Noncanonical Amino Acid Synthesis by Evolved Tryptophan Synthases

Thesis by

Patrick James Almhjell

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
(Defended October 11, 2022)

Patrick James Almhjell
ORCID: 0000-0003-0977-841X

ACKNOWLEDGEMENTS

No matter how fulfilling one's thesis work might be, it is an extraordinary physical and mental challenge to bring it to completion that requires kindness, support, and inspiration from many, many others. It is hard to do justice to the many people who have played this role during my time at Caltech, but I will attempt to do so here, and trust that I have better displayed my gratitude in one way or another in the event of any omissions.

First, to my advisor, Prof. Frances Arnold. I really could not have wished for a better mentor for my PhD. You allowed to me push my own boundaries while providing guidance and support, and taught me so much about science (and myself!) along the way. Thank you not only for sharing your expertise in all things scientific, but also for sharing your love of the outdoors, knowledge of auto repair terminology, and worldview with me over these past five years. You are an inspiration.

Thank you as well to the rest of my thesis committee: Prof. Kaihang Wang, Prof. Greg Fu, and Prof. Steve Mayo. Kaihang is infectiously excited about science in every conversation I've had with him, motivating those around him. Greg was very supportive at each milestone in my PhD and went out of his way to compliment my work, which was greatly encouraging for a young (and old) PhD student. Steve is an inspirational scientist and his words have always stuck with me, from my interview at Caltech to the end of my time here.

I have had the pleasure of overlapping with many wonderful people in the Arnold Lab who have shaped the way I look at science and at life. First, to the team TrpB folks: my mentor, Tina Boville, brought me up to speed in the lab and taught me a lot about balancing work and life, even as I tried my hardest not to do that (like coming to art night to work on slides because it was "kind of like art"—I have since learned to be better). David Romney taught me how to do the chemistry required for amino acid synthesis, and even though I could never match his wizardry, I would not be where I am without his mentorship. Ella Watkins-Dulaney was the best TrpB co-conspirator I could have asked to work with. And although I never overlapped with Andy Buller in the lab, his conversation with me at Lucky Baldwin's pub during my grad school interview weekend was the point that I knew I needed to come to Caltech. I am happy we were able to continue discussing TrpB and hype up PLP-dependent enzymes for ncAA synthesis at ACS Spring 2022.

Office 335—David Miller, Shilong Gao, Jenni Kennemur, Tyler Fulton, and Kadina Johnston—made lab feel like a home away from home. Anders Knight, Nat Goldberg, Nicholas Porter, and Bruce Wittmann, thank you for everything. I could name others, but at some point I am just naming everyone I've worked with—they've all had an incredible impact on my life!

My family has been very important in this journey. In particular, my grandparents, Barb and Whitey Almhjell, played a very important role in making me who I am today. On Lake Roosevelt and in their auto repair shop, I experienced the joys of learning and teaching. Not only did I learn about cars at the shop, I learned how to teach people about *their* cars, and how to translate odd car noises over the phone into real diagnoses. At Roosevelt, I learned about the diverse flora and fauna of the desert, how to water ski and wakeboard, and even how to feed wild javelina by hand. It's funny to think that the skills I learned from these experiences would one day help get our weird last name cited in a Nobel Prize release!

To my mom, Embla: thank you for your love and support, and especially for trusting me when I said that I wanted you to keep giving me updates even when I don't respond for days because I'm a bad son (and/or busy). And to my two step-parents who never made me feel like a step-son: thank you, Jeff, for many things, but perhaps most importantly for teaching me to not take myself so seriously! And thank you, Sheilah, for teaching me how to have fun, even when I'm working hard. To my siblings, Sadie, Macie, Eldon, and Rainy: you four mean the world to me. Thank you for understanding when I'm busy, and for being the best little sibs a big brother could ask for; I can't wait to see where your lives take you!

Thank you to Prof. Jeremy Mills, my undergraduate advisor. Jeremy trusted me and gave me extraordinary freedom in the lab—perhaps more than he should given a young undergraduate. Nonetheless, it gave me the opportunity to learn how to work in the lab, to read the literature, to solve problems, and really started this journey. Jeremy's support and belief in me was invaluable and cannot be appreciated enough.

My friends at Caltech not only kept me in good mental health during my PhD, but made it the best five years of my life. To Ella and Austin: I really don't know what I would have done without you two. From supporting me when various things break, to climbing and snowboarding and camping, to hanging out on the couch and giggling about things like "yee yee", you two were almost always a part of my most cherished moments. Thank you for everything. To Nicholas Sarai: thank you for facilitating some of the coolest experiences of my life, and teaching me there are few good reasons to miss a big storm. Your land-speed record on skis, your strength, and your scientific prowess are deeply enviable, and I am glad to be your friend.

Finally, to my partner Kadina. It would have been irresponsible to have expected (or even wished) that a relationship could be as fulfilling and easy as ours. That's a hard ask—but I guess the hardest part is finding somebody like you! You are the kindest person I know. You are the most driven and competent. You are intelligent, patient, creative, and fun. You push me to do new things, and you push me to do the same things but better. You complete me, and I am so thankful to have had you and our (good but mostly bad) dog Sierra during this part of my life. I can't wait for what's to come.

ABSTRACT

The β-subunit of tryptophan synthase (TrpB) is responsible for the final step of L-tryptophan biosynthesis in all of known biology. Recognized for this important role and its powerful chemistry, TrpB has more recently been used for the *in vitro* synthesis of tryptophan analogs and other noncanonical amino acids (ncAAs). This thesis describes some of these efforts as well as the application of TrpB for developing new methods in directed enzyme evolution. Chapter I first establishes important topical background. It begins with a general account of directed enzyme evolution by exploring the emergence of new catalytic functions in natural and laboratory settings, and how this information and chemical intuition can be used to create enzymes for desired reactions. This is followed by a description of the state of the field of ncAA synthesis, with a special focus on biocatalytic approaches using engineered enzymes. Chapters II and III examine targeted engineering campaigns to create enzymes that can efficiently synthesize valuable blue-fluorescent ncAAs such as 4-cyanotryptophan (Chapter II) and β-(1-azulenyl)-L-alanine (AzAla, Chapter III) from accessible starting materials. In Chapter IV, the native function of TrpB for L-tryptophan biosynthesis is used as a selection pressure to develop an *in vivo* continuous evolution system. Despite a selection pressure for only L-tryptophan synthesis the orthologous TrpB variants generated by this system have varying promiscuous activities for L-tryptophan analogs, paralleling the sequence-function diversity of natural enzyme homologs. Chapter V describes evSeq, an inexpensive and simple method for sequencing all protein variants generated during an engineering campaign, demonstrated by collecting ~800 TrpB sequence-function data points. Finally, in Chapter VI, directed evolution and chemical intuition are used to convert TrpB from a tryptophan synthase to a novel tyrosine synthase (TyrS). This enzyme can irreversibly and regioselectively alkylate simple phenol analogs to synthesize valuable tyrosine analogs, including the blue-fluorescent ncAA β-(1-naphthol-4-yl)-L-alanine (NaphAla) and 3-methyl-L-tyrosine at gram scales. Because TyrS synthesizes a primary metabolite, this transformation represents a noncanonical method for the biosynthesis of L-tyrosine. This is the first example of a feasible new route for *de novo* aromatic amino acid biosynthesis, which occurs through a universally conserved set of chemistry across all of life. In total, the work described here expands the fields of chemical synthesis and synthetic biology by presenting new enzymes—and methods for producing these enzymes—that are capable of synthesizing important amino acids *in vitro* and *in vivo*.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. **Almhjell, P. J.** & Arnold, F. H. Creating new enzymes with evolution and intuition. *Manuscript submitted*.

P.J.A. prepared the manuscript and all figures.

2. **Almhjell, P. J.**, Boville, C. E. & Arnold, F. H. Engineering enzymes for noncanonical amino acid synthesis. *Chemical Society Reviews* **47**, 8980–8997 (2018). doi: 10.1039/c8cs00665b

P.J.A. and C.E.B. prepared the manuscript and all figures. P.J.A. and C.E.B. conceived and prepared journal cover art.

3. Boville, C. E., Romney, D. K., **Almhjell, P. J.**, Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *The Journal of Organic Chemistry* **83**, 7447–7452 (2018) doi: 10.1021/acs.joc.8b00517

C.E.B. and D.K.R. participated in project conception. C.E.B., D.K.R., P.J.A., and M.S. designed and executed research. P.J.A. hypothesized and confirmed improved enzyme activity at lower temperatures. D.K.R. drafted the manuscript with input from all authors.

4. Watkins, E. J.,[†] **Almhjell, P. J.**,[†] & Arnold, F. H. Direct enzymatic synthesis of a deep-blue fluorescent noncanonical amino acid from azulene and serine. *ChemBioChem* **21**, 80–83 (2020) doi: 10.1002/cbic.201900497 ([†]denotes equal contribution)

P.J.A. and E.J.W. participated in the conception, design, and execution of the research. E.J.W. designed the screen for identifying improved enzyme variants. P.J.A and E.J.W. contributed equally to AzAla purification. P.J.A. performed enzyme kinetics. E.J.W. prepared the first manuscript and P.J.A. edited and constructed figures.

5. Rix, G., Watkins-Dulaney, E. J., **Almhjell, P. J.**, Boville, C. E., Arnold, F. H. & Liu, C. C. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities." *Nature Communications* **11**, 5644 (2020). doi: 10.1038/s41467-020-19539-6

All authors contributed to experimental design and data analysis. G.R. constructed the selection and performed all evolution experiments. P.J.A. analyzed results and performed enzyme kinetics. E.J.W.D. performed the panel HPLC-MS assays and indole conversion rate measurements on *Tm*TrpBs. C.E.B. performed *in vitro* characterizations of *Tm*TrpBs from variant set 1 and performed the thermal shift assay and substrate scope characterizations. G.R. and C.C.L. wrote the manuscript with input and contributions from all authors.

6. Wittmann, B. J., Johnston, K. E., **Almhjell, P. J.** & Arnold, F.H. evSeq: Cost-effective amplicon sequencing of every variant in a protein library. *ACS Synthetic Biology* **11**, 1313–1324 (2022). doi: 10.1021/acssynbio.1c00592

B.J.W. conceived of the project and performed initial design and execution of research and software development. B.J.W., K.E.J, and P.J.A. optimized the experimental workflow and software. K.E.J. wrote software for data visualization. P.J.A. optimized the graphical user interface installation and operation and constructed online documentation. B.J.W., K.E.J., and P.J.A. wrote the manuscript and prepared figures.

7. **Almhjell, P. J.**, Johnston, K. E., Porter, N. J., Kennemur, J. L., Ducharme, J. & Arnold, F. H. Directed evolution of tryptophan synthase for noncanonical L-tyrosine synthesis. *Manuscript submitted*.

P.J.A. conceived of the project, designed and executed research and wrote the manuscript with input from K.E.J., N.J.P., and J.L.K. Enzyme kinetics were performed by K.E.J., along with auxotroph complementation experiments. X-ray crystallography was performed by N.J.P, who also interpreted results. J.L.K. performed preparative-scale syntheses. J.D. assisted with rate measurements. All authors edited the manuscript.

PUBLISHED CONTENT NOT INCLUDED IN THESIS

8. Molina, R. S., Rix, G., Mengiste, A. A., Álvarez, B., Seo, D., Chen, H., Hurtado, J. E., Zhang, Q., García-García, J. D., Heins, Z. J., **Almhjell, P. J.**, Arnold, F. H., Khalil, A. S., Hanson, A. D., Dueber, J. E., Schaffer, D. V., Chen, F., Kim, S, Fernández, L. Á., Shoulders, M. D. & Liu, C. C. "*In vivo* hypermutation and continuous evolution." *Nature Reviews Methods Primers* **2**, 36 (2022) doi: 10.1038/s43586-022-00119-5

R.S.M., G.R., and C.C.L. drafted the manuscript with input and contributions from all authors.

TABLE OF CONTENTS

LIST OF FIGURES, TABLES, AND SCHEMES

Figure Number *Page*

Table Number *Page*

Scheme Number *Page*

ABBREVIATIONS

Å	Ångstrom
ACN	acetonitrile
AcOH	acetic acid
ADC	antibody-drug conjugates
AmpR	ampicillin resistance
aroAA	aromatic amino acid
AzAla	β-(1-azulenyl)-L-alanine
BCA	bicinchoninic acid assay
carb	carbenicillin
*Cf*TPL	TPL from *Citrobacter freundii*
COMM domain	communication domain of TrpB
CV	column volume
DE	directed evolution
DO	drop-out (media)
E(A-A)	enzyme amino-acrylate
$E(A_{in})$	enzyme internal aldimine
ee	enantiomeric excess
EIC	extracted ion count
epPCR	error-prone PCR
GOI	gene of interest
h	hour(s)
HPLC	high performance liquid chromatography
HPLC-MS	high performance liquid chromatography coupled mass spectrometry
IGP	indole-3-glycerol phosphate
IPTG	isopropyl β-D-thiogalactopyranoside
k	rate constant
kan	kanamycin
k_{cat}	maximum catalytic rate
K_M	Michaelis constant
KPi	potassium phosphate buffer (50 mM, pH 8.0)
LB	Lysogeny Broth (Luria-Bertani medium)
LCMS	liquid chromatography coupled mass spectrometry
L-DOPA	dihydroxy-L-phenylalanine
log	logarithmic or logarithm
M9–Tyr	M9 Minimal Media lacking Tyr
min	minute(s)
MSA	multiple sequence alignment
NaphAla	β-(1-naphthol-4-yl)-L-alanine
ncAA	noncanonical amino acid
NGS	next-generation sequencing
OD_{600}	optical density at 600 nm
PCR	polymerase chain reaction
PDB	Protein Data Bank

PET	polyethylene terephthalate
Pf	*Pyrococcus furiosus*
*Pf*TrpB	TrpB from *Pyrococcus furiosus*
Phe	L-phenylalanine
pK_a	acid dissociation constant
PLP	pyridoxal 5’-phosphate
RBF	round bottom flask
*Rma*NOD	nitric oxide dioxygenase from *Rhodothermus marinus*
RT	room temperature or retention time
Ser	L-serine
SIM	single-ion mode
StEP PCR	staggered extension process PCR
*St*TrpB	TrpB from *Salmonella typhimurium*
SSM	site saturation mutagenesis
TB	Terrific Broth
Thr	L-threonine
TIC	total ion count
T_m	melting temperature
Tm	*Thermotoga maritima*
*Tm*TrpB	TrpB from *Thermotoga maritima*
TOF	turnover frequency
TPL	Tyrosine phenol lyase
Trp	L-tryptophan
TrpA	Tryptophan synthase α-subunit
TrpB	Tryptophan synthase β-subunit
TrpS	Tryptophan synthase αββα dimeric complex
Tyr	L-tyrosine
TyrS	Tyrosine synthase
TTN	total turnover number
UV-vis	ultraviolet visible
V_max	maximum rate
V_0	initial velocity

Chapter I

USING CHEMICAL INTUITION AND EVOLUTIONARY INSIGHT TO CREATE NEW ENZYMES FOR NONCANONICAL AMINO ACID SYNTHESIS

Material from this chapter appears in: “**Almhjell P. J.** & Arnold F. H., Creating new enzymes with evolution and intuition. *Manuscript submitted*” and “**Almhjell P. J.**, Boville, C.E. & Arnold F. H., Engineering enzymes for noncanonical amino acid synthesis. *Chemical Society Reviews* **47**, 8980–8997 (2018). doi: 10.1039/c8cs00665b”.

ABSTRACT

The standard proteinogenic amino acids grant access to myriad chemistries that harmonize to create life. Outside of this limited set of building blocks are countless “noncanonical” amino acids (ncAAs), either found in nature or created by man. Interest in ncAAs has grown as research has unveiled their importance as precursors to natural products and pharmaceuticals, biological probes, and more. Despite their broad applications, synthesis of ncAAs remains a challenge, as poor stereoselectivity and low functional-group compatibility stymie effective preparative routes. The use of enzymes has emerged as a versatile approach to prepare ncAAs, and nature's enzymes can be engineered to synthesize ncAAs more efficiently and expand the amino acid alphabet. This chapter describes how enzymes can be successfully engineered for such a function through a deep understanding of chemical and biological principles. By gaining insight into how new catalytic functions have emerged in natural and laboratory settings, this information can be coupled with chemical intuition to rapidly create and optimize enzymes for desired reactions. Once these general concepts are established, the importance of ncAAs and the difficulties associated with their synthesis are examined in more depth. These concepts demonstrate the critical role of engineered enzymes for the synthesis of a growing collection of valuable ncAAs with applications in academic labs, industrial settings, and even within organisms.

1.1 Creating New Enzymes with Evolution and Intuition

1.1.1 The emergence of new enzyme function

Enzymes are nature's catalysts, performing the chemical reactions that living organisms use to extract materials and energy from their environments and create new life. Consider how sunlight and CO_2 are converted into sugars by plants, and how those sugars are later metabolized by the organisms that rely on plants for food. These reactions are all catalyzed by enzymes. Even the conversion of solubilized CO_2 in the blood (in the form of carbonate) to gaseous CO_2 that is exhaled and the metabolism of different pharmaceuticals to remove them from a human circulatory system are performed by these remarkable biological machines. Guided by evolution, enzymes have been optimized for their particular reactions and can exhibit extraordinary rate enhancements and selectivities compared to uncatalyzed reactions and other catalysts.

Enzymes are made of linear arrangements of the 20 canonical amino acids that fold into a complex three-dimensional structure. They often bind different cofactors to facilitate their reactions and are sometimes further modified by post-translational modifications which alter their structures and functions. Optimization of enzymes via evolution occurs due to the action of two main events: mutations occur in the DNA encoding the enzyme, followed by selection on its function, or *phenotype*. Mutations that improve function can become enriched in the population. This process enables organisms to co-opt and consume the resources around them; those that do this better have a better chance at passing their genes to the next generation. This process has also given rise to the great diversity of life that has adapted to occupy nearly every niche of the world.

The powerful design algorithm of evolution has been brought into the laboratory in the form of directed evolution, which can be used to engineer enzymes and other proteins to have properties useful for addressing human needs. Analogous to natural evolution, directed evolution uses mutagenesis, a method of creating genetic diversity, followed by some form of selection or screen to identify variant enzymes with improved properties. This cycle can be repeated until the property is sufficiently improved (**Figure 1-1a**). For the most part, directed evolution is conceptually and technically straightforward: once an enzyme is identified that displays a low level of activity for a desired function, mutagenesis and screening for improved function will often provide enhancements. However, that first step, the identification of enzymes with some initial activity, can be far from straightforward. Both in nature and in the laboratory, we often find ourselves wondering how exactly new functions arose over the course of natural evolution.

Similar to the tenet of cellular biology that "*omnis cellula e cellula*"—or "all cells come from cells"—enzymes are derived from earlier enzymes and other proteins, acquiring new functions over the course of evolution. What circumstances allow an enzyme that performs one function to evolve into another with a different function? This question is not completely answered, but this process has generally been observed to occur incrementally in nature, through small changes and chance—the right change or the right set of new conditions—over a long time (**Figure 1-1b**). We enzyme engineers, however, wish to create enzymes that address current, time-sensitive problems and thus do not have the luxury of waiting "evolutionary timescales" before a solution appears and is optimized. Furthermore, we might want to go in a different direction and create enzymes that serve us rather than support the organism that makes it. How, then, can we make faster, more directed jumps toward new catalytic activities? As we will see in this chapter, it is often about using one’s chemical intuition, the *how* and *why* a reaction might happen, and making the right changes under the right set of conditions.

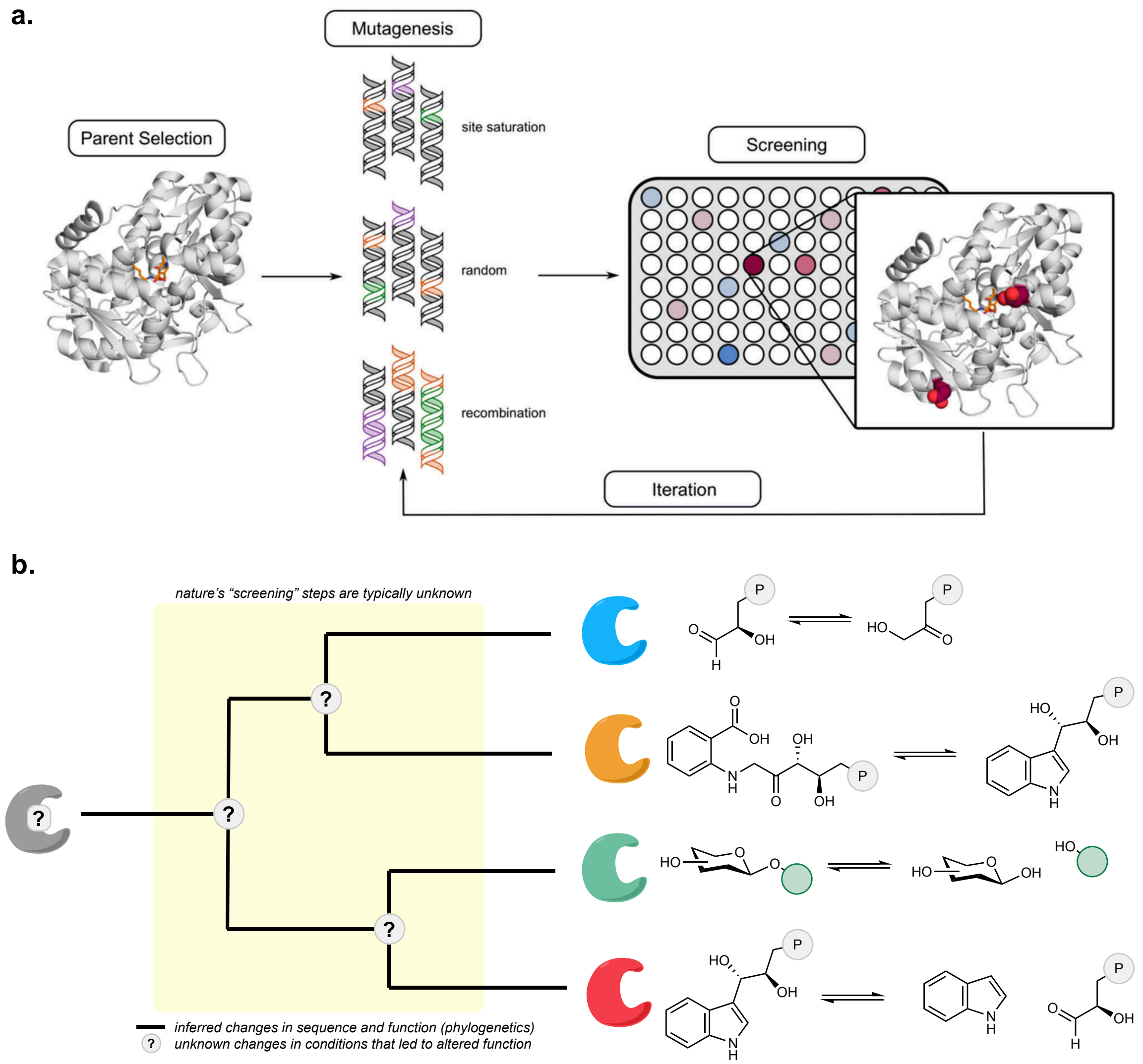

a.
Mutagenesis
Parent Selection
site saturation
random
recombination
Screening
Iteration
b.
nature’s “screening” steps are typically unknown
?
?
?
?
P
O
H
OH
HO
O
HO
P
O
OH
OH
N
H
O
OH
P
OH
N
H
HO
O
O
HO
HO
O
OH
HO
HO
P
OH
N
H
N
H
P
O
H
OH
inferred changes in sequence and function (phylogenetics)
unknown changes in conditions that led to altered function

Figure 1-1. Evolution in the laboratory and in nature. a. The process of directed enzyme evolution starts with finding a “parent” enzyme with some initial activity for the desired reaction. Mutations are made to the gene encoding this enzyme via mutagenesis techniques, and then these enzyme variants are screened for improved activity. An improved variant or variants can then be subjected to additional rounds of mutagenesis and screening until a sufficiently active enzyme is created. **b.** Natural evolution is often visualized with phylogenetic trees that map the branching changes in protein sequence from a progenitor. While excellent at representing speciation and diversification events, it does not tell us about the specific conditions that led to the observed diversification of enzyme functions.

1.1.2 Evolution is not just a thing of the past—it's happening right now

First, it is important to note that not all natural evolution requires millennia. Given a sufficiently strong selection pressure or advantage, enzymes and even entire organisms can adapt to a new environment in a matter of years or even months. Examples of rapid evolution can be attributed to changes in single enzymes within an organism, such as the appearance of resistance to an antibiotic or the emergence of bacterial strains which can consume plastic, a consequence of the plastic waste permeating our ecosystems.

Plastics, such as polyethylene terephthalate (PET), are organic polymers that were nonexistent in nature until recently (**Figure 1-2a**). They often contain very stable bonds that are uncommon—or at least unusually decorated—compared to those found in biology, which makes them difficult for enzymes to access and break down. This prevents organisms from converting these polymers into digestible monomeric units that can be used as a carbon source. Furthermore, some of these polymers assemble into materials with high crystallinity at all but extremely high temperatures, which reduces their accessibility to enzymes in the first place. These qualities mean that plastics accumulate in the environment. Nonetheless, other materials that *are* biodegradable share some of these qualities (**Figure 1-2b**). Wood, comprised of crystalline cellulose and highly cross-linked lignin polymers, is an example of a material that is difficult to break down into monomers, yet it is obviously biodegradable. Only a select few organisms, however, can efficiently perform this function, such as fungi and microbes in the guts of termites. Is plastic, a product of human engineering, fundamentally different from biological materials like cellulose? Or have organisms simply not had sufficient time to adapt to these new carbon sources?

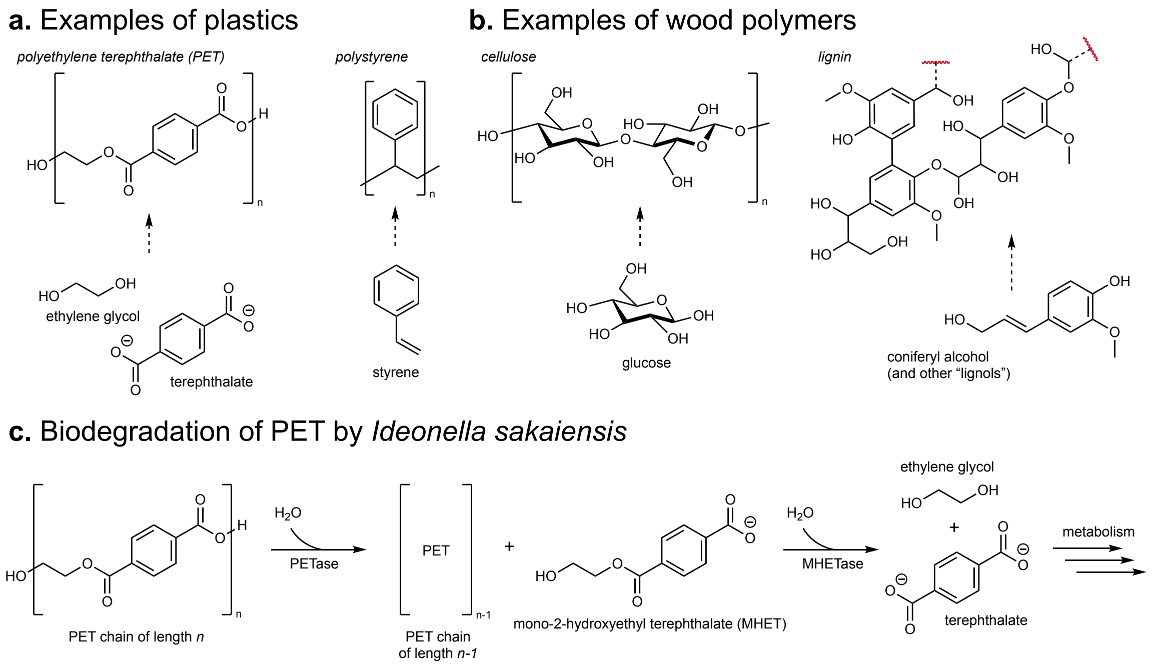


Figure 1-2. Organic polymers are potentially rich carbon sources if they can be broken down. a. Representative structures and monomeric units of plastics. **b.** Representative structures and monomeric units of the polymers found in wood. **c.** The bacterium *Ideonella sakaiensis*, recently isolated from a plastic recycling plant, can degrade PET into pieces that it can use as carbon sources.

Recently, researchers identified a strain of bacteria named *Ideonella sakaiensis*, isolated from a plastic bottle recycling facility, that is capable of hydrolyzing PET and metabolizing it as its primary carbon source to sustain growth (**Figure 1-2c**). There are two enzymes involved in this process, one that breaks down the PET polymer (a PETase) into its mono-(2-hydroxyethyl)terephthalate (MHET) monomers and one that then breaks down MHET (a MHETase) into two pieces. These pieces—ethylene glycol and terephthalate—can then enter a cell's metabolic cycle as carbon sources. While these new enzymes perform their functions well enough to sustain the growth of *I. sakaiensis,* they are slow compared to enzymes that perform other, similar hydrolysis reactions. Therefore, there is likely still room for improvement via natural or laboratory evolution, presenting an exciting opportunity for laboratory evolution to "compete" with natural evolution.

