has seen steady success over the past two decades. Despite the minimal rate enhancement achieved in the design of PZD2, this first example demonstrated that CPD methodologies could be used to confer new catalytic activities to inert protein scaffolds, paving the way for future development of enzyme design methods. However, it was clear that the use of a high-energy TS proxy, as well as approximation of this structure using a superrotamer library, was not ideal. As a result, algorithms for the incorporation of ideal QM-derived theozymes into the design procedure were developed. The first example of the application of these algorithms was the design of *de novo* Kemp eliminase enzymes, which catalyze perhaps the most widely studied model organic reaction in computational enzyme design.

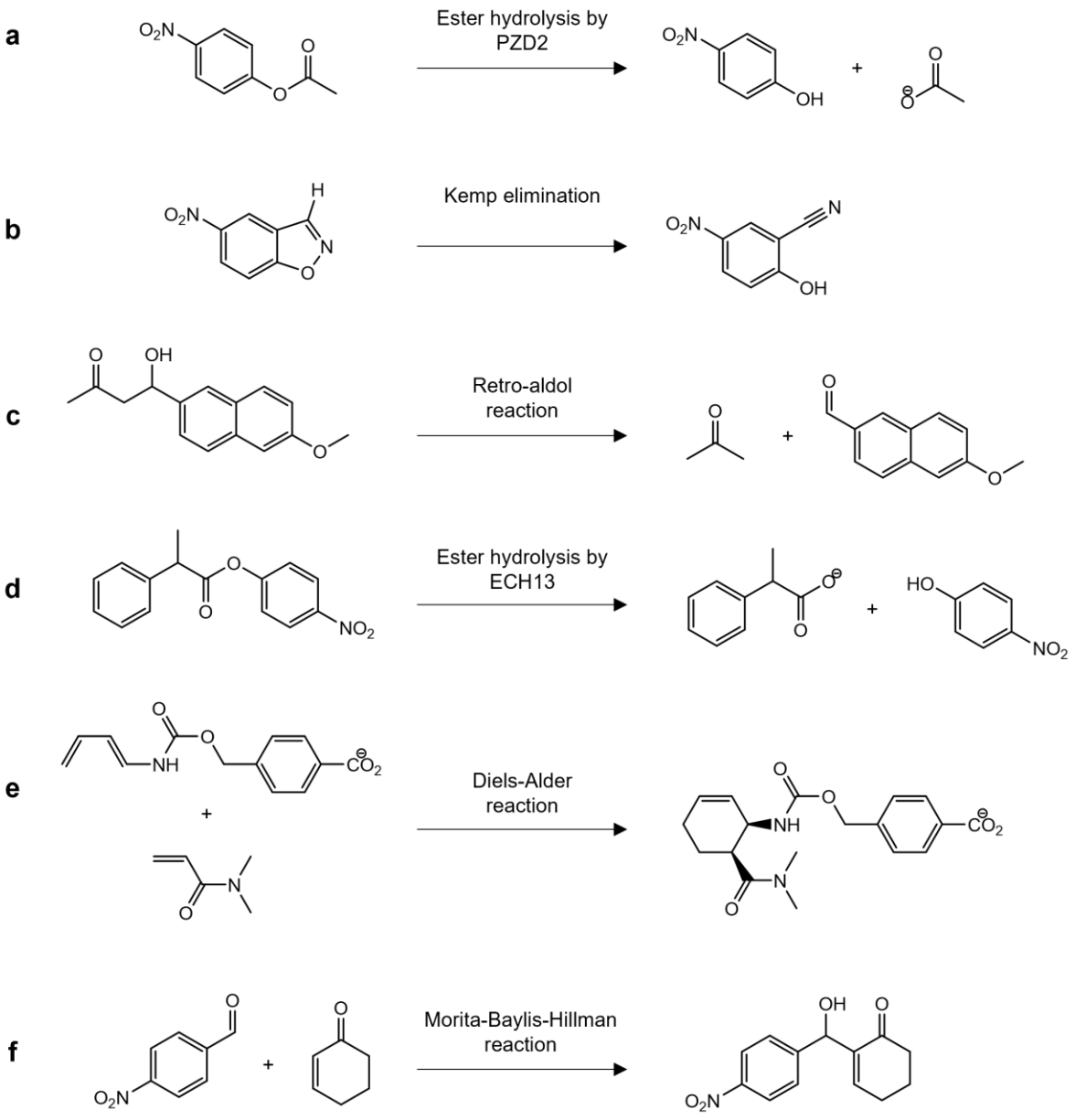

Figure 1.3. Chemical reactions catalyzed by computationally designed *de novo* enzymes. (a) Ester hydrolysis of *p*-nitrophenyl acetate catalyzed by PZD2. (b) Kemp elimination of 5-nitrobenzisoxazole cyanophenol catalyzed by KE59 and HG-2. (c) Retro-aldol decomposition of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone catalyzed by RA60 and RA95. (d) Ester hydrolysis of *p*-nitrophenyl-2-phenylpropanoate catalyzed by ECH13. (e) Diels-Alder [4 + 2] cycloaddition of 4-carboxybenzyl-*trans*-1,3-butadiene-1-carbamate and *N,N*-dimethylacrylamide catalyzed by DA_20_10. (f) Morita-Baylis-Hillman reaction between 4-nitrobenzaldehyde and 2-cyclohexenone catalyzed by BH32.

Table 1.1. Kinetic parameters of *de novo* designed and evolved enzymes.

Enzyme	Optimized[a]	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M (M^{-1} s^{-1})	k_{cat}/k_{uncat}	Ref.
Single-Step Bond-Breaking						
Kemp elimination						
KE59	Designed	0.29	1.8×10^3	1.6×10^2	2.5×10^5	22
KE59.13	Evolved	9.5	1.6×10^2	5.9×10^4	8.2×10^6	23
HG-2	Designed	–[b]	–[b]	1.2×10^2	–	24
HG-3	Improved	0.68	1.6×10^3	4.3×10^2	5.9×10^5	24,25
HG-3.17	Evolved	7.0×10^2	3×10^3	2.3×10^5	6.0×10^8	25
Multistep Bond-Breaking						
Retro-aldol						
RA60	Designed	0.00016	5.1×10^2	0.30	2.4×10^4	26,27
RA95	Designed	0.000033	5.4×10^2	0.19	4.8×10^3	27
RA95.5-8F	Evolved	11	3.2×10^2	3.4×10^4	1.7×10^9	28
Ester hydrolysis						
PZD2	Designed	0.00046	1.7×10^2	2.7	1.8×10^2	21
ECH13	Designed	0.018	57	3.1×10^2	4.4×10^3	29
Single-Step Bond-Forming						
Diels-Alder						
DA_20_00	Designed	0.000028	3.5×10^3(diene) 1.5×10^5(dienophile)	0.06[c]	4.4[d]	30,31
DA_20_10	Improved	0.00058	1.3×10^3(diene) 7.3×10^4(dienophile)	6.1[c]	94[d]	30,31
CE20	Evolved	0.030	2.9×10^2(diene) 1.9×10^4(dienophile)	5.4×10^2 [c]	4.8×10^2 [d]	31
Multistep Bond-Forming						
Morita-Baylis-Hillman						
BH32	Designed	0.00037	–[c]	–	–	32
BH32 Q128H	Improved	–[c]	–[c]	–	54	32
BH32.14	Evolved	0.0058	–[c]	–	–	33
Average reaction						
Natural enzymes		$1–10^2$	$10^1–10^3$	$10^3–10^6$	$10^7–10^{19}$	1,34

[a]Improved and evolved indicate enzymes optimized by rational design or directed evolution, respectively.
[b]Could not be determined as saturation with the substrate was not achieved within its solubility limit.
[c]Not reported in the article.
[d]Units are M^{-2} s^{-1}.
[e]Units are M.

1.3.1. Kemp elimination

Kemp elimination has been studied as a model for understanding the catalysis of proton abstraction from carbon and proceeds via a single TS. In 2008, Röthlisberger, Khersonsky, and Wollacott *et al.*[22] designed several enzymes for the Kemp elimination of 5-nitrobenzisoxazole

(Figure 1.3b) using a QM-derived theozyme composed of the TS bound by a general base to deprotonate the acidic proton and a hydrogen bond donor to stabilize the resulting negative partial charge of the isoxazolic oxygen. The newly developed matching algorithm[20] was employed to search for combinations of positions that could accommodate the idealized Kemp elimination theozyme in a library of template backbones. Positions surrounding the TS were subsequently redesigned in the repacking step for further stabilization. The most active tested design, KE59, displayed a rate enhancement approaching that of natural enzymes (Table 1.1). In 2012, the same idealized theozyme was used with the ligand placement algorithm[19] employed by the PHOENIX protein design software[35] to create the Kemp eliminase HG-3, which is the start of perhaps the most well-studied family of computationally designed enzymes (Table 1.1).[24] HG-3 is a single-point mutant (S265T) of the *in silico* design HG-2, engineered to reduce active site conformational heterogeneity observed in HG-2 through a combined experimental and computational approach involving X-ray crystallography and molecular dynamics (MD) simulations on the design model.

1.3.2. Retro-aldol reaction

While the successes of KE59 and HG-3 demonstrate that computational enzyme design can be used to create biocatalysts for new-to-nature model reactions, Kemp elimination represents a relatively simple single-step bond-breaking reaction. The retro-aldol bond-breaking reaction, in which a β-hydroxy carbonyl compound decomposes into an aldehyde or ketone and another carbonyl compound, is more challenging as it proceeds via multiple steps. In parallel to the design of the first *de novo* Kemp eliminases, Jiang and Althoff *et al.*[26] employed computational enzyme design to create several biocatalysts for the retro-aldol decomposition of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone to 6-methoxy-2-napthaldehyde and acetone (Figure 1.3c). However, as

this reaction proceeds via multiple steps, there are multiple TS structures to consider when creating the theozyme. While treatment of each TS structure individually would likely have been ideal, as each structure must be stabilized by the enzyme at different points along the reaction coordinate, computational limitations resulted in the implementation of a composite TS model whereby QM-derived models of various TS and intermediate structures along the reaction coordinate were superposed.

Using a matching algorithm,[20] ideal retro-aldol theozymes composed of the composite TS and one of four unique catalytic motifs applying different arrangements of catalytic residues were placed onto a library of protein scaffolds. Each motif comprised a catalytic lysine, which facilitates the reaction through enamine catalysis via a Schiff base intermediate, and a base (either a polar residue or ordered water molecule) to deprotonate the β-alcohol of the substrate at the carbon–carbon bond-breaking step. Following repacking of the active site to further stabilize the composite TS, the most active tested variant, RA60, displayed a rate enhancement of 2.4×10^4 (Table 1.1). To test the reproducibility of the computational procedures used in the design of RA60, Althoff, Wang, and Jiang *et al.*[27] subsequently created a new set of retro-aldolases, this time focusing on one catalytic motif and incorporating additional intermediate structures into the composite TS. The most active identified design was RA95, with a modest catalytic rate enhancement of 4.8×10^3 (Table 1.1).

1.3.3. Ester hydrolysis

Richter and Blomberg *et al.*[29] applied computational enzyme design to a second multistep bond-breaking reaction, creating enzymes able to catalyze ester hydrolysis, (Figure 1.3d) similar to the design of PZD2 by Bolon and Mayo. However, in this case, a QM-derived TS was placed

onto a library of protein scaffolds using a matching algorithm,[20] as opposed to placing a high-energy TS proxy. Three potential simplified catalytic motifs were employed in the theozyme, consisting of a catalytic Cys-His dyad and one backbone NH group, water molecule, or sidechain functional group to stabilize the oxyanion intermediate. The most active identified design was ECH13 (Table 1.1), displaying a catalytic rate enhancement one order of magnitude higher than that achieved by Bolon and Mayo[21] in the design of PZD2. However, the modest activity of ECH13 compared to those of *de novo* designs for simpler single-step reactions indicates that the need to improve upon current computational design procedures is especially true when it comes to more complex multistep reactions.

1.3.4. Diels-Alder and Morita-Baylis-Hillman reactions

Even more challenging than the design of *de novo* enzymes for bond-breaking reactions, such as Kemp elimination, retro-aldol decomposition, and ester hydrolysis, is the creation of enzymes able to catalyze bond-forming reactions. Two such reactions are the single-step Diels-Alder reaction and the multistep Morita-Baylis-Hillman (MBH) reaction. Bond-forming reactions such as this are extremely useful in organic syntheses, as complex organic molecules commonly need to be created from simpler precursors, often with a requirement for stereoselectivity. As such, the ability to design efficient *de novo* enzymes for these reactions would be particularly useful for industry. However, not only are there multiple substrates that need to be oriented in a way that is productive for catalysis and that can impart stereoselectivity, but in the case of the MBH reaction, there are also multiple TS structures along the reaction coordinate to be considered in TS stabilization.

Siegel and Zanghellini *et al.*[30] reported in 2010 the design of a *de novo* enzyme for the intermolecular Diels-Alder reaction, which is the reaction between a conjugated diene and an alkene (dienophile) to form unsaturated six-membered rings. Here, enzymes were designed for the well-studied model Diels-Alder reaction between 4-carboxybenzyl-*trans*-1,3-butadiene-1-carbamate and *N,N*-dimethylacrylamide, which occurs through pericyclic [4 + 2] cycloaddition to form a chiral cyclohexene ring[36] (Figure 1.3e). Using a composite TS constructed by overlaying various substrate, TS, and product structures, the authors created the design DA_20_00 (Table 1.1), which was further improved through point mutations predicted by MD to improve packing around the TS. This yielded DA_20_10, a highly stereoselective variant.

In 2013, Bjelic and Nivón *et al.*[32] subsequently applied computational enzyme design to create a biocatalyst for the MBH reaction, in which a nucleophile catalyzes the formation of a carbon–carbon single bond between the α position of an α,β-unsaturated carbonyl and an electrophilic carbon such as an aldehyde. The chosen model MBH reaction between 2-cyclohexanone and 4-nitrobenzaldehyde (Figure 1.3f) proceeds via two intermediates and four TS structures, such that a composite TS was modeled by superposing single structures derived from both short MD simulations and QM calculations. A single-point mutant of the most active design BH32 (BH32-Q128H) was engineered based on *in silico* mutagenesis predictions to enhance catalysis (Table 1.1). While the successful design of *de novo* enzymes for single- and multistep bond-forming reactions is promising, the catalytic rate enhancements presented here represent some of the lowest achieved in computational enzyme design, highlighting the difficulties associated with designing *de novo* enzymes for complex chemical reactions. Evidently, the greater the complexity of the reaction, the more difficult it is to design an enzyme capable of catalyzing it.

1.4. Challenges in computational enzyme design

While *de novo* enzymes have successfully been designed for model organic reactions including the Kemp elimination,[22,24] retro-aldol,[26,27] ester hydrolysis[29], Diels-Alder,[30] and Morita-Baylis-Hillman[32] reactions, researchers have been met with challenges when it comes to the successful implementation of current design methodologies. The success rate of active designs within those tested experimentally remains low, due to both the inability to express soluble protein and to detect the target reaction activity. A second limitation is the inaccurate prediction of designed active site structures. Even when a designed enzyme is able to catalyze the desired reaction, structural analyses of these designs often reveal inaccurate prediction of designed catalytic interactions.[24] Finally, as a result of the inaccuracies inherent to the computational methods being used, *de novo* designs have modest activities, with catalytic efficiencies (k_{cat}/K_M) being several orders of magnitude lower than those of natural enzymes.[1,3] This is most often due to low k_{cat} values and not because of high K_M values.[1,34]

As a result of the limitations of current computational enzyme design methods, *de novo* enzymes are often required to be optimized by directed evolution[37] to improve their catalytic efficiencies. Kemp eliminases,[22,23,25,38,39] retro-aldolases,[27,28,40] Diels-Alderases,[31] and enzymes catalyzing the Morita-Baylis-Hillman reaction[33] have been subjected to multiple rounds of directed evolution, yielding evolved enzymes with many new mutations (Table 1.2) and catalytic efficiencies similar to those found in nature (Table 1.1). In most cases, directed evolution does not significantly remodel the structure of the active site or overall enzyme, but instead contributes to widening of the active site,[23] improving shape complementarity,[39] or eliminating unproductive binding poses[25] or enzyme sub-states.[41] Interestingly, directed evolution tends to introduce a mix

of mutations both in the active site and at more distal positions (Figure 1.4), which is evident in the locations of mutations introduced into the *de novo* designs discussed above (Figure 1.5). Distant mutations have been shown to play large roles in the catalytic properties of evolved enzymes, including in their ability to exchange between various conformational sub-states.[41] As computationally designed positions are generally restricted to active site positions, the prediction of these beneficial distal mutations presents a major challenge in computational enzyme design.[42,43] It is unclear whether the ability to create highly active *de novo* enzymes without resorting to directed evolution will rely on the ability to identify designable distal sites, or if efficient *de novo* enzymes can be designed by focusing only on active site positions.

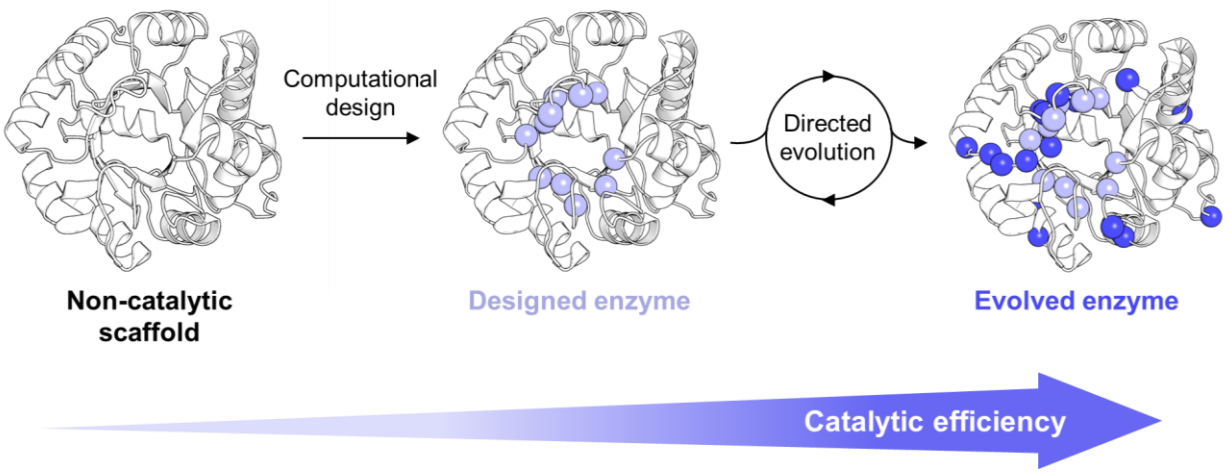

Figure 1.4. Evolutionary trajectory of a hypothetical enzyme. Directed evolution is used to optimize computational designs to improve their catalytic efficiencies by introducing a mix of both active site and distal mutations. Designed positions and evolved mutations are shown as light purple and dark purple spheres, respectively.

The challenges faced by enzyme designers can be attributed to various simplifications made during the computational design procedure. For example, while the limited combinatorial complexity of early computational enzyme design attempts has been improved in recent studies, the rigorous sampling of conformational degrees of freedom of the TS and residue sidechains is

limited by both computing power and the discretization applied by current methods. Limitations related to modelling of the TS are compounded when designing enzymes for multistep reactions, as multiple TS and intermediate structures are required to be stabilized by the enzyme at different stages of the reaction. While the design of *de novo* enzymes has been achieved for multistep reactions through the stabilization of a composite TS, these studies report low success rates and rate enhancements, indicating that the treatment of multistep intermediate and TS structures as a composite is not sufficient for the design of highly efficient biocatalysts for complex reactions.

Table 1.2. Mutations found in *de novo* designed and evolved enzymes.

Enzyme	PDB ID	Optimized[a]	No. mut.	Mutations[b]	Ref.
Single-Step Bond-Breaking					
Kemp elimination					
KE59	–	Designed	10	E51V, S81V, K110W, L131G, L157A, E159V, G178I, W180S, E210A, L231E	[22]
KE59.13	5UZJ	Evolved	22	K9E, L14R, L16Q, F21V, N33K, I48M, S69A, Y75G, A76V, V80A, T94D, I104V, F111I, Y151L, N160H, S179T, R181H, K190N, A208V, R222Y, S233T, L247Q	[23]
HG-2	3NYD	Designed	12	Q42M, T44W, R81G, H83G, T84M, N130G, N172M, A234S, T236L, E237M, T265S, W267F	[24]
HG-3	–	Improved	1	S265T	[24,25]
HG-3.17	4BS0	Evolved	18	V6I, Q37K, N47E, K50Q, G82A, M84C, S89N, Q90F, T105I, A125T, T142N, T208M, S265T, F267M, W275A, R276F, T279S, D300N	[25]
Multistep Bond-Breaking					
Retro-aldol					
RA60	–	Designed	12	N46W, V48K, N72T, Y74W, E87S, Y89S, T119Y, R121W, F133Y, Q135S, E176V, Y178V	[26,27]
RA95	4A29	Designed	11	E51V, K53E, L83T, K110S, E159L, N180M, L184F, L187G, E210K, S211L, G233S	[27]
RA95.5-8F	5AN7	Evolved	22	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, S110N, K135E, S151G, G178T, M180Y, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M	[28]

Ester-hydrolysis					
PZD2	–	Designed	3	F12A, L17H, Y70A	[21]
ECH13	3U13	Designed	8	F20L, F42Y, W43K, V44A, S45C, I60R, W63A, I100H	[29]
Single-Step Bond-Forming					
Diels-Alder					
DA_20_00	3I1C	Designed	13	E21A, E36Y, N120A, D121Y, Y144F, R146I, M148L, F173A, N175A, T195Q, E225K, D229A, N272A	[30,31]
DA_20_10	–	Improved	6	A21T, A74I, Q149R, A173C, S271A, A272N	[30,31]
CE20	4O5T	Evolved	24	V13M, A21T, I33V, T43I, K44N, P48L, K53E, S55R, R56S, G57D, R63H, I85S, A87I, K121N, R128C, E151G, G162R, A186C, K223N, D245V, S284A, A285N, E301D, L322S	[31]
Multistep Bond-Forming					
Morita-Baylis-Hillman					
BH32	3U26	Designed	12	F9S, V10L, L14N, E19A, T22S, I64L, E68L, H91S, H95S, Y128Q, L129A, H132F	[32]
BH32 Q128H	–	Improved	1	Q128H	[32]
BH32.14	–	Evolved	24	L10W, N14I, A19T, A20Y, S22V, L24F, T49A, Y56N, E70R, F87L, S95A, M120V, T122L, D123N, S124R, Q128L, A129S, M130T, F154S, E174K, Y177C, D180P, C186A, C212A	[33]

[a]Improved and evolved indicate enzymes optimized by rational design or directed evolution, respectively.
[b]Presented mutations are either designed, improved, or evolved mutations. Designed mutations are with respect to the WT template protein, while improved and evolved mutations are with respect to the designed enzyme.

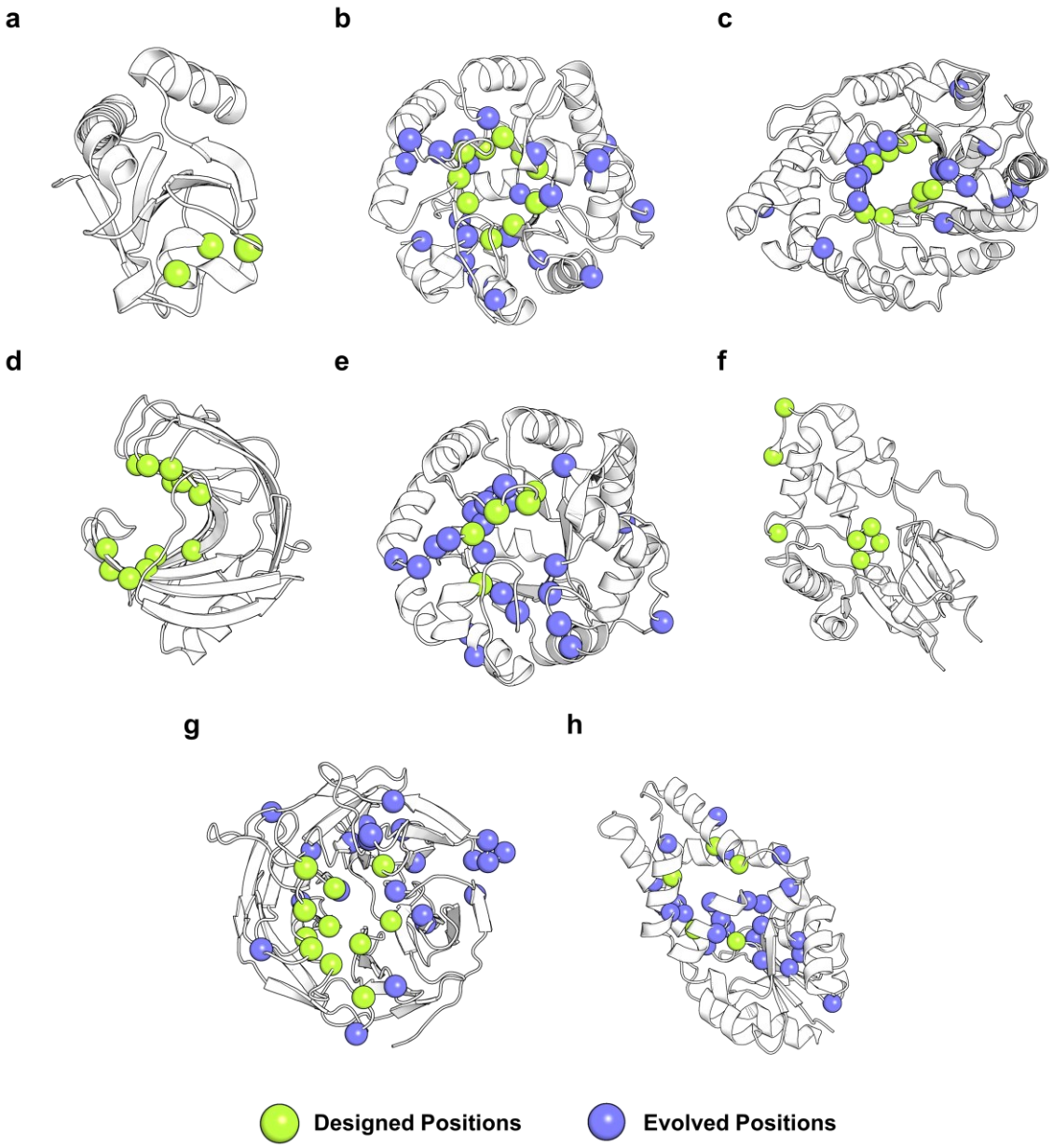

Figure 1.5. Computationally designed and evolved positions in *de novo* enzymes. Positions that were designed and mutated during directed evolution are shown as green and purple spheres, respectively. (a) *E. coli* thioredoxin scaffold (PDB ID: 2TRX[44]) used in the design of PZD2. (b) *S. solfataricus* indole-3-glycerolphosphate synthase scaffold (PDB ID: 1A53[45]) used in the design of KE59. (c) *T. auranticus* xylanase scaffold (PDB ID: 1GOR[46]) used in the design of HG-2. (d) *N.* flexuosa b-1,4-xylanase scaffold (PDB ID: 1M4W[47]) used in the design of RA60. (e) *S. solfataricus* indole-3-glycerolphosphate synthase scaffold (PDB ID: 1LBL[45]) used in the design of RA95. (f) *H. sapiens* mitochondrial deoxyribonucleotidase scaffold (PDB ID: 1Q91[48]) used in the design of ECH13. (g) *L. vulgaris* diisopropylfluorophosphatase scaffold (PDB ID: 1E1A[49]) used in the design of DA_20_00. (h) *P. horikoshii* haloacid dehalogenase scaffold (PDB ID: 1X42[50]) used in the design of BH32.

The need to stabilize multiple structures along a reaction pathway highlights another simplification made in current computational enzyme design procedures: the fixed backbone approximation. Single X-ray crystal structures, which have typically been used as design templates up to now, are not a good representation of the dynamic nature of proteins and their functions. Whether a single- or multistep reaction is being designed, the conformational changes that enzymes undergo during catalysis should be considered. In the simplest case, a single-step reaction proceeds through a minimum of four states over the course of the catalytic cycle, including the free enzyme, the Michaelis complex, the enzyme-bound TS, and the product-bound enzyme (Figure 1.6a). In addition to these four states, consideration of the intrinsic flexibility of the free enzyme is essential, as many alternate conformations of an enzyme are sampled in solution.[51] Not only is the ability of enzymes to undergo conformational changes essential for progression along a catalytic pathway,[52] but the presence of competing catalytically productive and unproductive conformational sub-states can also impede catalysis[53–55] (Figure 1.6b, c). The inability of current methodologies to model conformational changes undergone by enzymes during catalysis highlights the need for the continued development of computational enzyme design methodologies.