1.1.3 Directed evolution readily re-optimizes enzymes

Some of the most successful examples of directed laboratory evolution have involved asking an enzyme to perform its native function in a non-native way, such as under harsher conditions or on a substrate that is not its natural substrate. The enzyme carbonic anhydrase rapidly catalyzes the interconversion of CO_2, a poorly soluble gas, and bicarbonate, a highly soluble salt (**Figure 1-3a**). In fact, this reaction happens so fast—up to 1 million times per second—that it is limited only by the rate of diffusion of the substrate to the enzyme, which has afforded carbonic anhydrase the title of a "perfect" enzyme. Engineering the enzyme to perform the same transformation but for longer periods in hot and highly alkaline environments has enabled its application within carbon capture systems. These systems typically use a basic solvent, such as a solution of an amine, that helps absorb CO_2 and retain it as bicarbonate within water. The solvent is then cycled to a different chamber at a higher temperature to facilitate the desorption of bicarbonate from the solvent, releasing a stream of pure CO_2 that can then be captured and stored. Unfortunately, needing to use the same amine for absorption and desorption presents a problem: amines that are faster at absorbing CO_2 are typically slower at releasing it, requiring even higher temperatures and more energy input to make these systems viable for carbon capture. With the use of a prodigious catalyst like carbonic anhydrase to speed up the interconversion of CO_2 and bicarbonate, the kinetics of amine absorption become less important. Amines with lower desorption temperatures could thus be coupled with a catalyst to create a better system (**Figure 1-3b**). After engineering carbonic anhydrase to withstand the conditions of a carbon capture system—a pH of 10.0 with cycling between 25 °C for absorption and 87 °C for desorption—the catalyzed system captured roughly two-thirds of all CO_2 released from a power plant flue, up to 25-fold more CO_2 than without the enzyme, with no noticeable loss in enzyme activity over 60 hours.

A striking example of engineering an enzyme to perform on a substrate analog is found in the directed evolution of a transaminase to synthesize sitagliptin, an anti-diabetes drug that contains a chiral amine (**Figure 1-3c**). Transaminases are a broad class of enzymes that catalyze the interconversion of ketones and amines. While ketones lack chirality, the conversion of a ketone to an amine sets a new chiral center, which enzymes can perform with exquisite selectivity to afford only a single enantiomer. However, no transaminase had been known to work on a molecule as large and complex as prositagliptin, the achiral precursor that could be transaminated to sitagliptin. This was overcome by a clever strategy called substrate walking in which an enzyme is iteratively adapted to substrate analogs progressively more similar to the final target. Protein engineers were first able to increase the activity of a transaminase toward a fragment of prositagliptin that was closer in structure to the native substrate of the enzyme (**Figure 1-3d**). Once activity was improved toward this substrate, the enzyme then also displayed some activity toward the full prositagliptin molecule. This is called substrate promiscuity, where an enzyme can react with multiple substrates rather than just the one for which it was evolved, usually to a lower degree (**Figure 1-3d, purple region**). Once it was possible to reliably screen for activity on the desired prositagliptin substrate, the transaminase could be evolved to yield a nearly perfectly enantioselective and highly efficient catalyst for the synthesis of sitagliptin at industrial scales.

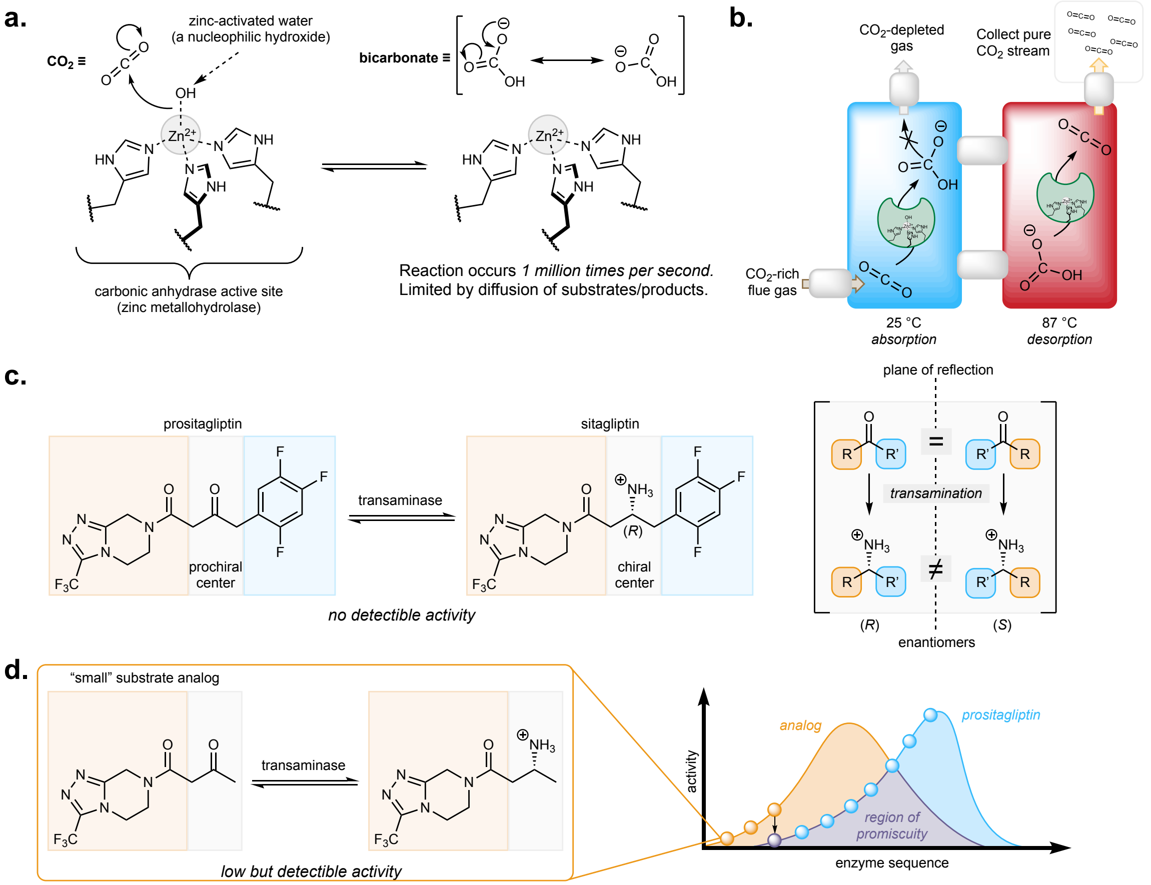
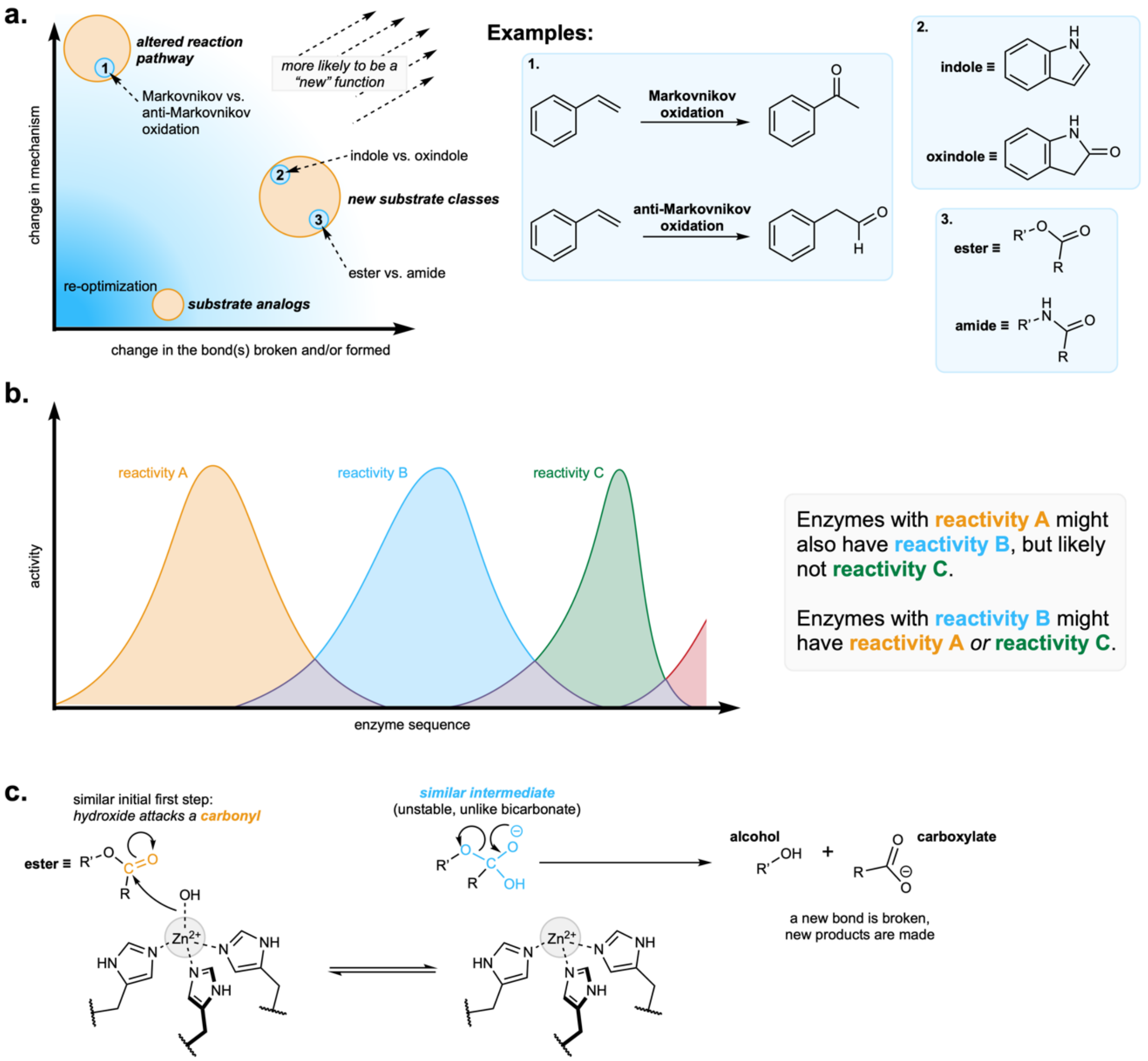


Figure 1-3. Re-optimizing enzymes to new conditions and substrate analogs. a. The reaction of carbonic anhydrase, a zinc metallohydrolase, which interconverts CO_2 and bicarbonate. **b.** Schematic of a carbonic anhydrase-catalyzed carbon capture system. **c.** Sitagliptin synthesis requires the installation of a chiral amine, which can theoretically be accomplished by a transaminase in a highly selective fashion, but no transaminase was identified to act on the precursor prositagliptin. **d.** A smaller prositagliptin analog was accepted by a transaminase (initial orange sphere), which could be evolved until an enzyme variant was identified with activity toward the full prositagliptin molecule (purple sphere). This could then be evolved into an industrially useful catalyst (blue spheres). This process is called a "substrate walk" and relies on substrate promiscuity—the ability of an enzyme to perform its reaction with a different substrate—and the selection of appropriate intermediate substrates.

1.1.4 But how can we evolve enzymes to perform *new* functions?

What qualifies as a "new" catalytic function is somewhat arbitrary. The "function" of an enzyme is often equated to the reaction it performs. However, reactions are distinguished not just by the specific products that are made, but also by the character of the transition states and intermediates, referred to as the catalytic "mechanism". Two reactions that make the same product but proceed through very different mechanisms might be considered to be more different than two reactions that create different products but do so through an identical mechanism. We will consider the two primary variables of an enzyme's function to be the specific mechanism and the nature of the bonds being broken and/or formed (**Figure 1-4a**). In the previous examples, the evolved enzymes still used their native mechanisms: the carbonic anhydrase was adapted to a new environment whereas the transaminase was adapted toward a more decorated substrate with the same reactive moiety—the same bonds being broken and formed. Here we will consider an enzyme to have a *new* catalytic function if it is working on a different substrate class or reactive moiety—even if the reaction is mechanistically similar—or if it is reacting through a completely new mechanism. Through chemical intuition and an in-depth understanding of a reaction, we can often visualize how an enzyme could in principle perform a whole new reaction.

As with an enzyme catalyzing the same reaction on different substrates, we say an enzyme is catalytically promiscuous when it can catalyze multiple different reactions (**Figure 1-4b**). This concept is often used to explain an observed reaction. Take, for example, the fact that carbonic anhydrase has promiscuous esterase activity, the ability to hydrolyze an ester into an alcohol and carboxylic acid (**Figure 1-4c**). We can intuit how this reaction occurs in a similar way to the native mechanism, but with a new bond-breaking step from the carbonate-like intermediate (compare **Figure 1-3a** to **Figure 1-4c**). However, when looking to create a new enzyme for a reaction that is not known to be catalyzed by an enzyme, we instead must hypothesize how a *new* reaction might be catalyzed by an *existing* enzyme *a priori* and provide the appropriate substrates (and sometimes even cofactors). This process can guide us to identify the initial activity required to begin directed evolution of a new enzyme function. The final section will discuss how this approach has been used in the laboratory evolution of tryptophan synthase.



Enzymes with reactivity A might also have reactivity B, but likely not reactivity C.

Enzymes with reactivity B might have reactivity A *or* reactivity C.

Figure 1-4. The appearance of new enzyme functions. a. A two-dimensional representation of enzyme functional differences, with the two primary axes being the character of the bonds that are broken and formed and the nature of the reaction mechanism. Specific examples are highlighted. **b.** A cartoon representation of catalytic promiscuity, the ability of an enzyme with a given reactivity to have (or not have) other reactivities. Such reactivities can be enhanced and tuned by directed evolution. **c.** Likely mechanisms of the promiscuous esterase activity of carbonic anhydrase, analogous to (but quite different from) its native reactivity.

1.1.5 Tryptophan synthase: a powerful biocatalyst for amino acid synthesis

Tryptophan synthase (TrpS) is an ancient enzyme responsible for the final two steps in the biosynthesis of L-tryptophan (Trp), one of the twenty amino acids used in protein synthesis. It is present in all organisms with the exception of animals, which instead must obtain it in some other way (it is *essential*), such as through diet or with the help of gut microbes. TrpS is commonly found as a heterodimeric complex of two subunits, the α-subunit (TrpA) and the β-subunit (TrpB), which are in some cases tethered together in a single protein chain. This close proximity is important. TrpA converts indole glycerol phosphate (IGP) into glyceraldehyde 3-phosphate (a three-carbon sugar that re-enters metabolism) and indole, a small and somewhat toxic molecule that can readily escape the cell or cause a stress response within (**Figure 1-5a**; TrpA). Escaping the cell would result in loss of the resources that went into making IGP in the first place, and a stress response is obviously undesirable. Before indole can do either of these things, however, it is quickly shuttled into the adjacent active site of TrpB where it is immediately converted to Trp.

Two key features contribute to the efficiency of this process. The first is that TrpB uses its pyridoxal 5'-phosphate (PLP) cofactor to convert L-serine (Ser) into a highly electrophilic intermediate—the amino-acrylate—that is poised to react with the nucleophilic indole substrate as soon as they meet, proceeding via electrophilic aromatic substitution (**Figure 1-5a**; TrpB). This intermediate can be degraded through a different reaction pathway, but this side activity is reduced by the second key feature: allostery. Allostery is a process of intermolecular interaction and communication in which a small molecule or protein affects the activity of another protein. In TrpS, the functions of TrpA and TrpB are each coordinated through different conformational states between the two subunits that occur during their reactions. Upon binding IGP, TrpA signals TrpB to convert Ser to the amino-acrylate; once done, TrpB signals TrpA to produce indole. The amino-acrylate intermediate can, theoretically, react with any other nucleophile that it might encounter, but it is sequestered in time and space to only appear when and where indole is available. Thus, the highly reactive intermediate and allosterically controlled structural changes work together to meet the biological demands of synthesizing Trp quickly and with minimal loss of resources.

1.1.6 Can we change the function of TrpB to make very different amino acids?

Outside of the demands of biology, however, one can envision TrpB to be capable of using its catalytic machinery for other reactions. The amino-acrylate is a potent electrophile, able to react with numerous different nucleophilic species. However, it is only given the chance to encounter indole in its native environment. When placed in an *in vitro* setting and engineered to favor the formation of the amino-acrylate without relying on allosteric signals from TrpA, TrpB became a highly general amino acid synthase that can take on a broad array of new catalytic activities (**Figure 1-5b**). Provided a competent nucleophile, even those significantly different from indole that undergo C–C bond formation via a different mechanism, variants of this enzyme can produce numerous other *noncanonical* amino acids (ncAAs) that are not part of the twenty used in protein synthesis but are important pharmaceuticals, building blocks, and biological probes. These important molecules are the focus of the second part of this chapter, **Section 1.2**.

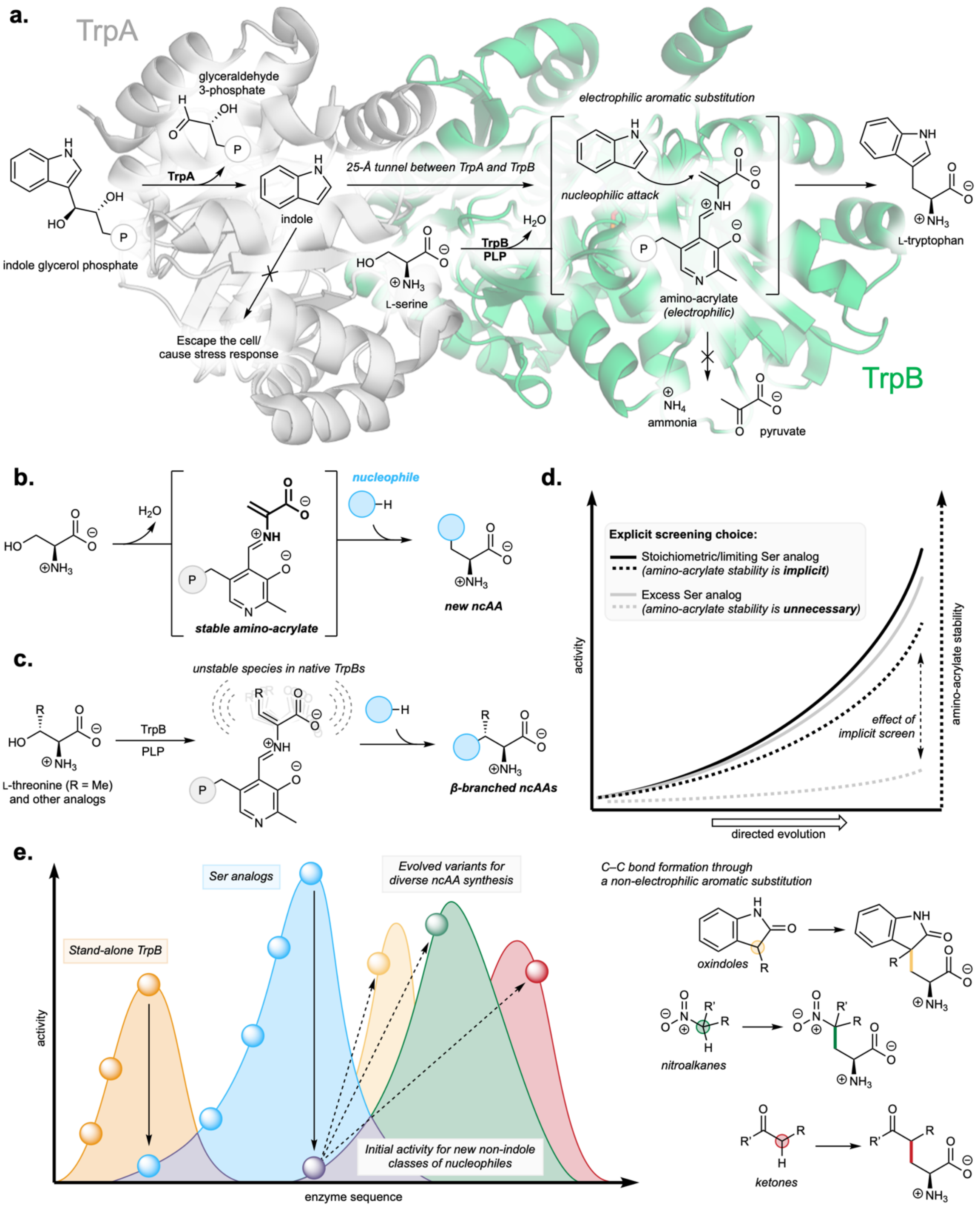


Figure 1-5. Directed evolution of tryptophan synthase. a. Tryptophan synthase (TrpS) converts indole glycerol phosphate (IGP) to indole and glyceraldehyde 3-phosphate within its TrpA subunit, then couples indole and L-serine (Ser) to make L-tryptophan in its TrpB subunit, avoiding non-productive outcomes. **b.** TrpB can work as a general noncanonical amino acid (ncAA) synthase, coupling a nucleophile with the electrophilic amino-acrylate intermediate. **c.** Ser analogs like L-threonine can also be used to generate an amino-acrylate analog, but directed evolution was required to stabilize this intermediate. **d.** The amino-acrylate could be implicitly stabilized by screening under conditions where the Ser analog was provided … *(continued on the following page)*

(continued from the previous page) … in stoichiometric (or limiting) quantities. The hypothetical curves demonstrate how using excess Ser might still improve activity with a given nucleophile to the same extent as limiting Ser but would not stabilize the amino-acrylate as much. **e.** Fitness landscapes for various activities, showing how improving amino-acrylate stability might then unlock new activities with less reactive nucleophiles that proceed through a different, non-native bond-forming mechanism.

Many of these new catalytic activities—particularly those providing non-tryptophan ncAAs—were not detectible, however, with the native enzyme. They needed to be coaxed out by directed evolution. Protein engineers applied the two concepts described above: they optimized the TrpB subunit to new conditions and adapted it to new substrates. Directed evolution to improve native Trp synthesis under new *in vitro* conditions meant that the reaction took place in the absence of TrpA, with indole already in the reaction mixture, and with TrpB no longer having to discriminate against other nucleophiles as indole was the only one. The enzyme also did not have to discriminate against Ser analogs and could use L-threonine (Thr) and other Ser analogs to generate amino-acrylate analogs (**Figure 1-5c**). However, this raised another potential issue: the amino-acrylate could still be degraded, and its analogs were even less stable.

How does one then increase the stability of these species? An apt adage in directed evolution is “you get what you screen for”. If you *want* a stable amino-acrylate you need to *screen for* a stable amino-acrylate, either explicitly or implicitly. One way to accomplish this implicitly is to use stoichiometric amounts of substrates in the screening reaction. Therefore, the only way to achieve 100% yield is to stabilize the amino-acrylate so that it is not degraded over the course of the reaction (**Figure 1-5d**). Indeed, when TrpB was evolved using stoichiometric amounts of substrates, greater amino-acrylate stability came right along with increased yield. Additional engineering using poorly reactive indole analogs (which required an even more stable amino-acrylate intermediate) yielded a remarkably efficient Trp-analog synthase.

At this point, the enzyme had been through very little *functional* change. Under the conditions used for directed evolution, the enzyme merely favored certain intermediates and disfavored non-productive pathways. Its "native" chemistry—synthesizing Trp—was relatively unchanged, it just did this in the absence of TrpA and without needing to discriminate against other nucleophiles and Ser analogs. However, by stabilizing the reactive intermediate, these new enzyme variants were capable of reacting with completely new substrate classes, those previously inaccessible with the native TrpB enzymes. With some chemical intuition, new nucleophilic species were identified—oxindoles, nitroalkanes, ketones, and more—that underwent a carbon–carbon (C–C) bond-forming reaction with the amino-acrylate, despite looking and behaving quite differently from the native aromatic indole substrate (**Figure 1-5e**). While the initial activities were typically low, they provided starting points for directed evolution that were absent from the native enzymes and could be further evolved into productive biocatalysts.

1.1.7 Conclusions for Section 1.1

Just as life adapts to new challenges and opportunities through the evolution of enzymes, directed enzyme evolution has provided a reliable and powerful approach to address human needs. As we saw in this section, enzymes can provide solutions to problems in fields as disparate as pharmaceutical manufacturing to industrial carbon capture. New technologies have certainly improved our ability to optimize enzymes, such as improvements in DNA synthesis and sequencing, analytical instruments and techniques for screening, and computational methods that can efficiently learn from collected data. (In fact, next-generation DNA sequencing uses enzymes that themselves have been the subject of directed evolution!) Our ability to identify enzymes with *new* functions—which we can then throw into the process of directed evolution—can start with chemical intuition and in some instances requires the connection of multiple steps to form a path from a known enzyme function to a new one. By appreciating the fundamentals of natural and laboratory enzyme evolution, we can improve our chances of quickly creating new enzyme functions when they are needed.

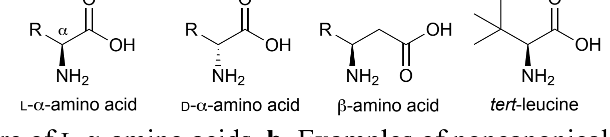
What if there simply is no enzyme that performs a desired reaction? We are currently limited to the enzymes that already exist, either through natural or laboratory evolution, which is an infinitesimal fraction of the possible protein sequences. Is there a way we can begin to create enzymes that can catalyze a reaction for which there is no good enzyme starting point? Advances in computational tools, such as machine learning and structure-based protein design, have begun to provide a glimmer of hope, but they also highlight the difficulties of this challenge. As stated above, a catalytic mechanism describes the transition states and intermediates that provide an energetically feasible path from substrates to products. To *design* an enzyme, one needs to know the nature of the transition states and intermediates—this is quite difficult—and then compose a protein sequence that will fold into a structure that stabilizes the transition states and accommodates any necessary intermediates—this is *extraordinarily* difficult. Nonetheless, computational advances have shown promise for relatively simple and well-characterized reactions, generating enzymes that could be further optimized through directed evolution. While it is clear that there is still much work to be done in this area, our ability to go from a hypothetical enzyme function to an observed one will only continue to improve with our understanding of enzymes and computational prowess. We predict that engineered enzymes will play increasingly important roles in future technology, thanks to the power of evolution.

1.2 Engineered Enzymes for Noncanonical Amino Acid Synthesis

The previous section discussed ncAA synthesis by TrpB as a general framework for exploring how we can access new functions of enzymes. This section will look more deeply at ncAAs themselves, addressing why they are so important across so many fields, why they are so difficult to synthesize, and how engineered enzymes—including TrpB—are perfectly poised to create efficient, sustainable routes to these valuable molecules.

1.2.1 General background

The twenty canonical L-α-amino acids (**Scheme 1-1a**) that serve as the primary basis of protein structure and function comprise only a small fraction of biologically and technologically important amino acids. Noncanonical amino acids (ncAAs), which are not naturally incorporated into proteins during translation, contain unusual side chains, D stereochemistry, or atypical backbone connectivity (**Scheme 1-1b**). These features in turn impart distinct chemical and biological properties, such as greater stability *in vivo*. These properties have elicited considerable interest, and ncAAs are used as therapeutics[1] and synthetic intermediates,[2] and are even encoded directly into proteins to confer useful new features.[3–5]



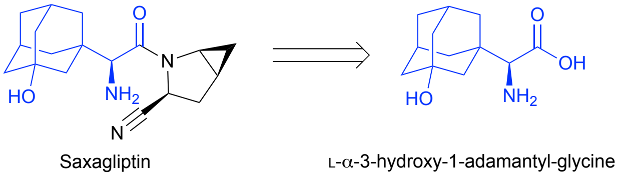
Scheme 1-1. a. Structure of L-α-amino acids. **b.** Examples of noncanonical amino acids (ncAAs).

Noncanonical amino acids are challenging to synthesize because they often contain a stereocenter at the α-carbon, which must be set in a precise configuration. In addition, the amine and carboxylic acid groups are reactive and often have to be protected. These problems are compounded when ncAAs have complex side chains that contain additional stereocenters or reactive functional groups. Nature circumvents such challenges by using enzymes, which bind and position substrates to accelerate a specific reaction, making enantiopure amino acids in aqueous media without the need for protecting groups. However, many enzymes that produce ncAAs in nature are not suitable for preparative ncAA synthesis due to low activity, poor endogenous expression levels and stability, need for allosteric activation, or limited substrate scope. Furthermore, many ncAAs are naturally synthesized via complicated multi-enzyme cascades, which may be difficult to identify and use for synthesis at scale. New strategies are needed for ncAA synthesis, and engineering new or improved enzymes offers some promising leads.

A number of enzymes have activities that could be used to make ncAAs, and protein engineering has been an indispensable tool for expanding this latent potential. Researchers have been able to requisition existing enzymes and engineer them to create high-yielding biocatalytic platforms that generate enantiomerically pure ncAAs. These enzymes will be described as ‘ncAA synthases’, since they are used to couple two molecules without requiring additional energetic input (e.g., ATP). Well-designed mutagenesis and screening strategies facilitate the engineering process, and iterative rounds of mutagenesis and screening (directed evolution) can produce greatly improved ncAA synthases; expanded substrate scopes, enhanced stability and heterologous expression, and increased yields are all achievable with the appropriate experimental design.

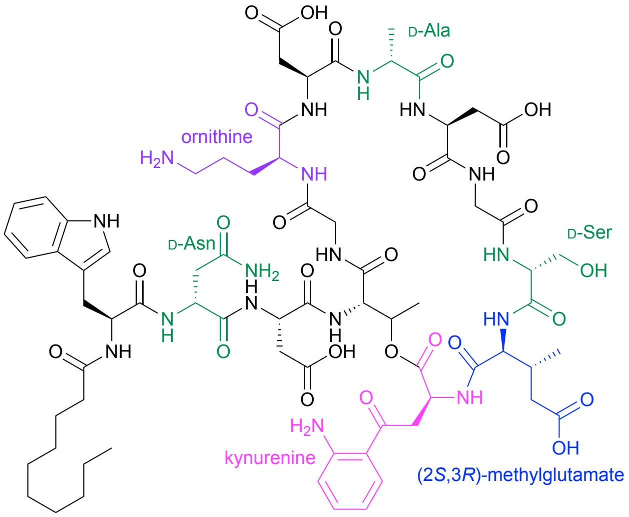
1.2.2 Applications of ncAAs in chemistry, medicine, and biology

Even with advances like photo-redox chemistry, metal-catalyzed cross-coupling, and asymmetric catalysis, the bottleneck for drug discovery is chemical synthesis.[2] Important pharmaceutical functionalities such as chiral amines, *N*-heterocycles, and unprotected polar groups are challenging to work with in synthesis, but the incorporation of ncAAs directly into synthetic pipelines can bypass many of these difficulties. For example, the diabetes medication saxagliptin (**Scheme 1-2**) contains a chiral amine as well as a challenging β-quaternary center that is important for the drug's activity.[1] Improved methods to synthesize this drug have focused on producing the ncAA L-α-3-hydroxy-1-adamantyl-glycine enzymatically for use as a building block in the synthesis of saxagliptin.[6] Other medicines contain alkaloids, pharmaceutically important natural products derived from amino acids. Alkaloids and similar compounds make up many essential medicines such as dopamine (heart failure), codeine and morphine (analgesic), vincristine and irinotecan (cancer), and quinine (antimalarial).[7] As biologically active and synthetically useful molecules, it is unsurprising that ncAAs are present in 12% of the 200 top-grossing drugs.[8] Incorporating ncAAs and ncAA-derived products into synthetic pipelines allows important pharmaceuticals to be synthesized more easily than ever before. However, few ncAAs are readily available, and improved synthetic and biocatalytic methods are needed to realize their full potential.



Scheme 1-2. The ncAA L-α-3-hydroxy-1-adamantyl-glycine is a building block of the diabetes drug saxagliptin.

Protein therapeutics including peptides and antibodies also make use of ncAAs. Therapeutic peptides have been used since the 1920s, when insulin was extracted from animal pancreases for diabetes treatment. Peptide drugs remain important to this day, with more than 60 approved for use.[1] However, most natural peptides are not suitable therapeutics because they are present only in low concentrations and are susceptible to proteolysis, limiting bioavailability. Incorporation of ncAAs with the D-configuration, unnatural backbones, or bulky side chains can reduce proteolysis, and modified side chains can tune biological specificity and pharmacokinetics.[1] For example, cyclic antimicrobial peptides such as daptomycin disrupt the membranes of infectious microorganisms. Daptomycin incorporates the ncAAs kynurenine, ornithine, and (2*S*,3*R*)-methylglutamate, as well as three D-amino acids (**Scheme 1-3**).



Scheme 1-3. Daptomycin, a cyclic peptide antibiotic, includes the ncAAs kynurenine (pink), ornithine (purple), (2*S*,3*R*)-methylglutamate (blue), as well as D-amino acids (green).

Antibodies and antibody-drug conjugates (ADCs) are another class of protein therapeutics that benefit from access to ncAAs. ADCs are versatile therapeutics composed of a chemotoxic agent coupled via an amino acid linker to an antibody that specifically targets a cellular component with limited side effects.[9] A common linker is a dipeptide composed of valine and the ncAA citrulline, which is cleaved in the lysosome to release the toxic 'payload' (**Figure 1-6**). The linker and payload are typically attached to the antibody via non-specific modifications of surface-exposed cysteine and lysine residues. With this non-specific conjugation, the payload may be attached at different positions and in different concentrations, resulting in a heterogeneous drug whose pharmacokinetics, safety, and efficiency are not well defined. Incorporation of ncAAs into ADCs can provide site-specific, bio-orthogonal attachment points for the linker, affording tunable and reproducible control over the payload concentration.[9]

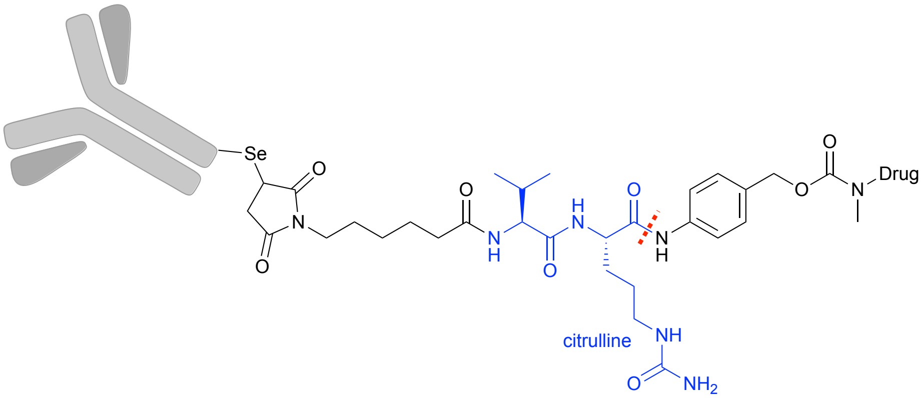


Figure 1-6. Antibody-drug conjugates (ADCs) target drugs to specific locations through the high-affinity interactions of the antibody to its antigen. In this simple example, the drug is attached to an antibody by site-specific incorporation of the ncAA selenocysteine (which can react with maleimides under different conditions than cysteine) and a cleavable valine-citrulline linker (blue). The drug is released following cleavage of the linker (indicated with a red dashed line).

Amino acid sequence dictates protein tertiary structure and affects protein function, localization, recognition, and post-translational modification. Consequently, incorporation of ncAAs at certain positions within a protein sequence can be used to modulate the physical and chemical properties of that protein. For example, global replacement of methionine with selenomethionine provides heavy atoms for X-ray crystallography,[10] while replacement of amino acids with their fluorinated analogues can influence the substrate specificity and stability of enzymes.[11] Furthermore, genetic code expansion permits the site-specific incorporation of ncAAs into proteins where promiscuous global replacement might be undesirable or impossible.[3,10] For example, ncAAs with side chains such as the environmentally sensitive fluorophore 7-hydroxycoumarin, the metal chelator 2,2'-bipyridine, and the metal-binding fluorophore 8-hydroxyquinoline have unique properties that can be used to probe biomolecular interactions or induce metal-dependent assembly or fluorescence.[12,13] Additionally, ncAAs with reactive unsaturated aliphatic, azido, and carbonyl side chains can be used as site-specific bio-orthogonal handles for chemical modification.[14] The ability to selectively manipulate proteins through the incorporation of ncAAs promotes the understanding and engineering of protein stability, activity, and mechanism.

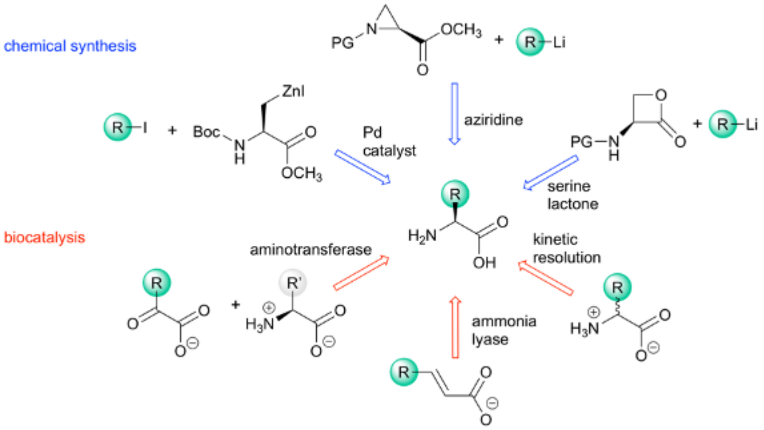
1.2.3 Methods for ncAA production

Although ncAAs are valuable chemical and biological tools, their applications are limited by inefficient routes of production. The most popular approaches, such as extraction from protein hydrolysate, fermentation, chemical synthesis, and biocatalysis fall short in terms of cost, yield, or scope.[15–17] Extracting amino acids from hydrolyzed proteins is excellent for large-scale production, especially when sourced from inexpensive industrial byproducts such as hair, meat, or plants. However, this is only suitable for naturally occurring ncAAs with unique physicochemical properties that enable purification, such as reactive side chains or extreme pH stability.

Amino acids are also produced on a large scale by microorganisms that convert sugars and other feedstocks into the desired products.[18] Bacterial strains of *Escherichia coli* or

Corynebacterium glutamicum have been extensively engineered for efficient metabolism and mitigated stress response to enhance yields.[19] Production by fermentation, however, requires an organism with the capacity to synthesize the ncAA. This is a major complication, since biosynthetic pathways for many ncAAs either give poor yields or are simply unknown.

Chemical synthesis can access numerous ncAAs by employing intermediates such as serine-derived lactones, hydantoins, or aziridines (**Scheme 1-4**, blue).[20,21] A significant benefit of chemical synthesis approaches is their broad applicability, allowing a variety of ncAAs to be produced from a single synthetic pipeline. Limitations are also apparent: chemical synthesis can be labor-intensive, utilize hazardous reagents and produce significant waste products, or generate racemic products that require further purification.



Scheme 1-4. Selected representative methods for synthesizing ncAAs using chemical synthesis (blue) or biocatalysis (red).

An increasingly useful approach to preparing ncAAs is biocatalysis, which can either replace or supplement chemical synthesis and fermentation with enzymes (**Scheme 1-4**, red).[16,22] Enzyme-catalyzed reactions benefit from mild reaction conditions and a broad range of biocatalysts that can be used in derivatization or bond-forming reactions. For example, aminotransferases are used to form chiral amines by transferring the amino group of one amino acid to a prochiral α-keto acid while setting the stereochemistry at the α-carbon. The process for manufacturing sitagliptin, a diabetes drug, incorporates an engineered aminotransferase that replaces two steps of the chemical synthesis route, as discussed in **Section 1.1.3**.[23] Other enzymes such as lyases capitalize on accessible, non-hazardous starting materials to synthesize diverse optically pure ncAAs. For example, ammonia lyases can catalyze the asymmetric amination of inexpensive, prochiral substrates such as fumarate or cinnamic acid derivatives to make optically pure ncAAs.[24,25] Other enzymes, such as tyrosine phenol lyase (TPL)[26] and tryptophan synthase (TrpS),[27,28] can generate more complexity from even simpler substrates by coupling a nucleophilic side chain to an amino acid backbone.

1.2.4 Modular ncAA synthesis through amino-acrylate intermediates

Ideally, an ncAA synthase would be modular, in that it would attach desired side chains to an amino acid backbone with perfect stereoselectivity. An advantage typically associated with chemical synthesis, modularity allows different pieces to be incorporated into a diverse array of products with the same technique (see **Section 1.2.3, Scheme 1-4**). The pyridoxal 5'-phosphate (PLP)-dependent enzymes tyrosine phenol lyase (TPL) and tryptophan synthase (TrpS) have this attractive feature. These enzymes catalyze the β-elimination of an L-amino acid substrate to form an electrophilic amino-acrylate intermediate (**Figure 1-7a**). Discussed in **Section 1.1.6**, the amino-acrylate is a versatile electrophile that allows diverse nucleophilic substrates to be incorporated as amino acid side chains to form new L-α-amino acids. Due to the range of acceptable nucleophiles, these enzymes are capable of C–C bond formation as well as C–N and C–S bond formation. These enzymes also act with perfect enantioselectivity, as the stereochemistry at the α-carbon is retained through proton abstraction and donation on the same face of the amino-acrylate by the active-site lysine (**Scheme 1-5**).[27]

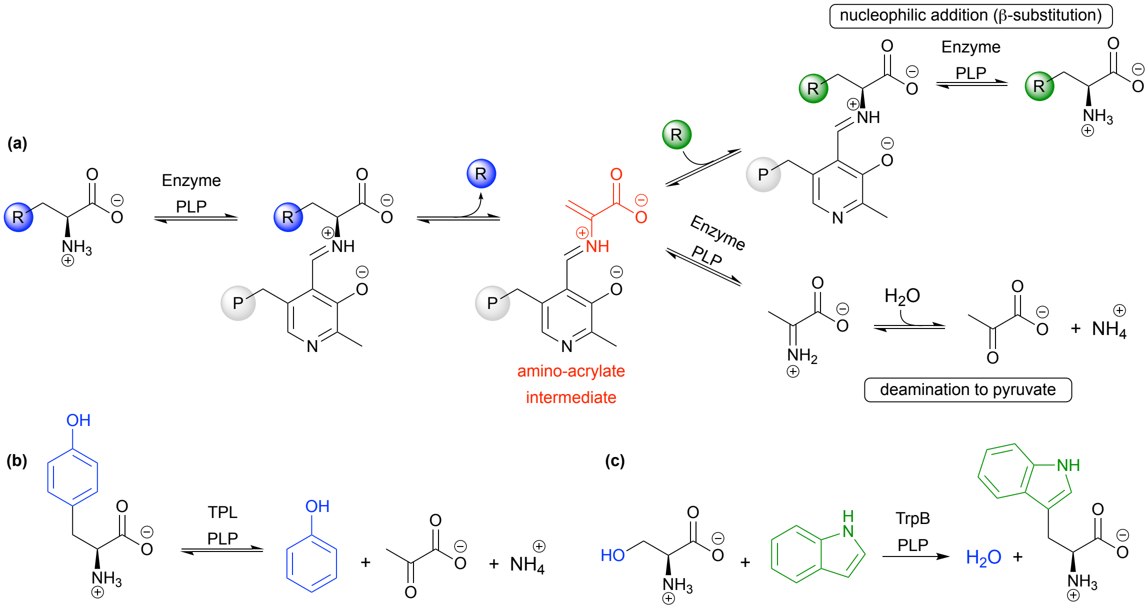
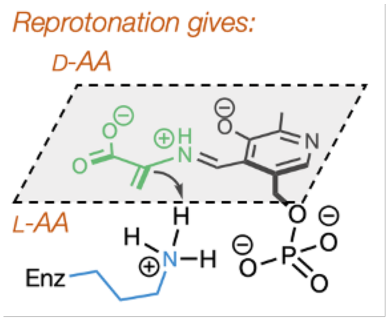


Figure 1-7. a. Generalized reaction for pyridoxal 5'-phosphate (PLP)-dependent ncAA synthesis. Simplified reactions that follow the overall scheme in **a** are shown for the PLP-dependent enzymes **b.** tyrosine phenol lyase (TPL) and **c.** the β-subunit of tryptophan synthase (TrpB).



Scheme 1-5. Stereoselective protonation of the amino-acrylate intermediate (AA = amino acid).

Tyrosine phenol lyase (TPL) catalyzes the degradation of L-tyrosine (Tyr) to phenol, pyruvate, and ammonia through a β-elimination reaction (**Figure 1-7b**). The reaction is readily reversible, and the addition of excess of ammonia and pyruvate shifts the equilibrium to favor Tyr production. This occurs by promoting the formation of the electrophilic amino-acrylate intermediate, which then reacts with phenol to form a C–C bond via an electrophilic aromatic substitution mechanism.[29] Phenol is nucleophilic at positions *para* and *ortho* to the electron-donating hydroxyl group, and the enzyme positions this substrate such that the C–C bond is formed exclusively at the *para* position.

One mechanistic limitation of TPL is that the reaction is under thermodynamic control. The forward and reverse reaction rates of the net reaction depend strongly on the concentrations of products and reactants, and excess reactants are needed to drive product formation. Ammonia lyases—another class of enzyme commonly employed in ncAA synthesis—also have this limitation, as do aminotransferases. The need for excess reagents is not a major issue when using TPL and ammonia lyases for preparative-scale synthesis, since ammonia and pyruvate are inexpensive and easy to exclude during purification. Nonetheless, it would be preferable to have the reaction under kinetic control such that product formation is effectively irreversible, improving atom economy and making *in vivo* applications more accessible. This type of reaction is possible when using TPL with specialized substrates. For example, *S*-(*o*-nitrophenyl)-L-cysteine can undergo rapid β-elimination in the presence of TPL, as the nitrothiophenol side chain acts as a good leaving group (**Figure 1-8**).[29] This subsequently forms the reactive amino-acrylate intermediate, which is attacked by phenol to produce Tyr. Furthermore, because TPL binds *S*-(*o*-nitrophenyl)-L-cysteine more tightly than Tyr, as long as *S*-(*o*-nitrophenyl)-L-cysteine is present in the reaction it preferentially undergoes β-elimination and inhibits Tyr degradation. This approach gives yields of ~70%.

Due to the reversible nature of enzymatic reactions under thermodynamic control, TPL suffers from inherently low substrate coupling efficiencies, with a high concentration of one or more substrates remaining upon reaching equilibrium. Although there are ways to circumvent this by using specialized substrates, an ideal biocatalyst would couple stoichiometric proportions of simple substrates at high rates and with quantitative yields. Additionally, these biocatalysts could have applications *in vivo*, as physiological concentrations of reactants could be sufficient to form products. To accomplish this, the enzymatic reaction should be under kinetic control. This is the case for tryptophan synthase (TrpS), which catalyzes the final steps of L-tryptophan (Trp) biosynthesis.

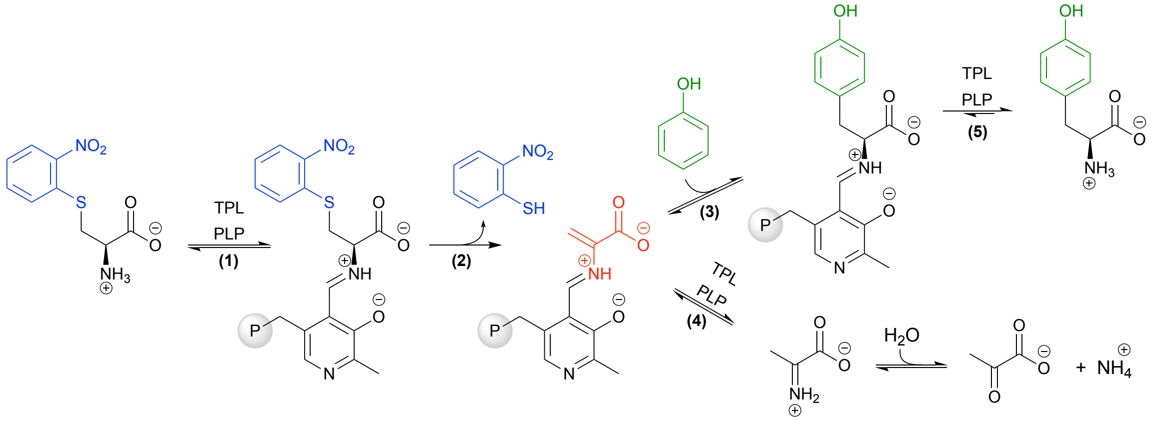


Figure 1-8. The use of a specialized substrate, *S*-(o-nitrophenyl)-L-cysteine, affords kinetic control over Tyr production with *Cf*TPL. This substrate binds in the active site (1) and rapidly forms the amino-acrylate (2), due to the good leaving group at the β-carbon (blue). The amino-acrylate (red) can either undergo nucleophilic attack by phenol (3, green) or deaminate to pyruvate and ammonia (4). Since TPL has a higher affinity for the substrate, *S*-(o-nitrophenyl)-L-cysteine, than for the Tyr product (5), TPL selectively binds the substrate, thereby reducing degradation of the Tyr product and providing higher yields.

As discussed in **Section 1.1.5**, TrpS a heterodimeric complex composed of an α-subunit (TrpA) that allosterically regulates the β-subunit (TrpB).[27] In the native reaction, indole glycerol-3-phosphate undergoes a retro-aldol reaction in TrpA to release indole. This induces TrpB to catalyze the β-elimination of L-serine (Ser), which generates the amino-acrylate intermediate (**Figure 1-9a**). Indole then diffuses through a hydrophobic tunnel connecting the subunits and attacks the amino-acrylate to form Trp (**Figure 1-9b**). Without any engineering, the wild-type TrpS enzyme can perform this C–C bond-forming reaction with an array of indole analogues *in vitro*, synthesizing substituted Trp analogues in a single step.[27] Numerous Trp derivatives have been made using this strategy. For example, the Goss group demonstrated that *Salmonella enterica* TrpS can use 7-chloroindole and Ser to form 7-chloroTrp, part of the antibiotic rebeccamycin.[30] This reaction occurs in a single step, whereas nature would require an additional Trp halogenase to add the chloro substituent to Trp. Additionally, non-indole nucleophiles have been used to form C–S and C–N bonds, demonstrating that TrpS can also be a platform for the production L-cysteine and L-β-aminoalanine ncAAs.[27]

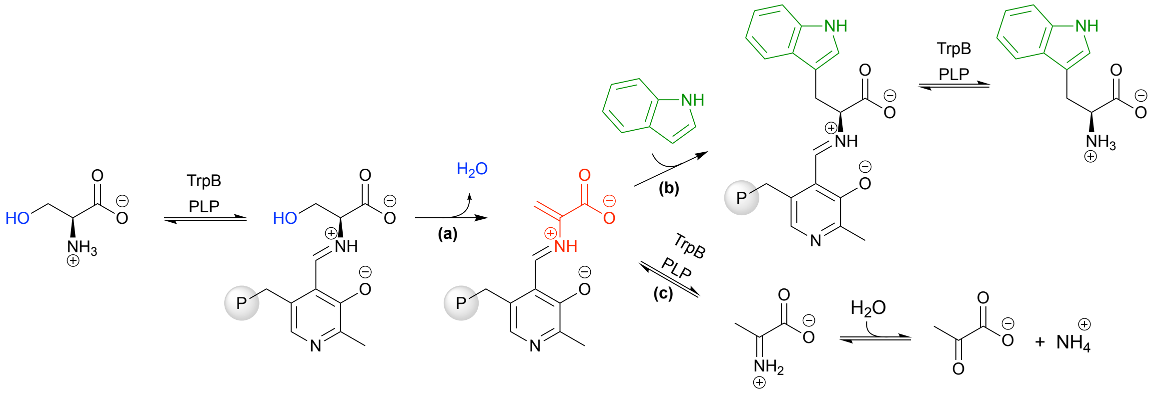


Figure 1-9. The reaction of TrpB is under kinetic control, and parallels that of TPL with specialized substrates (as shown in **Figure 1-8**). **a.** Formation of the reactive amino-acrylate intermediate (red) is favored by the β-elimination of L-serine and exclusion of water (blue) from the hydrophobic active site. **b.** Indole (green) is activated for nucleophilic attack within the active site to form tryptophan (Trp); little to no β-elimination of Trp occurs when the enzyme is provided with Trp as its only substrate, suggesting that this step is effectively irreversible. **c.** Nonproductive deamination of the amino-acrylate produces pyruvate and ammonia, a step that has been disfavored throughout the directed evolution of TrpB variants.

Although TrpS is an impressive biocatalyst, there are roadblocks for its application. Expression of the TrpS complex is metabolically challenging for the host cell, and the need for both the TrpA and TrpB subunits complicates purification and engineering. TrpB performs the synthetically interesting β-substitution reaction between indole and Ser to generate Trp, while TrpA generates indole *in situ* so that this toxic metabolite is not released into the cytosol. If the indole analogues are added exogenously, then TrpA is superfluous, but removing TrpA significantly decreases the activity of TrpB, due to the allosteric interactions between the subunits of TrpS.[28]

To overcome these limitations and access the full potential of TrpB as a stand-alone ncAA synthase, members of the Arnold lab have applied numerous rounds of directed evolution to hyperthermophilic TrpB enzymes, using the concepts described in **Section 1.1.6**. The engineering efforts benefitted from spectral features of the TrpB reaction. In particular, a red-shift in the absorbance of indole as it is converted to tryptophan allowed for high-throughput plate-reader based screening, and changes in the absorbance of the PLP cofactors identified the steady-state distributions of critical intermediates in the reaction. Dr. Andrew Buller (now an assistant professor at the University of Wisconsin, Madison) led the initial engineering of TrpB from *Pyrococcus furiosus* (*Pf*TrpB) as a stand-alone platform for Trp production in the absence of its allosteric partner TrpA (**Figure 1-10**, red). Without the unnecessary complication of TrpA, *Pf*TrpB was further engineered for high activity for β-methyl-Trp synthesis by using threonine[31] (**Figure 1-10**, orange). This variant generated a highly stable amino-acrylate intermediate that could react with poorly nucleophilic substrates (as discussed conceptually in **Section 1.1.6**). From here, Dr. David Romney (now co-founder and CTO of Aralez Bio) led engineering campaigns to create TrpB variants that could accept 6- and 7-substituted Trp analogues[32] (**Figure 1-10**, green) and 4- and 5-substituted Trp analogues via a transfer of beneficial mutations from *Pf*TrpB to TrpB from *Thermotoga maritima* (*Tm*TrpB, **Figure 1-10**, blue). Dr. Tina Boville (now co-founder and CEO of

Aralez Bio) unlocked activity with even more diverse β-substituted-Trp analogues by using β-alkyl-Ser analogues[33] as electrophiles instead of Ser (**Figure 1-10**, pink).

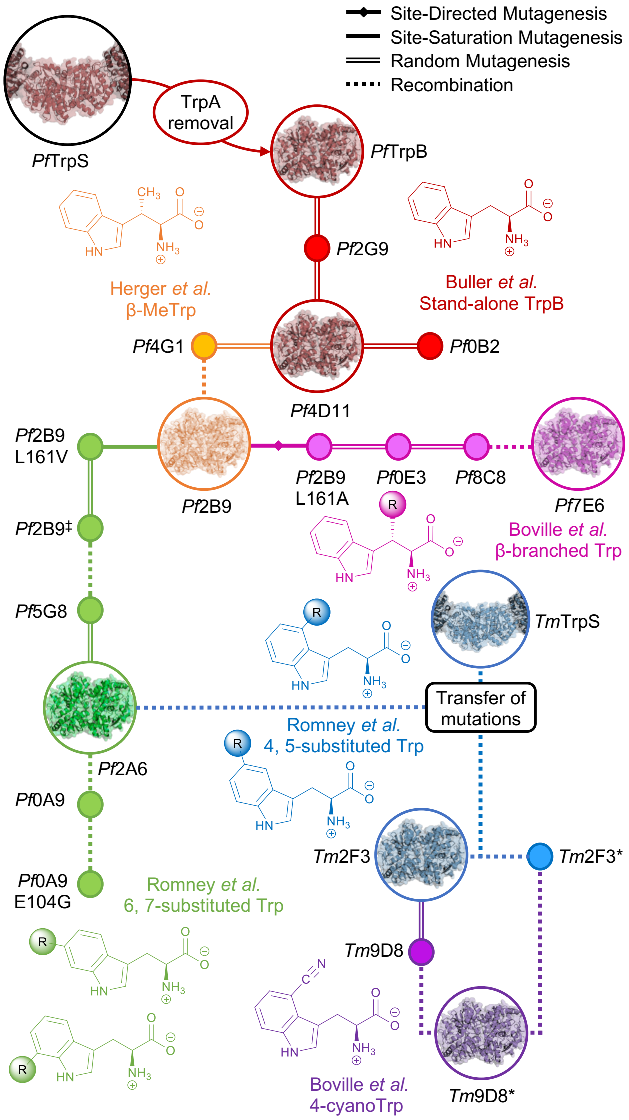


Figure 1-10. Evolutionary lineage of TrpB-based ncAA synthases. Starting with the isolated TrpB enzyme subunit from *Pyrococcus furiosus* (*Pf*TrpB), directed evolution has been used to improve standalone Trp production (red), as well as activity with L-threonine (orange), substituted indoles (green and blue), and bulkier β-branched L-serine analogues (pink). It has also been engineered to produce Trp analogues such as 4-cyanoTrp at lower temperatures (purple). Intermediate variants in the evolution are shown as nodes, with representative variants shown with their structural models. Nodes are connected by lines that represent the mutagenesis approach. Transfer of activating mutations to a homologous TrpB from *Thermotoga maritima* (*Tm*TrpB) resulted in variants that exhibited different substrate preferences than *Pf*TrpB variants; where *Pf*TrpB was adept at 6- and 7-substituted Trp production, *Tm*TrpB was adept at 4- and 5-substituted Trp production.

This was the state of TrpB engineering when I began my thesis work in the Arnold lab. No C–C bond formation with non-indole nucleophiles had been identified, such as those discussed in **Section 1.1.6**. On the other hand, nearly any reasonably substituted Trp analogue could be synthesized by a TrpB variant with ease, with the notable exceptions of 4-cyano-Trp (a bright blue fluorescent ncAA) and those with a trifluoromethyl substituent. Subsequent engineering to create a useful biocatalyst for 4-cyano-Trp production is represented in purple in **Figure 1-10** and is the subject of **Chapter II** of this thesis. The rest of this body of works represents ways in which noncanonical amino acid synthesis by TrpB has been expanded to new carbon nucleophiles as well as applied to develop new methods for protein engineering and evolution.

Chapter I Bibliography

Bibliography for Section 1.1†

(†Section 1.1 is adapted from an unreferenced textbook article. Relevant citations are provided here.)

I. sakaiensis discovery

- Yoshida, S. *et al.* A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **351**, 1196–1199 (2016).

Carbon capture with an engineered carbonic anhydrase

- Alvizo, O. *et al*. Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16436–16441 (2014).

Enzyme engineering for sitagliptin synthesis

- Savile, C. K. *et al*. Biocatalytic asymmetric synthesis of sitagliptin manufacture. *Science* **329**, 305–310 (2010).

Esterase activity of carbonic anhydrase

- Gould, S. M. Q. & Tawfik, D. S. Directed evolution of the promiscuous esterase activity of carbonic anhydrase II. *Biochemistry* **44**, 5444–5452 (2005).

TrpB engineering

See citations 28, 31–33 in ***Section 1.2****, and:*

- Romney, D. K., Sarai, N. S. & Arnold, F. H. Nitroalkanes as versatile nucleophiles for enzymatic synthesis of noncanonical amino acids. *ACS Catal.* **9**, 8726–8730 (2019).
- Dick, M., Sarai, N. S., Martynowycz, M. W., Gonen, T. & Arnold, F. H. Tailoring tryptophan synthase TrpB for selective quaternary carbon bond formation. *J. Am. Chem. Soc.* **141**, 19817–19822 (2019).
- Watkins-Dulaney, E. J. *et al.* Asymmetric alkylation of ketones catalyzed by engineered TrpB. *Angew. Chem. Int. Ed.* **60**, 21412–21417 (2021).
- Watkins-Dulaney, E., Straathof, S. & Arnold, F. Tryptophan synthase: Biocatalyst extraordinaire. *ChemBioChem* **22**, 5–16 (2021).