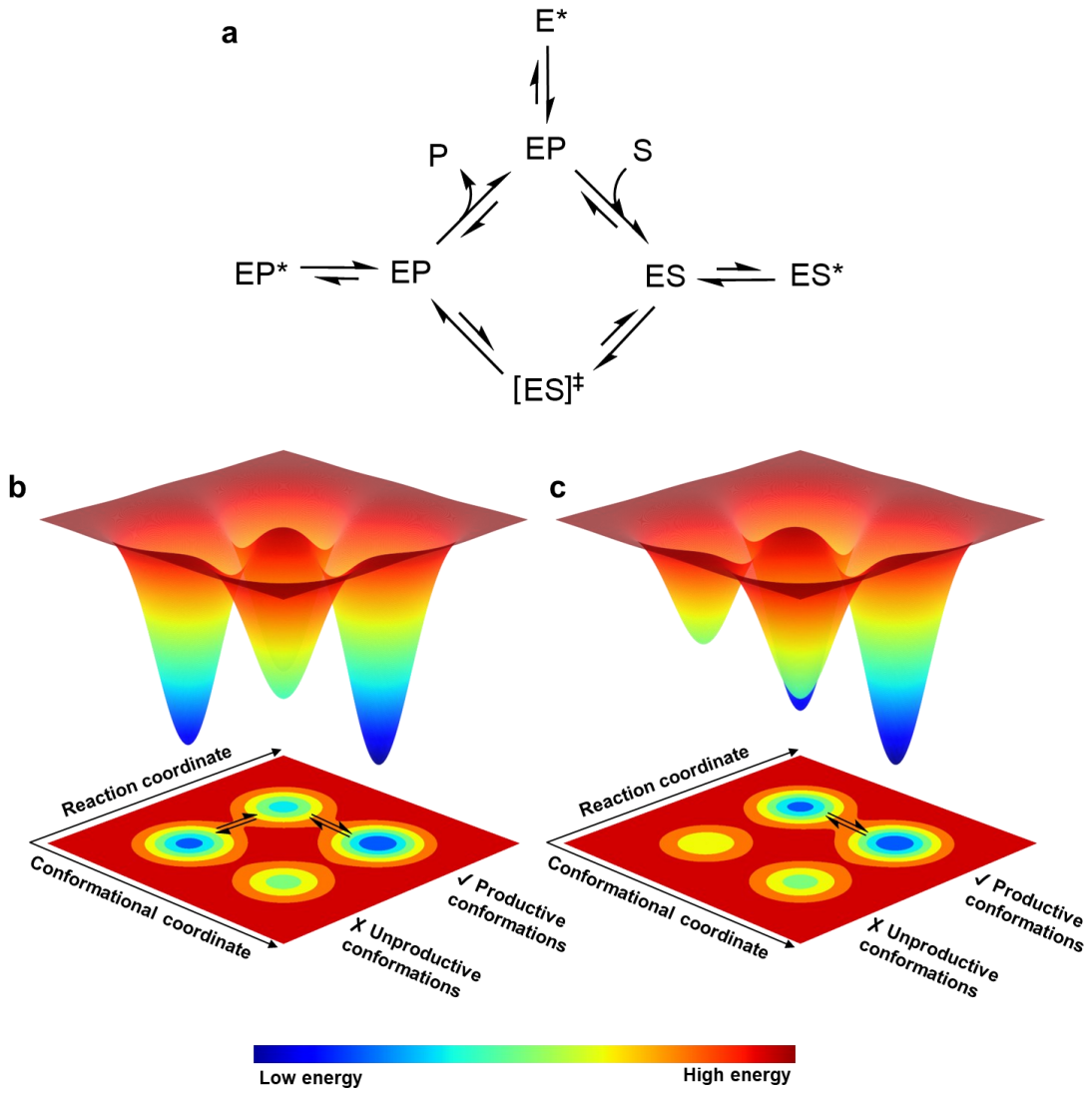

Figure 1.6. Computational enzyme design procedures lack methods for modelling the conformational changes undergone by enzymes in solution. (a) States adopted by an enzyme during the catalytic cycle. The enzyme (E) must bind the substrate (S) to form the Michaelis complex (ES), leading to catalysis by TS ($[ES]^{‡}$) stabilization. Following catalysis, the enzyme-product complex (EP) forms and product (P) is released. Potential competing sub-states include catalytically unproductive conformations of the free enzyme (E*) and unproductive binding poses of the substrate (ES*) or product (EP*) to the enzyme. An enzyme that can exchange between competing catalytically productive and unproductive sub-states (b) is expected to be less active than an enzyme that cannot access catalytically unproductive conformations (c).

1.5. A case study in computational enzyme design

One key family of artificial enzymes is the RA95 series of retro-aldolases, originally computationally designed by Althoff, Wang, and Jiang *et al.*[27] to catalyze the retro-aldol decomposition of the substrate 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) (Figure 1.7). This is a multistep reaction, making *de novo* retro-aldolases some of the most mechanistically complex computationally designed enzymes to date. The retro-aldolase (RA) reaction proceeds through enamine catalysis using a catalytic lysine residue, the first step being nucleophilic attack of the substrate ketone by a catalytic lysine to form a tetrahedral carbinolamine intermediate. Subsequent water elimination forms the imine or Schiff base intermediate. Next, deprotonation of the β-alcohol, with the Schiff base acting as an electron sink, triggers carbon–carbon bond cleavage to release 6-methoxy-2-naphthaldehyde (6-MNA) and form an enamine intermediate. The enamine finally tautomerizes back to an imine that is then hydrolyzed to release the covalently bound acetone product and the free enzyme to complete additional rounds of catalysis. Despite the relatively high success rates achieved during RA design (75%, compared to 2–44% achieved in other studies discussed above), the limitations of the implemented computational methodologies remain evident. In the following sections, the design and subsequent optimization of RA95 will be detailed, highlighting the challenges associated with the methods used.

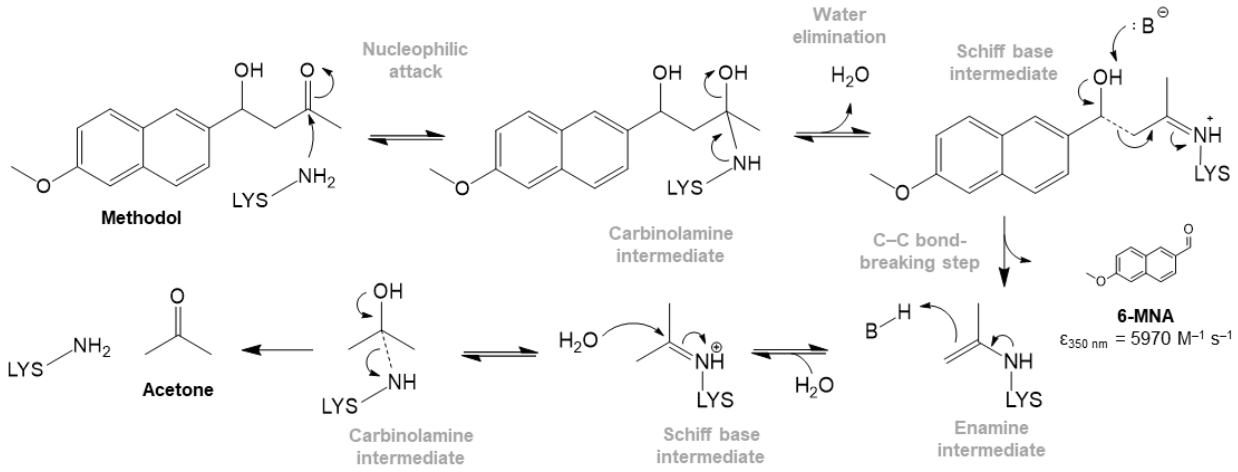

Figure 1.7. Multistep retro-aldol reaction mechanism. Computationally designed retro-aldolases catalyze the multi-step carbon–carbon bond cleavage reaction of the substrate 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) using a nucleophilic lysine residue. The products 6-methoxy-2-naphthaldehyde (6-MNA) and acetone are produced. Key steps and intermediate structures are labelled in grey.

1.5.1. Computational design of RA95

The *de novo* designed RA95 was created in the TIM barrel scaffold indole-3-glycerolphosphate synthase from *S. solfataricus* through the introduction of 11 active site mutations[27] predicted by RosettaDesign.[56] The theozyme was constructed using a composite TS intended to ensure that the designed enzyme would be capable of stabilizing multiple TS and intermediate structures in the multistep reaction. Specifically, the composite TS was created from TS structures of the water elimination and carbon–carbon bond-breaking steps, as well as carbinolamine and Schiff base intermediate structures (Figure 1.7). RA95 was computationally designed to possess a catalytic lysine (Lys210) to catalyze carbon–carbon bond cleavage through Schiff base formation. In addition, a catalytic water molecule was designed to facilitate proton transfers involving the β-alcohol and the hydroxyl group of the carbinolamine intermediate during the water elimination (Figure 1.8a) and carbon–carbon bond-breaking steps (Figure 1.8b). This

water molecule was oriented by a glutamate (Glu53) in the design model, and a large hydrophobic binding pocket was designed between two active site loops, L1 (residues 52–66) and L6 (residues 180–190), to accommodate the substrate naphthyl ring. Experimental characterization of RA95 revealed that it was able to catalyze methodol cleavage with a modest reported catalytic efficiency of 0.17 $M^{-1}s^{-1}$ (Figure 1.9a). While the enzyme was designed to be enantioselective for (*S*)-methodol, the preference for the (*S*)- over the (*R*)-configuration of the substrate is only 2.5:1. Solving the X-ray crystal structure of RA95 in complex with a covalent mechanism-based inhibitor[27] confirmed the general structure within the active site, including the designed Lys210, Glu53, and hydrophobic binding pocket (Figure 1.9b).

1.5.2. Evolution of an efficient retro-aldolase

Following its design, RA95 was heavily evolved in the lab through directed evolution (Figure 1.9a). The variant RA95.5, created through iterative cassette mutagenesis of active site residues,[27] contained six active site mutations relative to RA95 (V51Y, E53S, T83K, M180F, R182M, and D183N), resulting in a 74-fold increase in activity. This marks the appearance of the evolved catalytic lysine, Lys83, as well as one of the three additional catalytic tetrad residues, Tyr51. RA95.5 was previously crystallized in the presence of mechanism-based diketone inhibitor,[40] revealing that it is modified twice by the inhibitor, at both Lys83 and Lys210. However, mutagenesis studies show that Lys83 has a 30-fold catalytic advantage over Lys210. It was suggested that the other five favourable mutations preferentially enhance the small catalytic advantage of Lys83 over Lys210 by remodelling the active site in favour of the new substrate binding pocket. Mutations at positions 180, 182, and 183 induce a conformational change in loop

L6, creating the new binding pocket for the Lys83-bound substrate, while the new aromatic sidechains at positions 51 and 180 clash with the original Lys210 position in RA95.

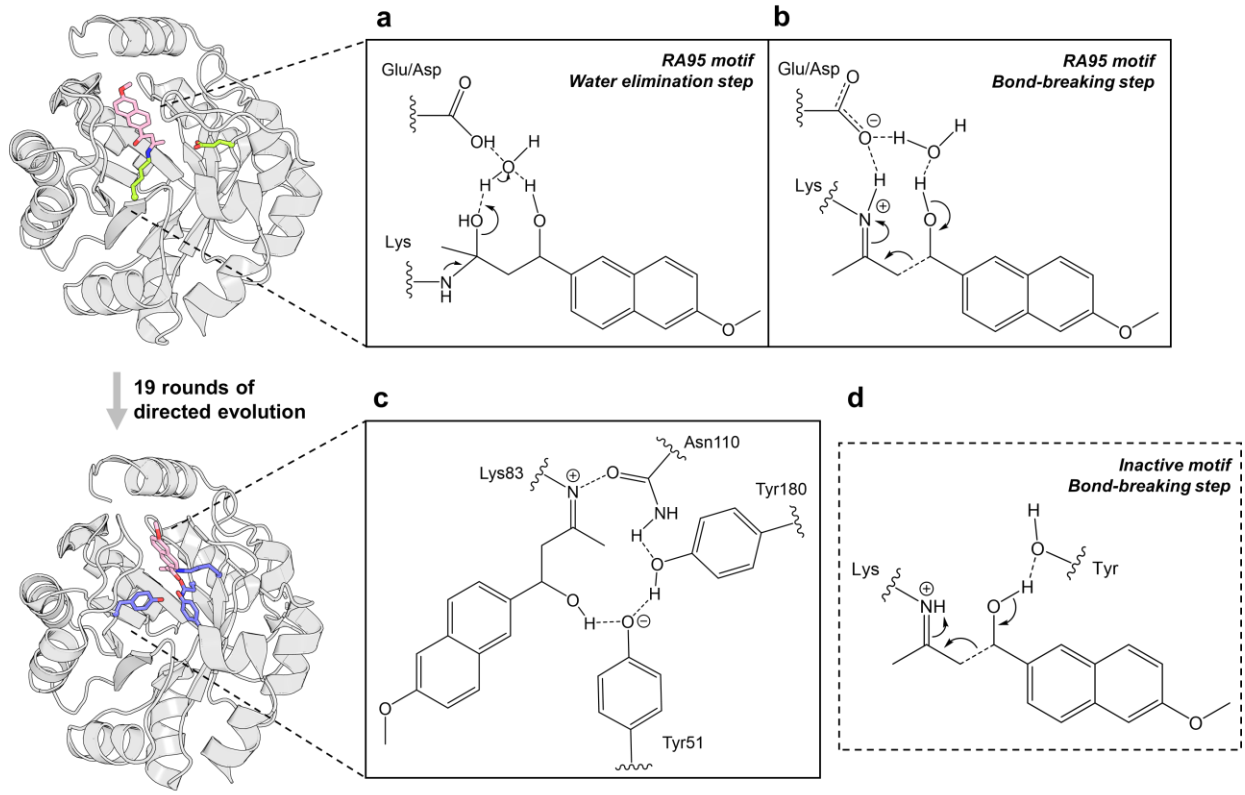

Figure 1.8. Catalytic motifs appearing in the RA95 series of retro-aldolases. The catalytic motif used in the design of RA95 possesses a catalytic lysine and a water molecule that facilitates the (a) water elimination and (b) carbon–carbon bond-breaking steps through proton shuttling on and off the substrate. The water molecule is oriented by a glutamate or aspartate, which additionally acts as a general acid-base. (c) The evolved catalytic motif in RA95.5-8F was identified after 19 rounds of directed evolution, consisting of the tetrad Lys83-Tyr51-Asn110-Tyr180. Hydrogen bonds between the catalytic tetrad residues and mechanism-based inhibitor used for crystallization are shown as black dashed lines. (d) Inactive catalytic motif tested in the first designs of *de novo* retro-aldolases, comprising a catalytic lysine and a tyrosine deprotonating the β-alcohol during the carbon–carbon bond-breaking step.

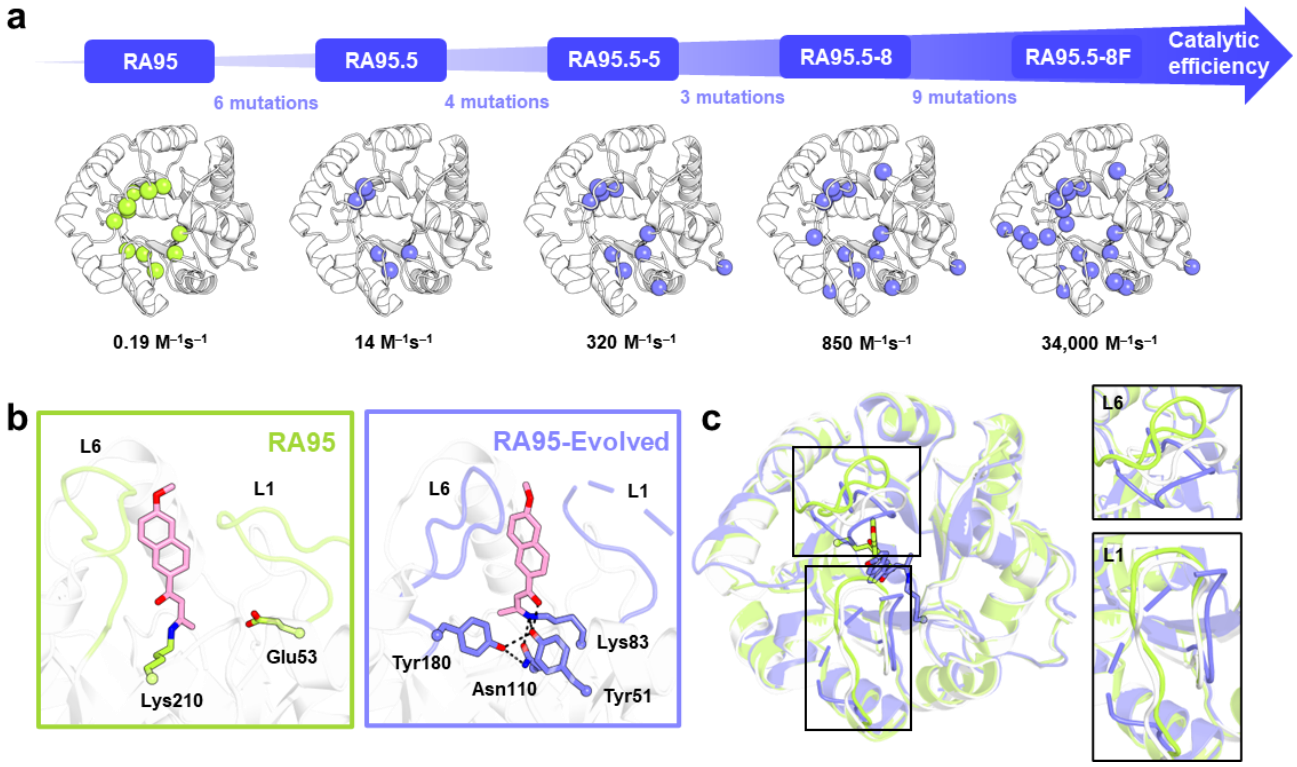

Figure 1.9. Computational design and directed evolution of RA95. (a) Evolutionary trajectory of the computationally designed *de novo* enzyme RA95. The 11 mutations introduced during computational design (green spheres) are shown on RA95. A total of 22 mutations (purple spheres) were introduced through 19 rounds of directed evolution to create RA95.5-8F. If a position was mutated multiple times along the evolutionary trajectory, the mutation is only shown in the variant where it was mutated for the last time. Reported catalytic efficiencies (k_{cat}/K_M) in $M^{-1}s^{-1}$ are displayed for each variant. (b) Active site configurations of RA95 (green) and RA95.5-8F (purple). RA95 was designed to have a catalytic lysine (Lys210) and a glutamate (Glu53) positioned to orient a catalytic water molecule. Through evolution, a catalytic tetrad was identified involving a new catalytic lysine (Lys83) and three additional residues participating in a hydrogen bond network (Tyr51, Asn110, and Tyr180). The mechanism-based diketone inhibitor used in crystallization is shown in pink. (c) Active site remodelling during evolution was accompanied by conformational changes of loops L1 and L6 (rectangles) to accommodate replacement of the catalytic lysine. The structures of RA95 (PDB ID: 4A29), RA95.5-8F (PDB ID: 5AN7), and wild-type (WT) template used for design of RA95 (PDB ID: 1A53) are shown in green, purple, and white, respectively.

Through random diversification of the full RA95.5 gene by error-prone PCR and DNA shuffling, the variant RA95.5-5 was created containing three additional active site mutations (S53T, S110N, and G178S) and three distal mutations (R23H, R43S, and T95M), allowing a 23-

fold increase in activity from RA95.5. One catalytic tetrad residue, Asn110, appears here. X-ray crystallography revealed that it is modified by inhibitor at only Lys83, and mutagenesis studies confirmed that Lys83 is the key catalytic lysine.[40] Notably, introduction of S110N in the first round of evolution of RA95.5 allowed a five-fold increase in activity. Further evolution created RA95.5-8, possessing an almost 3-fold activity increase from RA95.5-5 and containing two additional active site mutations (K135N and S178V) and three distal mutations (S43R, F72Y, and G212D). Interestingly, this includes the reversion of position 43 to its designed identity arginine, which was previously mutated to a serine in RA95.5.

Due to its conformational plasticity, RA95.5-8 has proven to be a promiscuous enzyme variant,[57] able to catalyze a variety of Schiff base mediated reactions including asymmetric Michael additions of carbanions to conjugated ketones,[58,59] the synthesis of γ-nitroketones,[60] and condensation reactions between electron-rich aldehydes and activated methylenes.[61,62] This variant was further evolved through an additional 6 rounds of directed evolution using a new ultrahigh-throughput droplet-based microfluidic screening system, introducing another 7 active site mutations (T53L, N90D, N135E, S178T, F180Y, K210L, L231M) and 6 distal mutations (R75P, S151G, A209, I213, S214F, R216P).[28,63] Notably, the original catalytic Lys210 is mutated here to a leucine, leaving Lys83 as the only possible catalytic lysine in the active site. The resulting highly evolved variant, RA95.5-8F, displays a reported catalytic efficiency of 34,000 $M^{-1}s^{-1}$, characteristic of natural class I aldolases (Figure 1.9a). Overall, evolution of RA95 allowed a 200,000-fold increase in activity through the introduction of a total of 22 mutations at both active site and distal positions, identified by 19 rounds of directed evolution.

Solving the crystal structure of the evolved variant, RA95.5-8F, revealed a new catalytic tetrad to be responsible for the increased activity, comprising a new catalytic lysine (Lys 83) and three additional hydrogen bonding residues (Tyr51, Asn110, and Tyr180) (Figure 1.9b). Interestingly, this structure is reminiscent of natural Schiff base-forming aldolases, which are characterized by networks of hydrogen bonds between the catalytic lysine, substrate, and other hydrogen bonding active site sidechains that may stabilize charge development and facilitate proton transfers during catalysis.[64] In RA95.5-8F, the catalytic tetrad is proposed to be involved in a network of hydrogen bonds that promotes catalysis by the lysine (Figure 1.8c). Their importance has been confirmed through the reversion of each residue of the tetrad individually, resulting in significant reductions in activity in each case. Tyr51 in particular has been proposed to be directly involved in catalysis, acting as a general base to deprotonate the β-alcohol of the substrate during the bond-cleavage step.[28]

Accompanying the replacement of the catalytic lysine in RA95.5-8F were large conformational changes in active site loops L1 and L6, required to avoid steric clashes with the new binding pose of the substrate. A section of L1 in the evolved variant is missing from the crystal structure, indicating that this loop is highly flexible or is present in multiple possible conformations in the crystal (Figure 1.9c). In addition to structural changes, significant modulations in enzyme function are observed during evolution, including the inversion of enantioselectivity of the enzyme from the (*S*)- to the (*R*)-enantiomer of the substrate, with an impressive preference of 480:1 after evolution. A shift in the overall rate-limiting step is also induced, from the carbon–carbon bond cleavage step (or an earlier step in the pathway) for earlier variants in the evolutionary trajectory,[65,66] to product release for RA95.5-8F. Specifically, the decomposition of the enamine

intermediate into a Schiff base following release of 6-MNA represents the slowest step in the reaction.[65,67]

1.5.3. Inaccuracies in the design of RA95

The RA95 series of retro-aldolases is particularly interesting as a case study in computational enzyme design for numerous reasons, many of these stemming from inaccuracies observed in the original RA95 design procedure. While the overall design features of RA95 were confirmed by solving the X-ray crystal structure, several structural and functional discrepancies were identified. Loops L1 and L6 displayed large disagreements with the design model, showing Cα atom root-mean-square deviation (RMSD) between 2 and 4 Å. While the mechanism-based inhibitor binds within the designed hydrophobic binding pocket, it is rotated 114º around its long axis relative to the design model, causing the carbonyl group of the inhibitor to point away from Glu53 and the water molecule designed to participate in catalysis (Figure 1.10). As a result, the enzyme retains full activity when Glu53 is mutated to alanine, confirming that this residue does not contribute to catalysis as designed. Either the designed function of this residue is not required for catalysis, or its function can be assumed by water. These differences between the RA95 design model and crystal structure, as well as the relatively low activity of RA95, highlight the structural and functional inaccuracies introduced by the design procedures used in the study.

The inaccuracies observed in the design of RA95 suggest that key components of the design procedure, such as selection of the catalytic motif or construction of the composite TS, were not optimal. This likely created an insufficient active site in RA95, leading to unexpected modifications to active site structure throughout the evolutionary trajectory towards RA95.5-8F. Unlike most cases of directed evolution in *de novo* designed enzymes, in which significant

rearrangement of the active site is not seen, RA95 undergoes significant active site remodelling. This included conformational remodelling of active site loops L1 and L6, replacement of the original catalytic Lys210 with the new Lys83, and creation of a novel catalytic tetrad proposed to enhance catalysis by the lysine through a network of hydrogen bonds. One tetrad residue, Tyr51, likely takes the place of the designed glutamate and catalytic water molecule in acting as a general base to deprotonate the β-alcohol of the substrate during the bond-breaking step of the reaction. Notably, a similar catalytic motif was tested in the design of the first *de novo* retro-aldolases by Jiang and Althoff *et al.*,[26] in which an active site tyrosine aids in catalysis through deprotonation of the substrate β-alcohol (Figure 1.8d). However, characterized designs containing this motif were not found to be active.

The unexpected structural differences between RA95 and RA95.5-8F explain the key functional changes accumulating during evolution, including the 200,000-fold activity increase, inversion of enantioselectivity from the (*S*)- to the (*R*)-enantiomer, and shift in rate-limiting step from the bond-cleavage to a product release step. However, while structural changes and their functional implications can be rationalized in retrospect, it is unclear whether these changes could have been predicted *a priori*. Why did the original *de novo* design procedure yield no designed sequences containing the catalytic lysine at position 83, which has been shown through directed evolution to catalyze the RA reaction more efficiently? Additionally, why were all original design sequences containing the Lys-Tyr catalytic dyad inactive, while this motif if highly efficient in RA95.5-8F? If the specific impacts of mutations introduced via directed evolution can be elucidated, we can determine if the active site structure of the evolved variant could have been designed computationally as opposed to that of RA95, and if not, how computational design procedures should be improved to more accurately predict efficient active sites.

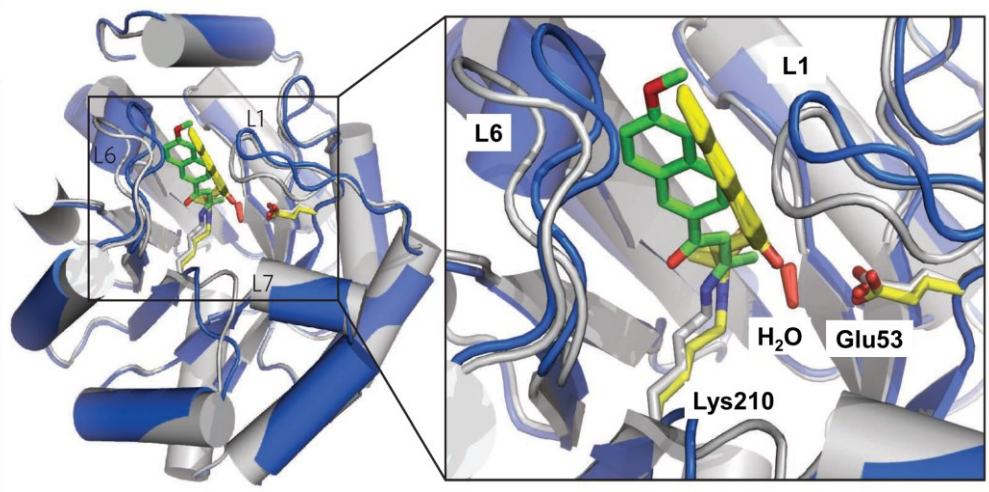

Figure 1.10. Structural prediction inaccuracy in the design of RA95. Overlay of the *in silico* design model of RA95 in blue (composite TS in yellow) and the X-ray crystal structure of RA95 in grey (covalent mechanism-based inhibitor in green). The designed Lys210-Glu53 catalytic motif, and catalytic water in the case of the model, are shown as sticks. Key flexible active site loops (L1 and L6) are identified. Figure reproduced from Giger and Caner *et al.*.[40]

1.6. Thesis objectives

In Chapter 1, methods in computational enzyme design were discussed, as well as current successes and limitations to design methodologies. Low success rates, inaccurate structure predictions, and modest catalytic rate enhancements represent challenges for *de novo* enzyme designers, causing the need for designs to be optimized via directed evolution through the introduction of both active site and distal mutations. It is problematic that many active site mutations are found to increase enzyme activity and thus should have been identified in the original computational design calculations. Additionally, the prediction of activity-enhancing distal mutations presents a major challenge in computational enzyme design, and it is often unclear how these beneficial distal mutations impact catalysis and whether they are essential for the increased catalytic activities seen in evolved enzymes. Through a greater understanding of how active site and distal mutations give rise to enhanced enzyme activity, we can progress towards the general

goal of designing highly active *de novo* enzymes without requiring directed evolution. Therefore, this thesis aims to investigate the impacts of active site versus distal mutations introduced by directed evolution on the structure and function of a *de novo* enzyme and use this information to guide the improvement of computational enzyme design methodologies (Figure 1.11).