Bibliography for Section 1.2

1. Blaskovich, M. A. T. Unusual amino acids in medicinal chemistry. *J. Med. Chem.* **59**, 10807–10836 (2016).
2. Blakemore, D. C. *et al.* Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
3. Young, D. D. & Schultz, P. G. Playing with the molecules of life. *ACS Chem. Biol.* **13**, 854–870 (2018).
4. Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **550**, 53–60 (2017).
5. Wang, K. *et al.* Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nat. Chem.* **6**, 393–403 (2014).
6. Park, E. S. & Shin, J. S. Biocatalytic cascade reactions for asymmetric synthesis of aliphatic amino acids in a biphasic reaction system. *J. Mol. Catal. B Enzym.* **121**, 9–14 (2015).

7. Ehrenworth, A. M. & Peralta-Yahya, P. Accelerating the semisynthesis of alkaloid-based drugs through metabolic engineering. *Nat. Chem. Biol.* **13**, 249–258 (2017).

8. McGrath, N. A., Brichacek, M. & Njardarson, J. T. A graphical journey of innovative organic architectures that have improved our lives. *J. Chem. Educ.* **87**, 1348–1349 (2010).

9. Thu, T., Nguyen, H., Bai, X. & Shim, H. Next generation antibody therapeutics: Bispecific antibodies and antibody-drug conjugates. *Biodesign* **3**, 154–161 (2015).

10. Liu, C. C. & Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **79**, 413–444 (2010).

11. Odar, C., Winkler, M. & Wiltschi, B. Fluoro amino acids: A rarity in nature, yet a prospect for protein engineering. *Biotechnol. J.* **10**, 427–446 (2015).

12. Liu, X. *et al.* Significant expansion of the fluorescent protein chromophore through the genetic incorporation of a metal-chelating unnatural amino acid. *Angew. Chem. Int. Ed.* **52**, 4805–4809 (2013).

13. Almhjell, P. J. & Mills, J. H. Metal-chelating non-canonical amino acids in metalloprotein engineering and design. *Curr. Opin. Struct. Biol.* **51**, 170–176 (2018).

14. Boutureira, O. & Bernardes, G. J. L. Advances in chemical protein modification. *Chem. Rev.* **115**, 2174–2195 (2015).

15. D'Este, M., Alvarado-Morales, M. & Angelidaki, I. Amino acids production focusing on fermentation technologies – A review. *Biotechnol. Adv.* **36**, 14–25 (2018).

16. Rudroff, F. *et al.* Opportunities and challenges for combining chemo- and biocatalysis. *Nat. Catal.* **1**, 12–22 (2018).

17. Völler, J. S. & Budisa, N. Coupling genetic code expansion and metabolic engineering for synthetic cells. *Curr. Opin. Biotechnol.* **48**, 1–7 (2017).

18. D'Este, M., Alvarado-Morales, M. & Angelidaki, I. Amino acids production focusing on fermentation technologies – A review. *Biotechnol. Adv.* **36**, 14–25 (2018).

19. Lütke-Eversloh, T., Santos, C. N. S. & Stephanopoulos, G. Perspectives of biotechnological production of L-tyrosine and its applications. *Appl. Microbiol. Biotechnol.* **77**, 751–762 (2007).

20. Ager, D. J. Synthesis of unnatural/nonproteinogenic α-amino acids. in *Amino Acids, Peptides, and Proteins in Organic Chemistry, Vol. 1 - Origins and Synthesis of Amino Acids* (ed. Hughes, A. B.) 495–526 (WILEY-VCH, 2009).

21. Brittain, W. D. G. & Cobb, S. L. Negishi cross-couplings in the synthesis of amino acids. *Org. Biomol. Chem.* **16**, 10–20 (2018).

22. Xue, Y.-P., Cao, C.-H. & Zheng, Y.-G. Enzymatic asymmetric synthesis of chiral amino acids. *Chem. Soc. Rev.* **47**, 1516–1561 (2018).

23. Slabu, I., Galman, J. L., Lloyd, R. C. & Turner, N. J. Discovery, engineering, and synthetic application of transaminase biocatalysts. *ACS Catal.* **7**, 8263–8284 (2017).

24. Gloge, A., Zon, J., Kovari, A., Poppe, L. & Rétey, J. Phenylalanine ammonia-lyase: The use of its broad substrate specificity for mechanistic investigations and biocatalysis - synthesis of L-arylalanines. *Chemistry—A European Journal* **6**, 3386–3390 (2000).

25. Raj, H. *et al.* Engineering methylaspartate ammonia lyase for the asymmetric synthesis of unnatural amino acids. *Nat. Chem.* **4**, 478–484 (2012).

26. Seisser, B. *et al.* Cutting long syntheses short: Access to non-natural tyrosine derivatives employing an engineered tyrosine phenol lyase. *Adv. Synth. Catal.* **352**, 731–736 (2010).

27. Phillips, R. S. Synthetic applications of tryptophan synthase. *Tetrahedron Asymmetry* **15**, 2787–2792 (2004).

28. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

29. Phillips, R. S., Chen, H. Y. & Faleev, N. G. Aminoacrylate intermediates in the reaction of *Citrobacter freundii* tyrosine phenol-lyase. *Biochemistry* **45**, 9575–9583 (2006).

30. Smith, D. R. M. *et al.* The first one-pot synthesis of L-7-iodotryptophan from 7-iodoindole and serine, and an improved synthesis of other L-7-halotryptophans. *Org. Lett.* **16**, 2622–2625 (2014).

31. Herger, M. *et al.* Synthesis of β-branched tryptophan analogues using an engineered subunit of tryptophan synthase. *J. Am. Chem. Soc.* **138**, 8388–91 (2016).

32. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

33. Boville, C. E. *et al.* Engineered biosynthesis of β-alkyl tryptophan analogues. *Angew. Chem. Int. Ed.* **57**, 14764–14768 (2018).

Chapter II

IMPROVED SYNTHESIS OF 4-CYANOTRYPTOPHAN AND OTHER TRYPTOPHAN ANALOGUES IN AQUEOUS SOLVENT USING VARIANTS OF TRPB FROM *THERMOTOGA MARITIMA*

Material from this chapter appears in: "Boville, C. E., Romney, D. K., **Almhjell, P. J.**, Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *The Journal of Organic Chemistry* **83**, 7447–7452 (2018) doi: 10.1021/acs.joc.8b00517".

C.E.B. and D.K.R. participated in project conception. C.E.B., D.K.R., P.J.A., and M.S. designed and executed research. P.J.A. hypothesized and confirmed improved enzyme activity at lower temperatures. D.K.R. drafted the manuscript with input from all authors.

ABSTRACT

The use of enzymes has become increasingly widespread in synthesis as chemists strive to reduce their reliance on organic solvents in favor of more environmentally benign aqueous media. With this in mind, we previously endeavored to engineer the tryptophan synthase β-subunit (TrpB) for production of noncanonical amino acids that had previously been synthesized through multistep routes involving water-sensitive reagents. This enzymatic platform proved effective for the synthesis of analogues of the amino acid tryptophan (Trp), which are frequently used in pharmaceutical synthesis as well as chemical biology. However, certain valuable compounds, such as the blue fluorescent amino acid 4-cyanotryptophan (4-CN-Trp), could only be made in low yield, even at elevated temperature (75 °C). Here, we describe the engineering of TrpB from *Thermotoga maritima* that improved synthesis of 4-CN-Trp from 24% to 78% yield. Remarkably, although the final enzyme maintains high thermostability (T_{50} = 93 °C), its temperature profile is shifted such that high reactivity is observed at ~37 °C (76% yield), creating the possibility for *in vivo* 4-CN-Trp production. The improvements are not specific to 4-CN-Trp; a boost in activity at lower temperature is also demonstrated for other Trp analogues.

2.1 Introduction

Noncanonical *α*-amino acids (ncAAs) resemble the building blocks of natural proteins but are not themselves used in protein synthesis. Despite this, ncAAs are prevalent precursors for functional synthetic compounds, including over 12% of the 200 top-selling pharmaceuticals.[1] However, ncAAs are challenging synthetic targets since they possess at minimum two reactive functional groups (the amine and carboxylic acid) and typically have at least one stereocenter. As a result, synthetic routes to ncAAs typically require multiple steps, most of which use organic solvents.[2,3] One of the most direct routes to ncAAs is to add a nucleophile to the β-position of a serine-derived lactone[4–6] or aziridine[7,8] (**Figure 2-1a**), but this approach has certain drawbacks, such as the need to presynthesize the water-sensitive electrophilic reactants.

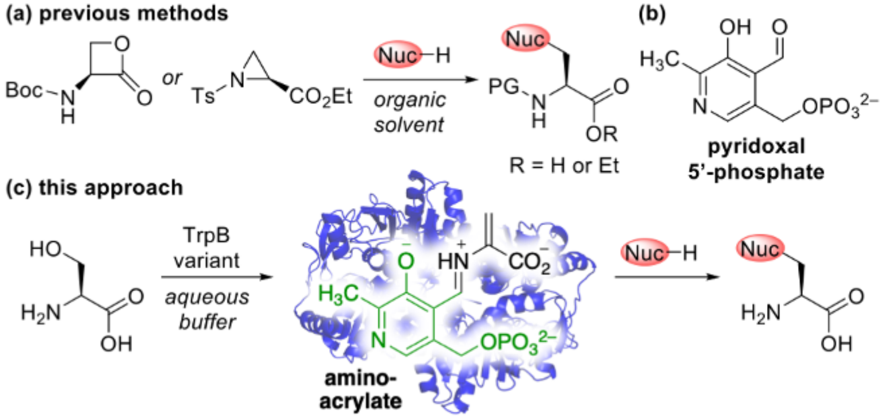
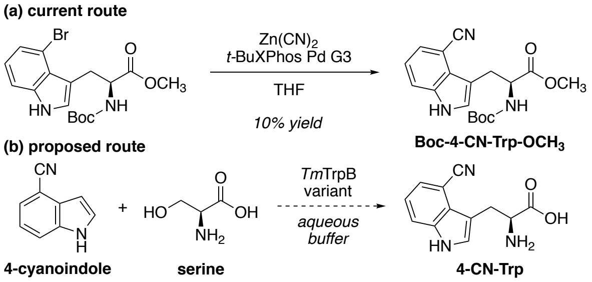


Figure 2-1. Amino acid synthesis by nucleophilic substitution at the β-position. a. Approach using preformed lactone or aziridine. **b.** Cofactor used by TrpB enzymes. **c.** Alternative approach in which an enzyme forms an amino-acrylate *in situ* from stable precursors like serine. (Boc, *tert*-butoxycarbonyl; Ts, 4-toluenesulfonyl; PG, protecting group.)

Enzymes are widely applied to the synthesis of ncAAs since they circumvent many of the limitations of chemical methods. Not only do these catalysts function in aqueous media, but they also exhibit chemoselectivity that obviates the need for protecting groups, thereby trimming synthetic steps. In addition, the reactions are often highly stereoselective.

Unfortunately, most enzymatic methods to synthesize ncAAs, such as those that rely on hydrolases or transaminases, require that the majority of the final product be synthesized in advance, usually by chemical means, with the enzyme only appearing at the end to set the stereochemistry. By contrast, enzymes like tryptophan synthase,[9–14] which uses the cofactor pyridoxal 5'-phosphate (PLP, **Figure 2-1b**), can form ncAAs by nucleophilic substitution at the β-position of readily available amino acids like serine. In this reaction scheme, the enzyme forms an active electrophilic species, the amino-acrylate (**Figure 2-1c**), directly in the active site, which is then intercepted by a nucleophilic substrate. These reactions can be run in aqueous conditions that would hydrolyze the serine-derived lactones or aziridines. Furthermore, the enzyme active site can bind the substrates to accelerate the reaction and control the regioselectivity of nucleophilic substitution.

The ncAA 4-cyanotryptophan (4-CN-Trp) was previously reported to exhibit blue fluorescence (λ_{max} ~ 405 nm) with a high quantum yield and long lifetime.[15] These properties, among others, make 4-CN-Trp an attractive small-molecule fluorophore for imaging studies *in vitro* and *in vivo*. However, the chemical synthesis requires multiple steps, including a low-yielding Pd-catalyzed cyanation reaction (**Scheme 2-1a**). We were excited to observe that an engineered variant of the β-subunit of tryptophan synthase (TrpB) from hyperthermophilic bacterium *Thermotoga maritima* could form 4-cyanotryptophan in one step from readily available 4-cyanoindole and serine (**Scheme 2-1b**).[16] We therefore wished to engineer this variant further to improve 4-CN-Trp production.



Scheme 2-1. Direct synthesis of 4-cyanotryptophan.

2.2 Results

2.2.1 Increasing activity with 4-cyanoindole

As the starting enzyme, we chose a variant designated *Tm*2F3 (**Table 2-1**), which is derived from *T. maritima* TrpB and has seven mutations. We selected this variant because in previous studies it exhibited high activity with other 4-substituted indole substrates.[16] In addition, this variant, like its wild-type progenitor, tolerated high temperatures (up to 75 °C), which accelerated the reaction. In the development of *Tm*TrpB-derived variants, we found that activating mutations were distributed throughout the protein sequence without any obvious patterns. One exception was the mutation I184F, which resides in the putative enzyme active site. Although incorporation of the I184F mutation increased the production of 4-CN-Trp with *Tm*2F3,[16] it was not beneficial for other 4-substituted indoles. Moving forward, we therefore decided to exclude the I184F mutation and instead perform global random mutagenesis on the *Tm*2F3 gene, with the option to revisit I184 at a later stage.

Table 2-1. Summary of *T. maritima* TrpB variants. [a]Measurements conducted in triplicate

Designation	Mutations	T_{50} (° C)[a]
*Tm*2F3	P19G, I69V, K96L, P140L, N167D, L213P, T292S	91.5 ± 0.8
*Tm*9D8	*Tm*2F3 + E30G, G228S	88.2 ± 0.7
*Tm*9D8*	*Tm*9D8 + I184F	92.7 ± 0.2

.

We observed from test reactions that conversion of 4-cyanoindole to 4-CN-Trp was accompanied by an increase in absorption at 350 nm. This spectral shift allowed us to screen the enzyme library rapidly by running reactions in 96-well plates and then monitoring the change in absorption at 350 nm using a plate reader. After screening 1,760 clones, we identified a new variant *Tm*9D8 (E30G and G228S) that appeared to exhibit a 2.5-fold increase in the yield of 4-CN-Trp. Strangely, when we retested *Tm*9D8 in vials, we found that it was no more active than the parent *Tm*2F3 (**Figure 2-2**). We hypothesized that although the plate and vial reactions were ostensibly conducted at the same temperature (75 °C), the reaction mixtures in the plate may have actually been at lower temperature due to the inherent difficulties in heating a 96-well plate uniformly. We therefore retested *Tm*9D8 and *Tm*2F3 at lower temperatures and found that *Tm*9D8 was almost 2-fold better than *Tm*2F3 at 50 °C and almost 5-fold better at 37 °C. Notably, *Tm*9D8 performed better at 37 °C than *Tm*2F3 did at 75 °C. This ability to function at lower temperature is not only advantageous for process development but also creates the possibility of synthesizing 4-CN-Trp *in vivo*.

We previously found that introduction of the mutation I184F into *Tm*2F3 improved the production of 4-CN-Trp.[16] We therefore constructed an 8-variant recombination library in which positions 30, 184, and 228 could either be the wild-type or the mutated residues. Screening this set would reveal if E30G and G228S were both responsible for the first-round improvement and whether I184F was still beneficial in this new variant. We found that the best variant, *Tm*9D8*, indeed retained all three mutations, boosting production of 4-CN-Trp to ~76–78% at both 37 and 50 °C (**Figure 2-2**). We also tested libraries in which positions 30, 184, and 228 were separately randomized to all 20 canonical amino acids; screening showed that glycine and serine were favored at positions 30 and 228, respectively (see Appendix A, **Figures A-1** and **A-2**). At position 184, leucine also improved activity compared to the native isoleucine (see Appendix A, **Figure A-3**), but rescreening showed this mutation was not as beneficial as phenylalanine. Thus, we adopted *Tm*9D8* for production of 4-CN-Trp.

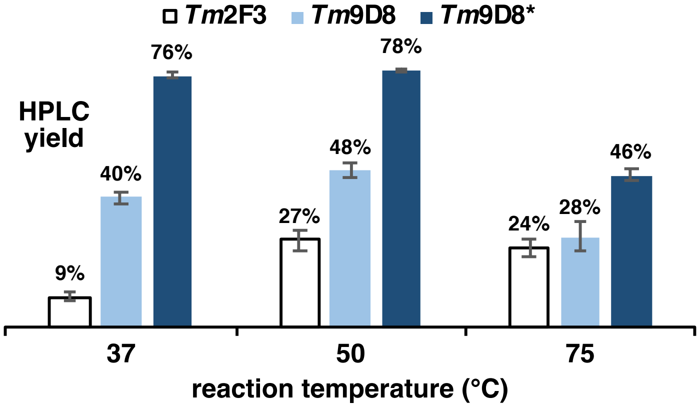
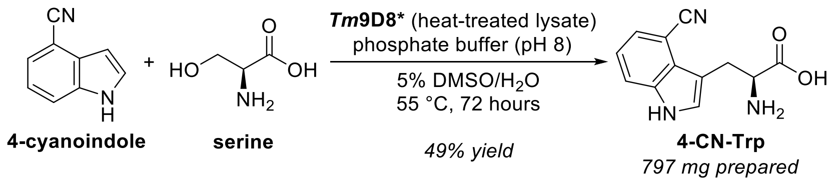


Figure 2-2. Production of 4-CN-Trp at different temperatures from equimolar 4-cyanoindole and serine (maximum of 1000 turnovers). Yields are averages of two replicates. Full data are reported in Appendix A, **Table A-1**.

2.2.2 Large-scale production of 4-CN-Trp

Although the final variant exhibits a low initial turnover frequency (0.95 ± 0.05 min^{-1} at 37 °C) and requires a relatively high catalyst loading (0.1 mol %) to achieve the yields in **Figure 2-2**, its expression level is such that enzyme from a 1-L culture can synthesize almost 800 mg of 4-CN-Trp at 55 °C (**Scheme 2-2**). Since the reaction is performed in aqueous media, most of the product precipitates directly from the reaction mixture and can be purified by simple wash steps. The new variant also retains excellent thermostability (T_{50} ~ 90 °C) (**Table 2-1**), which allows it to be prepared as heat-treated lysate, facilitating removal of cell debris, and used in the presence of organic solvents, improving solubility of hydrophobic substrates.



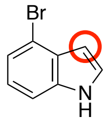
Scheme 2-2. Enzymatic preparation of 4-CN-Trp

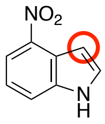
2.2.3 Activity with other substrates

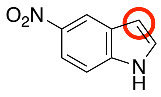
We tested the TrpB variants with other indole analogues to see how the mutations affected specificity (**Table 2-2**). To highlight the improved activity at lower temperature, reactions with *Tm*2F3 and *Tm*9D8 were screened at 50 °C, whereas reactions with *Tm*9D8* were screened at 37 °C. Although *Tm*2F3 already exhibits good activity (3,250 turnovers at 50 °C) with 4-bromoindole (**1**), the activity is improved in the later variants, with *Tm*9D8* performing a similar number of turnovers (3,750), but at lower temperature (37 °C). All variants, however, exhibited negligible activity with 4-nitroindole (**2**), suggesting that the active site is highly sensitive to the geometry of substituents at the 4-position.

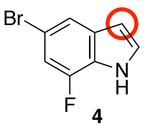
Previously, *T. maritima* TrpB variants had excelled in reactions with 5-substituted indoles.[16,17] However, 5-nitroindole (**3**) exhibited significantly inferior results with the later variants compared to *Tm*2F3. The later variants, however, exhibited significant improvements with 5,7-disubstituted substrates, providing almost quantitative conversion of **4** to product, even at 37 °C. With substrate **5**, the activity is improved almost an order of magnitude from the starting variant.

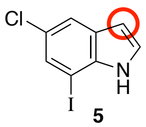
Table 2-2. HPLC yield of Trp analogues with TrpB variants[a]











variant	1	2	3	4	5
***Tm*2F3** :[b]	65%	2.6%	21%	34%	2.3%
***Tm*9D8** :[b]	71%	2.5%	5.3%	94%	20%
Tm*9D8 :[c]	75%	ND	7.5%	92%	18%

[a]Reactions had 0.02 mol % catalyst loading (maximum 5000 turnovers) and 1 equiv of serine relative to indole substrates. [b]Reactions run at 50 °C. [c]Reactions run at 37 °C. Red circles indicated site of C–C bond formation. Yields are averages of two replicates. Full data are reported in Appendix A, **Table A-3**.

2.3 Discussion

2.3.1 Effect of evolution on TrpB activity

4-Cyanoindole is an especially challenging substrate because its nucleophilicity is attenuated both electronically, due to the electron-withdrawing influence of the cyano group, and sterically, since substituents at the 4-position occlude the site of C–C bond formation. However, the new variant *Tm*9D8* exhibits improved activity with this substrate and even functions well at 37 °C. The high expression level of the protein (~40 mg *Tm*9D8* per L culture), the availability of the starting materials, and the convenient reaction setup and product recovery make this an effective method for laboratory preparation of 4-CN-Trp.

Mutations discovered to enhance activity with 4-cyanoindole also improved activity at lower temperatures for other structurally and electronically distinct substrates, such as 4-bromoindole (**1**) and disubstituted indoles **4** and **5**. In all cases, the final variant *Tm*9D8* gave higher yield at 37 °C than the starting variant *Tm*2F3 did at 50 °C. This general boost in activity at lower temperature is valuable because it not only facilitates process development but also enables future exploration of substrates that might be unstable in water at elevated temperature. The mutations, however, did not engender general tolerance for 4-

substitution since the enzyme showed negligible activity with 4-nitroindole (**2**). Surprisingly, activity with 5-nitroindole (**3**) decreased dramatically, even though *Tm*2F3 had originally been evolved for activity with this substrate.[16] These data suggest that the mutations have significantly reconfigured the active site compared to *Tm*2F3, although activity with the native substrate indole remains high for all variants (see Appendix A).

2.3.2 Role of mutations

The seven mutations in the parent protein *Tm*2F3 were all previously identified in a TrpB homologue from *Pyrococcus furiosus*, which had been evolved through global random mutagenesis and screening to accept 4-nitroindole as a nucleophilic substrate. Although the *P. furiosus* homologue has only 64% sequence identity to *T. maritima* TrpB,[17] we found that these seven mutations were activating in both protein scaffolds. Furthermore, the homologous variants had distinct substrate profiles, with the *T. maritima* variant performing better than *P. furiosus* with 4-CN-Trp, as well as indoles **1** and **3**–**5**.

To date, our efforts to solve crystal structures of *T. maritima* TrpB variants have been unsuccessful. We therefore constructed a homology model[18] based on a 1.65-Å crystal structure of *S. typhimurium* TrpB (PDB ID: 4HPX, 58% sequence identity)[19] with the PLP-bound amino-acrylate in the active site. From this model, it is apparent that of the ten mutations in *Tm*9D8* only two reside in the active site (I184F and G228S), with the other eight scattered throughout the protein structure (**Figure 2-3a**). The precise effects of these eight mutations are uncertain, but previous studies have suggested that they stabilize the closed state of the enzyme,[16,17] which is known to promote product formation.

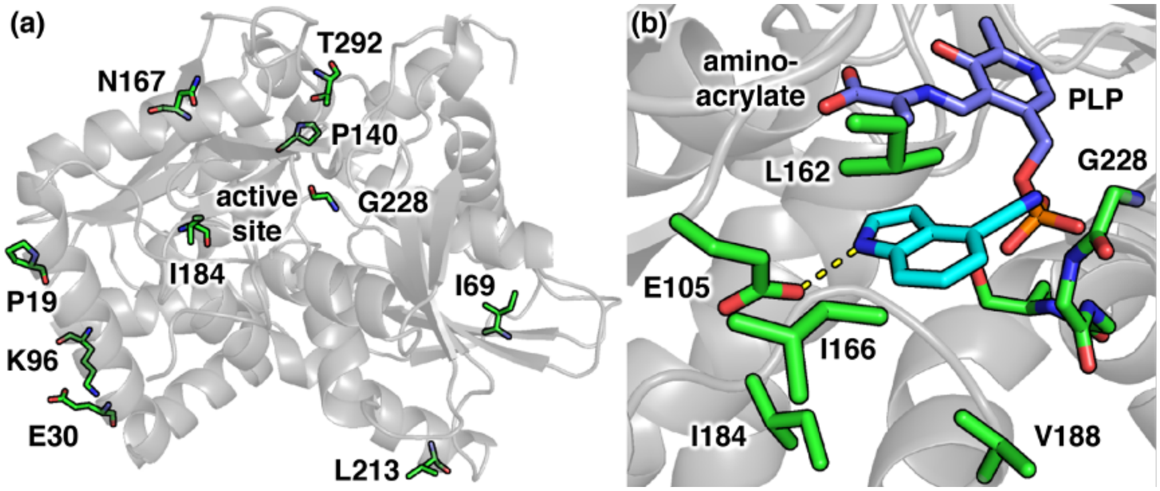


Figure 2-3. Homology model of the *T. maritima* TrpB showing **a.** the whole protein structure with the mutated sites and **b.** the active site with the PLP-bound amino-acrylate, 4-cyanoindole in a reactive binding pose, and residues predicted to interact with 4-cyanoindole highlighted.

The G228S mutation is striking not only because it is an active-site mutation, but also because it is predicted to occur at the beginning of a loop (<u>G</u>GGS) that binds the phosphate moiety of the PLP cofactor. To speculate on the role of this mutation, we modeled 4-cyanoindole in the putative binding pose necessary for C–C bond formation (**Figure 2-3b**). The side chains of residues L162, I166, and V188 extend into the active site and thus are expected to influence the positioning of the indole substrate through hydrophobic interactions. The active site also contains E105, a universally conserved residue that interacts with the endocyclic N–H of the native substrate, indole. It is immediately evident that the 4-cyano substituent would point directly toward the phosphate-binding loop and G228 in particular. Thus, the G228S mutation might reorganize the cofactor-binding site to create space for substituents at the 4-position. A survey of 5,738 TrpB homologues revealed that this GGGS sequence is almost universally conserved. This variant therefore serves as an example of how mutations of universally conserved residues can benefit reactions with non-natural substrates.

2.4 Conclusions

By applying global random mutagenesis to TrpB from *T. maritima*, we have engineered a variant with improved activity for the production of 4-CN-Trp directly from 4-cyanoindole and serine. Whereas the parent enzyme struggled to form 4-CN-Trp at 75 °C, this new variant exhibits considerable activity even at 37 °C, enabling production of 4-CN-Trp under mild conditions. The TrpB-catalyzed reactions occur in aqueous media with readily available starting materials and without the need for protecting groups. Thus, we believe that the TrpB platform will serve as a powerful tool to develop more efficient and direct routes to ncAAs that minimize use of organic solvents.

2.5 Experimental Procedures

2.5.1 General experimental methods

Chemicals and reagents were purchased from commercial sources and used without further purification. The proton NMR spectrum was recorded on a Bruker 400 MHz (100 MHz) spectrometer equipped with a cryogenic probe. Proton chemical shifts are reported in ppm (δ) relative to tetramethylsilane and calibrated using the residual solvent resonance (DMSO, δ 2.50 ppm). The NMR spectrum was recorded at ambient temperature (about 25 °C). Preparative reversed-phase chromatography was performed on a Biotage Isolera One purification system, using C-18 silica as the stationary phase, with CH_3OH as the strong solvent and H_2O (0.1% HCl by weight) as the weak solvent. Liquid chromatography/mass spectrometry (LCMS) was performed on an Agilent 1290 UPLC-LCMS equipped with a C-18 silica column (1.8 µm, 2.1 × 50 mm) using CH_3CN/H_2O (0.1% acetic acid by volume): 5% to 95% CH_3CN over 4 min; 1 mL/min. The optical purity of the products was determined by derivatization with *N*-(5-fluoro-2,4-dinitrophenyl)alanamide (FDNP-alanamide)[20] as described below.

2.5.2 Cloning, expression, and purification of *Tm*TrpB variants

*Tm*TrpB (UNIPROT ID P50909) was previously cloned into pET22b(+) between the *Nde*I and *Xho*I sites with a 6× C-terminal His-tag.[17] This study used the previously described variant *Tm*2F3[16] as the parent for subsequent evolution. All variants were expressed in BL21(DE3) E. cloni® Express cells. Cultures were started from single colonies in 5 mL Terrific Broth supplemented with 100 μg/mL ampicillin (TB_{amp}) and incubated overnight at 37 °C and 230 rpm. For expression, 2.5 mL of overnight culture were used to inoculate 250 mL TB_{amp} in a 1-L flask, which was incubated at 37 °C and 250 rpm for three hours to reach OD_{600} 0.6 to 0.8. Cultures were chilled on ice for 20 minutes and expression was induced with a final concentration of 1 mM isopropyl β-D-thiogalactopyranoside (IPTG). Expression proceeded overnight (approximately 20 hours) at 25 °C and 250 rpm. Cells were harvested by centrifugation at 5,000*g* for five minutes at 4 °C and stored at –20 °C.

Thawed cell pellets were resuspended in 9 mL of lysis buffer containing 50 mM potassium phosphate buffer, pH 8.0 (KPi buffer) with 1 mg/mL hen egg white lysozyme (HEWL), 200 μM PLP, 2 mM $MgCl_2$, 0.02 mg/mL DNase I. Pellets were vortexed until completely resuspended, and then cells were lysed with BugBuster® according to manufacturer's recommendations. Lysates were then heat treated at 75 °C for 10 minutes. The lysate was centrifuged for 15 minutes at 15,000*g* and 4 °C and the supernatant collected. Purification was performed with an AKTA purifier FPLC system (GE Healthcare) and a 1-mL Ni-NTA column. Protein was eluted by applying a linear gradient of 100 mM to 500 mM imidazole in 25 mM KPi buffer and 100 mM NaCl. Fractions containing purified protein were dialyzed into 50 mM KPi buffer, flash frozen in liquid nitrogen, and stored at –80 °C. Protein concentrations were determined using the Bio-Rad Quick Start™ Bradford Protein Assay.

2.5.3 Construction of random mutagenesis libraries

Random mutagenesis libraries were generated from the gene encoding *Tm*2F3 by adding 200 to 400 μM $MnCl_2$ to a Taq PCR reaction as reported previously.[16,21] PCR fragments were treated with *Dpn*I for two hours at 37 °C and purified by gel extraction. The purified library was then cloned into an empty pET22b(+) vector *via* Gibson assembly and transformed into BL21(DE3) E. cloni® Express cells.[22]

Forward primer (*Nde*I): `GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG`
Reverse primer (*Xho*I): `GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAG`

2.5.4 Construction of recombination libraries

Recombination libraries used primers with a degenerate codon to cause a 50/50 amplification of mutant and wild-type residues at a given site (E30G, I184F, G228S) (Appendix A, **Table A-4**). PCR with Phusion® Polymerase (NEB) produced four fragments of the *Tm*2F3 gene (*NdeI* to E30, E30 to I184, I184 to G228, G228 to *XhoI*). Fragments were treated with *Dpn*I for two hours at 37 °C and purified by gel extraction. The fragments were assembled by PCR with flanking primers that correspond to the *Nde*I and *Xho*I sites of the pET22b(+) vector. The assembled gene was then cloned into an empty pET22b(+) vector *via* Gibson assembly and transformed into BL21(DE3) E. cloni® Express cells.[22]

2.5.5 Construction of recombination libraries

Site-saturation libraries were generated using NEB Q5® site directed mutagenesis kit per manufacturer's instructions using *Tm*9D8 as the parent. Primers were designed using NEBaseChanger® software and incorporated the degenerate codons NDT (encoding for Ile, Asn, Ser, Gly, Asp, Val, Arg, His, Leu, Phe, Tyr, and Cys), VHG (encoding for Met, Thr, Lys, Glu, Ala, Val, Gln, Pro, and Leu), and TGG (Trp) at the residue of interest (Appendix A, **Table A-5**). Primers were mixed as reported previously.[23] Following PCR, samples were treated with KLD Enzyme Mix for five minutes, and transformed into BL21(DE3) E. cloni® Express cells.

2.5.6 Library expression and screening.

BL21(DE3) E. cloni® cells carrying variant plasmids were cultured in 96-well deep-well plates along with parent and negative controls as described previously.[16,21] Overnight cultures were grown by inoculating 300 µL TB$_{\text{amp}}$ with a single colony followed by incubation at 37 ºC and 250 rpm with 80% humidity. The following day 20 µL of the overnight culture were added to 630 µL TB$_{\text{amp}}$ and incubated at 37 ºC and 250 rpm with 80% humidity for 3 hours. Cells were then chilled on ice for 20 minutes and induced by addition of IPTG (final concentration 1 mM) followed by incubation at 25 °C and 250 rpm overnight (approximately 20 hours). Cells were pelleted by centrifugation at 5,000*g* for 5 minutes, and then decanted and stored at –20 °C. Cell plates were thawed and resuspended in 300 µL/well 50 mM KPi buffer with 1 mg/mL HEWL, 200 µM PLP, 2 mM $MgCl_2$, and 0.02 mg/mL DNase. Cells were lysed by a 30-minute incubation at 37 °C and heat treatment in a 75 °C water bath for 30 minutes (recombination and site saturation) to 180 minutes (random mutagenesis). Lysate was clarified by centrifugation at 5,000*g* for 10 minutes.

2.5.7 Random mutagenesis screen.

Reactions were performed in a UV-transparent 96-well assay plate with a total volume of 200 µL/well comprised of 40 µL heat-treated lysate, 5 mM 4-cyanoindole, and 50 mM serine with 5% (v/v) DMSO in 50 mM KPi buffer. Reactions proceeded in a 75 °C water bath for 24 hours. Plates were centrifuged briefly to collect condensation and assayed by measuring absorption at 350 nm.

2.5.8 Recombination and site saturation screen.

Reactions were performed in 96-well deep-well plates with a total volume of 200 µL/well comprised of 40 µL heat-treated lysate, 5 mM 4-cyanoindole, and 50 mM serine with 5% (v/v) DMSO in 50 mM KPi buffer. Reactions were sealed with Teflon sealing mats and incubated in a 75 °C water bath for 24 hours. Plates were briefly chilled on ice and centrifuged to collect condensation. Each well was charged with 500 µL 1 M aq. HCl and 500 µL ethyl acetate. The plate was sealed with a Teflon sealing mat followed by vigorous agitation to dissolve all precipitates and partition the product and substrate between the aqueous and organic phases, respectively. The plates were centrifuged for 2 minutes at 5,000*g* and then 200 µL of the aqueous phase was transferred to a 96-well UV-transparent assay plate. Activity was determined by measuring the absorption at 300 nm.

2.5.9 Calibration for measuring HPLC yield.

Using an authentic standard, mixtures of corresponding indole and tryptophan analogs in varied ratios (9:1, 3:1, 1:1, 1:3, and 1:9) were prepared in 1:1 1 M aq. HCl/CH_3CN with a total concentration of 1 mM. Each mixture was prepared in duplicate, then analyzed by LCMS. The ratios of the product and substrate peaks at 254 nm and 280 nm (reference 360 nm, bandwidth 100 nm) were correlated to the actual ratios by a linear relationship (see Appendix A, **Figure A-4**). The authentic standard for 4-cyanotryptophan was obtained from the gram-scale preparation described below. Authentic standards for 4-bromotryptophan, 5-nitrotryptophan, and 5-bromo-7-fluorotyptophan were synthesized as reported previously.[16]

2.5.10 Reactions for Figure 2-2 and Table 2-2.

A 2-mL glass HPLC vial was charged with the nucleophilic substrate as a solution in DMSO (10 µL, 400 mM). Next, serine (20 mM final concentration) and purified enzyme (either 4 µM or 20 µM final concentration) were added as a solution in 190 µL of 50 mM KPi buffer. Reactions were heated to 37 °C, 50 °C, or 75 °C for 24 hours. The reaction was then diluted with 800 µL of 1:1 1 M aq. HCl/CH_3CN and vortexed thoroughly. Finally, the reaction mixture was centrifuged at >20,000*g* for 10 minutes, and the supernatant was analyzed by HPLC. The identity of the product was confirmed by comparison to an authentic standard. The yield was determined by comparing the integrations of the HPLC peaks corresponding to product and starting material (see Appendix A for more details). Experiments were conducted at least in duplicate.

2.5.11 Approximation of initial turnover frequency.

Reactions with 4-cyanoindole were set up according to the procedure described above for **Figure 2-2**. The reactions were worked up after 1 hour and analyzed by HPLC. The integrations of 254-nm absorption peaks corresponding to product and starting material were used to calculate product formation. The reactions were conducted in triplicate. See Appendix A, **Table A-2** for full data.

2.5.12 Gram-scale preparation of 4-cyanotryptophan.

Heat-treated lysate was prepared following the protocol described above for preparing the enzymes for purification. In a 1-L Erlenmeyer flask, 4-cyanoindole (1.0 g, 7.0 mmol) and serine (810 mg, 7.7 mmol) were suspended in DMSO (17.5 mL) and 50 mM KPi buffer (250 mL). Heat-treated lysate from four 250-mL expression cultures was added, then the reaction mixture was heated in a water bath at 55 °C. After 72 hours, the reaction mixture was cooled on ice for 90 minutes. The precipitate was collected by filtration, washed twice with ethyl acetate and twice with water, then dried *in vacuo* to afford 4-CN-Trp as an off-white solid (797 mg, 49% yield).

The ^{1}H NMR spectrum was taken in a mixture of DMSO-d$_6$ and 20% DCl/D$_2$O and referenced to the residual DMSO peak (2.50 ppm). **^{1}H NMR** (400 MHz, DMSO-*d*$_6$) δ 7.57 (dd, J = 8.2, 1.0 Hz, 1H), 7.32 (s, 1H), 7.32–7.29 (m, 1H), 7.09–7.03 (m, 1H), 4.00 (dd, J = 5.8, 2.9 Hz, 1H), 3.30 (**AB**X, J_{AX} = 8.7 Hz, J_{BX} = 6.2 Hz, J_{AB} = 15.2 Hz, ν_{AB} = 85.2 Hz, 2H). The data were in concordance with the previous literature.[16]

2.5.13 Determination of T_{50} values.

A mastermix of 1 μM purified enzyme was prepared in 50 mM KPi buffer and 95 μL added to 12 PCR tubes. Ten test samples were incubated in a thermocycler for 60 minutes with a temperature gradient from 79 °C to 99 °C, while the two control samples were incubated at room temperature. All tubes were centrifuged for three minutes to pellet precipitated enzyme, and then 75 μL of the supernatant was transferred from each tube to a UV-transparent 96-well assay plate. Enzyme activity was determined by adding an additional 75 μL of 50 mM KPi buffer containing 1 mM indole and 1 mM serine to each well. Reactions were incubated for 45 min at 50 °C (*Tm*9D8 and *Tm*9D8*) or 75 °C (*Tm*2F3) and then briefly centrifuged to collect condensation. Activity was determined by measuring the absorption at 290 nm. Activity was correlated to incubation temperature, and the half-inactivation temperatures (T_{50}) were determined. Measurements were conducted in triplicate.

2.5.14 Determination of optical purity.

The optical purity of products was estimated by derivatization with FDNP-alanamide. In a 2-mL vial, a 200 μL reaction was carried out as described above. After 24 hours of incubation at 37 °C, 100 μL of 1 M aq. $NaHCO_3$ was added to the reaction, and 125 μL of the reaction mixture (up to 1.1 μmol product) was transferred into two 2-mL vials. FDNP-alanamide (33 μL of a 33-mM solution in acetone, 1.1 μmol) was added to each vial, followed by incubation at 37 °C and 230 rpm. After two hours, the reaction mixture was cooled to room temperature and then diluted with 1:1 1 M aq. HCl/CH_3CN (600 μL). The resulting solution was analyzed directly by LCMS. Each amino acid was derivatized with both racemic and enantiopure FDNP-alanamide for comparison. Absolute stereochemistry was inferred by analogy to L-tryptophan. All products were >99% ee.

2.5.15 Structural modeling.

A structure of TrpS from *S. enterica* has been reported (PDB ID: 4HPX), in which the β-subunit (*Se*TrpB) is in the closed state and contains benzimidazole and the Ser-derived amino-acrylate in the active site.[19] This structure served as a template for a homology model of wild-type TrpB from *T. maritima* (*Tm*TrpB, 58% sequence identity), which was constructed using the Swiss-Model program.[18] The homology model of *Tm*TrpB was aligned with the authentic structure of *Se*TrpB using PyMOL, which allowed the amino-acrylate and benzimidazole to be mapped directly into the homology model. Finally, the benzimidazole was replaced with a simulated structure of 4-cyanoindole, such that the indole moiety of 4-cyanoindole mapped onto the structure of benzimidazole.

Chapter II Bibliography

1. Smith, D. T., Delost, M. D., Qureshi, H. & Njardarson, J. T. Top 200 pharmaceutical products by retail sales in 2016, http://njardarson.lab.arizona.edu/sites/njardarson.lab.arizona.edu/files/2016Top200PharmaceuticalRetailSalesPosterLowResV2.pdf (accessed August 30, 2017).
2. Ager, D. J. Synthesis of unnatural/nonproteinogenic α-amino acids. In *Amino Acids, Peptides and Proteins in Organic Chemistry: Origins and Synthesis of Amino Acid*s (Hughes, A. B., Ed.); Wiley Online Library; **1**, 495–526 (2010).
3. Brittain, W. D. G., Cobb, S. L. Negishi cross-couplings in the synthesis of amino acids. *Org. Biomol. Chem.* **16**, 10–20 (2017).
4. Arnold, L. D., Kalantar, T. H. & Vederas, J. C. Conversion of serine to stereochemically pure β-substituted α-amino acids via β-lactones. *J. Am. Chem. Soc.* **107**, 7105–7109 (1985).
5. Arnold, L. D., Drover, J. & Vederas, J. C. Conversion of serine β-lactones to chiral α-amino acids by copper-containing organolithium and organomagnesium reagents. *J. Am. Chem. Soc.* **109**, 4649–4659 (1987).
6. Arnold, L. D., May, R. G. & Vederas, J. C. Synthesis of optically pure α-amino acids via salts of α-amino-β-propiolactone. *J. Am. Chem. Soc.* **110**, 2237–2241 (1988).
7. Tanner, D. Chiral aziridines—their synthesis and use in stereoselective transformations. *Angew. Chem. Int. Ed.* **33**, 599–619 (1994).
8. Ishikawa, T. Aziridine-2-carboxylates: Preparation, nucleophilic ring opening, and ring expansion. *Heterocycles* **85**, 2837–2877 (2012).
9. Corr, M. J., Smith, D. & Goss, R. One-pot access to L-5,6-dihalotryptophans and L-alknyltryptophans using tryptophan synthase. *Tetrahedron* **72**, 7306–7310 (2016).
10. Smith, D. R. M., Willemse, T., Gkotsi, D. S., Schepens, W., Maes, B. U. W., Ballet, S. & Goss, R. J. M. The first one-pot synthesis of L-7-iodotryptophan from 7-iodoindole and serine, and an improved synthesis of other L-7-halotryptophans. *Org. Lett.* **16**, 2622–2625 (2014).

11. Perni, S., Hackett, L., Goss, R. J., Simmons, M. J. & Overton, T. W. Optimisation of engineered *Escherichia coli* biofilms for enzymatic biosynthesis of L-halotryptophans. *AMB Express* **3**, 66 (2013).

12. Winn, M., Roy, A. D., Grüschow, S., Parameswaran, R. S. & Goss, R. J. M. A convenient one-step synthesis of L-aminotryptophans and improved synthesis of 5-fluorotryptophan. *Bioorg. Med. Chem. Lett.* **18**, 4508–4510 (2008).

13. Goss, R. J. M. & Newill, P. L. A. A convenient enzymatic synthesis of L-halotryptophans. *Chem. Commun.* **47**, 4924–4925 (2006).

14. Lee, M. & Phillips, R. S. Enzymatic synthesis of chloro-L-tryptophans. *Bioorg. Med. Chem. Lett.* **2**, 1563–1564 (1992).

15. Hilaire, M. R., Ahmed, I. A., Lin, C.-W., Jo, H., DeGrado, W. F. & Gai, F. Blue fluorescent amino acid for biological spectroscopy and microscopy. *Proc. Nat. Acad. Sci. U.S.A.* **114**, 6005–6009 (2017).

16. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

17. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A panel of TrpB biocatalysts derived from tryptophan synthase through the transfer of mutations that mimic allosteric activation. *Angew. Chem. Int. Ed.* **55**, 11577–11581 (2016).

18. SWISS-MODEL available online at https://swissmodel.expasy.org (accessed January 16, 2018)

19. Niks, D., Hilario, E., Dierkers, A., Ngo, H., Borchardt, D., Neubauer, T. J., Fan, L., Mueller, L. J. & Dunn, M. F. Allostery and substrate channeling in the tryptophan synthase bienzyme complex: Evidence for two subunit conformations and four quaternary states. *Biochemistry* **52**, 37, 6396–6411 (2013).

20. Brückner, H. & Gah, C. High-performance liquid chromatographic separation of DL-amino acids derivatized with chiral variants of Sanger's reagent. *J. Chromatogr. A* **555**, 81 (1991).

21. Buller, A. R., Brinkmann-Chen, S., Romney, D. K., Herger, M., Murciano-Calles, J. & Arnold, F. H. Directed evolution of the tryptophan synthase β-subunit for stand-

alone function recapitulates allosteric activation. *Proc. Nat. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

22. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A. & Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343–345 (2009).

23. Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T. & Acevedo, J. P. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2**, 83–92 (2013).

Chapter III

DIRECT ENZYMATIC SYNTHESIS OF A DEEP-BLUE FLUORESCENT NONCANONICAL AMINO ACID FROM AZULENE AND SERINE

Material from this chapter appears in: “Watkins E. J.,[†] **Almhjell P. J.**[†] & Arnold F. H., Direct enzymatic synthesis of a deep-blue fluorescent noncanonical amino acid from azulene and serine. *ChemBioChem* **21**, 80-83 (2020) doi: 10.1002/cbic.201900497” ([†]denotes equal contribution).

P.J.A. and E.J.W. participated in the conception, design, and execution of the research. E.J.W. designed the screen for identifying improved enzyme variants. P.J.A and E.J.W. contributed equally to AzAla purification. P.J.A. performed enzyme kinetics. E.J.W. prepared the first manuscript and P.J.A. edited and constructed figures.

ABSTRACT

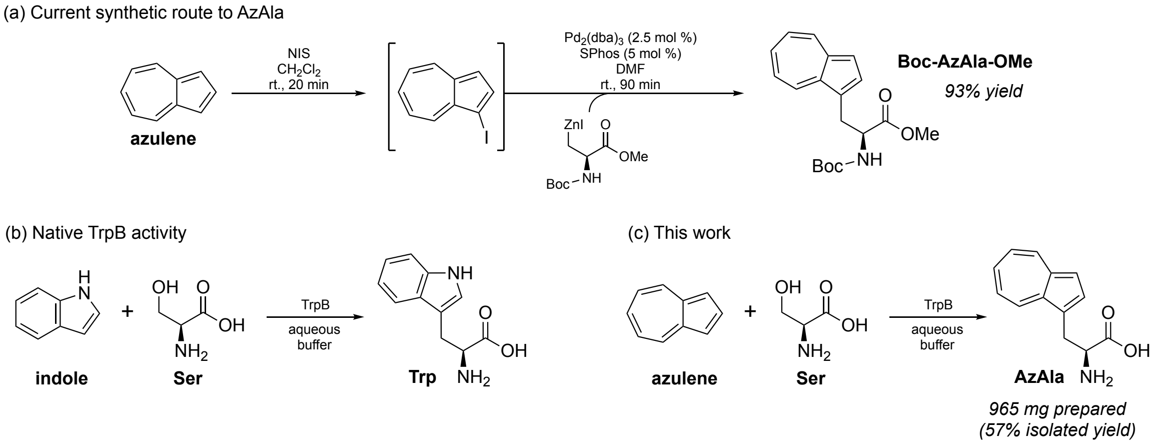
We report a simple, one-step enzymatic synthesis of the blue fluorescent noncanonical amino acid β-(1-azulenyl)-L-alanine (AzAla). Using an engineered tryptophan synthase β-subunit (TrpB), stereochemically pure AzAla can be synthesized at scale starting from commercially available azulene and L-serine. Mutation of a universally conserved catalytic glutamate in the active site to glycine has only a modest effect on native activity with indole but abolishes activity on azulene, suggesting that this glutamate activates azulene for nucleophilic attack by stabilization of the aromatic ion.

3.1 Introduction

Proteins and peptides can be imbued with new chemical and physical properties via the inclusion of noncanonical amino acids (ncAAs). These molecules resemble the natural building blocks of proteins but contain distinct structures and functional groups. When incorporated into proteins, ncAAs can serve as handles for chemical reactions or as spectroscopic probes to examine protein function, including reactivity, localization, and interaction with other biomolecules.[1–3] Unfortunately, applications of many potentially useful ncAAs are limited owing to their high cost and lack of availability. The paucity of available ncAAs also hinders engineering of aminoacyl tRNA synthetases (aaRS) necessary for site-specific, *in vivo* ncAA incorporation into proteins.

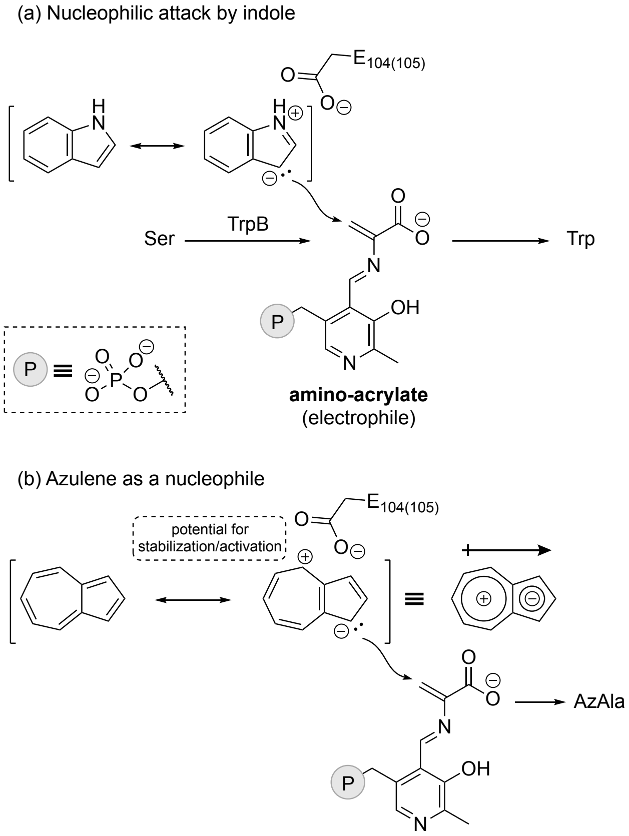
The ncAA β-(1-azulenyl)-L-alanine (AzAla, **Scheme 3-1**) is a tryptophan (Trp) isostere with unique fluorescent properties that make it a useful spectroscopic probe for investigating protein dynamics and protein-protein interactions. It can be incorporated into proteins in place of Trp without significantly disturbing tertiary structure or function.[4–8] In contrast to Trp, its spectroscopic properties are insensitive to the environment, making it ideal in contexts where local conditions and quenchers (e.g., methionine, histidine) could complicate analysis of fluorescent signals.[4] Its qualities have been leveraged for Förster resonance energy transfer (FRET) experiments to elucidate protein-protein interactions as well as vibrational energy transfer (VET) studies to probe anisotropic energy flow within proteins.[9,10] Recently, a method for synthesis of AzAla via Negishi cross-coupling was described (**Scheme 3-1a**).[11] Although proceeding with good yields on gram scale, the multi-step process is highly time sensitive and uses precious metal catalysts and organic solvents. A simpler route to AzAla would expand its applications in biochemical studies.

Researchers have begun to look to enzymes as complementary or alternative approaches for the synthesis of enantiomerically pure ncAAs.[12] Enzymes can perform enantio- and regioselective chemistry in the presence of reactive moieties such as primary amines, obviating the need for expensive and intricate chiral catalysts, chiral separations, and protecting groups. To this end, our lab previously reported the directed evolution of the tryptophan synthase β-subunit (TrpB) as a stand-alone biocatalytic platform for the synthesis of diverse tryptophan analogues (**Scheme 3-1b**).[13,14] Here we report a simple, efficient route for synthesis of AzAla from stable, commercially available starting materials using an engineered TrpB (**Scheme 3-1c**).



Scheme 3-1. Synthesis of AzAla and native TrpB activity. **a.** Current synthetic route to Boc-protected AzAla. **b.** TrpB natively catalyzes the condensation of indole and serine to form tryptophan. **c.** Single-step biocatalytic synthesis of AzAla from azulene and serine described in this work.

TrpB is a Type-II pyridoxal phosphate (PLP)-dependent enzyme that natively performs a conjugate addition reaction between indole and L-serine (Ser) to make Trp. During the catalytic cycle, Ser binds the PLP cofactor, and subsequent β-elimination and release of water forms an electrophilic amino-acrylate species (**Scheme 3-2a**). A highly conserved active-site glutamate stabilizes the accumulation of positive charge on the pyrrole ring of indole and helps facilitate the nucleophilic attack on the amino-acrylate to form a new C–C bond that produces Trp. The similarity of AzAla to Trp prompted us to investigate whether TrpB could accept azulene as a nucleophile in the place of indole. Unlike indole, azulene lacks heteroatoms that can help stabilize the accumulation of charge during nucleophilic attack. Despite this, azulene has a permanent dipole exemplified by its resonance structure of a cycloheptatrienyl cation (tropylium) fused to a cyclopentadienyl anion (Cp^-). We hypothesized that, analogous to indole, the buildup of electron density on the Cp^- could promote nucleophilic attack by azulene in the TrpB catalytic cycle, while the tropylium system could stabilize the resulting positive charge (**Scheme 3-2b**). Aromatic ions are common moieties in synthetic chemistry[15] and as enzyme inhibitors,[16] but there are few reports of enzymes that can interact productively with aromatic ions in their catalytic cycles. We were therefore unsure if the active site of TrpB could accommodate or activate such a substrate, or if the reactivity of azulene would be sufficient for nucleophilic attack.


(a) Nucleophilic attack by indole
E104(105)
Ser
TrpB
Trp
amino-acrylate
(electrophile)
(b) Azulene as a nucleophile
E104(105)
potential for stabilization/activation
AzAla
P
OH
N

Scheme 3-2. Parallels between indole and azulene in the TrpB reaction.

3.2 Results

We first examined the conversion of azulene and Ser to AzAla using a small panel of previously engineered TrpB variants from the thermophilic organisms *Pyrococcus furiosus* (*Pf*TrpB) and *Thermotoga maritima* (*Tm*TrpB). The variants were selected to provide an efficient sampling of the engineered TrpB evolutionary lineage beginning from wild-type TrpS and ending with stand-alone TrpB variants evolved for activity with different indole and serine analogues. Nearly every enzyme we tested demonstrated significant activity for this reaction, the exception being variants in which the highly conserved glutamate mentioned above was mutated to glycine.

The significant effect of the E104(105)G mutation suggested that this conserved catalytic residue may be playing an important role in the non-native azulene reaction. We explored this possibility by examining two engineered variants with and without the E104(105)G mutation: *Pf*5G8, which exhibits optimal activity at 75 °C,[13] and *Tm*9D8*, which exhibits optimal activity at lower temperatures such as 37 °C.[17] Challenging the enzymes with indole demonstrated that this mutation only modestly decreases the rate of Trp formation (**Figure 3-1**), with an additional slight decrease in the chemoselectivity of the reaction that leads to formation of trace amounts of isoTrp (a product of the N-alkylation of indole, shown in Appendix B, **Figure B-1** and described previously[13]). In contrast, the E104(105)G mutation exerts a profound effect on AzAla production, practically abolishing all activity with azulene. Product was only detected in low amounts when the reaction was performed with high catalyst loadings overnight (0.1 mol %, 16 h; Appendix B, **Figure B-2**). We speculate that the glutamate residue stabilizes the tropylium cation to facilitate nucleophilic attack from Cp^{-}. However, further mechanistic studies are required to elucidate the role of this mutation.

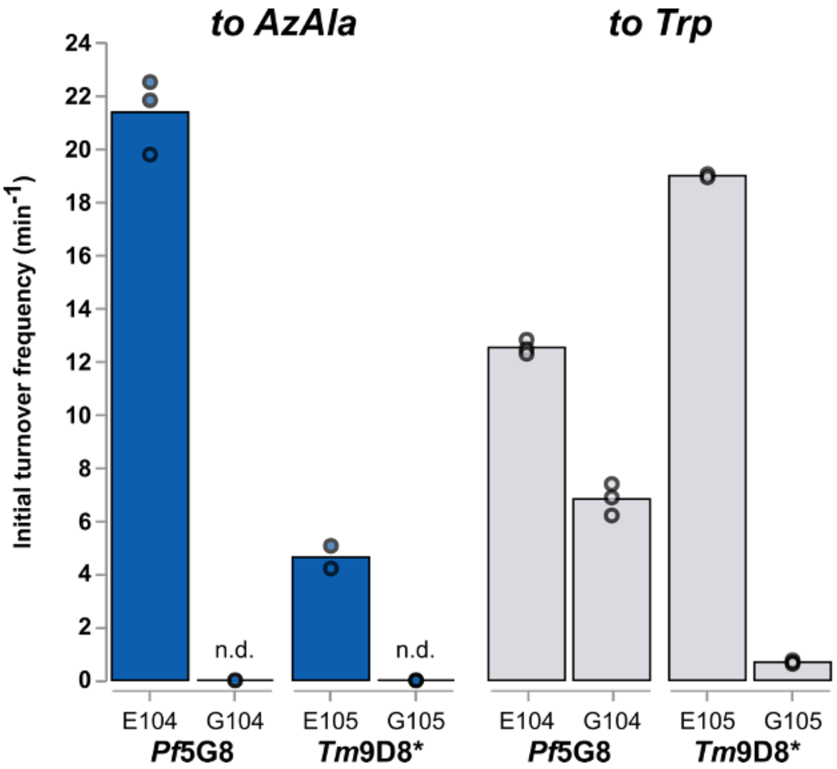


Figure 3-1. Effect of E104(105)G mutation on Trp and AzAla production. Mutation of a conserved glutamate residue affects the rate of AzAla production more significantly than that of native Trp production. Bars represent the average of two to three replicates, with replicates being shown as individual points. Reaction conditions can be found in Table B-1. n.d. = not detected by LC-MS.

Next, we wished to develop a biocatalytic method for the production of AzAla at scale. Although *Pf*5G8 catalyzed the reaction roughly 4.5-fold faster at its optimum temperature of 75 °C than *Tm*9D8* at 37 °C, azulene readily sublimes at high temperatures, making its containment at 75 °C difficult. We thus opted to use directed evolution to improve the activity of *Tm*9D8* at 37 °C to create a stand-alone biocatalyst for the production of AzAla *in vivo* and *in vitro*.