The RA95 series of *de novo* retro-aldolases will be used as a case study, as the significant remodelling of the RA95 active site observed throughout evolution is not seen in other families of *de novo* enzymes. In Chapter 2, the impacts of active site and distal mutations introduced by directed evolution on the structure, function, and dynamic properties of the *de novo* enzyme RA95 are evaluated. Computational recapitulation of RA95 series active sites is then attempted, with the goal of determining whether the highly active evolved active site could have been predicted during the original design of RA95. In Chapter 3, the insights into the impacts of evolved mutations on RA95 will be discussed to understand how this information can be used to address the limitations of current computational enzyme design methods, both in the RA95 series and in the broader context of enzyme design.

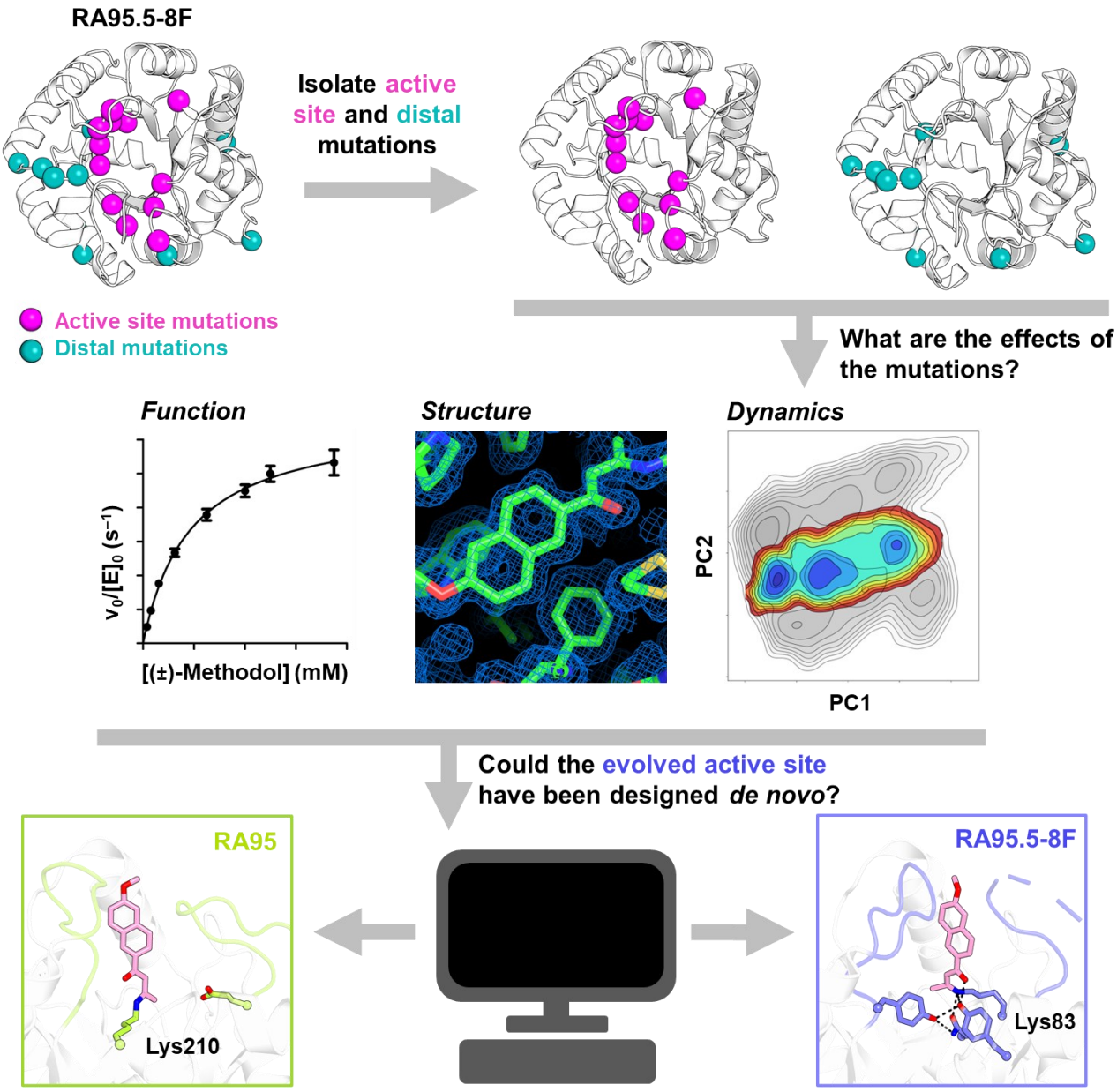

Figure 1.11. Overview of thesis objectives. Chapter 2 will focus on evaluating the individual impacts of active site and distal mutations introduced by directed evolution on the structure, function, and dynamics of RA95. Chapter 3 will discuss how this information can be used to guide the improvement of computational enzyme design methodologies, with the end goal of being able to design highly efficient *de novo* enzymes without requiring directed evolution.

Chapter 2. Structural and functional effects of active site and distal mutations on a computationally designed retro-aldolase

Serena E. Hunt,[1,2] Niayesh Zarifi,[1,2] Alec Martinez,[3] Michael C. Thompson,[3] and Roberto A. Chica[1,2]

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

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

[3]Department of Chemistry and Biochemistry, University of California, Merced, Merced, California, United States, 95343

2.1. Statement of Contribution

Protein crystallization was performed by Serena E. Hunt and Alec Martinez. X-ray data collection and processing was performed by Dr. Michael C. Thompson. Refinement of crystal structures was performed by Dr. Roberto A. Chica. Design of the RA95-Core and RA95-Shell variants was done by Niayesh Zarifi. Experimental design was done by Dr. Roberto A. Chica and Serena E. Hunt. All other experimental and computational experiments were completed by Serena E. Hunt. This chapter was written by Serena E. Hunt.

Acknowledgements

S.E.H. was the recipient of two Ontario Graduate Scholarships and a Collaborative Research and Training Experience (CREATE) program scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC).

2.2. Introduction

Enzymes are some of the most efficient known catalysts, able to catalyze chemical reactions with impressive rate enhancements, chemoselectivities, and specificities. These properties, along with their biodegradability, make them attractive catalysts for use in industrial applications such as the manufacturing of foods[68] and pharmaceuticals,[69] and as active ingredients in detergents.[70] However, nature's finite repertoire of enzymes limits their applicability in industry. For this reason, it is a general objective of protein engineering to be able to design efficient *de novo* enzymes on demand for any desired reaction, using general and robust computational methodologies. In 2001, Bolon and Mayo[21] demonstrated that this may be attainable, through the design of the first *de novo* enzyme using computational protein design (CPD) methods. The development of computational enzyme design algorithms throughout the past two decades[19,20] has since then brought us closer to reaching this goal. Using quantum mechanics (QM)-based ideal active site design, these methods have allowed the creation of *de novo* enzymes for model organic reactions including the Kemp elimination,[22,24] retro-aldol,[26,27] ester hydrolysis[29], Diels-Alder,[30] and Morita-Baylis-Hillman[32] reactions.

Despite the successes seen in computational enzyme design, current design methodologies possess inherent limitations, posing challenges in their implementation. Success rates, referring to the proportion of active designs of those tested experimentally, are generally low for *de novo* enzymes.[22,26,30] Additionally, even when a design is active, structural analysis often reveals discrepancies between the design model and crystal structure, with key catalytic and ligand-binding interactions being inaccurately predicted.[24] *De novo* designed enzymes display catalytic activities orders of magnitude lower than those of natural enzymes as a result,[71] highlighting the

need for continued development of robust computational enzyme design methods. To improve the activities of *de novo* enzymes, directed evolution has been implemented, allowing the rate enhancements of evolved enzymes to approach those seen in nature.[22,23,25,39] Throughout evolution, mutations are introduced at both active site positions and those more distant to the active site, both of which have been shown to be important in the observed enhancements to catalytic activity. While active site mutations have contributed to improved activity through enhanced shape complementarity to the TS,[39] widening of the active site,[23] and the introduction of new catalytic groups,[41] distal mutations have been shown to have more subtle effects, causing shifts in populations of conformational sub-states on the energy landscape of the enzyme.[41,72]

If we could replicate the effects of directed evolution computationally, we could come closer to reaching the goal of designing highly active artificial biocatalysts without requiring directed evolution. However, it is often unclear why the activity-enhancing mutations introduced through evolution were not identified in the original computational design calculations, especially since many of them include designed catalytic and ligand-binding residues. Additionally, while changes induced by beneficial distal mutations can be consistently rationalized using computational modelling techniques, the prediction of these mutations *a priori* presents a major challenge in computational enzyme design.[42] Thus, a better understanding of how active site and distal mutations introduced by directed evolution contribute to enhanced catalytic activity, including whether they are required for the design of efficient *de novo* enzymes, would guide the improvement of computational enzyme design procedures.

Here, we evaluate the contributions of active site and distal mutations introduced by directed evolution on the structure, function, and dynamics of the *de novo* enzyme RA95. We

observe that active site mutations alone contribute a 3,600-fold enhancement in catalytic activity, indicating that a much more efficient active site was identified by directed evolution. This suggests that distal mutations may not be required to design efficient artificial enzymes, and that we should be able to design a better active site than what was achieved in RA95 through the optimization of current computational design procedures. However, conformational and dynamical changes to flexible active site loops and other key backbone atoms, induced by distal mutations during evolution, prevent the computational recapitulation of the evolved active site on the original design template. Additional backbone and loop remodelling methods will likely need to be incorporated to increase accuracy of computational enzyme design methods and allow the design of highly active *de novo* biocatalysts without relying on directed evolution.

2.3. Results

2.3.1. Functional effects of mutations

The *de novo* enzyme RA95, one of the most mechanistically complex computationally designed enzymes to date,[35] catalyzes the retro-aldol decomposition of the methodol substrate (Figure 1.7) with a modest catalytic efficiency (k_{cat}/K_M) of 0.52 $M^{-1}s^{-1}$ (Table 2.1). RA95 has been evolved through 19 rounds of directed evolution, allowing a 200,000-fold increase in activity through the addition of 12 active site and 10 distal mutations over an evolutionary trajectory yielding the RA95.5, RA95.5-5, and RA95.5-8 intermediate variants (Figure 2.1a). This activity increase was accompanied by significant structural remodelling of the active site, including replacement of the original catalytic Lys210 with a new Lys83 (Figure 1.9b) and large conformational changes in active site loops L1 and L6 (Figure 1.9c). The resulting highly evolved

variant, RA95.5-8F, displays a catalytic efficiency of 12,000 $M^{-1}s^{-1}$ (Table 2.1), characteristic of natural class I aldolases. In this study, RA95.5-8F will be referred to as RA95-Evolved.

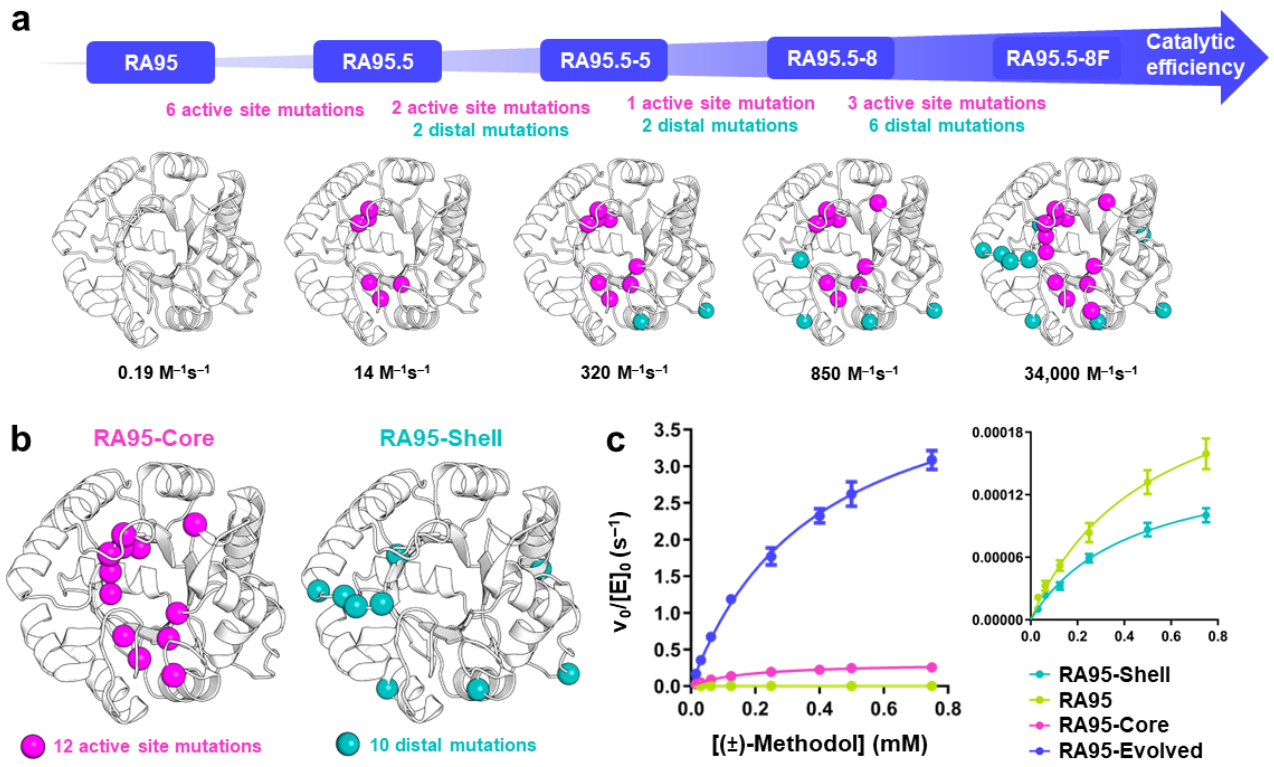

Figure 2.1. Generation of RA95-Core and RA95-Shell variants. (a) Evolutionary trajectory of the computationally designed *de novo* enzyme RA95. A combination of 12 active site mutations (magenta spheres) and 10 distal mutations (teal spheres) were introduced over 19 rounds of directed evolution to create RA95.5-8F, or RA95-Evolved. If a position was mutated multiple times along the evolutionary trajectory, the mutation is only shown in the variant where it was mutated for the last time. Reported catalytic efficiencies (k_{cat}/K_M) in $M^{-1}s^{-1}$ are displayed for each variant. (b) The RA variants RA95-Core and RA95-Shell are created by reverting the mutations identified by directed evolution to their designed identities at distal positions and active site positions, respectively. (c) Michaelis-Menten plots of normalized initial rates as a function of racemic methodol concentration show that RA95-Core is four orders of magnitude more active than RA95 and only one order of magnitude less active than RA95-Evolved. Data represent the mean ± SEM for triplicate measurements from $n = 3$ independent biological replicates.

To investigate the effects of active site and distal mutations introduced during directed evolution on the function of RA95, we created two new RA variants where the distal and active site mutations identified by directed evolution are reverted separately to their designed identities. These variants were termed RA95-Core and RA95-Shell, respectively (Figure 2.1b). Here, active site mutations were defined as those occurring within the first shell (i.e., within 4 Å of the inhibitor) and second shell (i.e., within 4 Å of the first shell) of the RA95-Evolved crystal structure, and the remaining were defined as distal mutations. This definition matches typical computational enzyme design protocols, where first and second shell positions are commonly chosen as design positions. Protein purifications of RA95-Core and RA95-Shell, as well as all other variants characterized in this study, were achieved with high purity (Figure 2.2).

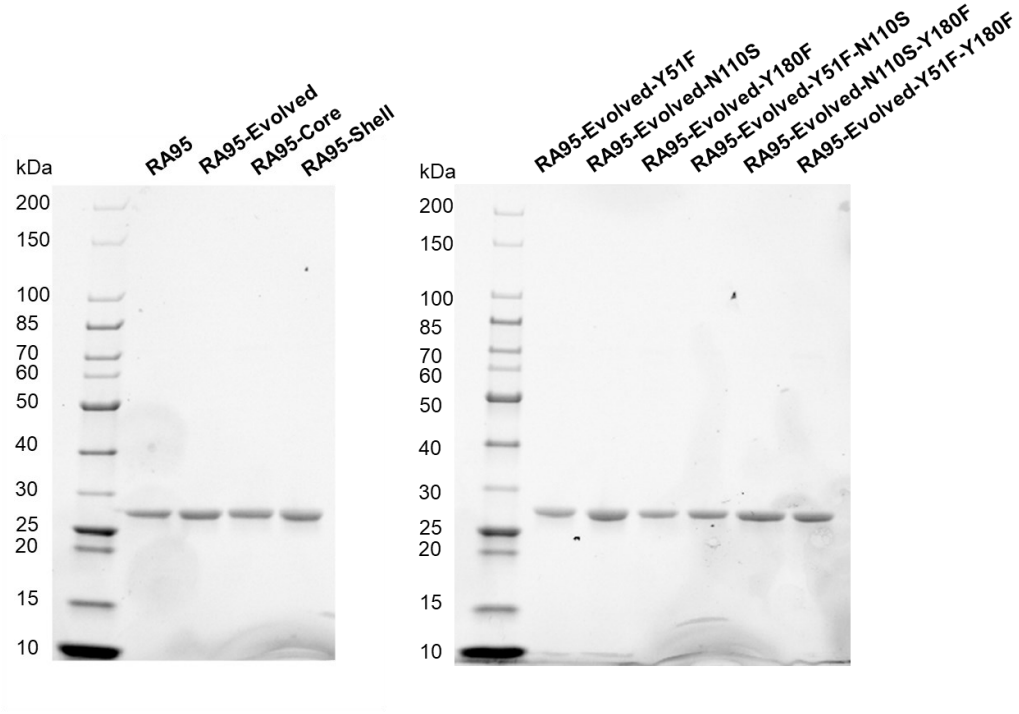

Figure 2.2. Representative SDS-PAGE gels for all purified enzymes. Samples were loaded with 3 μL of Unstained Protein Standard, Broad Range (10-200 kDa) (NEB) on 4–20% precast polyacrylamide gels. Each enzyme has a molecular weight between 29.5–29.8 kDa.

Experimental characterization of RA95-Core revealed that it catalyzes the cleavage of (±)-methodol with a catalytic efficiency of 1,900 $M^{-1}s^{-1}$, representing a 3,600-fold increase from RA95 (Figure 2.1c, Table 2.1). In addition, RA95-Core retains the highly perturbed catalytic lysine pK_a of the evolved variant, showing a sigmoidal pH-rate profile with an inflection point at 5.8 ± 0.1, compared to 5.6 ± 0.1 for RA95-Evolved (Figure 2.3, Table 2.1). This demonstrates that a much more effective active site was identified by directed evolution than what was computationally designed in RA95. As all but three of the mutations in RA95-Core (N90D, K135E, D183N) are found at positions that were optimized during the original computational design of RA95, it should be possible to design more efficient active sites *de novo* than what was achieved in RA95 through optimization of computational design procedures. These results support a previous study by Broom and Rakotoharisoa *et al.*[73] on the activity-enhancing features introduced during directed evolution of the Kemp eliminase HG3. The authors created an HG variant containing all active site mutations introduced by directed evolution of HG3, but not distal mutations. The resulting variant, HG4, had a catalytic efficiency >700-fold higher than that of the original design and only 20% lower than the evolved variant. Taken together, the combined results of RA95-Core and HG4 suggest that the significant challenge of designing distal mutations may not need to be tackled in order to design efficient *de novo* enzymes.

Table 2.1. Experimental characterization of RA variants.

Enzyme	# mut. from RA95	k_{cat} (s^{-1})[b]	K_M (μM^{-1})[b]	k_{cat}/K_M ($M^{-1}s^{-1}$)	Fold increase in k_{cat}/K_M from RA95	pK_a[c]	Tm[d] (°C)	ΔH_U[e] (kcal/mol)
RA95	-	$((2.7 \pm 0.4) \times 10^{-4})$	500 ± 200	0.52	-	-	83.66 ± 0.02	-257 ± 3
RA95-Evolved	22	4.6 ± 0.3	390 ± 50	12,000	23,000	5.6 ± 0.1	71.9 ± 0.1	-113 ± 5
RA95-Core	12	0.32 ± 0.01	170 ± 10	1,900	3,600	5.8 ± 0.1	68.7 ± 0.1	-93 ± 4
RA95-Shell	10	$((1.6 \pm 0.2) \times 10^{-4})$	400 ± 100	0.37	0.7	-	81.56 ± 0.03	-263 ± 6
RA95-Evolved-Y51F	22	0.059 ± 0.002	32 ± 6	1,820	3,500	5.55 ± 0.07	-	-
RA95-Evolved-N110S	21	0.237 ± 0.006	20 ± 3	11,800	23,000	5.1 ± 0.1	-	-
RA95-Evolved-Y180F	22	1.21 ± 0.06	20 ± 6	60,000	116,000	5.91 ± 0.05	-	-
RA95-Evolved-Y51F-N110S	21	0.0090 ± 0.0003	33 ± 5	272	520	-	-	-
RA95-Evolved-N110S-Y180F	21	0.039 ± 0.001	84 ± 9	463	890	-	-	-
RA95-Evolved-Y51F-Y180F	22	0.0042 ± 0.0001	31 ± 5	134	260	-	-	-

[a]The designed and evolved catalytic lysines are located at positions 210 and 83, respectively.
[b]Kinetic parameters were determined with (±)-methodol. k_{cat} and K_M were determined by fitting the data to the Michaelis-Menten model (Equation 1). Errors reflect the error of nonlinear regression fitting. $n = 3$ independent experiments.
[c]k_{cat}/K_M versus pH data were fitted to Equation 3 using nonlinear least squares regression. The apparent pK_a of the catalytic lysine (pK_{a1}) of each variant is presented, with the errors of nonlinear regression fitting provided.
[d]Thermal denaturation midpoint temperature determined through loss of CD signal at 222 nm. Errors of nonlinear regression fitting to a two-state transition model are provided.
[e]Enthaply of unfolding determined by thermal denaturation monitored by CD signal at 222 nm. Errors of nonlinear regression fitting to a two-state transition model are provided.

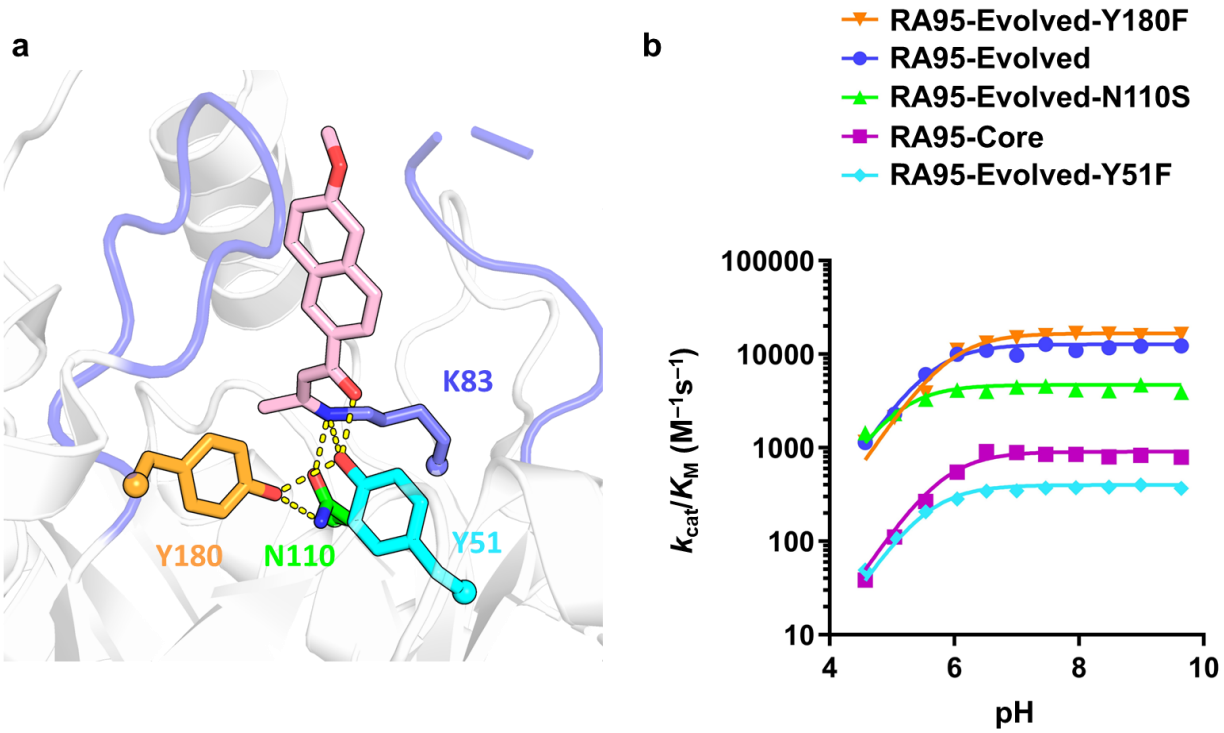

Figure 2.3. pH-rate profile determination of RA variants. (a) Representation of the RA95-Evolved active site highlighting the catalytic lysine and active site loops in purple, and the remaining three catalytic tetrad residues in orange, green, and cyan. These three tetrad residues were reverted to study their effects on the catalytic lysine pK_a. (b) pH-rate profiles of RA variants. Steady-state kinetic assays were carried out at 29ºC in Britton-Robinson buffer with 2.7% acetonitrile. Product (6-methoxy-2-napthaldehyde) formation was monitored spectrophotometrically at 350 nm (ε = 5,970 M^{-1} cm^{-1}). Triplicate measurements were completed with varying concentrations of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) at various pH. Initial reaction rates were fitted to the linear portion of the Michaelis-Menten model shown in Equation 2 by linear regression and k_{cat}/K_M values were deduced from the slope. Data on the pH-rate profiles represent the k_{cat}/K_M values determined at each pH. Error bars representing the errors of linear regression fitting to the linear portion of the Michaelis-Menten model are too small to be visible. These data were fitted to Equation 3 using nonlinear least squares regression. The apparent pK_a of the catalytic lysine (pK_{a1}) of each variant is presented in Table 2.1, with the errors of nonlinear regression fitting provided.

Of the 12 active site mutations, the most essential are arguably those forming the catalytic tetrad. The V51Y mutation was introduced along with the new catalytic lysine T83K in RA95.5, a variant early in the evolutionary trajectory of RA95 (Figure 2.1a) created by iterative cassette mutagenesis of only active site residues. Tyr51 was proposed to originally play a role in preferentially enhancing the small catalytic advantage of Lys83 over Lys210 in RA95.5 by clashing with the original Lys210 sidechain position.[40] Unlike in RA95-Evolved, the Tyr51 sidechain in RA95.5 is not oriented in a way that allows it to participate in catalysis through hydrogen bonding. The third catalytic tetrad residue to be introduced was S110N, appearing in the first round of evolution of RA95.5 and contributing a five-fold increase in activity towards the overall increase of 23-fold in RA95.5-8 (Figure 2.1a).[40] Asn110 is within hydrogen-bonding distance to the Lys83 sidechain amino group in RA95.5-8, indicating that it may have been the first additional tetrad residue to aid in facilitation of the reaction by Lys83. Finally, the catalytic tetrad was completed by the F180Y mutation introduced during the final step of the evolutionary trajectory, in which RA95-Evolved was created.[28] To gain a better understanding of how the tetrad impacts catalysis, we created point mutants where each of Tyr51, Asn110, and Tyr180 was reverted individually.