A single round of random mutagenesis and screening identified two variants containing the mutations F184S and W286R. These mutations were combined to yield the final variant *Tm*Azul, which had a three-fold improved rate of AzAla formation compared to *Tm*9D8*

(14.0 turnovers per minute; Appendix B, **Table B-1**). As we have not been successful in obtaining a crystal structure of *Tm*TrpS and its variants, we constructed a homology model of *Tm*9D8* to try to understand the locations of each mutation and their possible effects (**Figure 3-2**). The W286R mutation sits on a flexible loop, where its role in catalysis is difficult to infer. F184S sits directly in the active site and is one of only a handful of residues whose side chains are in close proximity to the azulene substrate. This residue may interact directly with the substrate during catalysis or adjust the active site to be more accommodating for AzAla synthesis.

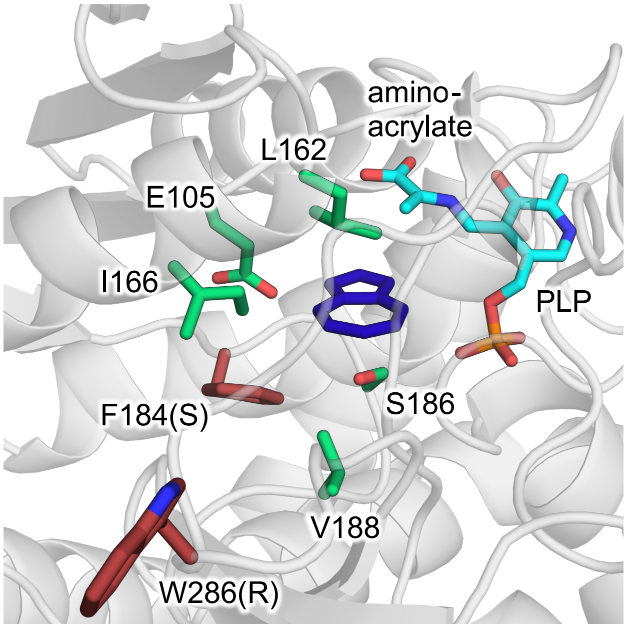


Figure 3-2. Homology model of *Tm*9D8* with azulene in the active site. Azulene (deep blue) is shown in a putative productive binding mode within a model of *Tm*9D8* with the amino-acrylate intermediate (cyan) formed in the active site (see Appendix B, **Section B.4** for model construction). Active-site residues in the native *Tm*TrpB enzyme (green) and the mutations found in this study (red, mutation in parentheses) are shown as sticks.

Because the *Escherichia coli* expression cultures were heat treated at 75 °C for three hours prior to screening, *Tm*Azul retains high thermostability, and highly pure enzyme can be obtained simply by heating the *E. coli* expression host at 75 °C for >1 h, pelleting the denatured *E. coli* proteins by centrifugation, and collecting the enzyme-bearing supernatant. To demonstrate scalability of this method, we synthesized AzAla on a gram scale using heat-treated lysate from a 1-L culture expressing the evolved *Tm*Azul variant. We found that increasing the concentration of DMSO in the reaction increased the reaction rate at this scale, presumably by keeping azulene from forming insoluble crystals in the reaction mixture due to the sparing solubility of azulene in aqueous buffer. We thus used 20% DMSO cosolvent, and the reaction progress was monitored by taking small aliquots of the reaction mixture and combining them with an equal volume of ethyl acetate to observe the relative ratio of azulene to AzAla in the organic and aqueous layers, respectively. After 48 hours, the product was purified by removing any remaining azulene by extraction with ethyl acetate, removing the aqueous solvent *in vacuo*, and precipitating AzAla from the remaining DMSO cosolvent by the addition of excess ethyl acetate. The crude precipitate was collected by filtration and then further purified by reverse-phased column chromatography to afford 965 mg of pure AzAla (57% isolated yield). Crude azulene can be recovered from ethyl acetate by gently evaporating off the solvent under a constant stream of nitrogen and reused. The enantiopurity of the isolated AzAla product is >99% ee.

In conclusion, we have described a new-to-nature reaction catalyzed by tryptophan synthase and identified a conserved residue that is critical for this non-native reaction. Based on comparisons to the native reaction, we suggest that E104(105) stabilizes the aromatic ions in azulene to facilitate nucleophilic attack from Cp^{-}. To improve access to this useful ncAA, we have engineered a highly active enzyme catalyst that synthesizes enantiomerically pure AzAla from commercially available starting materials in a single step. Given that an engineered synthetase/tRNA pair has been reported for this ncAA, *Tm*Azul also closes the gap for *in vivo* synthesis and incorporation of AzAla.[10]

Chapter III Bibliography

1. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79**, 413–444 (2010).
2. Neumann, H. Rewiring translation - Genetic code expansion and its applications. *FEBS Lett.* **586**, 2057–2064 (2012).
3. Agostini, F. *et al.* Biocatalysis with unnatural amino acids: Enzymology meets xenobiology. *Angew. Chem. Int. Ed.* **56**, 9680–9703 (2017).
4. Gosavi, P. M., Moroz, Y. S. & Korendovych, I. V. β-(1-Azulenyl)-L-alanine - A functional probe for determination of pK$_a$ of histidine residues. *Chem. Commun.* **51**, 5347–5350 (2015).
5. Loidl, G. *et al.* Synthesis of β-(1-azulenyl)-L-alanine as a potential blue-colored fluorescent tryptophan analog and its use in peptide synthesis. *J. Pept. Sci.* **6**, 139–144 (2000).
6. Sartori, E. *et al.* An oligopeptide doubly labelled with an azulene chromophore and a TEMPO radical. Azulene triplet generation by enhanced ISC from S2. *Chem. Phys. Lett.* **385**, 362–367 (2004).
7. Venanzi, M. *et al.* Structural properties and photophysical behavior of conformationally constrained hexapeptides functionalized with a new flourescent analog of tryptophan and a nitroxide radical quencher. *Biopolymers* **75**, 128–139 (2004).
8. Mazzuca, C. *et al.* Mechanism of membrane activity of the antibiotic trichogin GA IV: A two-state transition controlled by peptide concentration. *Biophys. J.* **88**, 3411–3421 (2005).
9. Moroz, Y. S., Binder, W., Nygren, P., Caputo, G. A. & Korendovych, I. V. Painting proteins blue: β-(1-Azulenyl)-L-alanine as a probe for studying protein-protein interactions. *Chem. Commun.* **49**, 490–492 (2013).

10. Baumann, T. *et al.* Site-Resolved Observation of Vibrational Energy Transfer Using a Genetically Encoded Ultrafast Heater. *Angew. Chem. Int. Ed.* **58**, 2899–2903 (2019).

11. Stempel, E., Kaml, R. F. X., Budisa, N. & Kalesse, M. Painting argyrins blue: Negishi cross-coupling for synthesis of deep-blue tryptophan analogue β-(1-azulenyl)-L-alanine and its incorporation into argyrin C. *Bioorganic Med. Chem.* **26**, 5259–5269 (2018).

12. Almhjell, P. J., Boville, C. E. & Arnold, F. H. Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997 (2018).

13. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc* **139**, 10769–10776 (2017).

14. Buller, A. R. et al. Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

15. Komatsu, K. & Kitagawa, T. Cyclopropenylium cations, cyclopropenones, and heteroanalogues - recent advances. *Chem. Rev.* **103**, 1371–1427 (2003).

16. Himmel, D. M. *et al.* Structure of HIV-1 Reverse transcriptase with the inhibitor β-thujaplicinol bound at the RNase H active site. *Structure* **17**, 1625–1635 (2009).

17. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima. J. Org. Chem.* **83**, 7447–7452 (2018).

Chapter IV

SCALABLE, CONTINUOUS EVOLUTION FOR THE GENERATION OF DIVERSE ENZYME VARIANTS ENCOMPASSING PROMISCUOUS ACTIVITIES

Material from this chapter appears in: “Rix, G., Watkins-Dulaney, E. J., **Almhjell, P. J.**, Boville, C. E., Arnold, F. H. & Liu, C. C. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nature Communications* **11**, 5644 (2020)”.

All authors contributed to experimental design and data analysis. G.R. constructed the selection and performed all evolution experiments. P.J.A. analyzed results and performed enzyme kinetics. E.J.W.D. performed the panel HPLC-MS assays and indole conversion rate measurements on *Tm*TrpBs. C.E.B. performed *in vitro* characterizations of *Tm*TrpBs from variant set 1 and performed the thermal shift assay and substrate scope characterizations. G.R. and C.C.L. wrote the manuscript with input and contributions from all authors.

ABSTRACT

Enzyme orthologs sharing identical primary functions can have different promiscuous activities. While it is possible to mine this natural diversity to obtain useful biocatalysts, generating comparably rich ortholog diversity is difficult, as it is the product of deep evolutionary processes occurring in a multitude of separate species and populations. Here, we take a first step in recapitulating the depth and scale of natural ortholog evolution on laboratory timescales. Using a continuous directed evolution platform called OrthoRep, we rapidly evolved the *Thermotoga maritima* tryptophan synthase β-subunit (*Tm*TrpB) through multi-mutation pathways in many independent replicates, selecting only on *Tm*TrpB's primary activity of synthesizing L-tryptophan from indole and L-serine. We find that the resulting sequence-diverse *Tm*TrpB variants span a range of substrate profiles useful in industrial biocatalysis and suggest that the depth and scale of evolution that OrthoRep affords will be generally valuable in enzyme engineering and the evolution of new biomolecular functions.

4.1 Introduction

Natural enzymes typically have many orthologs. While the primary activity of orthologous enzymes is largely the same,[1] promiscuous functions not under selective pressure can vary widely.[2,3] Such variation may be attributed to the deep and distinct evolutionary histories shaping each ortholog, including long periods of neutral drift, recalibration of primary activity, and adaptation to new host environments such as temperature. These rich histories act to produce extensive genetic diversity, which underpins different promiscuity profiles.[2]

Diversity in promiscuous functions across orthologs is of both fundamental and practical importance. An enzyme's reserve of promiscuous activities dictates what secondary reactions, environmental changes, or niches the enzyme can accommodate.[4,5] Diversity in promiscuous activities therefore contributes to the basic robustness of life and adaptation. An enzyme's reserve of promiscuous activities can also be mined in the application of enzymes for biocatalysis.[6,7] Ortholog diversity therefore expands the range of reactions at the disposal of enzyme engineers, supporting the growing role of "green" enzymatic processes in the chemical and pharmaceutical industries.[8–10]

Inspired by the remarkable ability of enzyme orthologs to encompass promiscuous activities, we asked whether we could extend the substrate scope of useful enzymes by evolving multiple versions of an enzyme in the laboratory, selecting only for its primary function. Although this idea has been explored before using classical directed evolution approaches, most notably through the generation of cryptic genetic variation with neutral drift libraries,[11–14] we recognized that our recently developed continuous evolution system, OrthoRep, may be considerably better poised for this challenge.[15,16] Classical directed evolution mimics evolution through an iterative procedure that involves diversifying a gene of interest (GOI) *in vitro* (e.g., through error-prone PCR), transforming the resulting GOI library into cells, and selecting or screening for desired activities, where each cycle of this procedure represents one step in an evolutionary search.[17] However, since each cycle is manually staged, classical directed evolution does not readily admit depth and scale during exploration of functional sequence space — it is difficult to carry out many iterations to mimic lengthy evolutionary searches (depth), let alone do so in many independent experiments (scale). Yet evolutionary depth and scale are precisely the two characteristics responsible for ortholog diversity in nature. Natural orthologs have diversified from their ancestral parent over great evolutionary timescales, allowing for the traversal of long mutational pathways shaped by complex selection histories (depth). Natural orthologs are also the result of numerous independent evolutionary lineages, since spatially separated species and populations are free to take divergent mutational paths and experience different environments (scale). Systems that better mimic the depth and scale of natural enzyme evolution, but on laboratory timescales, are thus needed for the effective generation of enzyme variants that begin to approach the genetic and promiscuity profile diversity of orthologs.

OrthoRep is such a system. In OrthoRep, an orthogonal error-prone DNA polymerase durably hypermutates an orthogonal plasmid (p1) without raising the mutation rate of the host *Saccharomyces cerevisiae* genome.[16] Thus, GOIs encoded on p1 rapidly evolve when cells are simply passaged under selection. By reducing the manual stages of classical directed evolution down to a continuous process where cycles of diversification and selection occur autonomously *in vivo*, OrthoRep readily accesses depth and scale in evolutionary search.[16,18] Here, we apply OrthoRep to the evolution of the *Thermotoga maritima* tryptophan synthase β-subunit (*Tm*TrpB) in multiple independent continuous evolution experiments, each carried out for at least 100 generations. While we only pressured *Tm*TrpB to improve its primary activity of coupling indole and serine to produce tryptophan, the large number of independent evolution experiments we ran (scale) and the high degree of adaptation in each experiment (depth) resulted in a panel of variants encompassing expanded promiscuous activity with indole analogs. In addition to the immediate value of these newly evolved *Tm*TrpBs in the synthesis of tryptophan analogs, our study offers a new template for enzyme engineering where evolutionary depth and scale is leveraged on laboratory timescales to generate effective variant collections covering broad substrate scope.

4.2 Results

4.2.1 Establishing a selection system for the evolution of *Tm*TrpB variants

To evolve *Tm*TrpB variants using OrthoRep, we first needed to develop a selection where yeast would rely on *Tm*TrpB's primary enzymatic activity for growth. *Tm*TrpB catalyzes the PLP-dependent coupling of L-serine and indole to generate L-tryptophan (Trp) in the presence of the tryptophan synthase α-subunit, *Tm*TrpA.[19] In *T. maritima* and all other organisms that contain a heterodimeric tryptophan synthase complex, TrpA produces the indole substrate that TrpB uses and the absence of TrpA significantly attenuates the activity of TrpB through loss of allosteric activation.[19,20] TRP5 is the *S. cerevisiae* homolog of this heterodimeric enzyme complex, carrying out both TrpA and TrpB reactions and producing Trp for the cell. We reasoned that by deleting the *TRP5* gene and forcing *S. cerevisiae* to rely on *Tm*TrpB instead, cells would be pressured to evolve high stand-alone *Tm*TrpB activity in order to produce the essential amino acid Trp in indole-supplemented media (**Figure 4-1**). This selection pressure would also include thermoadaptation, as yeast grow at mesophilic temperatures in contrast to the thermophilic source of *Tm*TrpB. Therefore, the selection on

*Tm*TrpB's primary activity would be multidimensional — stand-alone function, temperature, and neutral drift implemented when desired — and could result in complex evolutionary pathways that serve our goal of maximizing functional variant diversity across replicate evolution experiments. In addition, the multidimensional selection also serves practical goals as stand-alone activity is useful in biosynthetic applications (enzyme complexes are difficult to express and use *in vitro*) and activity at mesophilic temperatures is more compatible with heat-labile substrates, industrial processes where heating costs can compound, or *in vivo* applications in model mesophilic hosts (*e.g. S. cerevisiae* or *Escherichia coli*).

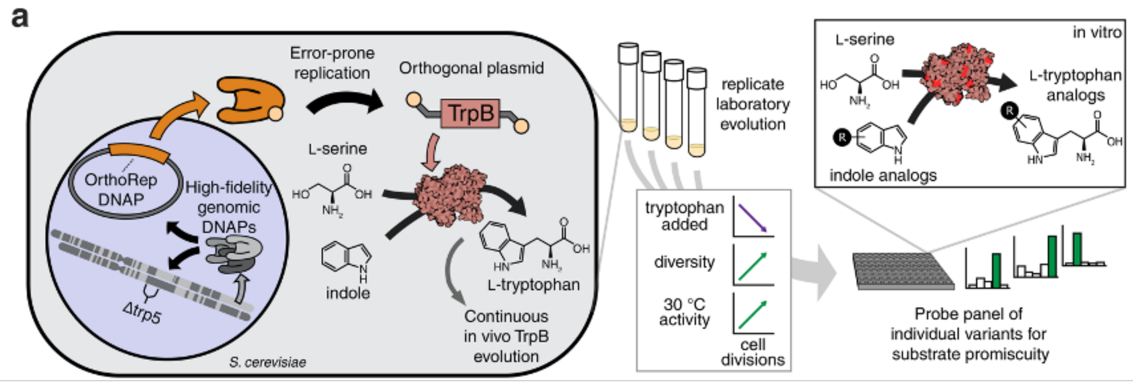


Figure 4-1. Pipeline for the use of OrthoRep continuous directed evolution to generate many diverse, functional *Tm*TrpB sequences. *Tm*TrpB variants are first evolved in replicate for Trp production in yeast. OrthoRep enables replicate evolution through error-prone replication of an orthogonal plasmid by an orthogonal polymerase, maintaining low error rates in genome replication. By encoding a *Tm*TrpB variant on this plasmid in a tryptophan synthase (TRP5) deletion mutant, *Tm*TrpB may be both continuously diversified and selected for through gradual reduction in Trp supplied in the growth medium. Evolved populations containing many diverse, functional individuals may then be randomly sampled and tested for activity with indole analogs.

To test this selection, we turned to a positive control *Tm*TrpB variant called *Tm*Triple. This variant was previously engineered to enable stand-alone activity, free from dependence on allosteric activation by TrpA, through a minimal set of three mutations.[7] We found that *Tm*Triple rescued TRP5 function in a Δ*trp5* strain in an indole-dependent manner, validating our selection (Appendix C, **Figure C-1**). Notably, *Tm*Triple, along with other TrpB variants tested, only supported complementation when expressed from a high-strength promoter (Appendix C, **Figure C-1**). This highlighted the opportunity for substantial adaptation and drift even in evolution experiments that start from already engineered *Tm*TrpB variants.

4.2.2 Continuous evolution of *Tm*TrpB with depth and scale

We encoded wild-type (wt) *Tm*TrpB, *Tm*Triple, as well as a nonsense mutant of *Tm*Triple, *Tm*TripleQ90*, onto OrthoRep's p1 plasmid, which is replicated by a highly error-prone orthogonal DNA polymerase. *Tm*TripleQ90* was included because reversion of the stop codon at position 90 in *Tm*TripleQ90* would act as an early indication that adaptation was occurring, giving us confidence to continue passaging our evolution experiments for several weeks to maximize evolutionary search depth. In all three OrthoRep Δtrp5 strains, the initial *Tm*TrpB sequences enabled only minimal indole-dependent complementation (Appendix C, **Figure C-1**). This was expected for wt *Tm*TrpB, which has low stand-alone enzymatic activity and *Tm*TripleQ90*, which has a premature stop codon, and was unsurprising for *Tm*Triple, since *Tm*Triple displayed indole-dependent complementation only when artificially overexpressed (Appendix C, **Figure C-1**).

To continuously evolve *Tm*TrpB, we passaged cells encoding wt *Tm*TrpB, *Tm*Triple, or *Tm*TripleQ90* on OrthoRep in the presence of 100 μM indole while reducing the amount of Trp in the medium over time. In total, six 100-mL and twenty 3-mL cultures were passaged, each representing a single independent evolutionary trajectory. Passages were carried out as 1:100 dilutions where Trp concentrations were decreased in the $(N+1)^{\text{th}}$ passage if cells grew quickly in the N^{th} passage, until Trp was fully omitted. All six of the 100-mL cultures, and four of the twenty 3-mL cultures fully adapted and were capable of robust growth in indole-supplemented media lacking Trp after 90–130 generations (13–20 passages) (**Figure 4-2**; Appendix C, **Table C-1**). Populations that did not achieve growth in the absence of Trp still adapted, but stopped improving at ~5 µM supplemented Trp, suggesting a suboptimal local fitness maximum that is easier to escape through the greater sequence diversity represented in larger populations. This could explain the different success rates in reaching full adaptation between the 3-mL and 100-mL populations. Cultures that did adapt fully were passaged for an additional ~40 generations without increasing selection stringency to allow for accumulation of further diversity through neutral drift.

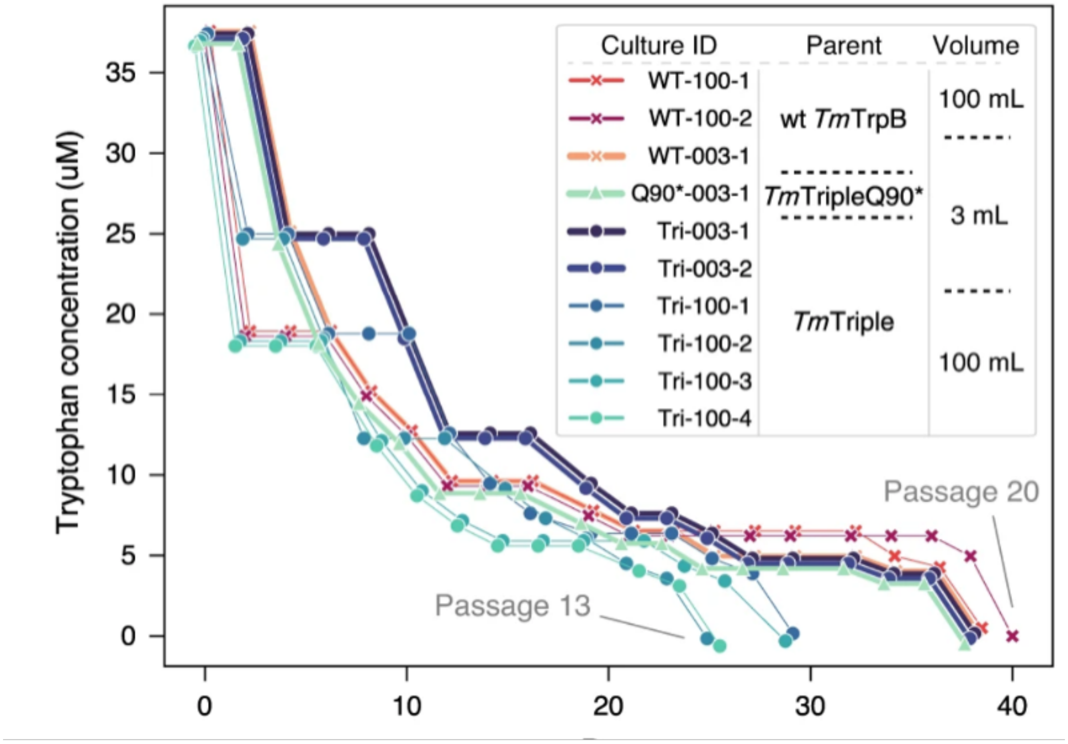
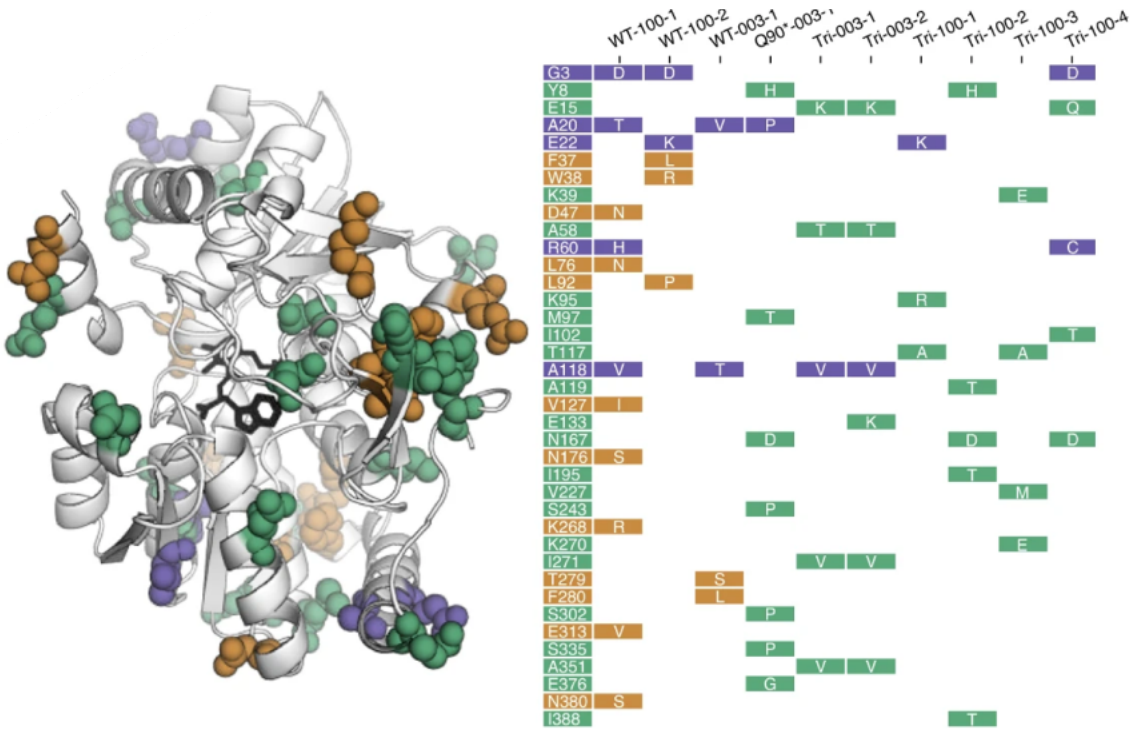


Figure 4-2. Selection trajectories for ten replicate cultures. Each culture evolved sufficient *Tm*TrpB activity to support cell growth without supplemented Trp. Each point represents a single 1:100 dilution (passage) into fresh indole-supplemented growth medium. Trp concentration of fresh media was reduced when high saturation was achieved in the previous passage. Plots are slightly offset from true values to allow for visibility of all selection trajectories.

For each of the ten fully adapted populations, we PCR amplified and bulk sequenced the *Tm*TrpB alleles on the p1 plasmid. Mutations relative to the parent *Tm*TrpB variant detected at >50% frequency in each population were deemed consensus mutations for that population, with the exception of reversion of the stop codon in populations evolving *Tm*TripleQ90*. This stop codon reversion occurred at 100% frequency in the relevant populations and was not counted in any subsequent analyses due to its triviality. An average of 5.6 (± 2.3 s.d.) and a range of 3–11 consensus amino-acid changes per population were observed (**Figure 4-3**; Appendix C, **Table C-2**). Some of these mutations occurred at residues previously identified as relevant in conformational dynamics (e.g., N167D and S302P).[20–22] Most mutations observed, however, have not been previously identified in laboratory engineering experiments, suggesting that even the consensus of these populations explored new regions of *Tm*TrpB's fitness landscape, doing so with diversity across replicates (**Figure 4-3**) that might translate to diversity in promiscuous activities across evolved variants.



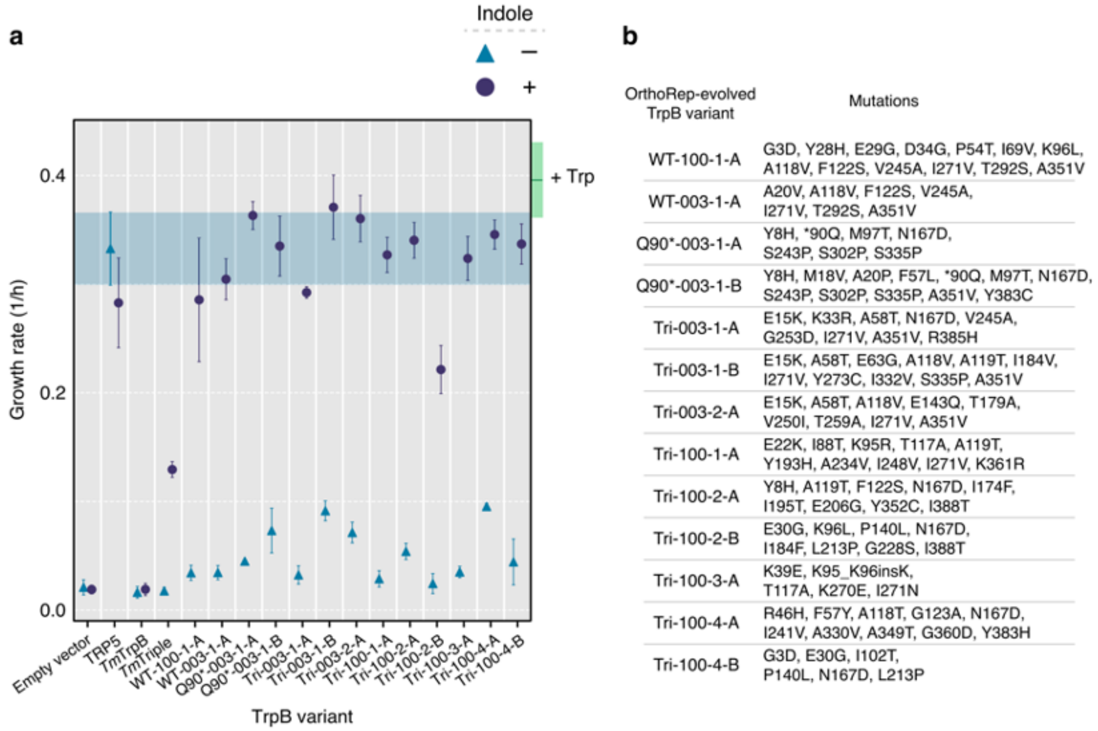
	WT-100-1	WT-100-2	WT-003-1	Q90*-003-1	Tri-003-1	Tri-003-2	Tri-100-1	Tri-100-2	Tri-100-3	Tri-100-4
G3	D	D								D
Y8				H				H		
E15					K	K				Q
A20	T		V	P						
E22		K					K			
F37		L								
W38		R								
K39									E	
D47	N									
A58					T	T				
R60	H									C
L76	N									
L92		P								
K95							R			
M97				T						
I102										T
T117							A		A	
A118	V		T		V	V				
A119								T		
V127	I									
E133						K				
N167				D				D		D
N176	S									
I195								T		
V227									M	
S243				P						
K268	R									
K270									E	
I271					V	V				
T279			S							
F280			L							
S302				P						
E313	V									
S335				P						
A351					V	V				
E376				G						
N380	S									
I388								T		

Figure 4-3. *Tm*TrpB homology model and table depicting consensus mutations of the ten cultures shown in Figure 4-1. Mutations are colored by their appearance in populations evolved from wt *Tm*TrpB (orange), *Tm*Triple or *Tm*TripleQ90* (green), or both (purple).

4.2.3 Evolved *Tm*TrpB variants improve Trp production *in vivo* and contain cryptic genetic variation

To ensure that evolved *Tm*TrpB variants and not potential host genomic mutations were primarily responsible for each population's adaptation, we cloned individual *Tm*TrpBs into a standard low copy yeast nuclear plasmid under a promoter that approximates expression from p1,[23,24] transformed a fresh Δtrp5 strain with the constructs, and tested for the ability of the variants to support indole-dependent growth in the absence of Trp (**Figures 4-4**; Appendix C, **Figure C-2**). Sixteen *Tm*TrpB mutants were tested, representing one or two individual variants from each of the ten fully adapted populations. We found that 12 of the 16 TrpB variants complemented growth to a similar degree as TRP5 when supplemented with 400 μM indole, demonstrating substantial improvement over their wt *Tm*TrpB and *Tm*Triple parents (**Figure 4-4a**).

Unsurprisingly, this set of clonal *Tm*TrpBs contained more sequence diversity than the consensus sequences of the ten populations from which they were taken. Together, the variants tested comprised a total of 85 unique amino-acid substitutions, with an average of 8.7 (± 2.1 s.d.) and a range of 5–13 non-synonymous mutations per variant (variant set 1; Appendix C, **Tables C-2 and C-3**). Since the 12 *Tm*TrpBs from this set exhibiting complementation were all similarly active in their primary activity yet mutationally diverse (**Figure 4-4b**), we may conclude that our scaled evolution experiments generated substantial cryptic genetic variation. We note that four of 16 *Tm*TrpB variants exhibited similar or lower Trp productivity compared to their parent (Appendix C, **Figure C-B**). We suspect that the multicopy nature of p1 in the OrthoRep system allowed for deleterious mutations that appeared toward the end of the experiment to be maintained for a period of time without experiencing purifying selection if they arose in the same cell as functional variants, explaining the presence of these low activity *Tm*TrpBs. Indeed, this multicopy "buffering" may have worked to our advantage by promoting genetic drift under selection, facilitating both greater adaptation and greater diversity of evolutionary pathways across replicates (see **Section 4.3**, *Discussion*). This may partly account for the high activity and high cryptic genetic variation present in the evolved *Tm*TrpBs.



OrthoRep-evolved TrpB variant	Mutations
WT-100-1-A	G3D, Y28H, E29G, D34G, P54T, I69V, K96L, A118V, F122S, V245A, I271V, T292S, A351V
WT-003-1-A	A20V, A118V, F122S, V245A, I271V, T292S, A351V
Q90*-003-1-A	Y8H, *90Q, M97T, N167D, S243P, S302P, S335P
Q90*-003-1-B	Y8H, M18V, A20P, F57L, *90Q, M97T, N167D, S243P, S302P, S335P, A351V, Y383C
Tri-003-1-A	E15K, K33R, A58T, N167D, V245A, G253D, I271V, A351V, R385H
Tri-003-1-B	E15K, A58T, E63G, A118V, A119T, I184V, I271V, Y273C, I332V, S335P, A351V
Tri-003-2-A	E15K, A58T, A118V, E143Q, T179A, V250I, T259A, I271V, A351V
Tri-100-1-A	E22K, I88T, K95R, T117A, A119T, Y193H, A234V, I248V, I271V, K361R
Tri-100-2-A	Y8H, A119T, F122S, N167D, I174F, I195T, E206G, Y352C, I388T
Tri-100-2-B	E30G, K96L, P140L, N167D, I184F, L213P, G228S, I388T
Tri-100-3-A	K39E, K95_K96insK, T117A, K270E, I271N
Tri-100-4-A	R46H, F57Y, A118T, G123A, N167D, I241V, A330V, A349T, G360D, Y383H
Tri-100-4-B	G3D, E30G, I102T, P140L, N167D, L213P

Figure 4-4. *In vivo* activity and diversity of individual *Tm*TrpB variants from OrthoRep-evolved populations. **a.** Evaluation of TRP5 complementation by evolved variants through a growth rate assay. Maximum growth rates over a 24-h period for Δtrp5 yeast strains transformed with a nuclear plasmid expressing the indicated *Tm*TrpB variant, grown in medium with or without 400 µM indole. Points and error bars represent mean ± s.d. for four biological replicates, respectively. Shaded area is the mean ± s.d. growth rate for the TRP5 positive control (i.e. plasmid expressing the endogenous yeast TRP5). Green box indicates the mean ± s.d. growth rate for all strains shown when Trp is supplemented. Growth rates for individual replicates in all three media conditions are shown in Appendix C, **Figure C-12**. Note that growth rates below ~0.15 per hour correspond to cultures that did not enter the exponential phase; in these cases, the reported growth rate is not meaningful and instead can be interpreted as no quantifiable growth. **b.** Parent populations from which OrthoRep-evolved variants shown in **a** are derived and all non-synonymous mutations present in each.

4.2.4 Evolved *Tm*TrpBs exhibit high primary and promiscuous activity *in vitro*

We further characterized the evolved *Tm*TrpBs *in vitro* to approximate conditions of industrial application, make kinetic measurements, and test whether promiscuous activity could be detected. Nine *Tm*TrpB variants were sampled from those that supported robust indole-dependent growth in the Δ*trp5* strain, cloned into an *E. coli* expression vector with a

C-terminal polyhistidine tag, and overexpressed. To mimic streamlined purification conditions compatible with biocatalytic application of *Tm*TrpBs, we generated heat treated *E. coli* lysates (1 hour incubation at 75 °C) and tested them for their ability to couple indole and serine to produce Trp at 30 °C. Three of the nine OrthoRep-evolved TrpBs, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A, demonstrated improved activity over the benchmark *Tm*Triple, as measured by total turnover number (TTN) (Appendix C, **Figure C-3**). Conveniently, each of these variants was evolved from a different starting point, meaning that wt *Tm*TrpB, *Tm*Triple, and *Tm*TripleQ90* were all viable starting points for reaching high activity *Tm*TrpBs. (See Appendix C, **Table C-3** for an explanation of variant naming conventions where each name designates the source of the variant. For example, WT-003-1-A designates variant A taken from the first replicate of a 3-mL evolution experiment starting from WT *Tm*TrpB.)

Since the benchmark *Tm*Triple against which we compared the evolved *Tm*TrpBs was engineered through classical directed evolution involving screening *E. coli* lysates, whereas our *Tm*TrpB variants were evolved in yeast but expressed in *E. coli* for characterization, it is likely that the high-activity evolved *Tm*TrpBs would compare even more favorably if normalized by expression. We therefore purified WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A by immobilized metal affinity chromatography (IMAC) and reevaluated their activity for coupling indole with serine to generate Trp. By TTN, all three variants showed a 4- to 5-fold increase in activity over *Tm*Triple at 30 °C (Appendix C, **Figure C-4**). At 75 °C, however, WT-003-1-A had only ~2-fold higher activity than *Tm*Triple, while the other two variants were less active than *Tm*Triple. Since the thermostability of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A had not been reduced dramatically ($T_{50} > 83.7$ °C; Appendix C,

Figure C-5), adaptation in these variants occurred at least partially by shifting the activity temperature profile. This is a practically valuable adaptation, since thermostable enzymes that operate at mesophilic temperatures allow for greater versatility in application without sacrificing durability and ease of purification through heat treatment.

Further testing of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A revealed that all three enzymes had at least a 22-fold higher k_{cat}/K_M for indole than did *Tm*Triple at 30 °C (Appendix C, **Table C-4** and **Figure C-6**). Finally, testing for production of Trp analogs revealed that these variants' improved performances with indole transferred to alternate substrates (Appendix C, **Figure C-4**), validating their utility as versatile biocatalysts and also the hypothesis that continuous evolution of *Tm*TrpB variants can uncover promiscuous activities for which they were not selected.

4.2.5 A diverse panel of evolved *Tm*TrpB variants encompasses a variety of useful promiscuous activities with indole analogs

Given the exceptional performance of WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A and their ability to transfer primary activity to new substrates as promiscuous activity, we decided to further sample the variant diversity generated across the multiple *Tm*TrpB evolution experiments. We cloned 60 randomly chosen *Tm*TrpBs from the ten continuous evolution populations into *E. coli* expression vectors for *in-vitro* characterization. These 60 *Tm*TrpBs represent extensive diversity, with an average of 9.3 (± 2.8 s.d.) non-synonymous mutations per variant and a total of 194 unique amino acid changes across the set; in addition, each sequence encoded a unique protein (variant set 2; Appendix C, **Tables C-2 and C-3**). Since each variant had multiple non-synonymous mutations (up to 16) accumulated through >100 generations of adaptation and neutral drift, the depth of OrthoRep-based evolution was indeed leveraged in their evolution. We visualized these sequences, together with the consensus sequences of the populations from which they were derived, as nodes in a force directed graph related by shared mutations (Appendix C, **Figure C-7**). With only one exception, all individual sequences cluster near the consensus sequence for their population, meaning that interpopulation diversity exceeded intrapopulation diversity. Thus, the scale of OrthoRep-based evolution was also leveraged in these variants—if fewer independent evolution experiments had been run, the reduction in diversity would not be recoverable from sampling more clones.

Preparations of *Tm*TrpBs WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A, the 60 new variants, and four top-performing TrpB benchmark variants from past classical directed evolution campaigns (including *Tm*Triple) were all tested for product formation with indole by UV absorption and nine indole analogs by high performance liquid chromatography-mass spectrometry (HPLC-MS) to detect substrate promiscuity (**Figure 4-5a**). The panel of 63 OrthoRep-evolved *Tm*TrpB variants exhibited an impressive range of activities (**Figure 4-5b**). First, we observed that a number of variants had primary activities with indole that surpass the benchmark *Tm*Triple in lysate, with initial velocities of Trp formation up to 3-fold higher than WT-003-1-A (**Figure 4-5b;** Appendix C**, Figure C-8**) whose $k_{\text{cat}}/K_{\text{M}}$ for indole is 1.37 x 10^5 M^{-1} s^{-1}, already 28-fold higher than *Tm*Triple's at saturating serine concentrations (Appendix C, **Table C-4** and **Figure C-6**). Second, direct comparison of some of the best panel variants to *Tm*Triple revealed dramatic general activity improvements for multiple indole analogs (**Figure 4-5b**). For example, across the three most versatile variants (Q90*-003-1-C, Tri-003-1-D, and Q90*-003-1-D) the maximum fold-improvement in product yields over *Tm*Triple were 37, 5, 19, and 50 using substrates **5-CN**, **7-CN**, **5-Br**, and **6-Br**, respectively (**Figure 4-5c**). Finally, with the exception of **6-Br** and **azulene**, at least one variant from the OrthoRep-evolved panel converted the indole analog substrates as well as or better than benchmark TrpBs *Pf*2B9, *Tm*Azul, and *Tm*9D8*, which had been deliberately engineered toward new substrate scopes, though at higher temperatures (**Figure 4-5b**; Appendix C, **Figure C-9**).[7,21,25,26]

The diverse properties represented in our 63 variants were not just limited to primary activity increases on indole and promiscuous activities for indole analogs. Multiple variants from the panel also exhibited substantial improvements in selectivity for differently substituted indoles, which could be useful when working with substrate mixtures that may be less expensive to use industrially. For example, we observed many *Tm*TrpBs with greater selectivity for **7-Br** over **5-Br** as compared to all four of the benchmark engineered TrpBs (**Figure 4-5d**). Another variant in the panel, Tri-100-1-G, stood out for having appreciable activity with nearly all substrates tested, including **6-CN** and **5-CF$_3$**, which are poorly utilized by most other TrpBs, likely due to electron-withdrawing effects of their respective moieties. Notably, the ability to accept **5-CF$_3$** as a substrate was unique to Tri-100-1-G: all other variants, as well as the benchmark TrpBs, showed no detectable product formation with this substrate (**Figure 4-5e**; Appendix C, **Figure C-9**). Repeating the reaction with purified enzyme in replicate confirmed the observed activity (Appendix C, **Figure C-10**). Tri-100-1-G may therefore be a promising starting point for new engineering efforts to access exotic Trp analogs. In short, despite having been selected for native activity with indole, OrthoRep-evolved *Tm*TrpBs have extensive and diverse activities on a range of non-native substrates, demonstrating the value of depth and scale in the evolution of enzyme variants.

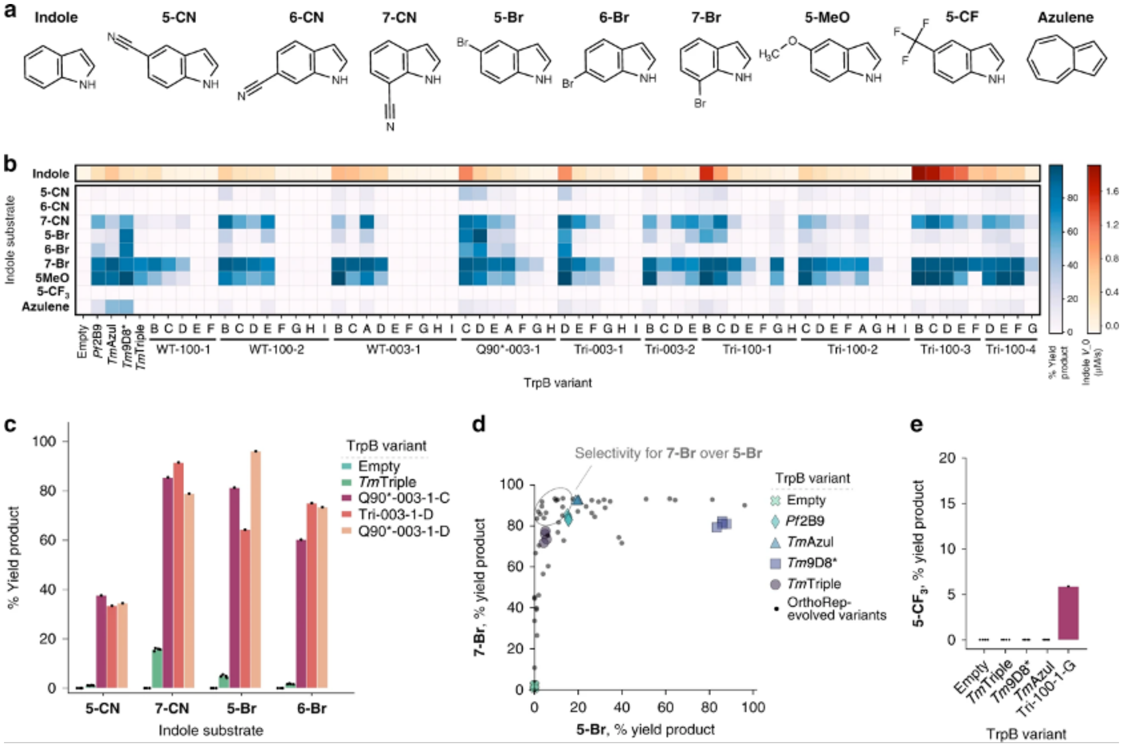


Figure 4-5. Promiscuous activities of a panel of evolved *Tm*TrpBs. a. Indole substrates used to test the substrate scope of a panel of *Tm*TrpB variants. **5-CN**, 5-cyanoindole; **6-CN**, 6-cyanoindole; **7-CN**, 7-cyanoindole; **5-Br**, 5-bromoindole; **6-Br**, 6-bromoindole; **7-Br**, 7-bromoindole; **5-MeO**, 5-methoxyindole; **5-CF$_3$**, 5-trifluoromethylindole. **b.** Heatmap of TrpB activities reported as yield of the Trp analog produced from indicated substrates where 100% yield corresponds to full conversion of the indole analog to the Trp analog. Reactions were carried out using heat treated (1 h at 75 °C) cell lysate, yield was measured by HPLC-MS, and V_0 is the initial rate of Trp formation from indole at saturating serine concentrations. Panel *Tm*TrpB variants are ordered first by the parental cultures from which they were derived, then by activity with indole. Reactions with OrthoRep-evolved variants other than WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A were performed in one replicate and all other reactions performed in quadruplicate. Empty designates expression vector without any TrpB encoded. **c.** Bar graph of selected indole analog activities from panel **b**. Points represent percentage HPLC yield for individual replicates, bars represent mean yield for multiple replicates or yield for a single replicate. **d.** Activities of all variants shown in **b** for reactions with substrates **7-Br** and **5-Br** to show selectivity. Individual replicates are shown for empty vector and benchmark TrpBs and only mean values are shown for OrthoRep-evolved variants tested in replicate (i.e. WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A) for clarity. **e.** Heat-treated lysate activity with **5-CF$_3$** for indicated TrpB variants.

4.2.6 Mutations in evolved *Tm*TrpBs may modulate conformational dynamics and fine tune the active site

Of the ~200 unique mutations in the OrthoRep-evolved *Tm*TrpBs that we characterized, there were some mutations whose effects could be rationalized from comparison to previous work. Since the *Tm*TrpBs had to evolve stand-alone activity, it is unsurprising that many of the mutations we observed have been implicated in the loss of allosteric regulation by TrpA. For example, Buller *et al.* previously examined a series of engineered variants from *Pyrococcus furiosus* TrpB (*Pf*TrpB) and found that evolution for stand-alone activity was facilitated by a progressive shift in the rate-limiting step from the first to the second stage of the catalytic cycle as well as stabilization of the 'closed' conformation of the enzyme.[27] That work implicated eight residues in this mechanism, seven of which correspond to homologous sites where we observed mutations in the evolved *Tm*TrpB variants (i.e., P14, M18, I69, K96, L274, T292, and T321). Another mutation, N167D, present in three of the ten consensus sequences for evolved populations (**Figure 4-3**), has also been implicated in stabilizing the closed state.[21] Additional mutations observed but not studied before (e.g., S277F, S302P, and A321T) could also reasonably alter the allosteric network linking *Tm*TrpB activity to its natural *Tm*TrpA partner, based on existing structures and molecular dynamics simulations on the homologous *Pf*TrpA/*Pf*TrpB complex.[22,27] Taken together, these mutations are likely implicated in converting allosteric activation by *Tm*TrpA into constitutive activity to establish stand-alone function of *Tm*TrpBs.

During the evolution of stand-alone activity, not only must allosteric activation by *Tm*TrpA be recapitulated by mutations in *Tm*TrpB, the surface of *Tm*TrpB that normally interacts with *Tm*TrpA must adjust to a new local environment. Consistent with this adaptation, all consensus sequences for the ten successfully evolved populations from which our *Tm*TrpB variants were sampled contain a mutation to at least one of a set of five residues located on the canonical TrpA interaction interface (**Figure 4-3**; Appendix C, **Figure C-11**). These mutations might improve solubility by increasing hydrophilicity (e.g. G3D, Y8H, and A20T) or form new intramolecular interactions that compensate for lost interactions with *Tm*TrpA, among other possibilities.

We also detected strong convergent evolution in a region near the catalytic lysine, K83, which directly participates in *Tm*TrpB's catalytic cycle through covalent binding of PLP and multiple proton transfers (Appendix C, **Figure C-12**).[19] For example, A118 was mutated in the consensus sequence of four of the ten fully adapted populations, while adjacent residues T117 or A119 were mutated in an additional three (**Figure 4-3**). Furthermore, the three populations in which these residues were not mutated contained other consensus mutations that are either part of the α-helix to which K83 belongs, or, like residues 117–119, within ~8 Å of this helix (**Figure 4-3**; Appendix C, **Figure C-12**). We hypothesize that the α-helix harboring K83 is a focal point of evolution, whereby mutations in its vicinity may finely adjust the positioning of K83 and the PLP cofactor to improve catalysis, perhaps as compensation for structural changes induced by thermo adaptation. Some OrthoRep-evolved variants also contained mutations to first- and second-shell active site residues (Appendix C, **Figure C-13**), which may directly modulate the activity of *Tm*TrpBs, although these mutations were rare. Taken together, we hypothesize that these mutations near the active site residues of TrpB were adaptive or compensatory.

The ~20 mutations considered above are rationalized with respect to their impact on *Tm*TrpB's primary catalytic activity. While substrate promiscuity changes may be influenced by these explainable mutations, previous literature suggests that substrate specificity is globally encoded by amino acids distributed across an entire enzyme.[28] Indeed, the majority of the ~200 mutations found in our panel of *Tm*TrpBs were far away from *Tm*TrpB's active site and not rationalizable based on the known structural and kinetic properties of TrpBs. We suspect that the cryptic genetic variation this majority of mutations encompasses contributes to the diversity in substrate scope across our variants.

4.3 Discussion

In this work, we showed how the depth and scale of evolutionary search available in OrthoRep-driven protein evolution experiments could be applied to broaden the secondary promiscuous activities of *Tm*TrpB while only selecting on its primary activity. The significance of this finding can be divided into two categories, one concerning the practical utility of the new *Tm*TrpB variants we obtained and the second concerning how this evolution strategy may apply to future enzyme evolution campaigns and protein engineering in general.

Practically, the new *Tm*TrpBs should find immediate use in the synthesis of Trp analogs. Trp analogs are valuable chiral precursors to pharmaceuticals as well as versatile molecular probes, but their chemical synthesis is challenged by stereoselectivity requirements and functional group incompatibility. This has spurred enzyme engineers to evolve TrpB variants capable of producing Trp analogs,[20,21,25,26] but the capabilities of available TrpBs are still limited. Compared to existing engineered TrpBs, our new panel of variants has substantially higher activity for the synthesis of Trp and Trp analogs at moderate temperatures from almost all indole analogs tested and also accepts indole analogs, such as **5-CF$_3$** (**Figure 4-5a**), for which benchmark TrpBs used in this study showed no detectable activity (**Figure 4-5e**). In fact, only one TrpB variant has shown detectable activity for this substrate in previous classical directed evolution campaigns.[21] In addition, at least one member of the panel accepted each of the nine indole analogs we used to profile promiscuity, suggesting that additional indole analogs and non-indole nucleophiles not assayed here will also be accepted as substrates.[29,30] Finally, the evolved *Tm*TrpBs are both thermostable and adapted for enzymatic activity at 30 °C. This maximizes their industrial utility, as thermostability predicts a protein's durability and can be exploited for simple heat-based purification processes, while mesophilic activity is compatible with heat-labile substrates, industrial processes where heating costs can compound, or *in vivo* applications in model mesophilic hosts (e.g. *S. cerevisiae* or *E. coli*).

Of more general significance may be the process through which the *Tm*TrpBs in this study were generated. Previous directed evolution campaigns aiming to expand the substrate scope of TrpB screened directly for activity on indole analogs to guide the evolution process,[21,26] whereas this study only selected for *Tm*TrpB's primary activity on indole. Yet this study still yielded *Tm*TrpBs whose secondary activities on indole analogs were both appreciable and diverse. Why?

A partial explanation may come from the high primary activities of OrthoRep-evolved *Tm*TrpBs, as validated by kinetic measurements showing that variants tested have $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values for indole well in the 10^5 M^{-1} s^{-1} range. Since OrthoRep drove the evolution of *Tm*TrpB in a continuous format for >100 generations, each resulting *Tm*TrpB is the outcome of many rounds of evolutionary improvement and change (evolutionary depth). This contrasts with previous directed evolution campaigns using only a small number of manual rounds of diversification and screening. Continuous OrthoRep evolution, on the other hand, allowed *Tm*TrpBs to become quite catalytically efficient with minimal researcher effort. We suggest that the high primary catalytic efficiencies also elevated secondary activities of *Tm*TrpB, resulting in the efficient use of indole analogs. However, this explanation is not complete, as evolved *Tm*TrpBs with similar primary activity on indole had differences in secondary activities (**Figure 4-5**). In other words, high primary activities did not uniformly raise some intrinsic set of secondary activities in *Tm*TrpB, but rather influenced if not augmented the secondary activities of *Tm*TrpB in different ways. We attribute this to the fact that we ran our evolution experiments in multiple independent replicates (evolutionary scale). Each replicate could therefore evolve the same primary activity through different mutational paths, the idiosyncrasies of which manifest as distinct secondary activities. A third explanation for the promiscuous profile diversity of these *Tm*TrpB variants is that each replicate evolution experiment had, embedded within it, mechanisms to generate cryptic genetic variation without strong selection on primary activity. Many of the clones we sampled from each *Tm*TrpB evolution experiment had novel promiscuity profiles but mediocre primary activity with indole (**Figure 4-5**). We believe this is because OrthoRep drove *Tm*TrpB evolution in the context of a multicopy plasmid such that non-neutral genetic drift from high-activity sequences could occur within each cell at any given point. Therefore, *Tm*TrpB sequences with fitness-lowering mutations could persist for short periods of time, potentially allowing for the crossing of fitness valleys during evolution experiments and, at the end of each evolution experiment, a broadening of the genetic diversity of clones even without explicitly imposed periods of relaxed selection. Since enzyme orthologs are capable of specializing toward different sets of secondary activities when pressured to do so,[2,3] non-neutral genetic drift from different consensus sequences across independent population should also access different secondary activities, further explaining the diversity of promiscuous activity profiles across clones selected from replicate evolution experiments. The combination of these mechanisms likely explains the variety of properties encompassed by the panel of *Tm*TrpBs.

Our approach to *Tm*TrpB evolution was inspired by the idea of gene orthologs in nature. Orthologs typically maintain their primary function while diversifying promiscuous activities through long evolutionary histories in different species.[2,31] We approximated this by evolving *Tm*TrpB through continuous rounds of evolution, mimicking long histories, and in multiple replicates, mimicking the spatial separation and independence of species. Such depth and scale of evolutionary search is likely responsible for the substrate scope diversity of the *Tm*TrpBs we report even as they were selected only on their primary activity. We recognize that the evolved *Tm*TrpBs represent lower diversity than natural orthologs. For example, the median amino acid sequence divergence between orthologous human and mouse proteins is 11%,[32] while the median divergence between pairs of variants from our experiment is 4.3% with a maximum of 8% (Appendix C, **Figure C-14**). Still, this level of divergence between functional variants is substantial for a laboratory protein evolution experiment and suggests that it is realistic to model future work on the processes of natural ortholog evolution (**Figure 4-6**). For example, it should be straightforward to scale our experiments further, to hundreds or thousands of independent populations each evolving over longer periods of time. This would better simulate the vastness of natural evolution. It should also be possible to deliberately vary selection schedules by adding competitive *Tm*TrpB inhibitors (such as the very indole analogs for which they have promiscuous activity), changing temperatures, or cycling through periods of weak and strong selection at different rates. Such evolutionary courses would approximate complexity in natural evolutionary histories. These modifications to OrthoRep-driven *Tm*TrpB evolution should yield greater cryptic genetic diversity, which may result in further broadening of promiscuous functions. The generation of cryptic genetic diversity at depth and scale should also be useful in efforts to predict protein folding and the functional effects of mutations via co-evolutionary analysis.[33,34] Indeed, catalogs of natural orthologs have proven highly effective in fueling such computational efforts, so our ability to mimic natural ortholog generation on laboratory timescales may be applicable to protein biology at large. Within the scope of enzyme engineering, we envision that the process of continuous replicate evolution, selecting only on primary activities of enzymes, will become a general strategy for expanding promiscuous activity ranges of enzymes as we and others extend it to new targets.

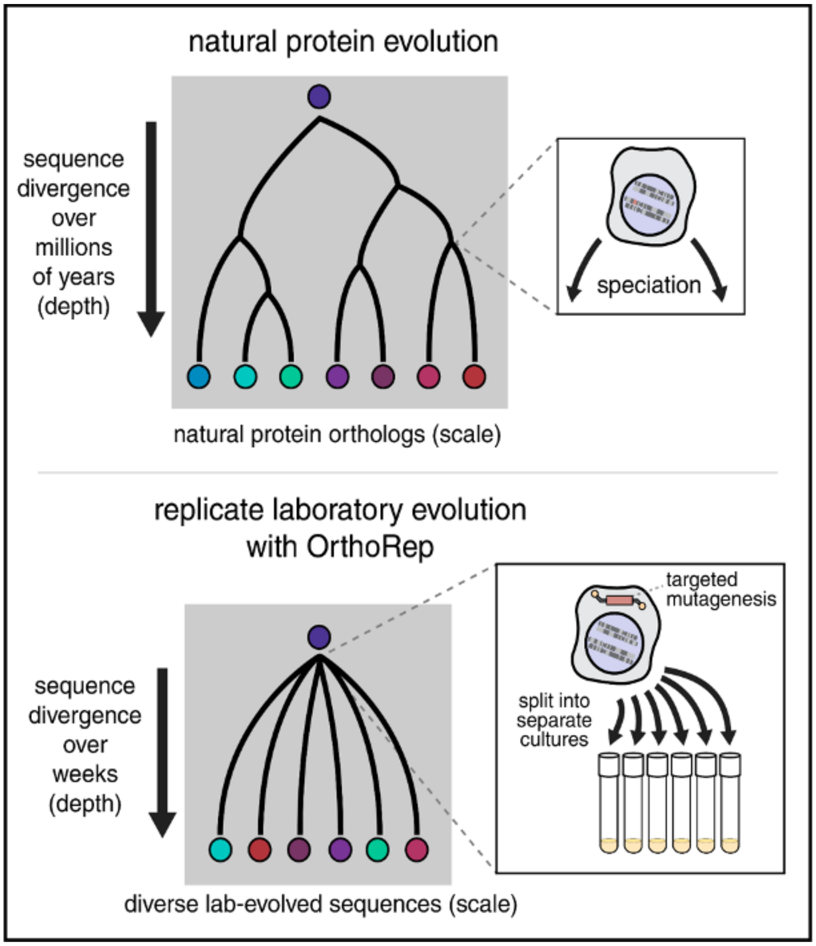

natural protein evolution
sequence divergence over millions of years (depth)
speciation
natural protein orthologs (scale)
replicate laboratory evolution with OrthoRep
targeted mutagenesis
split into separate cultures
sequence divergence over weeks (depth)
diverse lab-evolved sequences (scale)

Figure 4-6. Conceptual similarities between natural enzyme ortholog evolution and OrthoRep evolution. Splitting OrthoRep cultures into many replicates can be seen as a form of speciation occurring by spatial separation. Complex selection schedules may emulate varied selection histories of natural orthologs, generating substantial sequence divergence across replicates. Evolved OrthoRep cultures contain diverse populations of sequences akin to quasi species owing to high mutation rates[35].