Kinetic analysis of the point mutants revealed that reverting Tyr51, Asn110, and Tyr180 individually resulted in approximately 78-, 19-, and 4-fold decreases in k_{cat}, respectively (Figure 2.4, Table 2.1), which match the trends observed in previous mutagenesis studies using enantiopure substrate.[28] Additionally, reversion of each combination of two of these residues simultaneously revealed that the highest activity is preserved when Tyr51 is retained in the active site (Figure 2.4, Table 2.1). These results confirm that not only are each of the tetrad residues important for catalysis, but also that Tyr51 has the greatest impact. This matches with a previously proposed

mechanism of catalysis by RA95-Evolved, whereby Tyr51 is directly involved in the catalytic mechanism, acting as a general base to deprotonate the β-alcohol of the substrate during the bond cleavage step.[28] However, the activity-enhancing effects of active site mutations outside of the catalytic tetrad should not be ignored. The variant RA95.5, which contains only six active site mutations relative to RA95 including V51Y and T83K, has an activity approximately 30-fold lower than the mutant RA95-Evolved-N110S-Y180F, despite these mutants containing the same components of the catalytic tetrad.[40] The remaining 30-fold increase is due to the combination of the 10 distal mutations and remaining four active site mutations between RA95.5 and RA95-Evolved-N110S-Y180F.

To confirm whether the reductions in catalytic activity observed in the active site point mutants were due to involvement of these residues in the facilitation of the chemical reaction, and not simply their potential role in modulating the catalytic lysine pK_a, the pH-rate profiles of these points mutants were determined, revealing only small changes in pK_a (Figure 2.3, Table 2.1). This shows that the full catalytic tetrad is not necessarily required for the perturbed lysine pK_a seen in the evolved variant, and it is likely that the decrease in apparent pK_a during evolution is due simply to the increased hydrophobicity in the evolved active site. Many of the beneficial active site mutations introduced during evolution involved increases in sidechain size or hydrophobicity (V51Y, T83K, G178T, M180Y, R182M, and K210L), and it was suggested that many of the early RA designs were characterized by underpacking of the active site.[27]

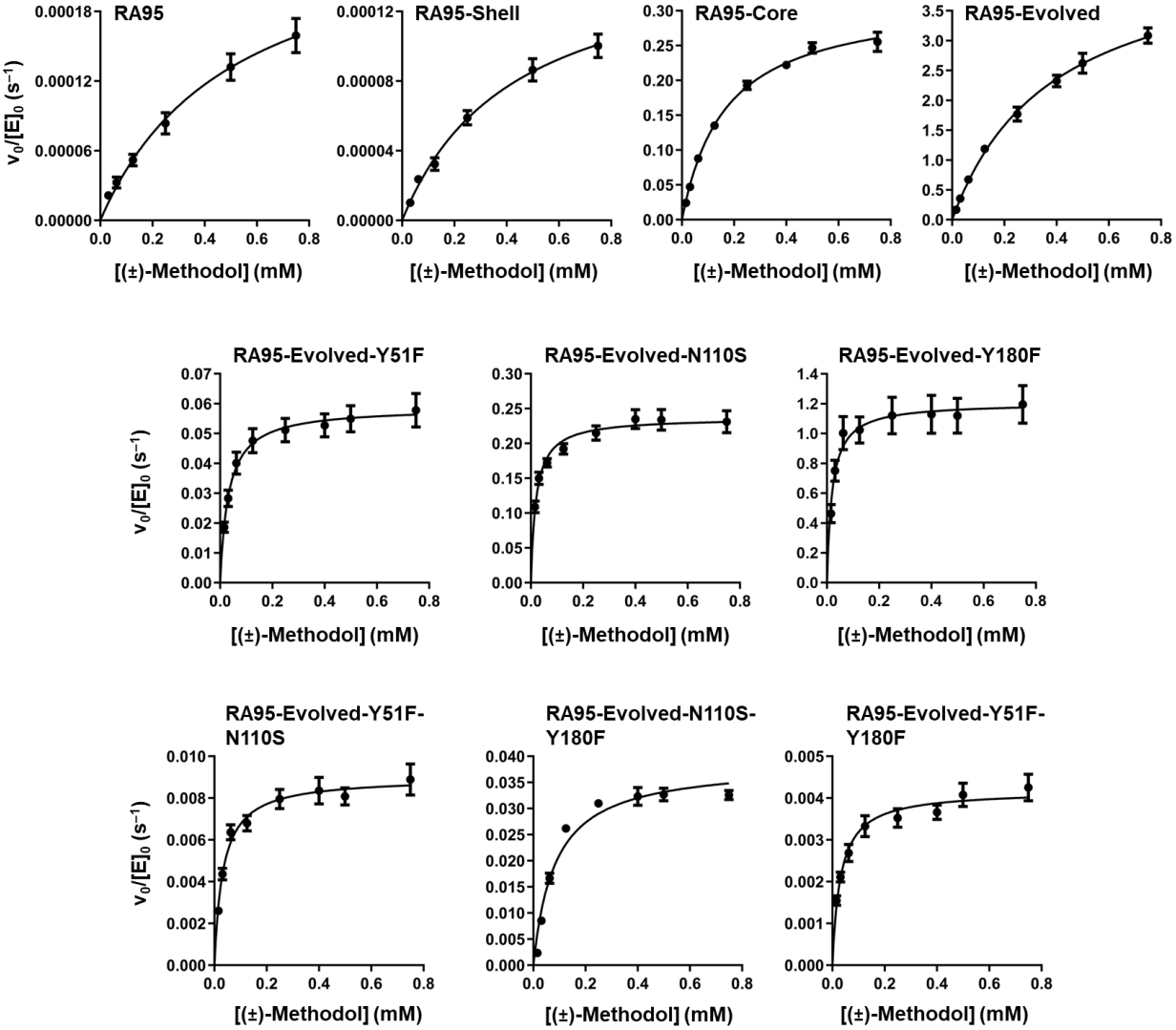

Figure 2.4. Steady-state kinetics for RA variants. Michaelis-Menten plots of normalized initial rates as a function of racemic 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) concentration are shown. Assays were carried out at 29ºC in 25 mM HEPES (pH 7.5), 100 mM NaCl, 2.7% acetonitrile. Product (6-methoxy-2-napthaldehyde) formation was monitored spectrophotometrically at 350 nm (ε = 5,970 M^{-1} cm^{-1}). Data represent the mean ± SEM for triplicate measurements from n = 3 independent biological replicates. k_{cat} and K_M were determined by fitting the data to the Michaelis-Menten model shown in Equation 1.

While active site mutations allow a 3,600-fold activity increase in RA95-Core compared to RA95, an additional 6-fold increase is seen in the evolved variant, indicating that the distal mutations introduced via directed evolution also play an important role. However, kinetic characterization of RA95-Shell shows no increase in activity compared to RA95 (Figure 2.1c, Table 2.1), indicating that the combined beneficial impacts of active site and distal mutations in RA95-Evolved are synergistic in nature. Given that the designed and evolved active sites are so different, it is expected that the distal mutations optimized by directed evolution for the evolved active site would not also be beneficial for the designed active site. This is also the case when considering thermal stability of the RA variants (Figure 2.5, Table 2.1). When adding the distal mutations to RA95, the melting temperature determined by following circular dichroism (CD) signal at 222 nm decreases from 83.66 ± 0.02°C in RA95 to 81.56 ± 0.03°C in RA95-Shell, indicating that the RA95 fold is slightly destabilized by the distal mutations. RA95-Core on the other hand is stabilized by the distal mutations, with melting temperature increasing from 68.7 ± 0.1°C to 71.9 ± 0.1°C in RA95-Evolved. This demonstrates how distal mutations are not only optimized for catalytic activity through evolution, but also play a role in thermal stability of the enzyme.

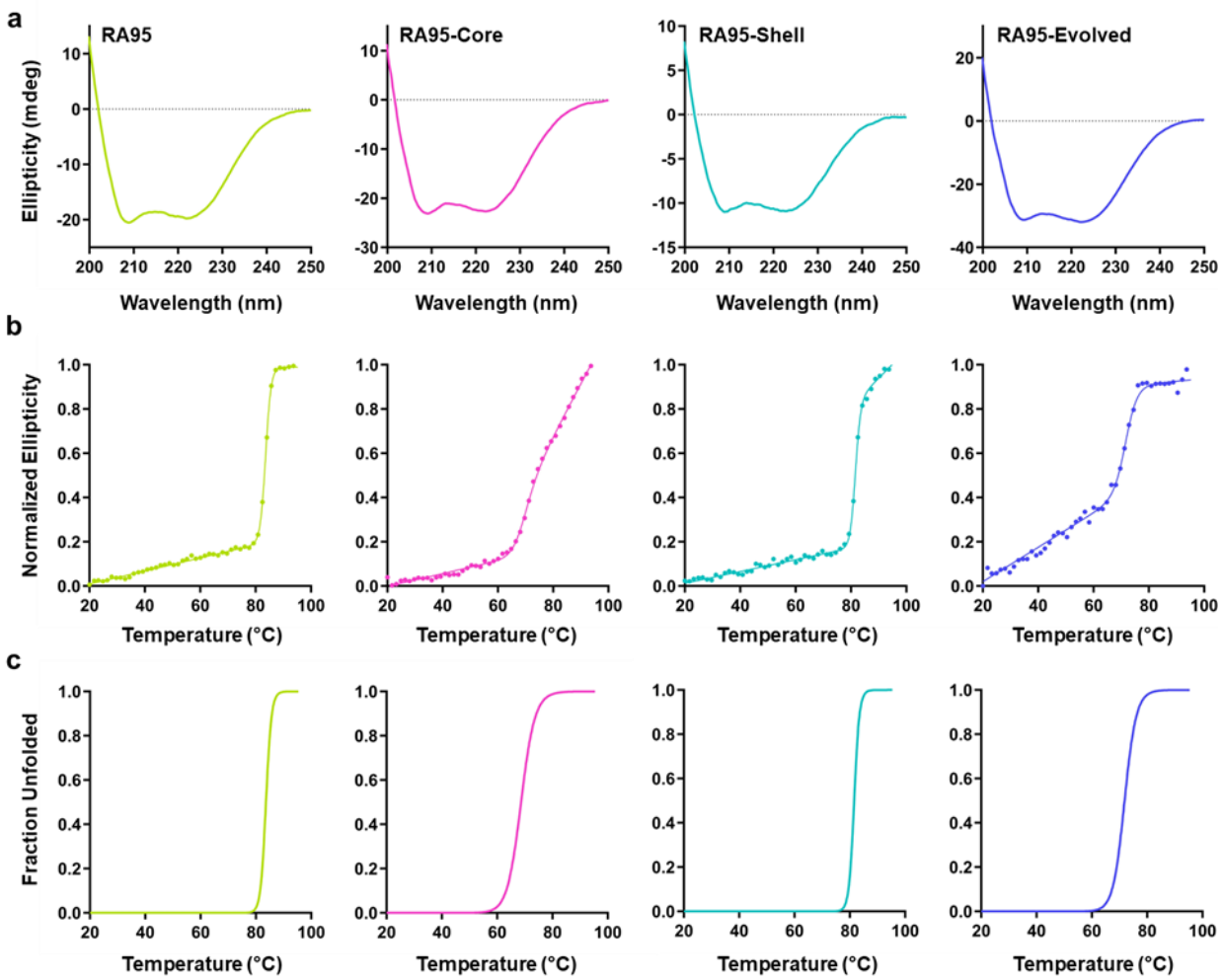

Figure 2.5. CD and thermal denaturation assays for RA variants. (a) Far-UV CD spectra. Scans at 20°C, sampled every 1 nm at a rate of 10 nm/min. Three scans were acquired and averaged for each sample. Each sample contained approximately 5 µM of enzyme. (b) Thermal denaturation monitored by CD at 222 nm. Samples were heated at a rate of 1°C per minute, and ellipticity at 222 nm was measured every 0.2°C (n = 1). Tm values were determined by fitting the data to a two-state transition model with correction for pre- and post-transition linear changes in ellipticity as a function of temperature using nonlinear least-squares regression. Reversibility of thermal denaturation was not assessed. (c) Fraction unfolded calculated from the fit in (b).

2.3.2. Structural effects of distal mutations

Despite the lack of an activity increase in RA95-Shell compared to RA95, distal mutations are activity-enhancing in RA95-Evolved, indicating that they must be contributing indirectly to activity. To elucidate any impacts of distal mutations on the structure of RA95, we further purified RA95 and RA95-Shell (Figure 2.6) and solved their X-ray crystal structures in the absence of any ligands. The unit cells for both variants corresponded to space group P 2 2_1 2_1 with one protein molecule in the asymmetric unit. They diffracted at resolutions of 1.77 Å and 1.89 Å for RA95-Shell and RA95, respectively (Table 2.2). A significant difference was observed in the third unit cell dimension between the two variants, either caused by differences in crystal packing or, more likely, the large conformational change seen in loop L1 of RA95-Shell (Figure 2.7a). While conformational changes are observed in both loops L1 and L6 throughout the evolutionary trajectory of RA95, which can be linked to the significant active site remodelling discussed previously, the observed conformation of L1 in RA95-Shell represents a new open conformation not seen in any other crystallized RA variant to date. Interestingly, the distal mutations inducing this change are not located on or near loops L1 or L6 (Figure 2.7b). Additionally, while a defined closed conformation of loop L1 is seen in RA95, loop L1 in RA95-Evolved displays a large amount of conformational heterogeneity in the crystal structure, resulting in residues 58–61 and residues 58–63 being unassigned in the inhibitor-bound and unbound structures, respectively. This suggests that the distal mutations create an open loop conformation in RA95-Shell, and upon addition of the active site mutations in RA95-Evolved, both open and closed loop conformations are present. As a result, it is possible that the role of distal mutations identified by directed evolution in RA95-Evolved involves active site accessibility, which may cause an increase in catalytic activity through improved substrate entry and/or product release.

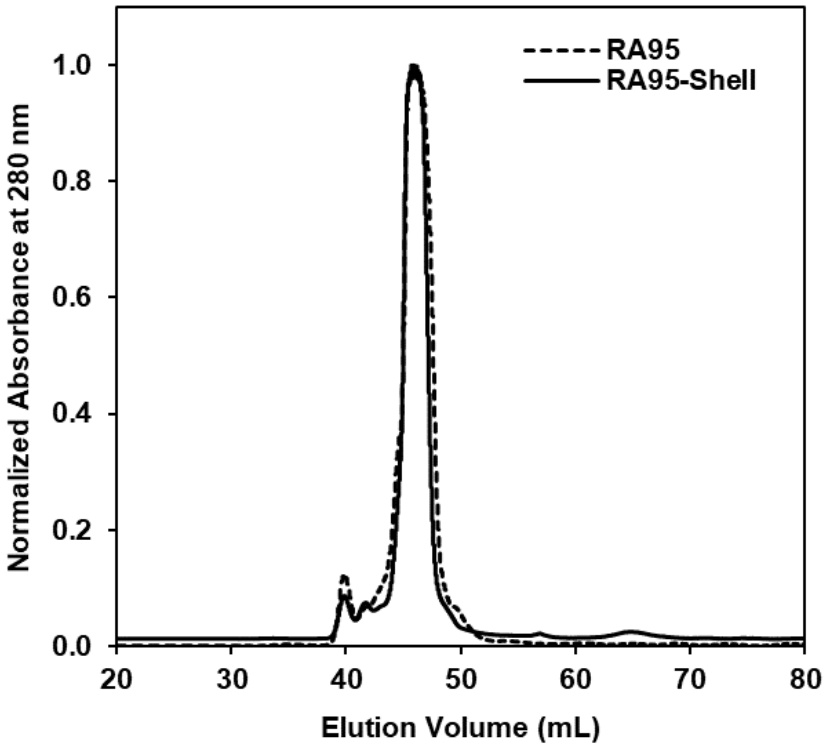

Figure 2.6. Fast protein liquid size-exclusion chromatography (FPLC) profiles for RA95 and RA95-Shell samples used for crystallography. Proteins were purified by gel filtration into 20 mM potassium phosphate, pH 7.4, 50 mM NaCl using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad).

Table 2.2. Crystallography data and refinement statistics.

	RA95	RA95-Shell
PDB ID	TBD[a]	TBD[a]
Crystallizations conditions	0.1 M sodium acetate pH 5.2 3.1 M NaCl 7 mg/mL protein	0.1 M sodium acetate pH 4.4 19% PEG 3000 6 mg/mL protein
Protein buffer	20 mM potassium phosphate pH 7.4 50 mM NaCl	20 mM potassium phosphate pH 7.4 50 mM NaCl
Data collection[b]		
Temperature (K)	277	280
Resolution (Å)	49.00–1.89	51.66–1.77
Space group	P 2 21 21	P 2 21 21
Cell params.		
a b c (Å)	44.377 65.156 97.995	41.034 64.804 85.638
α β γ (°)	90 90 90	90 90 90
Chains per asymm. unit	1	1
R_{pim}	0.093 (0.464)	0.098 (0.648)
$CC_{1/2}$	0.988 (0.576)	0.991 (0.391)
I/σI	4.9 (0.9)	7.0 (1.0)
Completeness (%)	98.6 (97.3)	100.0 (100.0)
Multiplicity	6.4 (6.7)	6.3 (5.9)
Wilson B-factor ($Å^2$)	18.820	17.570
# unique reflections	23151 (1137)	22968 (1127)
Refinement		
R work/free	0.1588/0.1969	0.1869/0.2184
No. atoms		
Protein	2109	2073
Ligand	1	0
Water	109	77
Averaged B-factors ($Å^2$)		
Protein	30.53	30.46
Ligands	29.68	–
Water	34.19	32.22
RMSD		
bond lengths (Å)	0.011	0.003
bond angles (°)	1.007	0.525
Molprobity statistics		
Ramachand. outliers (%)	0.00	0.00
Ramachand. allowed (%)	1.63	0.82
Ramachan. favored (%)	98.37	99.18
Rotamer outliers (%)	0.00	0.00
MolProbity clashscore	2.09	0.95

[a]To be determined.
[d]Highest resolution shell is shown in parentheses.

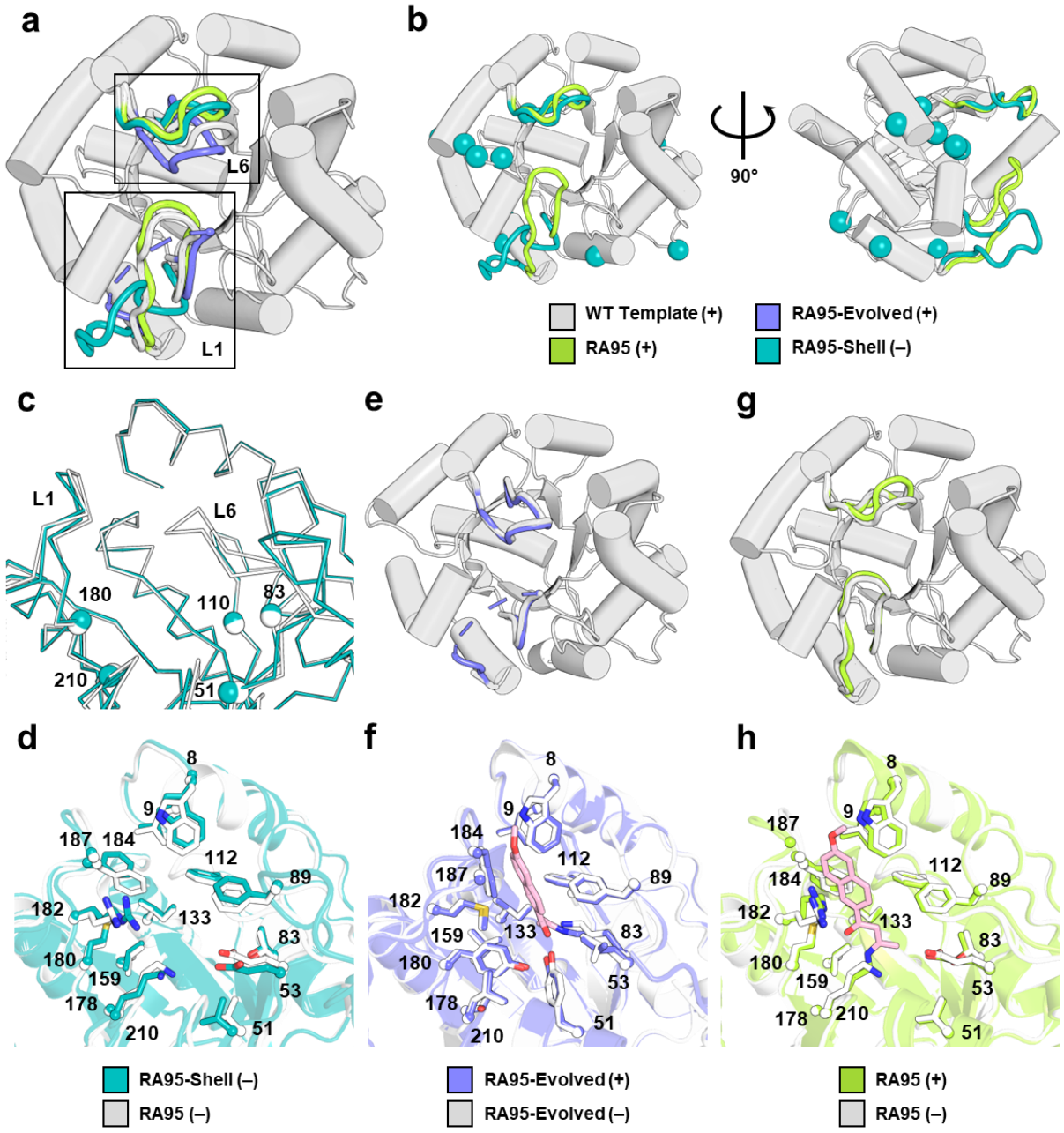

Figure 2.7. Structural impacts of distal mutations. (a) Superposition of crystal structures for the WT template protein (PDB ID: 1A53, grey), inhibitor-bound (+) RA95 (PDB ID: 4A29, green), inhibitor-bound (+) RA95-Evolved (PDB ID: 5AN7, purple), and unbound (–) RA95-Shell (teal). Loops L1 and L6 are indicated and are coloured while a representative RA structure in grey is shown for the remainder of the protein. (b) Distal mutations (shown as teal spheres) are not located on or near L1 or L6. (c) Superposition of ribbon representations of the unbound (–) RA95 and RA95-Shell active sites in white and teal, respectively. Cα carbons of positions 51, 83, 110, 180, and 210 are shown as spheres and loops L1 and L6 are indicated. (d) Superposition of unbound (–) active site structures of RA95 (white) and RA95-Shell (teal). (e) Active site loop preorganization

and (f) active site sidechain preorganization for RA95-Evolved. (g) Active site loop preorganization and (h) active site sidechain preorganization for RA95. In (e) through (h), the inhibitor-bound structure (+) is coloured and is superposed with the unbound structure (–) in grey/white. The mechanism-based diketone inhibitor in the bound structures is shown as pink sticks.

Not only do distal mutations play a role in inducing large conformational changes in flexible active site loops, but they also cause more subtle effects on active site backbone conformation. Comparison of unbound structures of RA95 and RA95-Shell reveals that the addition of distal mutations causes unexpected changes to Cα carbon positions in the active site, despite being far from the positions being mutated (Figure 2.7c). This is most relevant when considering residue positions that played key roles during the evolution of RA95, specifically those of the catalytic tetrad in RA95-Evolved (Tyr51, Asn110, Lys83, Tyr180) and the original catalytic lysine in RA95 (Lys210). Most significantly, the Cα carbons at positions 51 and 210 have shifted 0.8 and 0.7 Å, respectively, when comparing the RA95 and RA95-Shell structures. Given that Tyr51 appeared early in the evolutionary trajectory of RA95, it is possible that the shift in Cα carbon position caused by distal mutations allows it to adopt an orientation conducive to its role in catalysis in RA95-Evolved, and that the shift in position 210 played a role in the preferential enhancement of Lys83 as a more efficient catalyst than Lys210. Positions 83, 110, and 180 also shift by 0.3, 0.2, and 0.4 Å, respectively, with the addition of distal mutations. The effects of these shifts in active site backbone positions can be seen when comparing the overall active site structures of RA95 and RA95-Shell, whose structural differences are reflected in an active site root-mean-square deviation (RMSD) of 1.302 Å (Figure 2.7d). Taken together, these results highlight the subtle effects that distal mutations can have on protein structure and suggest that the evolved active site arrangement may not have fit within the unevolved protein backbone, only being made possible by the addition of distal mutations optimized by directed evolution.

Given the effects of distal mutations on backbone conformation throughout the protein, it is not surprising that they also play a role in preorganization of both the active site and overall protein structure. The inhibitor-bound RA95-Evolved, which contains distal mutations optimized by directed evolution, has backbone and sidechain conformations similar to those of the unbound structure, with overall RMSD of 0.853 Å. The inhibitor-bound and unbound structures of RA95 are less similar, with RMSD of 1.145 Å. This trend extends to the active site, where RA95-Evolved is highly preorganized (Figure 2.7f) with an active site RMSD of 0.565 Å. On the other hand, RA95 has an active site RMSD of 1.355 Å, indicating a decrease in preorganization compared to the evolved variant (Figure 2.7h). Active site loop L6 in RA95-Evolved (Figure 2.7e) also appears to be preorganized, whereas loop L6 in RA95 (Figure 2.7g) must undergo a shift away from the active site to accommodate inhibitor binding. This is reflected in the Cα carbon RMSD of loops L1 and L6, which increases from 0.279 Å in RA95-Evolved to 0.759 Å in RA95. Together, these results suggest that the optimization of distal positions by directed evolution is beneficial for preorganization of the active site and the overall protein.

While X-ray crystal structures clearly demonstrate structural changes induced by distal mutations in RA95, this only provides information on a static structure. To determine the effects of distal mutations on changes in the conformational ensemble of the enzyme, we studied the backbone B-factors of various unbound and inhibitor-bound RA variants. Atomic B-factors, which are a measure of an atom's average displacement in a crystal, can be affected by both conformational heterogeneity and crystalline disorder. Since a number of structure-dependent factors, both chemical (e.g., pH, ligands) and physical (e.g., resolution, crystal packing, temperature) contribute to B-factors, the Z-scores of the atomic B-factors were calculated. The Z-

score analysis, which normalizes the B-factors relative to the mean value in each structure, allows comparison across the crystal structures of different enzyme variants.

B-factor analysis shows that when comparing unbound RA95 and RA95-Evolved structures, loop L6 becomes more rigid through evolution (Figure 2.8), matching the increased active site loop preorganization observed in the static X-ray crystal structure after evolution. This trend is less pronounced in the inhibitor-bound structures, where L6 is only slightly more rigid in the evolved variant. This is expected as L6 interacts directly with the inhibitor in the crystal structure. If the residues surrounding the missing section of loop L1 in the RA95-Evolved crystal structure are not considered, the region of highest flexibility in the inhibitor-bound structures of both RA95 and RA95-Evolved is a helical region made up of residues 217–255 (Figure 2.8b). Interestingly, a hotspot of distal mutations is located on one of the two loops connecting this helix to the β-barrel, containing five of the ten distal mutations in RA95-Shell (A209P, G212D, I213F, S214F, and R216P). Flexibility in this helix appears to be correlated with increased rigidity in loop L6 in the RA structures. Furthermore, while this helix shows increased flexibility in both inhibitor-bound RA structures, its flexibility is not significantly increased in the wild-type (WT) template protein used in the initial computational design of RA95 (Figure 2.8b). These results suggest that its flexibility may be related to RA catalysis.

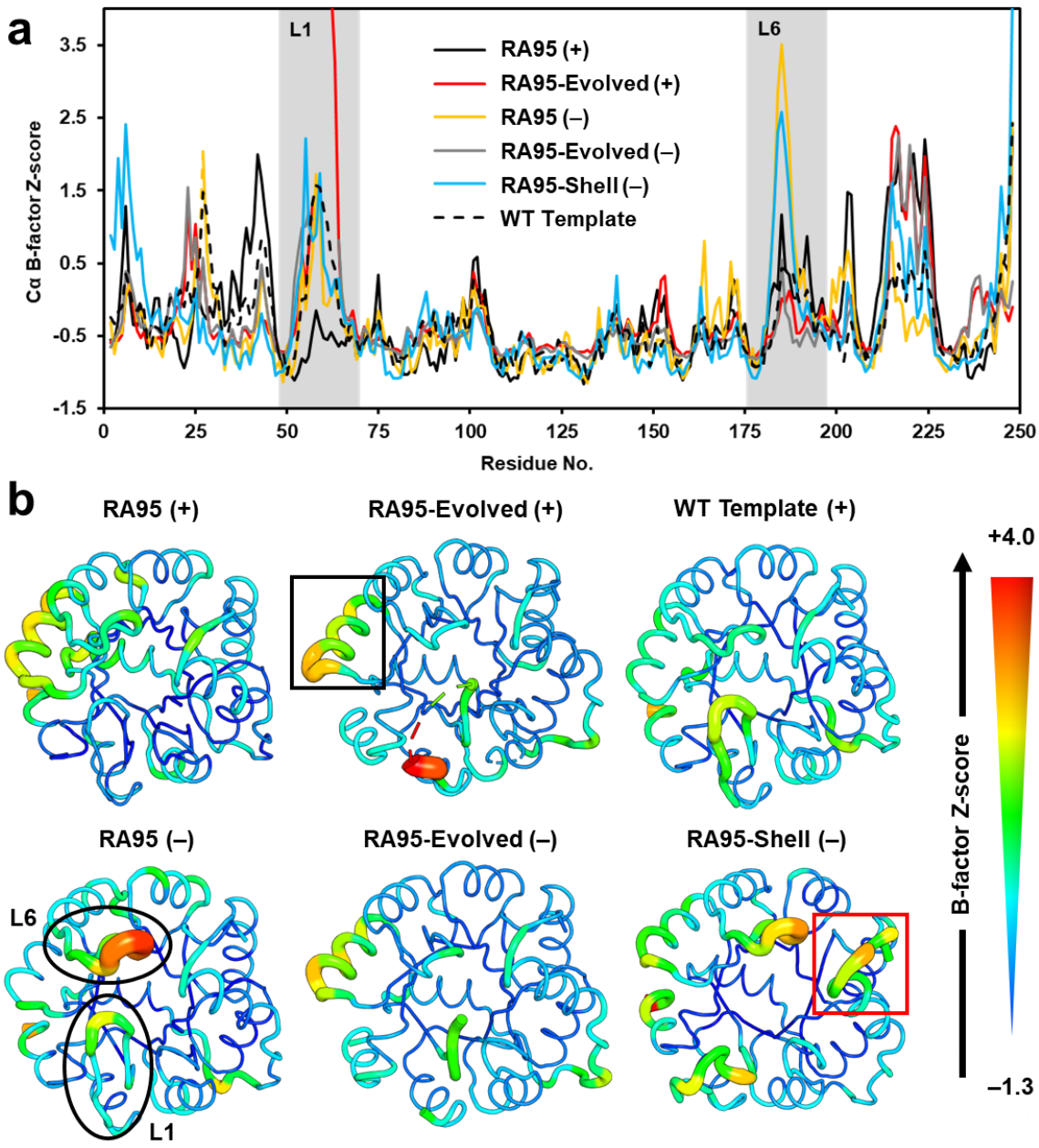

Figure 2.8. Conformational heterogeneity in RA variants. (a) Cα B-factor Z-score as a function of residue position for inhibitor-bound (+) RA95 (PDB ID: 4A29), unbound (–) RA95, inhibitor-bound RA95-Evolved (PDB ID: 5AN7), unbound (–) RA95-Evolved (PDB ID: 5AOU), WT template protein (PDB ID: 1A53), and unbound RA95-Shell. Loops L1 and L6 are indicated. Positive and negative Z-scores represent increased flexibility and rigidity relative to the average residue in the protein, respectively. (b) Backbone atom B-factor Z-score plotted on the crystal structure of each RA variant. Thickness of the putty plot increases with B-factor Z-score, indicating increased flexibility. L1 and L6 are circled on RA95 (–). The helix formed by residues 217–225 and the N-terminal region formed by residues 1–10 are boxed in black on RA95-Evolved (+) and in red on R95-Shell (–), respectively.

As seen in the differences between the conformational ensembles of unbound RA95 and RA95-Shell (Figure 2.8), adding distal mutations to RA95 causes a decrease in flexibility of loop L6, which decreases further in unbound RA95-Evolved. This suggests that interactions involving mutated distal or active site positions may be contributing to its flexibility. Asp212, one of the mutated distal positions at one end of the residue 217–255 helix discussed above, forms a hydrogen bond with a loop L6 residue in both RA95-Shell and RA95-Evolved (Arg182 in RA95-Shell and Lys191 in RA95-Evolved). In RA95 however, Asp212 is replaced with a glycine, and no obvious interactions replace this hydrogen bond. Perhaps these hydrogen bonds contribute to the increased rigidity of loop L6 in the unbound RA95-Shell and RA95-Evolved structures compared to RA95. This could also be a reason for the increased loop conformational preorganization in the evolved variant compared to RA95 (Figure 2.7e, f). Another region whose conformational ensemble is modified by the addition of distal mutations is the N-terminal region made up of residues 1–10, which is more flexible in RA95-Shell than in any other structure, including the evolved variant (Figure 2.8b). This indicates that the addition of active site mutations in the evolved variant rigidifies this area. However, as there are neither any active site nor distal mutations on or near the N-terminus in the protein fold, there are no clear interactions contributing to these changes in flexibility. This highlights the synergistic and sometimes unclear impacts of point mutations on distant areas of a protein structure.

2.3.3. Dynamical effects of mutations

Solving the crystal structure of RA95-Shell revealed that distal mutations contribute to large conformational changes in flexible active site loops in RA95. Subsequently, analysis of the changes to conformational ensembles of RA variants showed that these mutations can also affect

dynamics throughout the protein. To gain a better understanding of the effects of distal and active site mutations on protein conformational dynamics in the RA series, we performed long-time-scale molecular dynamics (MD) simulations on RA95, RA95-Evolved, RA95-Shell, and RA95-Core. Long-time-scale MD has previously been performed on multiple variants along the evolutionary trajectory of RA95 bound to the Schiff base intermediate of the retro-aldol reaction. By evaluating active site conformational dynamics, it was found that distal mutations progressively stabilize the catalytically competent arrangement, causing a population shift towards catalytically active conformational sub-states of the enzyme during evolution.[41] Here, we perform MD on unbound structures to evaluate conformational dynamics in the absence of any bound ligands. The structural differences observed along the MD trajectories were analyzed using principal component analysis (PCA), which is a statistical method used to reduce the dimensionality of complex datasets. It aims to define artificial variables, called principal components (PC), that explain as much of the variation in the data as possible.[74] The first two principal components (PC1 and PC2) describe the most variation and are used to visualize general trends in the data using a standard 2-dimensional scatterplot.

PCA analysis of the MD trajectories based on Cα contacts reveals population shifts in conformational sub-states occurring during directed evolution (Figure 2.9). The greatest variation in the dataset comprising the trajectories of all four enzymes was found to be created by residues 59–62, which fall on loop L1. This loop interconverts between open and closed conformational sub-states, which can be indicated, for example, by the Cα distance between residues 58 and 185 located on loops L1 and L6, respectively (Figure 2.9c). In this way, PC1 is able to distinguish between states with a closed conformation of L1 as seen in the crystal structure of RA95 (those with short distances between loops L1 and L6), from those with an open conformation of L1 as

seen in the crystal structure of RA95-Shell (those with long distances between loops L1 and L6). By comparing the PCA plots of RA95 and RA95-Evolved (Figure 2.9a), it can be seen that evolution flattens the conformational landscape, causing a shift from one major conformational sub-state in RA95, to three distinct populations in RA95-Evolved. This shift increases the accessibility of sub-states where loop L1 is in an open conformation. Thus, the proportion of MD snapshots having a closed loop L1 conformation decreases from 45% in RA95 to 37% in RA95-Evolved (Figure 2.9b). Here, the closed L1 conformation is defined by having a loop distance within two standard deviations of the mean distance of the closed sub-state shown in the PCA plot.

This result is in contrast to a similar analysis reported by Bunzel *et al*.,[75] where long-time-scale MD was performed on designed and evolved variants of a Kemp eliminase series that used the same WT protein scaffold in its initial design calculations as the one used in the design of RA95. It was found that while active site loops L1 and L6 similarly interconverted between open and closed sub-states, the populations of these states were shifted towards the closed state through evolution as opposed to the open state. However, these MD simulations were performed on structures bound to the TS of the Kemp elimination reaction. While loop closure is essential for active site desolvation during catalysis, resulting in an evolutionary preference for the closed state when bound to the TS as reported by Bunzel *et al*.,[75] it is likely that loop opening facilitates ligand binding. Thus, open states may be preferred in the unbound enzyme to promote substrate binding, resulting in a population shift towards the open state during evolution as seen in the unbound RA structures. Additionally, the retro-aldol reaction substrate and TS are much larger than those of the Kemp elimination reaction. This may be reflected in the preference for the open state in RA variants after evolution, as the larger substrate requires more space to enter the active site.

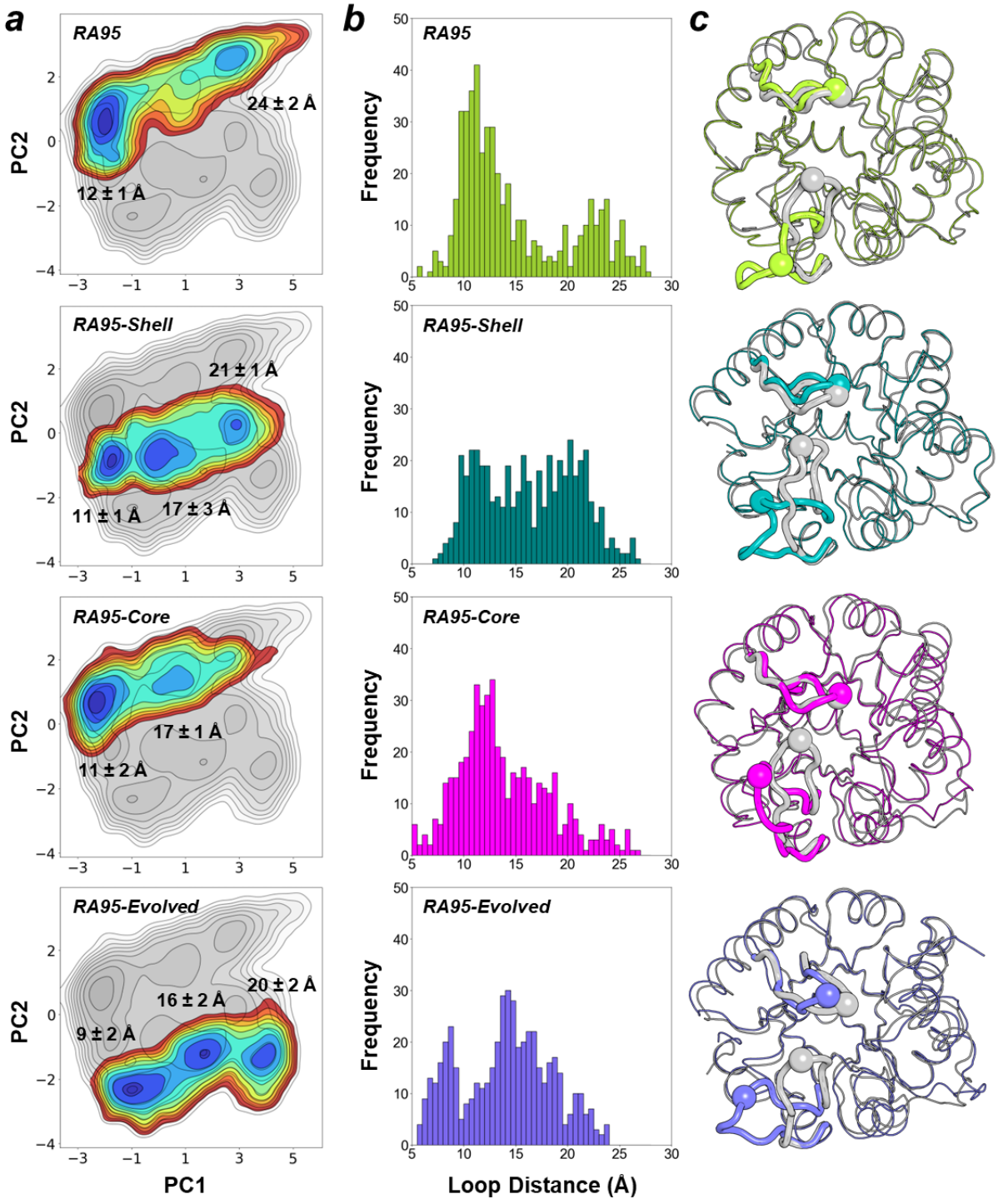

Figure 2.9. Representation of MD trajectories for the designed, core, shell, and evolved apo enzymes. (a) Trajectories projected into the two most important principal components (PC1 and PC2) based on Cα contacts. For each substate, the mean and standard deviation (in Å) of the distance between the Cα carbons of residues 58 and 185 are presented for 200 structures within the substate. PC1 differentiates structures with closed active site loops (low PC1 values, smaller

distances between residues 58 and 185) from those with open active site loops (higher PC1 values, larger distances between residues 58 and 185). (b) Histograms showing the distance between the two active site loops L1 and L6, described by the distance between Cα carbons of residues 58 and 185, of 500 structures along each MD trajectory. (c) An overlay of a representative structure from the most closed substate (grey) and the most open substate (coloured) for each trajectory. The Cα carbons of residues 58 and 185 are show as spheres. Active site loops L1 and L6 are shown as thicker regions of the cartoon structure.

In addition to open conformations of loop L1 being more prevalent after evolution, it is evident that this shift is caused by the addition of distal mutations. Upon the addition of distal mutations to RA95 (to create RA95-Shell) or RA95-Core (to create RA95-Evolved), the conformational landscape broadens, increasing the accessibility of sub-states where loop L1 is open or partially open (Figure 2.9a). 45% of MD snapshots are in the closed conformation in RA95, which decreases to only 28% in RA95-Shell. Similarly, the proportion of snapshots in the closed conformation decreases from 62% in RA95-Core to 37% in RA95-Evolved (Figure 2.9b). On the other hand, the addition of active site mutations to RA95 (to create RA95-Core) or RA95-Shell (to create RA95-Evolved) causes population shifts towards more closed loop L1 conformations. In the case of RA95, active site mutations virtually eliminate the open conformation (loop distance of 24 ± 2 Å) but create a new sub-state where loop L1 is in a partially open conformation (loop distance of 17 ± 1 Å) in RA95-Core (Figure 2.9a). These results demonstrate the effects that distal mutations can have on enzyme conformational dynamics, causing population shifts on the conformational landscape that favour sub-states that are likely more conducive to catalysis.

2.3.4. Structure recapitulation by ensemble-based design

We showed that distal mutations introduced during directed evolution of RA95 likely contribute to increased catalytic activity through structural and dynamical changes to the enzyme.

However, as the impacts of distal mutations are difficult to predict, it would be beneficial if highly efficient active sites could be created *de novo* without requiring distal mutations. Given that RA95-Core possesses an activity 3,600-fold higher that RA95 and only 6-fold lower than RA95-Evolved through the addition of only active site mutations, we investigated whether this active site could have been computationally predicted during *de novo* design, and more importantly, whether it could have been predicted instead of the RA95 active site. First, positive control calculations were performed in which the active site sequences of RA95 and RA95-Evolved were computationally optimized on their respective inhibitor-bound crystal structure backbones. In the case of RA95, the theozyme was composed of the catalytic lysine Lys210 and the TS for the carbon–carbon bond-breaking step of the reaction. This resulted in RA95 models with predicted rotameric configurations in good agreement with the crystal structure using both the (*R*)- and (*S*)-configurations of the TS (Figure 2.10a, b), the only residue with a different rotamer being Arg182. This can be attributed to structural differences between the inhibitor in the crystal structure and the TS in the design model, which has one additional chiral carbon relative to the inhibitor (Figure 2.11). As a result, the TS does not align perfectly with the inhibitor and thus clashes with the Arg182 sidechain location from the crystal structure.

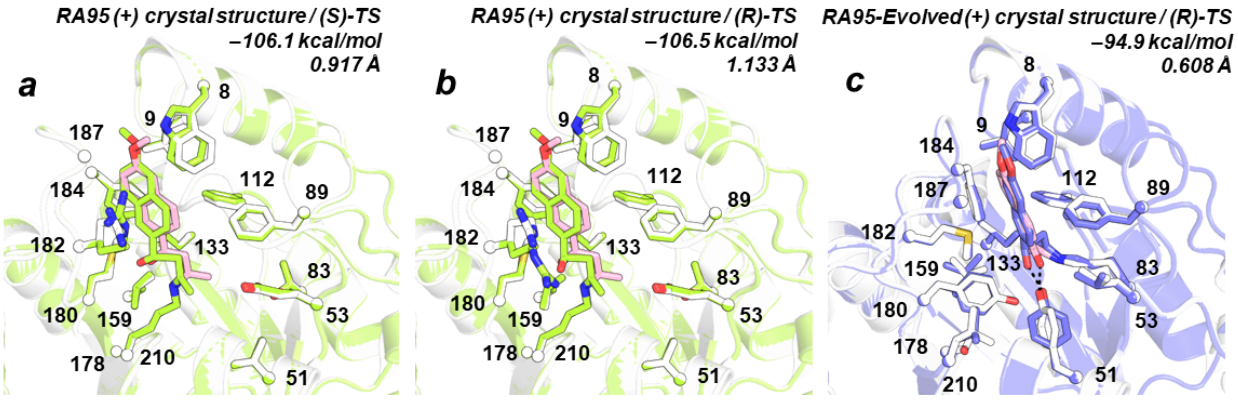

Figure 2.10. Computational recapitulation of RA95 and RA95-Evolved active sites on design templates derived from crystal structures. The RA95 crystal structure (white) with bound diketone inhibitor, 1-(6-methoxy-2-naphthalenyl)-1,3-butanedione (pink), is overlaid with RA95 design models (green) obtained using the (a) (*S*)- or (b) (*R*)-configuration of the TS and the inhibited (+) RA95 crystal structure (PDB ID: 4A29) as the design template. (c) Similarly, the RA95-Evolved crystal structure (white) with inhibitor (pink) is overlaid with the RA95-Evolved design model (purple) obtained using the (*R*)-configuration of the TS and the inhibited (+) RA95-Evolved crystal structure (PDB ID: 5AN7) as the design template. PHOENIX energies of design models after active site repacking are provided in kcal/mol, as well as active site sidechain atom RMSD from the overlaid crystal structure in Å. Sidechains of all designed residues are shown, with alpha carbons represented by spheres. The key catalytic contact between the TS and Tyr51 is shown in black dashed lines.

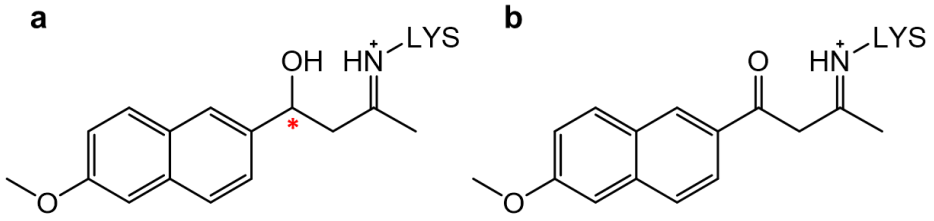

Figure 2.11. The TS contains a chiral carbon missing from the diketone inhibitor. (a) TS for the carbon–carbon bond-breaking step of the RA reaction mechanism. (b) Structure of the covalent diketone inhibitor, 1-(6-methoxy-2-naphthalenyl)-1,3-butanedione, used during crystallization of RA variants. The chiral carbon present only in the TS is highlighted with a red asterisk.

In the case of RA95-Evolved, the catalytic tetrad had to be considered when constructing the theozyme to be used in design calculations. While mutagenesis experiments showed that all four residues in the tetrad contribute to increased catalysis, the inclusion of all four in the catalytic motif would lower the chances of successfully placing the theozyme in the template backbone. Given that reversion of Tyr51 was shown to cause the greatest reduction in activity and that pK_a does not appear to be affected by reversion of this residue (Figure 2.3), only Tyr51 was chosen to be included in the theozyme, along with the catalytic lysine Lys83. The positive control calculations yielded a computational model of RA95-Evolved in excellent agreement with the crystal structure using the (*R*)-configuration of the TS (Figure 2.10c), but the theozyme with the (*S*)-configuration could not be placed on the backbone. Importantly, the remaining two catalytic tetrad residues, Asn110 and Tyr180, are accurately predicted with the (*R*)-TS, despite not being specifically defined in the calculations. These results match the experimentally determined enantioselectivities of RA95 and RA95-Evolved, which are approximately equal for both enantiomers in RA95 (2.5:1 for the (*S*)-enantiomer) and highly selective for the (*R*)-enantiomer in RA95-Evolved (480:1). This is encouraging, as it indicates that the computational protocol can predict enantioselectivity in RA95-Evolved. Taken together, these control calculations demonstrate that the chosen combination of rotamer library, energy function, search algorithm, and theozyme are sufficient for the accurate prediction of both RA95 and RA95-Evolved if the correct backbone template is provided.

On the other hand, when the crystal structure backbone templates are replaced by the *S. solfataricus* indole-3-glycerolphosphate synthase backbone, which was the WT template protein used in the design of RA95 (PDB ID: 1A53), neither the RA95 nor the RA95-Evolved active sites could be placed on the backbone template. This was largely due to steric clashes between the TS

and the backbone of loop L6 in the case of RA95, or the TS and the backbone of loop L1 in the case of RA95-Evolved. This indicates that not only is the 1A53 backbone template not well-suited to accommodate the RA95-Evolved active site sequence, but it is also not suited for the accurate prediction of the RA95 active site, and this is due to conformational changes of flexible active site loops. This is expected for RA95-Evolved, as the original calculations performed in the design of RA95 did not identify any sequences with the catalytic lysine at position 83.[27] In fact, the inability to accurately recapitulate the RA95 active site on 1A53 may also be expected, as similar structural inaccuracies are seen when comparing the RA95 crystal structure with the original design model. Compared to the design model TS, the naphthyl group of the crystallized inhibitor is rotated 114º around its long axis (Figure 1.10).[40] While this could be due in part to structural differences between the TS and inhibitor, it also reflects the conformational differences of loops L1 and L6 in the crystal structure compared to the design model.

It is evident from these results that the use of a single fixed backbone template in the design calculations is not sufficient to accurately model the conformational changes undergone by the protein backbone during evolution. In attempt to address this approximation, we employed a method known as ensemble refinement, which is the use of molecular dynamics simulations restrained by crystallographic diffraction data to generate an ensemble of protein backbones.[76] In a previous study by Broom and Rakotoharisoa *et al.*,[73] it was shown that the use of a backbone ensemble derived from ensemble refinement in design calculations can improve the structural prediction accuracy for an engineered Kemp eliminase, HG4. Optimizing the RA95 and RA95-Evolved sequences on an ensemble of backbones derived from the 1A53 diffraction data showed that the theozymes could be placed on a number of the ensemble members. However, the subsequent active site repacking step resulted in steric clashes involving the TS and several amino

acid sidechains, resulting in energy scores greater than 0 kcal/mol in all cases (Figure 2.12a, b, c). This signifies that even when the active site loops adopt conformations that allow theozyme placement, subtle changes to the overall backbone configuration prevent the recapitulation of RA95 and RA95-Evolved active site structures on 1A53.

Use of the inhibitor-bound and unbound RA95 crystal structures as the design template also did not allow accurate recapitulation of the RA95-Evolved active site, even when using ensembles of backbones derived from diffraction data. The unbound RA95 ensemble allowed theozyme placement but was severely destabilized by repacking (Figure 2.12d), while the inhibitor-bound ensemble did not allow theozyme placement on any ensemble member. This is in contrast with results from Broom and Rakotoharisoa *et al.*,[73] where accurate prediction of the highly active Kemp eliminase variant HG4 was achieved using an ensemble of backbones derived from a lower activity variant, but not from the original design template. This is because the use of a backbone ensemble derived from crystallographic data of a low-activity variant likely allows sampling of catalytically competent backbone sub-states during computational design. Based on this, we would expect recapitulation of RA95-Evolved to be possible on an ensemble of backbones derived from RA95, but not from 1A53. However, recapitulation of RA95-Evolved is not possible on either backbone ensemble, likely indicating that catalytically competent sub-states of RA95 are not catalytically competent in the context of the evolved active site. This is expected given the significant active site remodelling that occurred during RA95 evolution, which is not observed during the evolution of most other *de novo* enzymes including the HG series of Kemp eliminases.

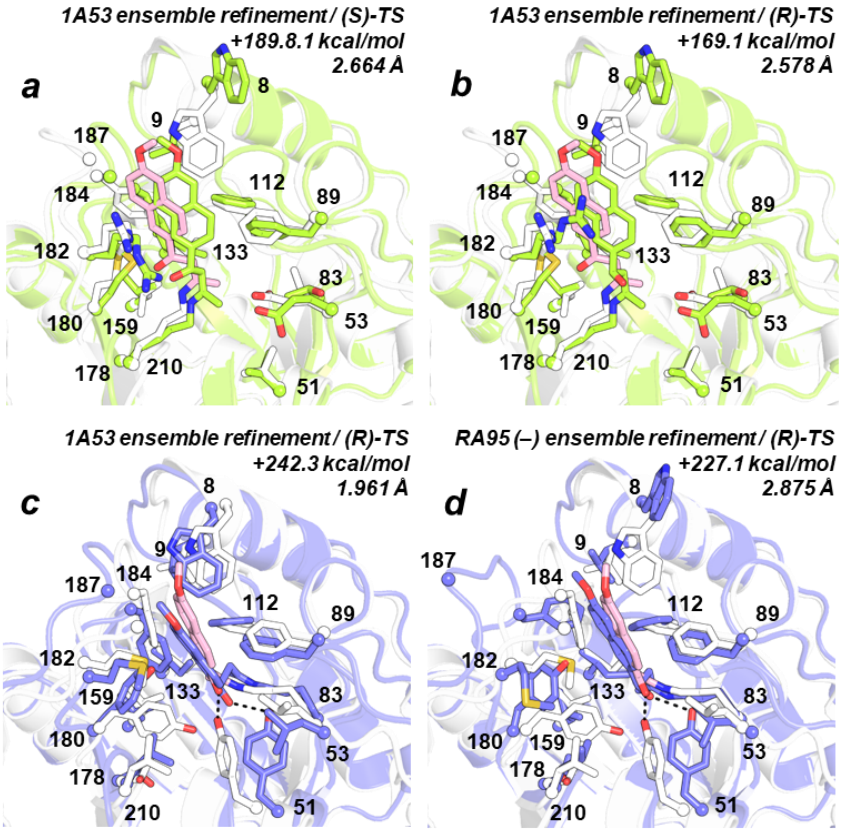

Figure 2.12. Computational recapitulation of RA95 and RA95-Evolved active sites on design templates derived from ensemble refinement. The RA95 crystal structure (white) with bound diketone inhibitor, 1-(6-methoxy-2-naphthalenyl)-1,3-butanedione (pink), is overlaid with RA95 design models (green) obtained using the (a) (*S*)- or (b) (*R*)-configuration of the TS and indole-3-glycerolphosphate synthase from *S. solfataricus* (PDB ID: 1A53) as the design template. Similarly, the RA95-Evolved crystal structure (white) with inhibitor (pink) is overlaid with the RA95-Evolved design model (purple) obtained using the (*R*)-configuration of the TS and (c) indole-3-glycerolphosphate synthase from *S. solfataricus* (PDB ID: 1A53) or (d) uninhibited (–) RA95 as the design template. In each case, the design model shown was obtained using the ensemble refinement-derived template that gave the best energy after active site repacking. PHOENIX energies of design models after active site repacking are provided in kcal/mol, as well as active site sidechain atom RMSD from the overlaid crystal structure in Å. Sidechains of all designed residues are shown, with alpha carbons represented by spheres. The key catalytic contact between the TS and Tyr51 is shown in black dashed lines.

Evidently, the backbone flexibility allowed by ensemble refinement is not sufficient to account for the variations in backbone configuration accumulating during evolution of RA95. To investigate whether removing the restraints imposed by crystallographic diffraction data would allow more accurate prediction of the RA active sites, we generated backbone ensembles from unconstrained MD starting from the 1A53 and the unbound RA95 crystal structures. Optimizing the RA95 (Figure 2.13a, b) and RA95-Evolved (Figure 2.13c) sequences on the MD-derived 1A53 ensemble yielded stable models, evidenced by negative energy scores. However, the active site sidechain RMSD values exceeding 2 Å in all cases indicate that these models are highly structurally inaccurate. Similar results were seen for the prediction of RA95-Evolved on the MD-derived unbound RA95 ensemble (Figure 2.13d). This shows that while removing the MD constraints imposed by crystallographic diffraction data allowed remodelling of the backbones in ways that could stabilize the RA95 and RA95-Evolved sequences, the resulting backbone configurations are clearly not good representations of those present in the crystal structures. Overall, these results suggest that backbone and loop remodelling that occurred during evolution of RA95 prevents accurate recapitulation of the RA95-Evolved active site on the original design template, indicating that the evolved active site could not have been designed *de novo* using current computational methods.

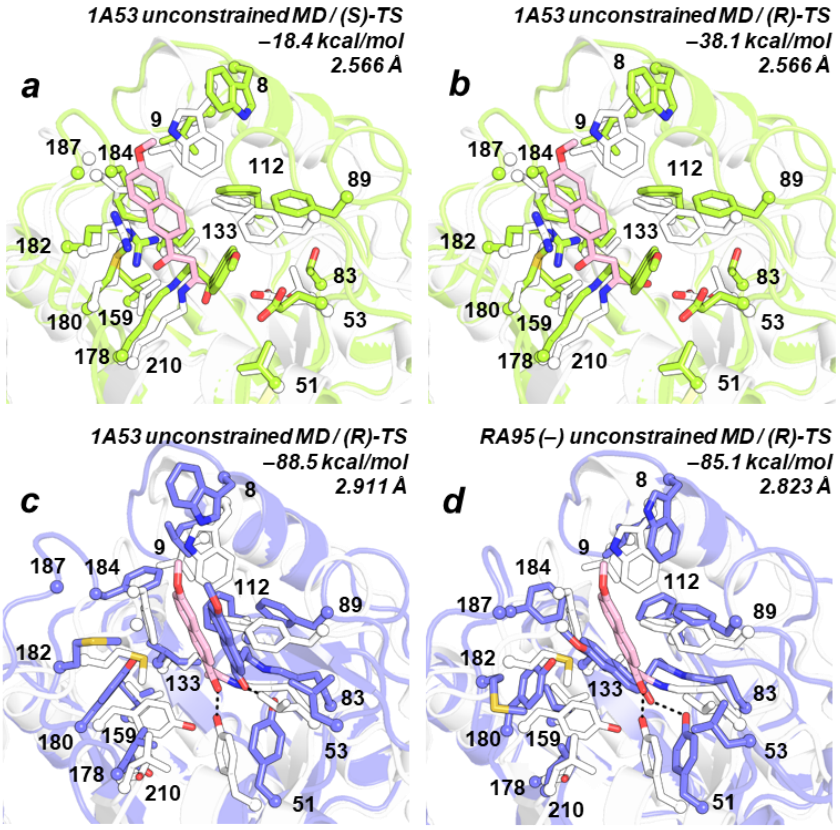

Figure 2.13. Computational recapitulation of RA95 and RA95-Evolved active sites on design templates derived from unconstrained molecular dynamics (MD) simulations. The RA95 crystal structure (white) with bound diketone inhibitor, 1-(6-methoxy-2-naphthalenyl)-1,3-butanedione (pink), is overlaid with RA95 design models (green) obtained using the (a) (*S*)- or (b) (*R*)-configuration of the TS and indole-3-glycerolphosphate synthase from *S. solfataricus* (PDB ID: 1A53) as the design template. Similarly, the RA95-Evolved crystal structure (white) with inhibitor (pink) is overlaid with the RA95-Evolved design model (purple) obtained using the (*R*)-configuration of the TS and (c) indole-3-glycerolphosphate synthase from *S. solfataricus* (PDB ID: 1A53) or (d) uninhibited (–) RA95 as the design template. In each case, the design model shown was obtained using the MD-derived template that gave the best energy after active site repacking. PHOENIX energies of design models after active site repacking are provided in kcal/mol, as well as active site sidechain atom RMSD from the overlaid crystal structure in Å. Sidechains of all designed residues are shown, with alpha carbons represented by spheres. The key catalytic contact between the TS and Tyr51 is shown in black dashed lines.

2.4. Discussion

In this study, we evaluated the impacts of active site and distal mutations introduced by directed evolution on the structure, function, and dynamics of the *de novo* enzyme RA95. In

contrast to most other cases of directed evolution of *de novo* designed enzymes, where activity is increased through subtle structural changes such as widening of the active site[23] or elimination of unproductive binding poses,[25] RA95 undergoes significant active site remodelling over the course of evolution, making the RA95 series a unique case study in computational enzyme design. While numerous studies have investigated the determinants of catalysis within the RA95 series active sites, most notably through the identification of the evolved catalytic tetrad, few have studied the impacts of distal mutations alone. Compared to the arguably clear activity-enhancing effects of active site mutations, as evidenced by the 3,600-fold activity increase of RA95-Core relative to RA95, distal mutations can have more subtle impacts. Here, we showed that distal mutations have synergistic impacts on the structure and dynamics of RA95, causing population shifts within the conformational landscape and enhancing sub-states that are more conducive to catalysis. In this case, the changes to backbone and active site loop configurations created by distal mutations during evolution are large enough to prevent computational prediction of the evolved active site on the original design template protein.

Despite our inability to accurately recapitulate the RA95-Evolved active site, the successful engineering of the highly active RA95-Core variant, created by mutating only active site positions in RA95, suggests that efficient artificial enzymes could be designed *de novo* if computational design procedures were improved. However, it is unlikely that the evolved active site could have been predicted on 1A53 using current design methods, as they do not account for the structural changes in RA95-Evolved induced by distal mutations. Indeed, Althoff, Wang, and Jiang *et al.*[27] did not identify any sequences with the catalytic lysine at position 83 in their original design calculations. Additionally, despite one catalytic motif tested by Jiang and Althoff *et al.*[26] comprising a catalytic lysine and a tyrosine acting as a general base, analogous to the motif used

in the recapitulation of RA95-Evolved, no active designs resulted from calculations with this motif. While Tyr51 was introduced early in the evolutionary trajectory of RA95 during mutagenesis of only active site residues, it was not able to be oriented for catalysis until distal mutations had been introduced, likely remodelling the backbone in the process. These findings support the fact that the 1A53 backbone configuration is not compatible with the catalytic motif found in the RA95-Evolved variant.

Results from this study suggest avenues for improvement of current computational design methodologies. Current methods lack protocols for incorporating backbone conformational changes that mimic those induced by distal mutations during evolution, which is demonstrated by the inability to predict the RA95-Evolved active site structure on the 1A53 backbone template. To design an artificial enzyme with a desired active site, a backbone configuration that is compatible with that active site must be allowed. While ensembles of backbones generated from crystallographic data have previously been used to reach this goal,[73] this method was not sufficient to approximate the conformational changes occurring during evolution of RA95. Other ways of achieving sufficient backbone remodelling should be considered, including alternative methods for the generation of backbone ensembles to be used in design calculations, as well as flexible-backbone design algorithms to be incorporated into the design procedure.[77] For example, backrub movements[78] and kinematic closure (KIC) algorithms,[79] the latter shown to accurately model loop conformations, have both been used to treat backbone flexibility in the past. Recent advances in machine learning methods may also prove useful in designing custom protein backbones for design when remodelling of existing backbones is not sufficient. It is likely that remodelling of design templates will be required for the design of highly efficient *de novo* biocatalysts, and we expect that the RA95 series of retro-aldolases could aid in benchmarking these remodelling efforts.

2.5. Materials and Methods

2.5.1. Protein expression and purification

Codon-optimized and His-tagged (C-terminus) RA genes (Supplementary Table 3) cloned into the pET-29b(+) vector via *Nde*I and *Xho*I were obtained from Twist Bioscience. Enzymes were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth (LB) supplemented with 50 µg/mL kanamycin. Cultures were grown at 37ºC with shaking (220 rpm) to an optical density of approximately 0.6 at 600 nm. Protein expression was then induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and cells were incubated for 16 hours at 18°C with shaking (220 rpm). Cells were harvested by centrifugation, resuspended in 10 mL lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer, pH 8.0), and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography using Ni–NTA agarose (Qiagen) pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Columns were washed twice with lysis buffer. Bound proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and exchanged into 25 mM HEPES (pH 7.5) supplemented with 100 mM NaCl using Econo-Pac 10DG desalting pre-packed gravity-flow columns (Bio-Rad). Purified samples were quantified by measuring the absorbance at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPAsy ProtParam tool (https://web.expasy.org/protparam/).

2.5.2. Steady-state kinetics

Steady-state kinetic assays were carried out at 29ºC in 25 mM HEPES (pH 7.5) supplemented with 100 mM NaCl. Triplicate 200 µL reactions with varying concentrations of

freshly-prepared racemic 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) (Achemica) dissolved in acetonitrile (2.7% final concentration) were initiated by the addition of ~180 μM RA95, 0.1 μM RA95-Evolved, 2 μM RA95-Core, 120 μM RA95-Shell, 1 μM RA95-Evolved-Y51F/RA95-Evolved-N110S-Y180F, 0.2 μM RA95-Evolved-N110S/RA95-Evolved-Y180F, or 15 μM RA95-Evolved-Y51F-N110S/RA95-Evolved-Y51F-Y180F. Product (6-methoxy-2-napthaldehyde) formation was monitored spectrophotometrically at 350 nm (ε = 5,970 M^{-1} cm^{-1})[40] in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax 384Plus plate reader (Molecular Devices). Path lengths for each well were calculated ratiometrically using the difference in absorbance of 25 mM HEPES (pH 7.5) supplemented with 100 mM NaCl and 2.7% acetonitrile at 900 and 975 nm (27°C).[80] Linear phases of the kinetic traces were used to measure initial reaction rates. k_{cat} and K_M were determined by fitting the data to the Michaelis-Menten model shown in Equation 1, where v_0 is the reaction rate in [M/s], [S] is the substrate concentration in [M], [E_0] is the enzyme concentration in [M], k_{cat} is the turnover number in [s^{-1}], and K_M is the Michaelis constant in [M].

$$v_0 = \frac{k_{cat}[E_0][S]}{K_M + [S]} \qquad \textbf{Equation 1}$$

2.5.3. pH-rate profile determination

Steady-state kinetic assays for pH-rate profile determination were carried out at 29ºC in Britton-Robinson buffer (0.04 M boric acid, 0.04 M phosphoric acid, 0.04 M acetic acid) at varying pH values. Triplicate 200 μL reactions with varying concentrations of freshly-prepared racemic 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (methodol) (Achemica) dissolved in acetonitrile (2.7% final concentration) were initiated by the addition of the enzyme. Product (6-methoxy-2-

napthaldehyde) formation was monitored spectrophotometrically at 350 nm ($\varepsilon = 5{,}970\ M^{-1}\ cm^{-1}$)[40] in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax 384Plus plate reader (Molecular Devices). Path lengths for each well were calculated ratiometrically using the difference in absorbance of the Britton-Robinson buffer supplemented with 2.7% acetonitrile at 900 and 975 nm (27°C)[80]. Linear phases of the kinetic traces were used to measure initial reaction rates. Initial reaction rate data were fitted to the linear portion ([S] ≪ K_M) of the Michaelis-Menten model shown in Equation 2 and k_{cat}/K_M was deduced from the slope. k_{cat}/K_M data were fitted to Equation 3 using nonlinear least squares regression, where $(k_{cat}/K_M)_{obs}$ is the observed catalytic efficiency at a given pH, $(k_{cat}/K_M)_{max}$ is the maximum catalytic efficiency on the pH-rate profile, p*K*a1 is the first inflection point on the pH-rate profile, and p*K*a2 is the second inflection point on the pH-rate profile.

$$v_0 = \frac{k_{cat}[E_0][S]}{K_M} \qquad \textbf{Equation 2}$$

$$\left(\frac{k_{cat}}{K_M}\right)_{obs} = \frac{\left(\frac{k_{cat}}{K_M}\right)_{max}}{1 + 10^{pKa1-pH} + 10^{pKa2-pH}} \qquad \textbf{Equation 3}$$

2.5.4. Circular dichroism and thermal denaturation assays

CD measurements were performed using a Jasco J-815 spectrometer using 300 µL aliquots of each RA sample at a concentration of approximately 5 µM in 10 mM sodium phosphate buffer (pH 7.0) in a 1-mm path-length quartz cuvette (Jasco). For structural characterization of protein folds, CD spectra were acquired from 200 to 250 nm at 20°C, sampled every 1 nm at a rate of 10 nm/min. Three scans were acquired and averaged for each sample. For thermal denaturation assays, samples were heated at a rate of 1°C per minute, and ellipticity at 222 nm was measured

every 0.2°C. Melting temperatures were determined by fitting the CD data to a two-state transition model with correction for pre- and post-transition linear changes in ellipticity as a function of temperature.[81] Data were fitted to Equation 4 through Equation 6 using nonlinear least-squares regression, where T is the temperature in degrees Kelvin, Tm is the melting temperature in degrees Kelvin, θ_F is the ellipticity when 100% folded, θ_U is the ellipticity when 100% unfolded, c_F is the linear correction for pre-transition changes in ellipticity, c_U is the linear correction for post-transition changes in ellipticity, ΔH_U is the enthalpy of unfolding in [cal/mol], k is the folding constant, F is the fraction folded, U is the fraction unfolded, and θ is the ellipticity at temperature T.

$$\theta = F\,[\theta_F + c_F T - \theta_U - c_U T] + \theta_U + c_U T \qquad \textbf{Equation 4}$$

$$k = \exp\left[\left(\frac{\Delta H_U}{1.987\ T}\right)\left(\frac{T}{Tm} - 1\right)\right] \qquad \textbf{Equation 5}$$

$$F = \frac{k}{1 + k} \qquad \textbf{Equation 6}$$

2.5.5. Crystallization

Purified protein samples to be used for crystallography were further subjected to purification by gel filtration in 20 mM potassium phosphate buffer (pH 7.4) and 50 mM NaCl using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad). Samples were concentrated to 10–15 mg/mL using Amicon Ultracel-3K centrifugal filter units (EMD Millipore) and quantified by measuring the absorbance at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPAsy ProtParam tool (https://web.expasy.org/protparam/). Crystals were obtained by the hanging-drop vapour diffusion

method at 293 K in drops prepared by mixing 1 μL of protein solution with 1 μL of the mother liquor and sealing the drop inside a reservoir containing an additional 500 μL of the mother liquor solution. The mother liquor solution that yielded the crystals of RA95 (RA95) used for X-ray data collection contained 0.1 M sodium acetate (pH 5.2) and 3.1 M NaCl with a protein solution concentration of 7 mg/mL. The mother liquor solution that yielded the crystals of RA95-Shell used for X-ray data collection contained 0.1 M sodium acetate (pH 4.4) and 19% wt/vol PEG 3,000 with a protein solution concentration of 6 mg/mL.

2.5.6. X-ray data collection and processing

Crystals were mounted in polyimide loops and sealed using a MicroRT tubing kit (MiTeGen). Single-crystal X-ray diffraction data were collected on beamline 8.3.1 at the Advanced Light Source. The beamline was equipped with a Pilatus3 S 6M detector (Dectris) and was operated at a photon energy of 11111 eV. Crystals were maintained at 277 K throughout the course of data collection. Multiple data sets were collected for each variant either from different crystals or from different regions of larger crystals. X-ray data were processed using the Xia2 software,[82] which performed indexing, integration, and scaling with XDS and XSCALE,[83] followed by merging with Pointless.[84]

2.5.7. Structure determination

Initial phase information for calculation of electron density maps was obtained by molecular replacement using the program Phaser,[85] as implemented in the PHENIX suite.[86] Several different RA-series enzymes were used as molecular replacement search models. All members of the RA-series of enzymes crystallized in the same crystal form, containing one copy of the

molecule in the crystallographic asymmetric unit. Next, we performed iterative steps of manual model rebuilding followed by refinement of atomic positions, atomic displacement parameters, and occupancies using a translation-libration-screw (TLS) model, a riding hydrogen model, and automatic weight optimization. All model building was performed using Coot[87] and refinement steps were performed with *phenix.refine*[88] within the PHENIX suite. Further information regarding model building and refinement are presented in Table 2.2. Time-averaged ensembles were generated with *phenix.ensemble_refinement*[76] implemented in PHENIX. To prepare the structures for ensemble refinement, low-occupancy conformers were removed, and occupancies adjusted to 100% using *phenix.pdbtools*. Hydrogen atoms were then added using *phenix.ready_set*. This procedure yielded ensembles containing 125, 72, 100, and 72 templates from the crystal structures of inhibitor-bound RA95 (PDB ID: 4A29), WT template protein (PDB ID: 1A53), unbound RA95, and unbound RA95-Shell, respectively.

2.5.8. Unconstrained molecular dynamics

Long-time-scale MD simulations in explicit water were performed using the AMBER 20 package[89] using the Amber 99SB force field (ff99SB).[90] Unbound crystal structures of RA95 (RA95) and RA95-Shell were used for MD. The unbound crystal structure of RA95-Evolved (PDB ID: 5AOU) was used with missing residues modelled using MODELLER.[91] The unbound structure of RA95-5.8F-Core (RA95-Core) was generated from the unbound crystal structure of RA95 by introducing the active site mutations using the TRIAD protein design software (Protabit LLC, Pasadena, CA). Amino acid protonation states were predicted using the H++ server (http://biophysics.cs.vt.edu/H++). Prior to molecular dynamics, the structures were parameterized using the LEaP program from the amber suite (http://ambermd.org/). The protein was surrounded

in a cubic box of explicit TIP3P[92] water molecules, and the electrostatic charge neutralized by addition of Na+ and Cl− counter-ions. Long-range electrostatic effects were considered using the particle mesh Ewald method,[93] and water molecules were treated using the SHAKE algorithm.[94] The Langevin equilibration scheme was used to control and equalize the temperature with a 2 fs time step at a constant pressure of 1 atm and temperature of 300 K, and an 8 Å cutoff was applied to Lennard-Jones and electrostatic interactions. Production trajectories were run for 1000 ns and were analyzed using the *cpptraj* module included in the AMBER 20 package. PCA analysis was done with the *pyEMMA* software.[95] Ensembles of MD-derived backbone templates were created using snapshots taken at a chosen interval such that the ensemble sizes matched those created by ensemble refinement described above. This resulted in ensembles containing 72 and 100 templates from the MD trajectories of WT template protein (PDB ID: 1A53) and unbound RA95, respectively.

2.5.9. Computational enzyme design

All calculations were performed with the TRIAD protein design software (Protabit, Pasadena, CA, USA) using a Monte Carlo[96] with simulated annealing search algorithm for rotamer optimization. The crystal structures of the WT template used for RA95 design (PDB ID: 1A53), RA95 bound to inhibitor (PDB ID: 4A29), and RA95-Evolved bound to inhibitor (PDB ID: 5AN7) were obtained from the Protein Data Bank (PDB)[97]. Structures of unbound RA95, unbound RA95-Shell, and ensembles of templates were obtained from the refinement of crystallographic data or from snapshots taken from long-time-scale MD simulations as described above. Hydrogen atoms were added using the *addH.py* application in TRIAD. The TS structure for the RA carbon–carbon bond-breaking step was built from the TS lowest energy QM model provided in the Supplementary

Materials of Althoff, Wang, and Jiang *et al*..[27] Residue positions surrounding the TS (Table 2.3). were mutated to Gly during ligand placement. The 2002 Dunbrack backbone-independent rotamer library[10] with expansions of ± 1 standard deviation around chi 1 and chi 2 was used to provide sidechain conformations, with the exception of the catalytic lysine position. A library of lysine–TS conformations was generated by allowing sampling of the conformations provided in Table 2.4. TS pose energies were calculated using the PHOENIX energy function,[24] which consists of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field[98] with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal/mol and an equilibrium donor-acceptor distance of 2.8 Å,[99] an electrostatic energy term modeled using Coulomb's law with a distance-dependent dielectric of 10, an occlusion-based solvation potential with scale factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial,[100] and a secondary structural propensity term.[101] During the energy calculation step, TS–tyrosine interaction energies in the case of RA95-Evolved were biased to favor interactions that satisfy contact geometries listed in Table 2.5. These energy calculation and biasing steps are discussed by Lassila *et al.*.[19]

Following ligand placement, the lowest energy TS pose found on each template was selected as a starting point for repacking of the RA95 or RA95-Evolved sequence. Residues that were converted to Gly in the ligand placement step were allowed to sample all conformations of the amino acid found at that position in the sequence being used for repacking. The identities and conformations of the catalytic residues were fixed, as well as the conformation of the TS. Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above. The single lowest energy repacked structure on each backbone template was used for analysis. To compare energies of models obtained on the various templates, the energy

difference between each repacked structure and the corresponding all-Gly structure obtained during ligand placement was calculated, and these energies are reported throughout the figures and text.

Table 2.3. Residue positions optimized during active site recapitulation.

Enzyme	Ligand placement[a]	Repacking[b]
RA95	8, 9, 51, 53, 83, 89, 112, 133, 159, 178, 180, 182, 184, 187	W8, L9, V51, E53, T83, F89, F112, I133, L159, G178, M180, R182, F184, G187
RA95-Evolved	8, 9, 53, 89, 112, 133, 159, 178, 180, 182, 184, 187, 210	W8, L9, L53, F89, F112, I133, I159, T178, Y180, M182, F184, G187, L210

[a]Positions that were mutated to Gly during ligand placement in the recapitulation of RA95 and RA95-Evolved. Catalytic residues (K210 for RA95; K83 and Y51 for RA95-Evolved) were allowed to sample alternate rotamers. All other residues were kept fixed.
[b]Positions and amino acid types that were allowed to sample alternate rotamers during repacking. All other residues were kept fixed.

Table 2.4. Lysine and TS rotamers used during the ligand placement step of RA active site recapitulation.

Enzyme[a]	Measure[b]	Atom 1[c]	Atom 2[c]	Atom 3[c]	Atom 4[c]	Values
RA95	Torsion[d]	N	CA	CB	CG	185, 190, 195
	Torsion[d]	CA	CB	CG	CD	157, 162, 167
	Torsion[d]	CB	CG	CD	CE	179, 184, 189
	Torsion[d]	CG	CD	CE	NZ	245, 150, 155
	Torsion[e]	CD	CE	NZ	C9	140, 150, 160
	Torsion[f]	CE	NZ	C9	C11	170, 180, 190
	Torsion[f]	NZ	C9	C11	C7	81, 91, 101
	Torsion[g]	C9	C11	C7	C14	-187, -177, -167
	Torsion[g]	C11	C7	C14	C1	99, 109, 119
	Torsion[i]	C13	C8	O2	C4	180
RA95-Evolved	Torsion[d]	N	CA	CB	CG	202, 107, 212
	Torsion[d]	CA	CB	CG	CD	46, 51, 56
	Torsion[d]	CB	CG	CD	CE	110, 115, 120
	Torsion[d]	CG	CD	CE	NZ	161, 166, 171
	Torsion[e]	CD	CE	NZ	C9	200, 210, 220
	Torsion[f]	CE	NZ	C9	C11	-10, 0, 10
	Torsion[f]	NZ	C9	C11	C7	-99, -89, -79
	Torsion[g]	C9	C11	C7	C14	-187, -177, -167
	Torsion[h]	C11	C7	C14	C1	50, 60, 70
	Torsion[i]	C13	C8	O2	C4	180

[a]Lys210 and Lys83 are the catalytic lysines in RA95 and RA95-Evolved recapitulation, respectively.
[b]Torsions are given in degrees. Only torsions are used to define the catalytic lysine contact as it is treated as a covalent bond.
[c]Atom names are shown in Figure 2.14. Atoms of the catalytic lysine, tyrosine, and TS are shown in red, green, and blue, respectively.
[d]Lysine rotamer torsions were allowed to vary within ± 5º around the ideal angle taken from the crystal structure of PDB ID: 4A29 (for RA95) or PDB ID: 5AN7 (for RA95-Evolved).
[e]Torsions were defined as freely rotatable by Jiang and Althoff *et al.*.[26] Ideal angles were taken from the crystal structure of PDB ID: 4A29 (for Lys210) or PDB ID: 5AN7 (for Lys83).
[f]Torsions were allowed to vary within ± 10º around the ideal angle based on the QM calculation for the carbon–carbon bond-breaking step according to Jiang and Althoff *et al.*.[26]
[g]Torsions were allowed to vary within ± 20º around the ideal angle based on the QM calculation for the carbon–carbon bond-breaking step according to Jiang and Althoff *et al.*.[26]
[h]The ideal angle based on the QM calculation for the carbon–carbon bond-breaking step did not allow active site recapitulation. The ideal angle was instead taken from the crystal structure of PDB ID: 5AN7.
[i]Ideal angle based on QM calculation according to Jiang and Althoff *et al.*.[26] was not allowed to vary.

Table 2.5. Geometric constraints used to define catalytic tyrosine contact during the ligand placement step of RA95-Evolved active site recapitulation.

Contact	Measure	Atom 1[c]	Atom 2[c]	Atom 3[c]	Min	Max
Tyr51	Distance[a]	OH	O1	-	2.0	3.2
	Angle[b]	OH	O1	C7	99.5	139.5
	Angle[b]	CZ	OH	O1	99.5	139.5

[a]Given in Å.
[b]Given in degrees.
[c]Atom names are shown in Figure 2.14.

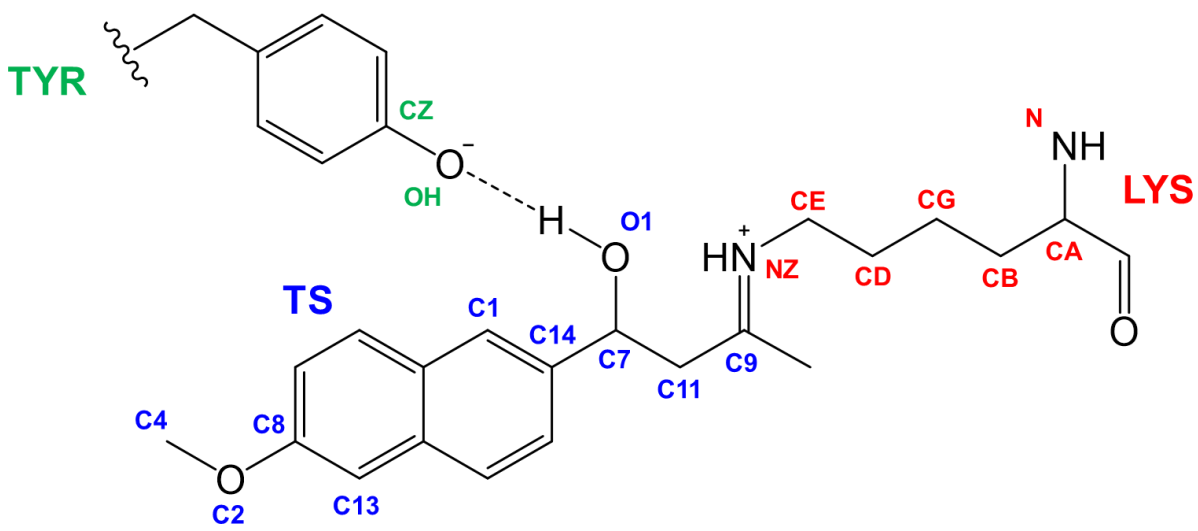

Figure 2.14. Atoms defining the geometric constraints used for the ligand placement step of RA active site recapitulation. Key atoms of the catalytic lysine, tyrosine, and TS are shown in red, green, and blue, respectively.

Chapter 3. Conclusions and perspectives

3.1. Summary

To design efficient *de novo* enzymes without requiring directed evolution, current computational enzyme design methodologies must be improved. A better understanding of how mutations introduced by directed evolution contribute to increased enzymatic activity will guide the future development of design methods, helping us to address the major challenges in computational enzyme design including low success rates, inaccurate structure prediction, and low activities of *de novo* designs. Indeed, many of the mutations introduced through directed evolution of computationally designed enzymes are located at active site positions that were previously optimized by design calculations, making it unclear why the evolved amino acid identities were not found during the design step. On the other hand, the prediction of distal mutations introduced by directed evolution presents a major challenge in enzyme design, as design calculations tend to focus solely on the active site. Therefore, this thesis focused on the evaluation of structural, functional, and dynamical impacts of active site and distal mutations found by directed evolution of the *de novo* enzyme RA95, an enzyme that presents an important case study in design due to the significant structural remodelling that was observed during evolution.

We found that active site mutations alone create a functionally optimized active site in RA95-Core without the need for distal mutations, showing activity within one order of magnitude of the fully evolved variant. This suggests not only that computational design methods can be improved to create much more efficient enzymes than what was previously achieved in RA95, but also that the challenge of predicting activity-enhancing distal mutations may not need to be addressed in the process. However, it was determined that the active site structure of the evolved

variant could not be computationally recapitulated on the original template protein that was used in the design of RA95, and this was a result of conformational changes in the backbone occurring during evolution, due in large part to distal mutations. This indicates that the optimized active site identified through directed evolution could not have been designed *de novo* instead of RA95 using current design methodologies, highlighting the need for improvement of computational methods. Overall, this work suggests that backbone remodelling is necessary to design *de novo* enzymes with native-like activities in the future. Thus, computational enzyme design procedures require the incorporation of improved methods for approximating backbone flexibility, allowing backbone configurations that are compatible with the desired efficient active site.

3.2. Future directions

3.2.1. Computational recapitulation of RA95-Evolved through backbone remodelling

To successfully recapitulate the RA95-Evolved active site structure on the original design template, backbone remodelling needs to be incorporated into design. The inability to stabilize the evolved sequence on the 1A53 backbone, even on backbone ensemble members where active site loops did not clash with the TS, demonstrates that more subtle movements of backbone atoms throughout the active site are also required. Key backbone Cα carbons, such as those at positions 51, 83, and 210, undergo shifts during evolution. In the cases of positions 51 and 210, Cα carbons have shifted 1.4 Å and 1.1 Å, respectively, between the RA95 and RA95-Evolved structures. Interestingly, the position 51 and 210 Cα carbons in 1A53 have similar positions to those in RA95, and those in RA95-Shell appear to have intermediate positions between the corresponding atoms in RA95 and RA95-Evolved (Figure 3.1a), demonstrating how distal mutations alone contribute to a portion of the backbone changes seen during evolution. Unsurprisingly, ensemble refinement

of 1A53 does not capture enough backbone flexibility to allow these backbone atoms to sample positions similar to those in RA95-Evolved (Figure 3.1a), which is likely contributing to the inability to stabilize the evolved sequence on the 1A53 ensemble. While ensemble refinement of RA95 captures more flexibility in position 210, position 51 remains largely static (Figure 3.1b). This suggests that position 51, whose tyrosine residue was included in the catalytic motif for recapitulation of RA95-Evolved, may be a key factor limiting the ability to accurately predict the RA95-Evolved structure on various design templates.

To confirm whether the backbone movements induced by distal mutations could allow recapitulation of RA95-Evolved, we attempted to stabilize its sequence on a backbone ensemble derived from ensemble refinement of RA95-Shell. While this ensemble refinement did not capture significantly more flexibility of the position 51 and 210 Cα carbons (Figure 3.1c), these atoms in the RA95-Shell crystal structure have positions closer to those in RA95-Evolved than in RA95. This likely contributed to our ability to stabilize the RA95-Evolved sequence with a negative energy score on a RA95-Shell backbone ensemble member (Figure 3.1d). Despite yielding a structurally inaccurate model, evidenced by an active site sidechain RMSD of 3.185 Å, the RA95-Shell ensemble is the only ensemble refinement-derived backbone ensemble able to stabilize the RA95-Evolved sequence. This suggests that if we were able to approximate the backbone remodelling induced by distal mutations *in silico*, we would have a better chance of successfully recapitulating the RA95-Evolved active site on the 1A53 or RA95 backbone structures.

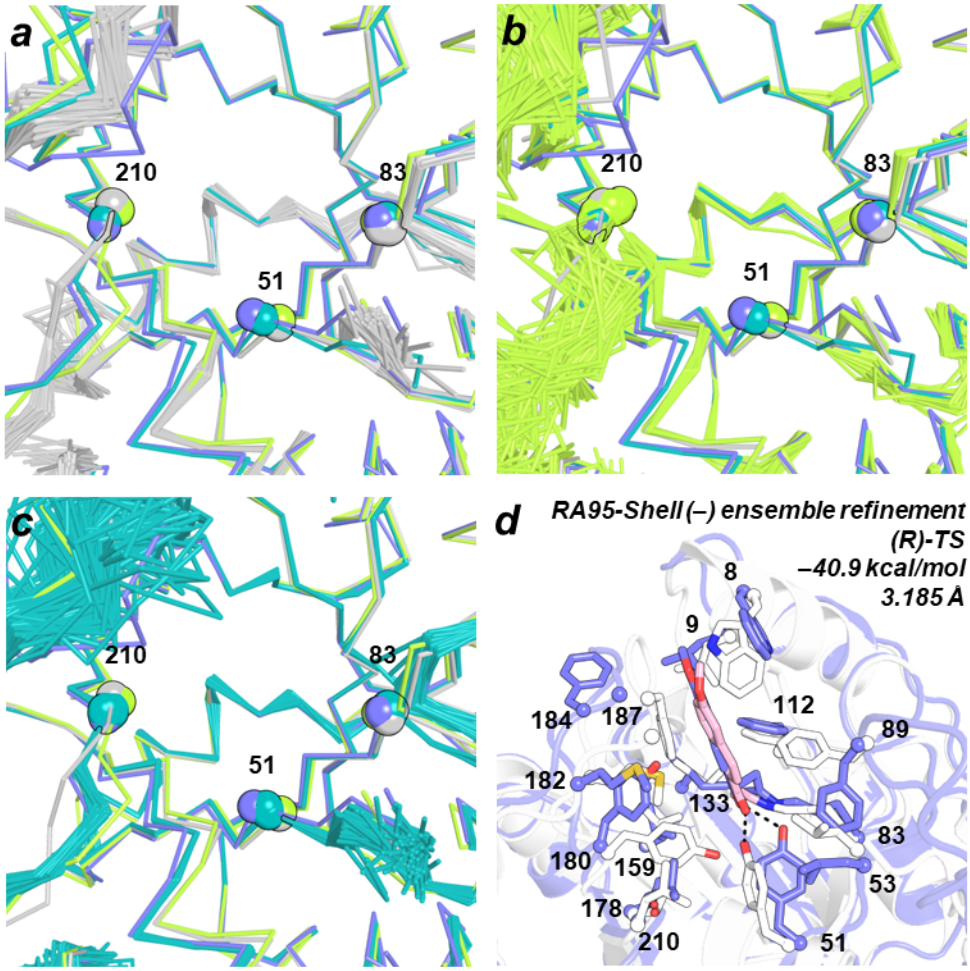

Figure 3.1. Ensemble refinement of 1A53 and RA95 does not capture enough flexibility to allow recapitulation of the RA95-Evolved active site. (a), (b), and (c) show superpositions of the ribbon representations of the WT template protein (PDB ID: 1A53) in grey, inhibitor-bound RA95 (PDB ID: 4A29) in green, inhibitor-bound RA95-Evolved (PDB ID: 5AN7) in purple, and unbound RA95-Shell in teal. The ensemble refinement-derived backbone ensembles for the WT template, RA95, and RA95-Shell are shown in (a), (b), and (c), respectively. The Cα carbons of key positions (51 for Tyr51 in RA95-Evolved, 83 for the catalytic Lys83 in RA95-Evolved, and 210 for the catalytic Lys210 in RA95) are shown as spheres for all single structures and ensemble members. Loop L1 (residues 56–62) is hidden for clarity. (d) Computational recapitulation of the RA95-Evolved active site on an ensemble refinement-derived RA95-Shell design template. The RA95-Evolved crystal structure (white) with diketone inhibitor (pink) is overlaid with the RA95-Evolved design model (purple) obtained using the (*R*)-configuration of the TS. The design model shown was obtained using the ensemble refinement-derived template that gave the best energy after active site repacking. The PHOENIX energy of the design model after active site repacking is provided in kcal/mol, as well as the active site sidechain atom RMSD from the overlaid crystal structure in Å. Sidechains of all designed residues are shown, with alpha carbons represented by spheres. The key catalytic contact between the TS and Tyr51 is shown in black dashed lines.

This analysis highlights a major limitation of this work, which is the lack of a solved crystal structure for RA95-Core. Given that the prediction of beneficial distal mutations presents such a challenge in computational enzyme design, and that active site mutations alone in RA95 create a highly efficient active site, the purpose of recapitulating the evolved active site structure is to determine whether the RA95-Core sequence could have been designed *de novo*. However, in doing so we assume that the active site structures of RA95-Core and RA95-Evolved are the same, which is unlikely given the effects that distal mutations have been shown to have on backbone conformation. It is possible that the RA95-Core backbone is more similar to that of RA95 than that of RA95-Evolved, which would mean that backbone remodelling would not be needed to recapitulate its active site on the original design template 1A53. However, this is also unlikely as the RA95-Evolved active site sequence could not be stabilized on the 1A53 or RA95 backbones. An RA95-like backbone structure in RA95-Core is also not supported by the previous results of Broom and Rakotoharisoa *et al.*,[73] who found that the active site of the "Core" Kemp eliminase variant HG4, created by introducing only active site mutations into a low-activity design, was highly structurally similar to that of the fully evolved variant. Together, this indicates that the RA95-Core active site is similar enough to RA95-Evolved that conformational differences between 1A53 and RA95-Core prevent its recapitulation on 1A53. The RA95-Core active site backbone likely has a structure in between those of the designed and evolved variants, similar to what is observed in the key Cα carbon positions in RA95-Shell (Figure 3.1).

Evidently, ensemble refinement is not sufficient to allow recapitulation of RA95-Evolved on the 1A53 or RA95 backbones. Additional methods of addressing the fixed backbone approximation should be considered to allow successful *de novo* prediction of the evolved active site. This could be achieved either by optimizing sequences on pre-generated ensembles of fixed

backbone templates, as was done here with ensembles generated from ensemble refinement and MD, or through the incorporation of backbone flexibility directly into the design procedure using flexible-backbone design algorithms.[77] Aside from ensemble refinement, various other methods have been used for the computational generation of backbone ensembles. These include the use of unconstrained MD simulations, solution NMR structures, and a method known as PertMin,[102] which involves the perturbation of atomic coordinates followed by energy minimization. However, it was previously found that backbone ensembles derived from unconstrained MD and from solution NMR have low prediction accuracy because they do not represent physically valid models of the desired protein fold.[102] This matches with the results obtained in this study, where unconstrained MD-derived backbone ensembles were able to stabilize the desired RA sequences but with low structural prediction accuracy. In general, backbone ensembles that can achieve high prediction accuracy in CPD must have ensemble members that are highly structurally similar to the input crystal structure.[102]

Other ensemble generation methods that have been shown to produce physically valid backbone models are those employing backrub movements. Backrubs are a mode of local backbone motion underlying commonly observed structural heterogeneity in protein crystal structures[78] and have been useful in modelling conformational changes in response to mutations.[103] In addition to generating backbone ensembles, backrubs can also be incorporated into flexible-backbone design algorithms within the design procedure. While physically valid modes of protein flexibility such as backrubs may be helpful in remodelling the backbone in subtle ways during computational design, more significant conformational changes are needed to approximate loop movements such as those occurring in the active sites of RA variants. One method to be considered is the local loop reconstruction method known as kinematic closure (KIC), which has been shown

to model loop conformations with sub-Angstrom accuracy.[79] Perhaps the combination of a loop remodelling KIC algorithm and a more subtle backrub algorithm could allow backbone remodelling in such a way that the RA95-Evolved active site could be recapitulated on the original design template.

While numerous methods exist for the potential incorporation of backbone remodelling into computational enzyme design calculations, it is possible that the backbone movements required to allow recapitulation of RA95-Evolved on 1A53 are too specific to be achieved. This raises the question of whether it would be more effective to design custom protein backbone templates for design when remodelling efforts are not sufficient. For example, Yeh and Norn *et al*.[104] recently used computational enzyme design to create *de novo* luciferase enzymes using non-native protein backbones as design templates, as opposed to native protein scaffolds taken from the PDB. A deep-learning-based "family-wide hallucination" approach was used to generate large numbers of ideal protein scaffolds with diverse binding pocket sizes and shapes based on the desired native fold of the nuclear transport factor 2 (NTF2)-like protein superfamily. However, structural prediction accuracy of the designed luciferases could not be verified as X-ray crystal structures of designs were not determined. Additionally, the identification of three active designs required screening of almost 8,000 variants, suggesting that the implementation of custom protein backbones as design templates in this study was suboptimal.

In the study discussed above from Yeh and Norn *et al*.,[104] machine learning methods were used to generate libraries of non-native protein scaffolds for use as backbone templates in design calculations. In the context of protein design, machine learning methods can learn from patterns in large amounts of pre-existing biological data to make predictions related to relationships

between protein sequence, structure, and function. In this regard, recent advances in machine learning will likely prove useful in computational enzyme design. The neural network AlphaFold[105] is one example of a machine learning method that tackles the protein folding problem, taking an amino acid sequence as input and predicting the corresponding protein fold with atomic accuracy. It does this by leveraging both multiple sequence alignment (MSA) statistics and pairwise representations of protein structures describing relationships between residues.[105] On the other hand, ProteinMPNN[106] addresses the protein design problem by predicting an amino acid sequence that will fold to a protein fold of interest. Unlike the protein design software Rosetta and TRIAD, which treat sequence design as an energy optimization problem, ProteinMPNN uses a neural network to predict protein sequences from the N- to C- terminus using protein backbone features as inputs, including backbone atom distances and dihedral angles.[106] AlphaFold is increasingly being used to evaluate the accuracy of *de novo* designs, while sequence design by ProteinMPNN can result in better stability and solubility than RosettaDesign. Another structure prediction method, RoseTTAFold (RF), has been implemented as the denoising network in a generative diffusion model RFdiffusion,[107] which can generate diverse protein structures by denoising random input data. While training inputs are generated by noising structures taken from the PDB, denoised structures have little overall structural similarity to the training inputs.

Machine learning methods such as AlphaFold,[105] ProteinMPNN,[106] and RFdiffusion[107] could be used to generate custom protein backbones to be used as templates during design, or perhaps to generate custom sections of native backbones that need to be remodelled. Returning to the challenge of recapitulating the RA95-Evolved active site structure on the original design template for RA95, it may be useful to replace the active site loops of the template with custom loops generated by machine learning. A method such as RFdiffusion could be used to create a loop

or a library of loops, followed by the use of ProteinMPNN to design sequences that would fold into the diffused loop structure. Better yet, loops could be generated following active site placement on the template, ensuring that the generated loops are complementary to the desired active site. Whether the recapitulation of RA95-Evolved is achieved or not, machine learning methods have opened the door to the design of *de novo* enzymes through the creation of a protein backbone around an active site, as opposed to the traditional placement of desired active sites on the backbone template. Indeed, it is possible that recapitulation of RA95-Evolved on 1A53 will not be possible and given that the RA95-Evolved active site represents only one possible solution that could have been identified by directed evolution, this may be unsurprising. In fact, it is unclear whether the generation of designed sequences with the catalytic lysine at position 83, as is the case in RA95-Evolved, would have been useful during the design of RA95. Althoff, Wang, and Jiang *et al.*,[27] who designed various retro-aldolases in parallel to RA95, were unable to differentiate active and inactive designs computationally, even in those with the same catalytic lysine position on the same scaffold. Nevertheless, the methods of backbone and loop remodelling proposed above will be generally useful in the field of computational enzyme design.

3.2.2. Prediction of activity-enhancing distal mutations in retro-aldolases

While the majority of the increase in catalytic activity of RA95 during evolution can be attributed to active site mutations, the addition of distal mutations led to an additional 6-fold increase in activity. Thus, if the active site mutations in RA95-Evolved identified by directed evolution could be successfully designed *de novo* through improvements to computational enzyme design procedures, the next challenge to be tackled would be the prediction of activity-enhancing distal mutations. In many cases, distal mutations have been found to alter networks of noncovalent

interactions within enzymes, causing a redistribution of pre-existing conformational states that favour catalysis to different degrees, or altering the conformational dynamics of flexible regions such as active site loops and lids.[42] This matches with the results obtained in this study, where conformational changes were seen in flexible active site loops upon the addition of distal mutations to RA95, resulting in a population shift of conformational states towards those with more accessible active sites. However, the feasibility of predicting these mutations is unclear. This challenge is two-fold, as target distal positions must be selected followed by the identification of amino acid identities that will be beneficial at these positions. Progress has been made towards the former problem through the treatment of distal mutations in a way that is analogous to allosteric regulation through ligand binding at allosteric sites. Allosteric regulation is often computationally analyzed using graph-based approaches, whereby enzymes are represented by weighted graphs with nodes corresponding to residues, and these methods have been used to identify key distal positions that contribute to shifts in the conformational landscapes of enzymes.[42,72]

One such graph-based method is the Shortest Path Map (SPM), which relies on the construction of graph representations based on residue-by-residue correlated movements and inter-residue distances obtained from long-time-scale MD simulations. This generates a map that identifies pairs of residues with higher contributions to the hypothetical allosteric communication pathway in the enzyme. The application of SPM has been shown to predict distal mutation hotspots targeted by directed evolution in various unrelated enzymes including the RA95 series of retro-aldolases,[41] as well as Monoamine Oxidase from *Aspergillus niger* (MAO-N)[108] and Tryptophan synthase (TrpS) from *Pyrococcus furiosus*.[43] In fact, the SPM of the template protein used in the design of RA95 identifies the ten distal mutation positions in RA95-Evolved either exactly or at positions adjacent to the SPM (Figure 3.2). This suggests that the SPM method could be used for

the rational selection of target positions for the introduction of activity-enhancing distal mutations during enzyme design. A design strategy can be imagined where an enzyme with an efficient active site, such as RA95-Core, could be designed *de novo*, followed by the introduction of activity-enhancing distal mutations at positions chosen using SPM to create an optimized enzyme similar to RA95-Evolved. However, the challenge of identifying amino acid identities at chosen distal positions that would yield activity enhancements remains. For now, positions identified by SPM could be used for the design of smart libraries for the evolution of low-activity designs.

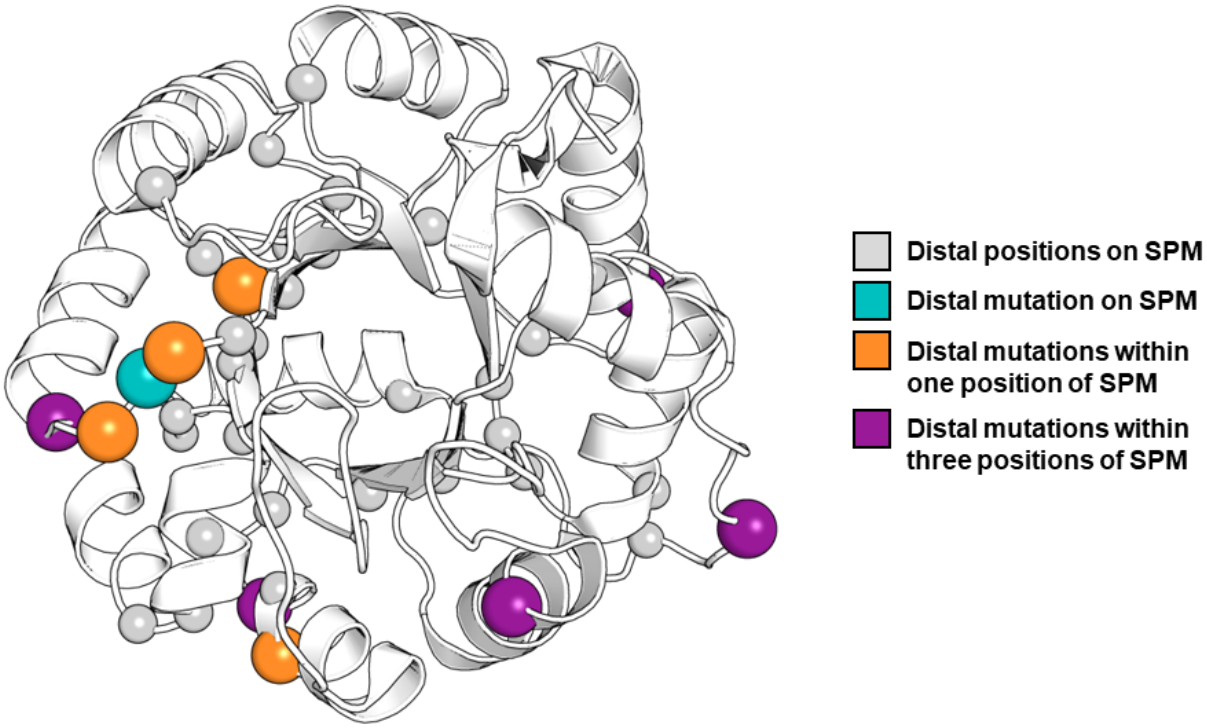

Figure 3.2. Distal mutations in RA95-Evolved are identified on or adjacent to the shortest path map (SPM) of the original design template. The protein backbone shown is the template protein used in the original design of RA95 (PDB ID: 1LBL, which is the same protein as that in 1A53, bound to a different ligand). Distal positions on the SPM of 1LBL are shown in small grey spheres. Distal mutations in RA95-Evolved located on the SPM (I213F), within one position from the SPM (F72Y, A209P, G212D, S214F), and within two or three positions away in sequence from the SPM (R23H, R75P, T95M, S151G, R216P) are shown in large teal, orange, and purple spheres, respectively. Figure created using SPM data from Romero-Rivera *et al.*.[41]

3.2.3. Optimization of RA95 activity by expanding the catalytic motif

If recapitulation of the evolved RA active site is not possible on the original design template after backbone remodelling, this would indicate that it could not have been designed *de novo*. It is possible that the design template is simply not compatible with the catalytic lysine being placed at position 83. In this case, it would be useful if the low-activity RA95 design could be improved through computational design, as this would be more efficient than directed evolution. It has been suggested by Broom and Rakotoharisoa *et al.*[73] that computational enzyme design using a crystallographically derived backbone ensemble from a low-activity enzyme could eliminate the need for directed evolution to improve the design. In this study, the active site structure of a highly active Kemp eliminase variant could be recapitulated using ensemble refinement of a low-activity variant, but not that of the original design template. This could be because catalytically competent sub-states, which are likely more prevalent in the backbone ensemble of a low-activity enzyme than in an inactive protein scaffold, can be sampled during the design procedure. In the same way, ensemble refinement of RA95 may allow the design of more active RA enzymes using position 210 as the location of the catalytic lysine.

Using an ensemble of backbones generated from ensemble refinement of RA95 as design templates, theozyme placement followed by repacking and experimental characterization of resulting sequences should be attempted to increase the activity of RA95. However, an improved catalytic motif should be implemented, as experimental validation of RA95 revealed that the original designed catalytic motif was predicted highly inaccurately. Not only was the designed glutamate, positioned to orient a catalytic water molecule, found to have no impact on catalytic activity, but the degree of accuracy in prediction of the remaining active site sidechain

conformations is also unclear. Given that computational recapitulation of RA active sites in this study was carried out using a catalytic dyad made up of a lysine and tyrosine, it would be interesting to determine whether a tyrosine analogous to Tyr51 in RA95-Evolved would be compatible with the catalytic lysine of RA95 in position 210. Thus, placement of a tyrosine residue in the designed active site, positioned to act as a general base during the bond cleavage step, should be attempted to increase the activity of RA95. While a similar catalytic motif, comprising a catalytic lysine and a tyrosine positioned to deprotonate the β-alcohol of the substrate, did not yield any active designs during the design of RA95, this motif has not been tested on any backbone templates derived from enzymes that already possess RA activity, such as RA95.

3.2.4. Composite TS structures in *de novo* RA design

Up to now, we have discussed modification of the backbone template and the catalytic motif in the context of improving the design of *de novo* retro-aldolases. A final consideration that should be mentioned in the case study of the RA95 series is the choice of TS to be incorporated into the ideal theozyme. Retro-aldolases, catalyzing the multistep carbon–carbon bond cleavage of methodol, represent some of the most mechanistically complex enzymes to have been designed to date. As such, the success of RA designs can provide insight into the correct treatment of a multistep reaction mechanism in the construction of a theozyme. While a single TS, that of the carbon–carbon bond-breaking step, is used in recapitulation of the RA active sites in this work, a composite TS was implemented for the original design of RA enzymes. During design of the first *de novo* retro-aldolases, only one of the six intermediate and one of the five TS structures occurring during the RA reaction mechanism were incorporated into the composite TS, specifically the carbinolamine intermediate and carbon–carbon bond-breaking TS (Figure 1.7). Through

superposition of each QM-derived intermediate and TS model, it was determined that these two structures together approximate most of the atomic motion along the reaction pathway.[26] To create a simplified composite TS model, the carbinolamine alcohol and terminal methyl from the carbinolamine intermediate were superimposed onto the carbon–carbon bond-breaking TS. In the subsequent design of RA95, models of the Schiff base intermediate and water elimination step TS were incorporated into the composite TS model.[27]

While the composite TS models employed in the *de novo* RA designs described above yielded active designs, this strategy is not optimal. Firstly, intermediate and TS models of many of the reaction steps were not incorporated into the composite TS, meaning that the design procedure did not consider the ability of the enzyme to stabilize the structures along the complete reaction pathway. In the case of the first RA designs, where the composite TS was based on the superposition of additional chemical groups on the bond-breaking step TS, it was suggested that the resulting low success rates were a result of too great of a focus on the bond-breaking step TS, resulting in too little space to accommodate the carbinolamine alcohol.[27] While subsequent composite TS models incorporated additional steps, the second half of the reaction mechanism was still disregarded, as the assumption was made that steps following the first product release have identical TS structures except for the naphthyl group now being absent. It is likely that the inclusion of the latter half of the reaction in the composite TS would benefit design as the rate-limiting step in the evolved enzyme has been shown to be related to product release.[65,67]

In addition to the inclusion of a limited number of intermediate and TS structures, the employed composite TS strategy is not ideal because it assumes that the same static protein backbone structure will allow efficient stabilization of all structures along the RA reaction

coordinate. As such, it is likely that the treatment of each intermediate and TS structure individually will be required to achieve the accurate design of *de novo* enzymes for multistep reactions, and the incorporation of backbone remodelling strategies discussed above will be helpful in this regard. Incorporation of backbone flexibility into the design procedure will allow modelling of conformational changes undergone by enzymes to effectively stabilize different structures during the catalytic cycle. However, the most appropriate approach for incorporating multiple intermediate and TS structures into the design procedure is unclear, as different potential design sequences are expected to stabilize various structures to different degrees. Should sequences be ranked assuming that the stabilization of all intermediate structures is equally important, or are some steps during the reaction more important? If it is the latter, how should the most important steps be chosen? RA designs could aid in answering these questions and in the benchmarking of new strategies for the computational design of biocatalysts for multistep reactions.

References

1. Wolfenden, R. & Snider, M. J. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).
2. Edwards, D. R., Lohman, D. C. & Wolfenden, R. Catalytic proficiency: The extreme case of S–O cleaving sulfatases. *J. Am. Chem. Soc.* **134**, 525–531 (2012).
3. Mak, W. S. & Siegel, J. B. Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
4. Bornscheuer, U. T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol.* **5**, 137–143 (2001).
5. Zanghellini, A. De novo computational enzyme design. *Curr. Opin. Biotechnol.* **29**, 132–138 (2014).
6. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).
7. Dahiyat, B. I. & Mayo, S. L. Protein design automation. *Protein Sci.* **5**, 895–903 (1996).
8. Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Mol. Biol.* **5**, 470–475 (1998).
9. Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656–659 (2006).
10. Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
11. Street, A. G. & Mayo, S. L. Computational protein design. *Struct.* **7**, R105–R109 (1999).
12. Schmidt Am Busch, M., Lopes, A., Mignon, D. & Simonson, T. Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *J. Comput. Chem.* **29**, 1092–1102 (2008).
13. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* **124**, 2723–2729 (2002).
14. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542 (1992).
15. Desmet, J., Spriet, J. & Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins: Struct. Funct. Genet.* **48**, 31–43 (2002).
16. Koehl, P. & Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Mol. Biol.* **2**, 163–170 (1995).

17. Koehl, P. & Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275 (1994).

18. Tantillo, D. J., Jiangang, C. & Houk, K. N. Theozymes and compuzymes: Theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).

19. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 16710–16715 (2006).

20. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).

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

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

23. 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).

24. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3790–3795 (2012).

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

26. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).

27. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).

28. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).

29. Richter, F. *et al.* Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).

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

31. Preiswerk, N. *et al.* Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8013–8018 (2014).

32. 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).

33. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).

34. Bar-Even, A. *et al.* The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochem.* **50**, 4402–4410 (2011).

35. Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.* **31**, 904–916 (2009).

36. Gouverneur, V. E. *et al.* Control of the exo and endo pathways of the Diels-Alder reaction by antibody catalysis. *Science* **262**, 204–208 (1993).

37. Arnold, F. H. & Volkov, A. A. Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.* **3**, 54–59 (1999).

38. Khersonsky, O. *et al.* Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. *J. Mol. Biol.* **396**, 1025–1042 (2010).

39. Khersonsky, O. *et al.* Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.* **407**, 391–412 (2011).

40. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).

41. Romero-Rivera, A., Garcia-Borràs, M. & Osuna, S. Role of conformational dynamics in the evolution of retro-aldolase activity. *ACS Catal.* **7**, 8524–8532 (2017).

42. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comput. Mol. Sci.* **11**, e1502 (2021).

43. Maria-Solano, M. A., Kinateder, T., Iglesias-Fernández, J., Sterner, R. & Osuna, S. In silico identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase. *ACS Catal.* **11**, 13733–13743 (2021).

44. Katti, S. K. & LeMaster, D. M. Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J. Mol. Biol.* **212**, 167–184 (1990).

45. Hennig, M., Darimont, B. D., Jansonius, J. N. & Kirschner, K. The catalytic mechanism of indole-3-glycerol phosphate synthase: Crystal structures of complexes of the enzyme from Sulfolobus solfataricus with substrate analogue, substrate, and product. *J. Mol. Biol.* **319**, 757–766 (2002).

46. Lo Leggio, L. *et al.* Substrate specificity and subsite mobility in T. aurantiacus xylanase 10 A. *FEBS Lett.* **509**, 303–308 (2001).

47. Hakulinen, N., Turunen, O., Janis, J., Leisola, M. & Rouvinen, J. Three-dimensional structures of thermophilic beta-1,4-xylanases from Chaetomium thermophilum and Nonomuraea flexuosa. *Eur. J. Biochem.* **270**, 1399–1412 (2003).

48. Rinaldo-Matthis, A. *et al.* Crystal structures of the mitochondrial deoxyribonucleotidase in complex with two specific inhibitors. *Mol. Pharmacol.* **65**, 860–867 (2004).

49. Scharff, E. I., Koepke, J., Fritzsch, G., Lücke, C. & Rüterjans, H. Crystal structure of diisopropylfluorophosphatase from Loligo vulgaris. *Struct.* **9**, 493–502 (2001).

50. Arai, R. Crystal structure of the probable haloacid dehalogenase PH0459 from Pyrococcus horikoshii OT3. *Protein Sci.* **15**, 373–377 (2006).

51. Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *eLife* **4**, e07574 (2015).

52. Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638–1642 (2006).

53. Jiménez-Osés, G. *et al.* The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **10**, 431–436 (2014).

54. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).

55. Otten, R. *et al.* How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).

56. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).

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

58. Garrabou, X., Beck, T. & Hilvert, D. A promiscuous de novo retro-aldolase catalyzes asymmetric Michael additions via Schiff base intermediates. *Angew. Chem.* **127**, 5701–5704 (2015).

59. Garrabou, X., Macdonald, D. S., Wicky, B. I. M. & Hilvert, D. Stereodivergent evolution of artificial enzymes for the Michael reaction. *Angew. Chem. Int. Ed.* **57**, 5288–5291 (2018).

60. Garrabou, X., Verez, R. & Hilvert, D. Enantiocomplementary synthesis of γ-nitroketones using designed and evolved carboligases. *J. Am. Chem. Soc.* **139**, 103–106 (2017).

61. Garrabou, X., Macdonald, D. S. & Hilvert, D. Chemoselective Henry condensations catalyzed by artificial carboligases. *Chem. Eur. J.* **23**, 6001–6003 (2017).

62. Garrabou, X., Wicky, B. I. M. & Hilvert, D. Fast Knoevenagel condensations catalyzed by an artificial Schiff-base-forming enzyme. *J. Am. Chem. Soc.* **138**, 6972–6974 (2016).

63. Obexer, R., Pott, M., Zeymer, C., Griffiths, A. D. & Hilvert, D. Efficient laboratory evolution of computationally designed enzymes with low starting activities using fluorescence-activated droplet sorting. *Protein Eng. Des. Sel.* **29**, 355–366 (2016).

64. Choi, K. H., Shi, J., Hopkins, C. E., Tolan, D. R. & Allen, K. N. Snapshots of catalysis: The structure of fructose-1,6-(bis)phosphate aldolase covalently bound to the substrate dihydroxyacetone phosphate. *Biochem.* **40**, 13868–13875 (2001).

65. Zeymer, C., Zschoche, R. & Hilvert, D. Optimization of enzyme mechanism along the evolutionary trajectory of a computationally designed (retro-)aldolase. *J. Am. Chem. Soc.* **139**, 12541–12549 (2017).

66. 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).

67. De Raffele, D., Martí, S. & Moliner, V. QM/MM theoretical studies of a de novo retro-aldolase design. *ACS Catal.* **9**, 2482–2492 (2019).

68. Motta, J. F. G., Freitas, B. C. B. D., Almeida, A. F. D., Martins, G. A. D. S. & Borges, S. V. Use of enzymes in the food industry: A review. *Food Sci. Technol.* **43**, e106222 (2023).

69. Meghwanshi, G. K. *et al.* Enzymes for pharmaceutical and therapeutic applications. *Biotechnol. Appl. Biochem.* **67**, 586–601 (2020).

70. Maurer, K. Detergent proteases. *Curr. Opin. Biotechnol.* **15**, 330–334 (2004).

71. St-Jacques, A. D., Gagnon, O. & Chica, R. A. Chapter 4. Computational Enzyme Design: Successes, Challenges, and Future Directions. in *Catalysis Series* (eds. Williams, G. & Hall, M.) 88–116 (Royal Society of Chemistry, 2018).

72. Maria-Solano, M. A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J. & Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun.* **54**, 6622–6634 (2018).

73. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).

74. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 498–520 (1933).

75. Bunzel, H. A. *et al.* Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **13**, 1017–1022 (2021).

76. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).

77. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.* **380**, 757–774 (2008).

78. Davis, I. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. The backrub motion: How protein backbone shrugs when a sidechain dances. *Struct.* **14**, 265–274 (2006).

79. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).

80. Thermo Fisher Scientific. *Microplate-Based Pathlength Correction Method for Photometric DNA Quantification*. (2015).

81. Greenfield, N. J. Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. *Nat. Protoc.* **1**, 2733–2741 (2006).

82. Winter, G. xia2: An expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).

83. Kabsch, W. XDS. *Acta. Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

84. Evans, P. Scaling and assessment of data quality. *Acta. Crystallogr. D Biol. Crystallogr.* **62**, 72–82 (2006).

85. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

86. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta. Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

87. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta. Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

88. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta. Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

89. Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R. & Cerutti, D. S. AMBER 2020. (2020).

90. Maier, J. A. *et al.* Ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

91. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

92. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

93. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

94. Miyamoto, S. & Kollman, P. A. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).

95. Scherer, M. K. *et al.* PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).

96. Voigt, C. A., Gordon, D. B. & Mayo, S. L. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803 (2000).

97. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B. & Meyer Jr., E. F. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324 (1977).

98. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).

99. Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10172–10177 (1997).

100. Chica, R. A., Moore, M. M., Allen, B. D. & Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20257–20262 (2010).

101. Shortle, D. Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci.* **12**, 1298–1302 (2003).

102. Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins* **82**, 771–784 (2014).

103. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **380**, 742–756 (2008).

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

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

106. Dauparas, J., Anishchenko, I., Bennett, N. R., Bai, H. & Ragotte, R. J. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

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

108. Curado-Carballada, C., Feixas, F., Iglesias-Fernandez, J. & Osuna, S. Hidden conformations in Aspergillus niger monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed.* **58**, 3097–3101 (2019).

Supplementary Information

Supplementary Table 1. Amino acid sequences of RA variants.

Enzyme	Sequence[a]
RA95	PRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIAVYERKSPSGLDVERDPI EYAKFMERYAVGLSITTEEKYFNGSYETLRKIASSVSIPILMSDFIVKESQIDDAYNLGADTVLLI VKILTERELESLLEYARSYGMEPLILINDENDLDIALRIGARFIGIMSRDFETGEINKENQRKLIS MIPSNVVKVAKLGISERNEIEELRKLGVNAFLISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**Y**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILM**N**DFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**Y**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Core	PRYLKGWLEDVVQLSLRRPSVRASRQRPIISLNERILEFNKRNITAIIA**Y**Y**L**RKSPSGLDVERDPI EYAKFMERYAVGLSI**K**TEEKYF**D**GSYETLRKIASSVSIPILM**N**DFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYARSYGMEPLILINDENDLDIALRIGARFI**T**I**Y**S**MN**FETGEINKENQRKLIS MIPSNVVKVA**L**LGISERNEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Shell	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIAVYERKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSITTEEKYFNGSYE**M**LRKIASSVSIPILMSDFIVKESQIDDAYNLGADTVLLI VKILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFIGIMSRDFETGEINKENQRKLIS MIPSNVVKV**P**KL**DFF**E**P**NEIEELRKLGVNAFLISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-Y51F	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**F**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILM**N**DFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**Y**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-N110S	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**Y**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILMSDFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**Y**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-Y180F	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**Y**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILM**N**DFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**F**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-Y51F-N110S	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**F**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILMSDFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**Y**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-Y51F-Y180F	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**F**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILM**N**DFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**F**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH
RA95-Evolved-N110S-Y180F	PRYLKGWLEDVVQLSLRRPSV**H**ASRQRPIISLNERILEFNKRNITAIIA**Y**Y**L**RKSPSGLDVERDPI EYAK**Y**ME**P**YAVGLSI**K**TEEKYF**D**GSYE**M**LRKIASSVSIPILMSDFIVKESQIDDAYNLGADTVLLI V**E**ILTERELESLLEYAR**G**YGMEPLILINDENDLDIALRIGARFI**T**I**F**S**MN**FETGEINKENQRKLIS MIPSNVVKV**PL**L**DFF**E**P**NEIEELRKLGVNAF**M**ISSSLMRNPEKIKELIEGSLEHHHHHH

[a]Mutations from RA95 are highlighted in bold and underlined. All sequences contain a 6xHis-tag at the C-terminus.

Supplementary Table 2. Mutations of RA variants relative to RA95.

Enzyme	Mutations from RA95[a]
RA95	-
RA95-Evolved	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, S110N, K135E, S151G, G178T, M180Y, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Core	V51Y, E53L, T83K, N90D, S110N, K135E, G178T, M180Y, R182M, D183N, K210L, L231M
RA95-Shell	R23H, F72Y, R75P, T95M, S151G, A209P, G212D, I213F, S214F, R216P
RA95-Evolved-Y51F	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, S110N, K135E, S151G, G178T, M180Y, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Evolved-N110S	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, K135E, S151G, G178T, M180Y, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Evolved-Y180F	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, S110N, K135E, S151G, G178T, M180F, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Evolved-Y51F-N110S	R23H, V51F, E53L, F72Y, R75P, T83K, N90D, T95M, K135E, S151G, G178T, M180Y, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Evolved-N110S-Y180F	R23H, V51Y, E53L, F72Y, R75P, T83K, N90D, T95M, K135E, S151G, G178T, M180F, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M
RA95-Evolved-Y51F-Y180F	R23H, V51F, E53L, F72Y, R75P, T83K, N90D, T95M, S110N, K135E, S151G, G178T, M180F, R182M, D183N, A209P, K210L, G212D, I213F, S214F, R216P, L231M

[a]The designed and evolved catalytic lysines are located at positions 210 and 83, respectively.

Supplementary Table 3. DNA sequences of RA variants.

Enzyme	Sequence
RA95	CCGCGTTACTTGAAAGGATGGCTTGAAGATGTGGTTCAATTGTCGTTACGCCGCCCATCGGTCCGT GCCAGTCGTCAGCGTCCCATTATCTCCCTGAACGAGCGTATCTTGGAGTTTAACAAGCGTAATATT ACGGCTATCATCGCCGTGTATGAGCGTAAGTCGCCTTCCGGTCTGGACGTTGAACGCGATCCAATT GAGTACGCCAAATTTATGGAGCGTTATGCGGTGGGTTTGTCGATTACGACTGAAGAGAAGTATTTC AACGGCTCATACGAAACTTTGCGTAAGATTGCGTCGTCGGTCAGCATCCCCATCCTGATGTCGGAT TTCATCGTAAAAGAGAGCCAGATCGACGATGCATACAATCTGGGTGCTGACACAGTTCTTCTGATT GTGAAGATCTTAACAGAACGTGAGTTAGAGTCCTTGCTTGAATACGCGCGTAGCTACGGCATGGAA CCTTTGATTCTTATCAACGACGAAAATGATCTTGATATCGCGTTACGTATTGGTGCGCGTTTCATC GGGATTATGTCGCGCGATTTCGAGACCGGTGAGATCAACAAGGAAAATCAACGCAAGCTTATTAGC ATGATCCCTTCCAATGTTGTGAAGGTTGCAAAATTGGGCATTTCGGAGCGCAACGAGATCGAAGAG CTGCGTAAATTGGGAGTCAATGCATTCTTGATCTCCAGCTCTCTGATGCGCAATCCTGAGAAAATC AAGGAGTTAATTGAAGGTAGCCTGGAGCACCACCATCACCACCATTAA
RA95-Evolved	CCTCGCTATTTGAAGGGGTGGCTGGAAGACGTAGTACAACTCTCATTGCGTCGCCCATCAGTTCAT GCGAGTCGTCAACGCCCGATTATCTCATTGAACGAACGCATTCTTGAGTTCAATAAGCGTAATATC ACCGCCATCATCGCGTATTACCTTCGCAAGAGTCCTAGTGGTCTTGACGTAGAACGCGATCCGATT GAGTACGCCAAGTACATGGAACCGTACGCGGTAGGTTTAAGCATTAAGACCGAAGAGAAGTATTTT GACGGCTCCTACGAAATGTTACGCAAAATCGCCTCTAGTGTTAGTATTCCAATCCTCATGAACGAT TTCATCGTTAAGGAATCACAGATCGACGATGCCTACAATTTAGGTGCAGACACCGTGTTGCTTATT GTCGAAATTTTAACCGAACGCGAGTTAGAGTCTCTCCTTGAATACGCACGTGGCTACGGTATGGAG CCCTTAATTCTGATTAATGATGAGAATGATCTCGATATCGCGCTGCGCATCGGCGCTCGCTTCATC ACAATCTATTCCATGAATTTTGAGACGGGTGAAATTAATAAAGAGAATCAACGCAAGTTAATTAGC ATGATTCCGAGCAACGTGGTGAAGGTGCCCCTTCTGGACTTCTTCGAGCCCAACGAGATTGAAGAG TTACGTAAGCTCGGCGTGAATGCGTTCATGATTTCCTCTAGCCTTATGCGTAATCCCGAGAAAATT AAAGAGTTAATCGAGGGGTCCTTAGAGCATCATCACCACCACCACTGA
RA95-Core	CCACGTTACTTAAAAGGCTGGTTGGAAGATGTGGTACAACTTTCGTTACGCCGCCCTAGCGTGCGT GCGAGCCGCCAACGTCCAATCATTTCCTTGAATGAGCGCATTCTCGAATTCAATAAGCGTAACATT ACAGCAATCATTGCTTATTACTTGCGTAAGTCGCCGAGTGGATTGGATGTGGAGCGCGACCCGATT GAGTACGCCAAGTTTATGGAGCGCTATGCCGTGGGGTTATCGATTAAGACAGAGGAGAAGTACTTC GACGGGTCCTACGAAACGCTGCGCAAGATCGCTTCATCGGTTTCCATCCCCATCTTAATGAATGAC TTTATCGTTAAGGAAAGTCAAATCGACGATGCATATAATCTGGGTGCGGATACTGTCCTGTTAATC GTTGAGATCCTTACAGAACGCGAGTTGGAGTCTTTGTTGGAGTACGCACGTTCCTATGGTATGGAG CCATTGATCCTTATCAACGACGAGAATGACTTGGACATTGCGTTACGCATCGGTGCTCGCTTTATC ACTATCTATTCTATGAACTTTGAGACCGGAGAGATCAATAAAGAGAACCAACGTAAATTGATCAGT ATGATTCCTAGTAACGTGGTGAAAGTGGCCCTTCTGGGGATTTCGGAGCGCAATGAAATTGAGGAG CTCCGCAAGTTAGGTGTCAATGCATTTATGATTTCCAGCAGCCTGATGCGCAATCCGGAGAAGATC AAGGAGTTAATTGAAGGCAGCCTTGAGCACCACCATCACCATCACTGA
RA95-Shell	CCGCGTTATTTAAAAGGATGGCTGGAAGATGTGGTTCAGTTGAGTCTTCGGCGACCCTCAGTTCAT GCTTCCCGTCAGCGACCAATTATTTCGCTCAATGAACGTATCCTGGAGTTTAATAAGCGCAATATT ACGGCAATTATTGCCGTGTACGAGCGCAAGAGTCCGAGCGGCCTCGATGTAGAACGAGATCCAATC GAGTATGCCAAGTATATGGAACCCTATGCGGTAGGTCTCAGCATAACCACTGAGGAAAAGTATTTC AATGGCAGCTATGAGATGCTGCGCAAGATCGCCTCTTCCGTGTCCATACCCATATTAATGTCGGAC TTCATAGTTAAAGAAAGCCAAATTGATGATGCGTATAATCTTGGGGCGGACACAGTACTTCTGATT GTGAAAATATTAACCGAACGAGAACTGGAATCGTTATTGGAGTATGCCCGTGGATATGGGATGGAG CCTCTTATTTTGATAAACGATGAAAATGATCTTGATATAGCTCTCCGCATTGGGGCACGCTTTATT GGGATAATGAGTCGTGATTTTGAAACTGGGGAAATCAATAAGGAAAACCAGCGTAAGCTTATCAGC ATGATCCCCAGTAATGTGGTGAAAGTTCCTAAATTAGATTTCTTTGAGCCCAATGAAATTGAGGAG CTCCGTAAACTTGGTGTGAATGCGTTTCTTATCTCTTCAAGCCTGATGAGAAATCCGGAGAAAATA AAAGAACTGATCGAAGGTTCCTTAGAGCATCACCATCATCATCACTGA
RA95-Evolved-Y51F	CCCCGCTACCTGAAGGGATGGTTAGAAGATGTCGTTCAATTATCACTTCGCCGTCCGAGCGTCCAT GCGTCTCGTCAACGGCCAATAATAAGTTTGAACGAGCGTATCTTGGAGTTTAACAAGCGTAACATA

	ACTGCCATAATCGCCTTCTACTTACGCAAGTCACCGTCGGGCCTGGATGTCGAGCGCGATCCGATC GAGTATGCCAAGTATATGGAGCCATACGCTGTCGGACTGTCTATCAAGACCGAAGAGAAGTACTTT GACGGCTCTTACGAGATGCTGCGTAAGATTGCTTCCAGTGTCTCAATACCGATACTGATGAACGAC TTCATCGTTAAGGAAAGTCAAATTGATGACGCATACAATCTGGGTGCAGATACCGTACTGTTGATA GTGGAAATCCTTACTGAGAGAGAGCTCGAAAGTCTCTTGGAGTATGCACGCGGGTATGGGATGGAA CCGCTTATTTTGATTAATGACGAAAACGACCTTGACATTGCGCTTCGTATTGGCGCGCGATTTATT ACCATATATTCTATGAATTTCGAGACGGGTGAGATTAACAAAGAAAATCAACGTAAGCTGATATCT ATGATCCCCAGCAACGTCGTGAAGGTCCCCTTGTTAGATTTCTTCGAACCTAATGAGATTGAAGAA CTTCGTAAGCTCGGCGTTAATGCTTTCATGATATCTTCGAGTCTGATGCGAAATCCGGAGAAGATA AAGGAGCTGATTGAGGGGTCCCTGGAGCATCATCATCACCATCACTAA
RA95-Evolved-N110S	CCACGATACCTGAAGGGTTGGTTAGAGGACGTAGTCCAATTGTCATTACGTAGACCATCAGTTCAT GCTAGCAGACAACGCCCCATCATAAGTCTTAACGAGCGCATTTTAGAGTTCAACAAAAGAAACATC ACTGCGATCATTGCGTACTACCTTCGGAAGTCTCCATCGGGGTTAGATGTGGAACGTGACCCTATT GAGTATGCTAAGTACATGGAACCTTACGCCGTAGGACTTAGTATTAAGACTGAGGAGAAATATTTC GACGGATCCTACGAGATGCTCCGTAAGATAGCGTCGTCCGTGTCGATTCCGATCCTCATGTCTGAT TTCATCGTAAAGGAATCTCAGATTGATGACGCGTACAATCTCGGGGCAGACACCGTGCTTCTCATC GTGGAGATATTAACAGAGCGTGAGCTTGAGAGTCTTCTGGAGTACGCTAGAGGGTATGGAATGGAG CCCTTGATTCTGATTAACGACGAAAACGATTTGGACATCGCTCTGCGAATCGGAGCACGGTTCATC ACCATATATTCGATGAACTTTGAAACTGGTGAAATCAACAAGGAAAACCAGCGCAAGCTGATTTCT ATGATACCAAGCAACGTAGTAAAGGTTCCACTTCTCGATTTCTTTGAGCCGAATGAGATTGAGGAA CTGCGAAAGCTTGGGGTTAATGCATTTATGATCTCAAGCAGCCTGATGCGCAATCCCGAAAAGATA AAGGAACTGATCGAGGGGTCGCTGGAGCATCATCACCACCACCATTGA
RA95-Evolved-Y180F	CCACGCTATTTGAAGGGTTGGCTCGAAGATGTTGTGCAGCTGTCTCTGCGTCGTCCATCAGTGCAT GCTTCCCGTCAAAGACCTATCATCAGCCTGAATGAAAGAATATTGGAGTTCAATAAACGAAACATC ACGGCCATCATCGCATATTATCTGCGGAAGAGCCCTAGCGGGTTAGACGTGGAACGTGACCCAATT GAGTACGCAAAATACATGGAGCCGTACGCTGTGGGGCTGAGCATCAAGACTGAAGAGAAGTACTTT GATGGGTCGTATGAGATGTTGCGCAAAATCGCGTCGAGTGTTTCCATACCTATATTAATGAATGAT TTCATTGTAAAAGAGAGTCAAATCGACGACGCATACAACCTGGGTGCCGACACCGTGTTGTTGATT GTCGAAATTCTGACAGAGAGAGAGTTGGAATCTCTCCTTGAGTATGCAAGAGGGTACGGTATGGAG CCATTAATATTGATCAATGACGAAAACGACTTAGACATCGCGTTGCGCATAGGTGCCCGATTCATC ACCATCTTCTCTATGAATTTCGAAACAGGCGAGATAAATAAGGAAAACCAACGTAAGTTGATTAGC ATGATACCTTCCAATGTAGTAAAAGTGCCTTTACTGGACTTCTTCGAGCCAAACGAGATAGAAGAG CTGAGAAAACTTGGAGTTAATGCCTTCATGATAAGTTCCAGCCTGATGCGCAACCCAGAGAAAATA AAGGAGTTAATCGAGGGCAGTCTGGAGCATCATCATCACCACCATTAG
RA95-Evolved-Y51F-N110S	CCTCGTTACCTGAAAGGGTGGTTGGAAGACGTCGTGCAATTATCTCTGCGTCGCCCTAGTGTTCAT GCGTCTCGCCAGCGACCGATCATCTCATTAAATGAGCGGATTCTGGAATTCAACAAGCGTAATATT ACCGCAATCATCGCGTTTTACCTGCGTAAAAGCCCTTCAGGTTTGGATGTAGAGCGTGACCCTATT GAGTATGCGAAATACATGGAGCCGTATGCTGTTGGACTGAGTATTAAGACGGAAGAGAAGTATTTC GACGGATCGTACGAGATGTTACGTAAAATCGCCAGCAGTGTTTCGATCCCGATTCTTATGTCTGAC TTCATTGTTAAAGAGAGTCAAATCGACGACGCATATAACTTAGGCGCAGACACTGTACTTTTGATC GTTGAGATACTGACCGAGCGGGAATTAGAGAGCTTGCTGGAGTATGCAAGAGGGTACGGCATGGAA CCCCTTATTCTGATAAACGATGAGAACGACTTAGACATTGCGCTTCGCATCGGAGCGCGCTTCATT ACAATCTATAGTATGAACTTTGAGACAGGCGAAATTAACAAGGAGAACCAAAGAAAGCTTATCTCT ATGATCCCATCGAATGTAGTCAAGGTCCCATTACTGGATTTCTTTGAACCTAATGAAATAGAAGAA CTTCGAAAGCTTGGCGTTAACGCCTTCATGATATCGTCATCACTTATGCGCAATCCAGAGAAGATT AAAGAGCTCATTGAGGGCTCGTTGGAGCACCACCACCATCACCATTGA
RA95-Evolved-Y51F-Y180F	CCCAGATACCTCAAGGGCTGGTTGGAAGATGTGGTGCAGTTAAGCTTGCGACGGCCCAGTGTGCAC GCCTCTCGGCAACGCCCCATTATTTCACTTAATGAACGCATCTTGGAATTCAATAAGCGTAATATT ACCGCGATAATTGCCTTTTACCTTAGAAAATCGCCGTCAGGGTTAGATGTTGAGCGTGATCCTATT GAATATGCGAAATATATGGAACCATACGCAGTGGGTCTGAGTATTAAAACGGAAGAGAAATATTTT GATGGTAGCTACGAGATGCTCCGCAAAATTGCTTCCAGCGTGTCGATTCCGATACTCATGAACGAT TTTATCGTGAAAGAGAGTCAGATTGATGATGCGTATAACCTCGGGGCAGACACCGTGCTGCTTATT GTGGAGATCCTCACCGAACGTGAACTGGAATCTCTGTTGGAGTATGCCCGGGGTTACGGAATGGAA CCATTGATCTTAATCAATGACGAGAATGATCTGGATATAGCGCTTCGCATCGGCGCGCGTTTCATT

	ACGATCTTTAGCATGAATTTCGAGACTGGCGAAATTAATAAAGAGAACCAGCGCAAACTGATTTCT ATGATCCCAAGTAACGTGGTTAAAGTGCCATTGCTTGACTTCTTTGAGCCTAATGAGATAGAAGAA CTTCGGAAGCTGGGCGTTAACGCTTTTATGATCAGCAGCTCTCTGATGCGTAATCCAGAGAAAATT AAAGAGCTGATAGAAGGCTCACTGGAGCATCATCATCATCATCATTAA
RA95-Evolved-N110S-Y180F	CCTCGTTATTTGAAAGGTTGGCTCGAAGACGTCGTTCAGTTATCCCTGCGTCGCCCAAGCGTTCAC GCTTCCCGTCAGAGACCTATCATAAGCCTCAACGAGCGTATCCTTGAATTTAATAAACGGAATATT ACTGCGATAATTGCTTATTACTTAAGAAAGAGTCCAAGTGGATTAGACGTGGAGCGCGATCCCATT GAATATGCGAAGTACATGGAGCCCTATGCCGTGGGTTTATCCATTAAAACGGAAGAGAAATATTTC GATGGATCATACGAAATGCTCCGAAAGATTGCCTCAAGCGTCAGTATTCCGATTCTCATGTCGGAT TTCATCGTCAAGGAAAGCCAGATCGATGACGCTTATAACCTGGGTGCTGATACGGTGCTGTTGATC GTGGAGATTTTGACAGAGCGCGAACTCGAAAGCCTGCTGGAATACGCGAGAGGATATGGCATGGAA CCACTGATTCTCATTAACGATGAAAACGACCTTGACATAGCACTGCGTATCGGGGCACGTTTTATC ACAATATTTAGCATGAACTTTGAAACAGGTGAGATTAATAAGGAAAACCAGCGAAAGTTAATCTCC ATGATCCCGAGCAACGTTGTCAAAGTTCCTCTCTTGGACTTCTTCGAACCTAACGAAATCGAGGAA CTGCGAAAATTGGGCGTCAATGCGTTTATGATTAGTTCATCCTTAATGCGCAACCCTGAAAAGATC AAGGAGTTAATCGAAGGTAGTTTAGAACATCATCATCATCATCATTAA