Chapter IV Bibliography

1. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).

2. Khanal, A., McLoughlin, S. Y., Kershner, J. P. & Copley, S. D. Differential effects of a mutation on the normal and promiscuous activities of orthologs: Implications for natural and directed evolution. *Mol. Biol. Evol.* **32**, 100–108 (2015).

3. Baier, F. *et al.* Cryptic genetic variation shapes the adaptive evolutionary potential of enzymes. *Elife* **8**, 1–20 (2019).

4. Aharoni, A. *et al.* The ‘evolvability’ of promiscuous protein functions. *Nat. Genet.*

37, 73–76 (2005).

5. Tawfik, O. K. and D. S. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).

6. Leveson-Gower, R. B., Mayer, C. & Roelfes, G. The importance of catalytic promiscuity for enzyme design and evolution. *Nat. Rev. Chem.* **3**, 687–705 (2019).

7. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A panel of TrpB biocatalysts derived from tryptophan synthase through the transfer of mutations that mimic allosteric activation. *Angew. Chem. Int. Ed.* **55**, 11577–11581 (2016).

8. Devine, P. N. *et al.* Extending the application of biocatalysis to meet the challenges of drug development. *Nat. Rev. Chem.* **2**, 409–421 (2018).

9. Truppo, M. D. Biocatalysis in the pharmaceutical industry: The need for speed. *ACS Med. Chem. Lett.* (2017).

10. Almhjell, P. J., Boville, C. E. & Arnold, F. H. Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997 (2018).

11. Zheng, J., Payne, J. L. & Wagner, A. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science* **365**, 347–353 (2019).

12. Gupta, R. D. & Tawfik, D. S. Directed enzyme evolution via small and effective neutral drift libraries. *Nat. Methods* **5**, 939–942 (2008).

13. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2**, 7–10 (2007).

14. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).

15. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

16. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, continuous evolution of genes at mutation rates above genomic error thresholds. *Cell* **175**, 1946–1957 (2018).

17. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

18. Zhong, Z. *et al.* Automated continuous evolution of proteins *in vivo*. *ACS Synth. Biol.* (2020).

19. Dunn, M. F. Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Arch. Biochem. Biophys.* **519**, 154–166 (2012).

20. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

21. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

22. Maria-Solano, M. A., Iglesias-Fernández, J. & Osuna, S. Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution. *J. Am. Chem. Soc.* **141**, 13049–13056 (2019).

23. Zhong, Z., Ravikumar, A. & Liu, C. C. Tunable expression systems for orthogonal DNA replication. *ACS Synth. Biol.* **7**, 2930–2934 (2018).

24. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A highly characterized yeast toolkit for modular, multipart assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).

25. Watkins, E. J., Almhjell, P. J. & Arnold, F. H. Direct enzymatic synthesis of a deep-blue fluorescent noncanonical amino acid from azulene and serine. *ChemBioChem* **21**, 80–83 (2020).

26. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *J. Org. Chem.* **83**, 7447–7452 (2018).

27. Buller, A. R. *et al.* Directed evolution mimics allosteric activation by stepwise tuning of the conformational ensemble. *J. Am. Chem. Soc.* **140**, 7256–7266 (2018).

28. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8**, 1–10 (2017).

29. Dick, M., Sarai, N. S., Martynowycz, M. W., Gonen, T. & Arnold, F. H. Tailoring

tryptophan synthase TrpB for selective quaternary carbon bond formation. *J. Am. Chem. Soc.* **141**, 19817–19822 (2019).

30. Romney, D. K., Sarai, N. S. & Arnold, F. H. Nitroalkanes as versatile nucleophiles for enzymatic synthesis of noncanonical amino acids. *ACS Catal.* **9**, 8726–8730 (2019).

31. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.* **4**, 617–623 (2008).

32. Makałowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407–9412 (1998).

33. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, (2011).

34. Stiffler, M. A. *et al.* Protein structure from experimental evolution. *Cell Syst.* **10**, 15-24 (2020).

35. Eigen, M., McCaskill, J. & Schuster, P. Molecular quasi-species. *J. Phys. Chem.* **92**, 6881–6891 (1988).

Chapter V

EVSEQ: COST-EFFECTIVE AMPLICON SEQUENCING OF EVERY VARIANT IN A PROTEIN LIBRARY

Material from this chapter appears in: “Wittmann, B. J., Johnston, K. E., **Almhjell, P. J.** & Arnold, F.H. evSeq: Cost-effective amplicon sequencing of every variant in a protein library. *ACS Synthetic Biology* **11**, 1313–1324 (2022). doi: 10.1021/acssynbio.1c00592”.

B.J.W. conceived of the project and performed initial design and execution of research and software development. B.J.W., K.E.J, and P.J.A. optimized the experimental workflow and software. K.E.J. wrote software for data visualization. P.J.A. optimized the graphical user interface installation and operation and constructed online documentation. B.J.W., K.E.J., and P.J.A. wrote the manuscript and prepared figures.

ABSTRACT

Widespread availability of protein sequence-fitness data would revolutionize both our biochemical understanding of proteins and our ability to engineer them. Unfortunately, even though thousands of protein variants are generated and evaluated for fitness during a typical protein engineering campaign, most are never sequenced, leaving a wealth of potential sequence-fitness information untapped. Primarily, this is because sequencing is unnecessary for many protein engineering strategies; the added cost and effort of sequencing is thus unjustified. It also results from the fact that, even though many lower cost sequencing strategies have been developed, they often require at least some sequencing or computational resources, both of which can be barriers to access. Here, we present every variant sequencing (evSeq), a method and collection of tools/standardized components for sequencing a variable region within every variant gene produced during a protein engineering campaign at a cost of cents per variant. evSeq was designed to democratize low-cost sequencing for protein engineers and, indeed, anyone interested in engineering biological systems. Execution of its wet-lab component is simple, requires no sequencing experience to perform, relies only on resources and services typically available to biology labs, and slots neatly into existing protein engineering workflows. Analysis of evSeq data is likewise made simple by its accompanying software (found at github.com/fhalab/evSeq, documentation at fhalab.github.io/evSeq), which can be run on a personal laptop and was designed to be accessible to users with no computational experience. Low-cost and easy to use, evSeq makes collection of extensive protein variant sequence-fitness data practical.

5.1 Introduction

Engineered proteins are valuable tools across the biological and chemical sciences and have revolutionized industries ranging from food to fuels, pharmaceuticals, and textiles by providing green and efficient protein solutions to challenging chemical problems.[1] Over the course of a protein engineering campaign, hundreds to thousands or more protein variants will be constructed and have their fitnesses (level of, e.g., thermostability, catalytic activity, substrate binding, etc.) evaluated. Notably, sequence information is typically not gathered alongside the functional information, even though it could provide useful biochemical insight.[2–4] This is largely because many engineering strategies can be applied without sequencing. For example, during a typical directed evolution (DE) experiment, often only the best-performing variant or variants are sequenced in each round of mutagenesis and screening; sequencing every variant is viewed as an unnecessary expense. Given the massive amount of functional data gathered during a typical DE campaign, however, if sequencing *were* performed for the variants generated during these experiments, the resultant large datasets of sequence-fitness information could be revolutionary for biological, biochemical, and biocatalytic research. This is especially true for data-driven protein engineering strategies such as machine learning (ML), the development of which has benefitted tremendously from large sequence-fitness datasets made available by strategies like deep mutational scanning (DMS) and in databases like ProtaBank.[5–16]

Unfortunately, the standard sequencing strategy employed during DE—Sanger sequencing—is too expensive for sequencing all variants tested during a round of evolution.[17] Sanger sequencing is ubiquitous due to ease of sample preparation and ready availability of sequencing providers. However, the cost of Sanger sequencing scales linearly with the number of samples (Appendix D, **Figure D-1**). Thus, while the cost of sequencing just the top variants in a round of DE is minor, sequencing the hundreds or thousands of variants generated over the full engineering endeavor is not. Ideally, any new approach to sequencing during a protein engineering campaign would be comparable in cost, effort, and accessibility to that of sequencing just the top variants by Sanger sequencing. Here we present a collection of standardized and accessible protocols, components, and software that accomplishes this goal. This collection, which we call every variant sequencing (evSeq), democratizes barcode sequencing strategies and expands on services made available by multiplexed next-generation sequencing (NGS) providers to allow amplicon sequencing of a region of interest within every variant produced during a round of DE at a cost of cents per variant.[18,19] Sample preparation for evSeq is simple, and the method requires no experience with NGS to perform, relies only on resources and services typically available to biology labs, and slots neatly into existing protein engineering experimental workflows. The accompanying software for analysis of evSeq data (found at github.com/fhalab/evSeq, documentation at fhalab.github.io/evSeq) was designed to be accessible to users with no computational experience and can be run on a personal laptop.

In this paper, we detail the underlying strategies, protocol, and potential applications of evSeq. We begin by describing the strategies employed by evSeq to extend multiplexed NGS for sequencing protein variant libraries in a way that reduces both cost and effort. We then describe the wet-lab protocol of evSeq sample preparation, focusing on how it can be completed without disrupting an existing protein engineering workflow. Next, we discuss the features of the evSeq software before finally presenting two case studies that highlight potential applications of evSeq. In particular, we highlight how (1) the sequence-fitness data from evSeq can provide valuable information about the quality of variant libraries and the functional screen as well as how mutations modulate protein activity, and how (2) the data generated from evSeq can be used to implement ML for protein engineering. We designed evSeq for use as a routine procedure in many protein/enzyme assays (especially DE and protein engineering experiments leveraging mutagenesis strategies that target specific sites or a segment of the sequence). This tool brings cost-effective, easy-to-use sequencing to all protein engineers, regardless of experience with NGS and access to sequencing and computational resources. We believe that widespread adoption of evSeq—and the resultant datasets generated—will be invaluable for future ML-guided protein engineering and will help us better understand protein sequence-fitness relationships.

5.2 Results

5.2.1 evSeq uses inline barcoding to expand on commercially available multiplexed next-generation sequencing.

Unlike Sanger sequencing, which outputs a single chromatogram that represents the population of DNA in a sequenced sample, NGS outputs millions of individual DNA reads that represent a random draw from the population of DNA in the sequenced sample.[18] Confidence in NGS sequencing results is largely determined by the sequencing “coverage”, which for the purposes of this paper is defined as the number of returned reads that map to a specific nucleotide on a reference sequence. Higher coverage enables more confident identification of mutations relative to a reference sequence as the increased redundancy allows distinguishing between true sequence mutations and errors that arise during library preparation, clustering, or sequencing.

A single NGS run is roughly three orders of magnitude more expensive than a Sanger sequencing run, but because the run outputs millions of reads, this cost can be spread over multiple samples using a technique known as "multiplexed NGS" (Appendix D, **Figure D-1**). In multiplexed NGS, each submitted sample is tagged with a "molecular barcode"—a unique piece of DNA that encodes the sample's original identity—before all samples are sequenced together in the same NGS run.[19–25] Post sequencing, the barcodes are used to assign individual reads to individual samples. For instance, barcodes can be used to distinguish reads coming from samples belonging to specific plates and wells.[26] Importantly, multiplexed NGS can be outsourced just like Sanger sequencing (making it accessible to all laboratories regardless of sequencing experience), and sequencing providers typically charge tens of dollars per sample in a multiplexed sequencing run, yielding on the order of 10^4–10^5 individual sequences per sample (assuming the run is performed on an Illumina MiSeq instrument).

The level of coverage granted by a set number of reads depends on the length of the DNA sample being sequenced, the length of the NGS read used to sequence it, and whether those reads are paired-end. NGS reads are short (300 bp or less on Illumina systems), and so reads must be spread across a longer sample to sequence it in full. The expected coverage (average coverage per nucleotide) obtained for a DNA sample thus depends both on its length and the read length used for sequencing. For example, with the ~10^5 reads returned by a commercial MiSeq multiplexed sequencing run, a 3 Mb genome could be sequenced with 150 bp paired-end reads to an expected coverage of ~10x, whereas a 20 kb plasmid could be sequenced to an expected coverage of ~1500x.

Because shorter samples can be sequenced at higher coverage for a given number of reads, it can be advantageous to sequence only the region of interest of a sample. This is exemplified by amplicon sequencing, a strategy in which a researcher sequences a PCR product (an amplicon) that targets a specific region of interest in the DNA.[27] For instance, continuing the example from above, with ~10^5 total 150 bp paired-end reads, a 300 bp PCR product could be sequenced to an expected coverage of ~100,000x.

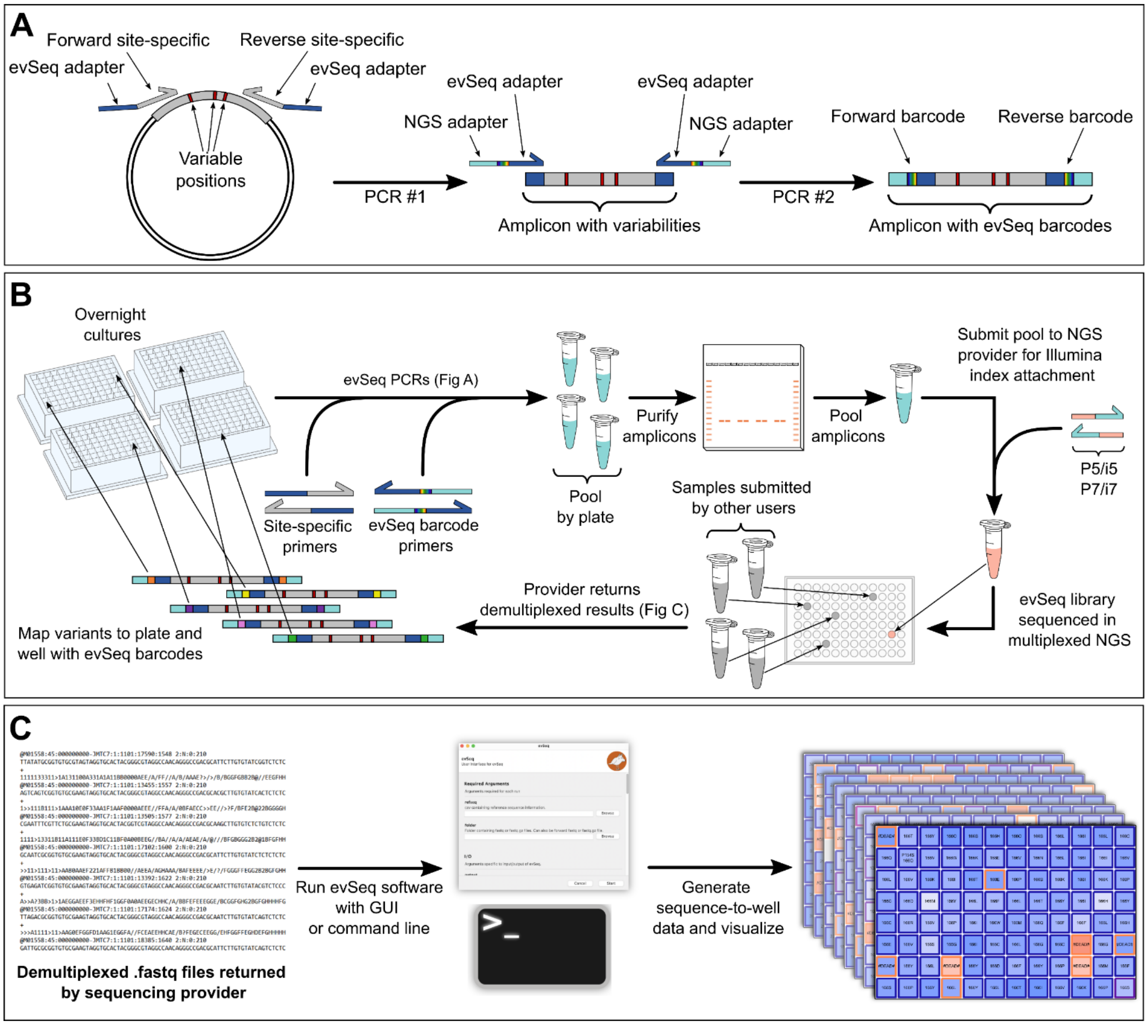
Many mutagenesis methods employed in protein engineering (e.g., site-saturation[28] and tile-based mutagenesis[29] strategies) target mutations to a specific position or region in the sequence of a protein, and thus the variants produced can be sequenced with amplicon sequencing to high coverage.[20] Notably, however, even though increasing coverage yields more confident results, it comes with diminishing returns, and it is generally held that coverage in the tens is more than sufficient for effective reference-based identification of mutations (Appendix D, **Figure D-1**).[30] Indeed, clinical sequencing of human genomes targets 30x coverage or greater to minimize false base calls. Given this reference, it is clear that the ~100,000x coverage that would be returned from a multiplexed sequencing run for a 300 bp amplicon is far higher than necessary for effective identification of mutations—2,000 amplicons could be sequenced in the same run and still yield clinical-grade coverage.

evSeq achieves cost-effectiveness by relying on the facts that (1) at tens of dollars per sample, the cost of sending a single sample to an outsourced multiplexed NGS run is comparable to the total cost of Sanger sequencing the top variants in a round of DE, (2) amplicon sequencing can be used to identify mutations in protein variants from many protein engineering library types, and (3) enough reads are returned for a single sample in a commercial multiplexed

NGS run to sequence hundreds of amplicons. Specifically, the evSeq protocol (**Figure 5-1** and Appendix D, **Section D.2.3**, *evSeq Library Preparation/Data Analysis Protocol*) works by focusing all reads of a single multiplexed NGS sample to specific regions on hundreds of protein variants, achieving sequencing depths of 10^1–10^3 at the approximate cost and level of accessibility of using Sanger sequencing of just the top variants in a round of DE (Appendix D, **Figure D-1**).

The evSeq library preparation protocol begins with PCR amplification of the region of interest in each variant (i.e., the position/region where mutations were made) and appending inline DNA barcodes to the resultant amplicons that encode their original plate-well position (**Figure 5-1a**).[26,31,32] This is a one-pot, two-step, plate-based PCR procedure that uses two sets of primer pairs. Each primer in the first set of primers ("inner" primers) consists of a user-specified 3' "seed" region that binds to the regions flanking the region of interest as well as a 5' predefined universal adapter (Appendix D, **Section D.1.1**, *Inner Primer Design*). Each primer in the second set of primers ("outer" primers) consists of (1) a 3' region that matches the adapter of the inner primers, (2) a central 7-nucleotide barcode where each barcode pair between forward and reverse outer primers is unique to a plate-well position, and (3) a 5' sequence matching the Illumina Nextera transposase adapters (Appendix D, **Section D.1.2**, *Outer Primer Design,* Appendix D, **Section D.1.3**, *Barcode Design*, **Tables D-1** and **D-2**). We designed 96 unique forward and 96 unique reverse outer primers for evSeq which, because both forward and reverse outer primers contain a barcode, can be combined to encode up to 96^2 = 9,216 possible plate-well positions (Appendix D, **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes,* and Appendix D, **Tables D-3–10**. Note that we also provide a pre-filled IDT order form for the outer primers on the GitHub associated with this work—see Appendix D, **Section D.2.1**, *Ordering Barcode Primers from IDT* for details. While we recommend using these pre-tested barcodes, users can also design their own to, e.g., further expand the number of available combinations.). Importantly, this set of outer primers can be used to sequence any target region from any gene with evSeq, and so only needs to be ordered once, constituting a one-time initial setup cost in the range of a few hundred dollars (the exact cost will vary based on oligo provider and any institutional agreements set up with said provider). Once outer primers are ordered, only a new inner primer pair is needed for each new region of interest to be targeted by evSeq.

Once all barcoded amplicons have been produced, they are pooled and sent to a sequencing provider, who will then use the transposase adapters installed with the outer primers as a handle to perform a third and final PCR to barcode the *pool* of amplicons *once again* with a pair of sample-specific Illumina indices (**Figure 5-1b**). At this point each amplicon in the pool has one pair of sample-specific Illumina barcodes and one pair of plate-well-specific inline evSeq barcodes. This complete evSeq library is sequenced as a single sample in a multiplexed NGS run along with samples from other users (whether or not they are also evSeq samples). Post sequencing, the sequencing provider uses the sample-specific barcodes to identify those sequences belonging to the evSeq pool and returns them to the user (i.e., the provider “demultiplexes” the run, separating evSeq sequences from those of other users). The user then uses the evSeq software to analyze the returned sequences, assigning them to corresponding plate-well positions using the evSeq barcodes and identifying the mutations in the variants relative to a reference (**Figure 5-1b** and **5-1c**).


A
Forward site-specific evSeq adapter
Reverse site-specific evSeq adapter
Variable positions
PCR #1
evSeq adapter
NGS adapter
evSeq adapter
NGS adapter
Amplicon with variabilities
PCR #2
Forward barcode
Reverse barcode
Amplicon with evSeq barcodes
B
Overnight cultures
evSeq PCRs (Fig A)
Site-specific primers
evSeq barcode primers
Pool by plate
Purify amplicons
Pool amplicons
Submit pool to NGS provider for Illumina index attachment
P5/i5
P7/i7
Samples submitted by other users
evSeq library sequenced in multiplexed NGS
Provider returns demultiplexed results (Fig C)
Map variants to plate and well with evSeq barcodes
C
Demultiplexed .fastq files returned by sequencing provider
Run evSeq software with GUI or command line
Generate sequence-to-well data and visualize

```
@M01558:45:000000000-JMTC7:1:1101:17590:1548 2:N:0:210
TTATATGCGGTGTGCGTAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCATTCTTGTGTATCGGTCTCTC
+
111113331l>1A131100A331A1A11BB0000AEE/A/FF//A/B/AAAE?>/>/B/BGGFGBB2B@//EEGFHH
@M01558:45:000000000-JMTC7:1:1101:13455:1557 2:N:0:210
AGTCAGTCGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCACGCTTGTGTATCACTCTCTC
+
1>>111B111>1AAA10E0F33AA1F1AAF0000AEEE//FFA/A/0BFAECC>>EE//>?F/BFE2B@22BGGGGH
@M01558:45:000000000-JMTC7:1:1101:13505:1577 2:N:0:210
CGAATTTCGTTCTGCGAAGTAGGTGCACTACGGTCGTAGGCCAACAGGGCCGACGCAAGCTTGTGTCTCTGTCTCTC
+
1111>13311B11A111E0F33BD1C11BF0A00BEEG//BA//A/A/AEAE/A/@///BFGBGGG2B2@1BFGFHH
@M01558:45:000000000-JMTC7:1:1101:17102:1600 2:N:0:210
GCAATCGCGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCATTCTTGTGTATCTCTCTCTC
+
>>11>111>11>AAB0AAEF221AFFB1BB00//AEEA/AGHAAA/BAFEEEE/>E/?/FGGGFFEGG2B2BGFGHH
@M01558:45:000000000-JMTC7:1:1101:13392:1622 2:N:0:210
GTGAGATCGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCAATCTTGTGTATACGTCTCCC
+
A>>A?3BB>1>1AEGGAEEF3EHHFHF1GGF0A0AEEGECHHC/A/BBFEFEEEGGE/BCGGFGHG2BGFGHHHHFG
@M01558:45:000000000-JMTC7:1:1101:17174:1624 2:N:0:210
TTAGACGCGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCAATCTTGTGTATCAGTCTCTC
+
>>>A1111>11>AAG0EFGGFD1AAG1EGGFA//FCEAEEHHCAE/B?FEGECEEGG/EHFGGFFEGHDEFGHHHHH
@M01558:45:000000000-JMTC7:1:1101:18385:1640 2:N:0:210
GATTGCGCGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAACAGGGCCGACGCATTCTTGTGTATCAGTCTCTC
```

Figure 5-1. Overview of evSeq library preparation and processing. a. In the first stage of the PCR, a region of interest is amplified with primers that include a 3' site-specific region (gray) with 5' adapter sequences (dark blue). The second PCR stage adds molecular barcodes (rainbow) with primers that bind to the adapter regions and add adapters for downstream NGS processing (light blue). **b.** To avoid costly DNA isolation steps, evSeq uses liquid cultures of cells harboring mutated DNA (e.g., an "overnight culture" of *E. coli*) as template during the one-pot two-step barcoding PCR described in **a**. Each plate is pooled individually and gel purified. Purified pools are then adjusted for concentration differences and pooled together before being sent to a sequencing provider, who then appends another set of barcodes as well as sequence elements necessary for Illumina NGS sequencing. This sample is now pooled with those of other users and a multiplexed sequencing run is performed. After sequencing, the sequencing provider uses the barcodes that they attached to separate ("demultiplex") the evSeq reads from reads of other users; the provider returns evSeq reads in .fastq files. **c.** The .fastq files returned by the NGS provider are inputs to the evSeq software, which uses the evSeq forward/reverse barcode pair to map each read to a specific plate and well based on known barcode combinations. The software also processes the mapped reads (see Appendix D and evSeq documentation for more details) to, among other things, assign variant identities to each well and return interactive HTML visualizations.

5.2.2 evSeq library preparation fits into existing protein engineering and sequencing workflows and was designed to be resource efficient.

A typical procedure for evaluating protein variants involves (1) arraying colonies of an organism (e.g., *Escherichia coli*) that harbor a plasmid encoding a protein variant into the wells of a (usually 96-well) microplate, (2) growing the resulting cultures to stationary phase (colloquially, an "overnight culture"), (3) using the overnight culture to inoculate a fresh culture that will be used to express the protein variants, and (4) evaluating the fitnesses of expressed protein variants. The expression stage (step 3) typically involves downtime where the experimentalist must wait until the culture reaches sufficient density before inducing protein expression and then again as expression takes place. evSeq library preparation can be performed easily in either of these time windows. The evSeq library preparation protocol begins with the barcoding PCR described at the end of the previous section; this one-pot, two-step, plate-based PCR was designed to be compatible with outsourced sequencing workflows, minimize preparation time, and minimize laboratory resource usage (Appendix D, **Section D.2.3**, *evSeq Library Preparation/Data Analysis Protocol*). For instance, use of inline barcodes is a known, effective strategy for expanding the number of samples that can be multiplexed without having to modify the Illumina indices used during multiplexed sequencing.[31,32] Because evSeq library preparation uses inline barcodes, it grants the outsourced sequencing provider maximal flexibility in choice of Illumina indices. In other words, evSeq library preparation is decoupled from preparation of the Illumina library that will eventually be sequenced, allowing the evSeq library to be run just as any other sample would be that is submitted to a sequencing provider.

As mentioned in the previous section, use of a two-step PCR reduces the number of primers that must be ordered per new sequencing region of interest. Because evSeq relies on 96 unique forward barcodes and 96 unique reverse barcodes, a single-primer PCR would require ordering 192 new barcoding primers for each new target region evaluated in each library. In a two-primer protocol, however, the inclusion of a universal adapter on the inner primers allows the same 192 outer primers to be used regardless of target position in the variant—only two unique primers (forward and reverse inner) must be purchased for each new target region, and only if existing inner primers from previously targeted regions are not already compatible. Additionally, the evSeq PCR directly uses liquid from the overnight culture as a source of template DNA (**Figure 5-1b** and Appendix D, **Section D.2.3**, *evSeq Library Preparation/Data Analysis Protocol*); the template DNA is released from lysed cells during the initial heating step of the PCR, avoiding a costly and time-intensive DNA isolation/purification step and allowing researchers to use materials already prepared as part of the protein expression workflow.[32]

The remaining steps of evSeq library preparation were, like the PCR stage, also designed to be resource and time efficient. After completion of the PCR, the resulting barcoded amplicons are pooled by plate and purified via gel extraction. Pooling prior to purification goes against standard practice for multiplexed NGS library preparation, which is to purify samples individually, quantify their DNA concentration, then combine them in equimolar quantities to ensure more equal read distribution across samples after sequencing.[33] However, because individual plates in protein engineering libraries tend to contain variants from the same region of the same protein scaffold (e.g., as would be typical for variants from a comprehensive site-saturation library), it is assumed that the variation in PCR reaction yield will be minor within plates and that, as a result, the same plate can be pooled prior to quantification with only minor effects on read distribution. Using this "pooling first" strategy, only as many purifications as there are *plates* must be performed as opposed to as many as there are *variants*, thus enabling faster processing of evSeq amplicons while reducing resource usage. As will be shown in later sections, the distribution of reads returned using pooling first is perfectly acceptable for confidently identifying variant sequences.

Once all pooled plates have been purified, the concentrations of the individual purified pools are measured. The pools are then normalized by molarity and combined into a final evSeq library, which is in turn submitted as a single sample to a sequencing provider. As described in the previous section, the provider will perform a final PCR on the evSeq library to add sample-specific barcodes before sequencing it as a single sample in a multiplexed sequencing run. Outsourcing the sequencing stage has two main benefits: First, it allows evSeq to be performed by research groups with no prior sequencing experience and no direct access to sequencing equipment—groups need only be familiar with PCR, a ubiquitous technology in protein engineering laboratories. Second, to be cost effective, multiplexed sequencing should be run with tens of samples at least (Appendix D, **Figure D-1**). By outsourcing the sequencing stage, groups that do not frequently produce evSeq libraries need not wait until enough libraries have accumulated to run sequencing—a single outsourced submission, for instance, can be run along with those of other research groups with a variety of different sequencing needs.

The final stage of the evSeq workflow is data analysis using the evSeq software (github.com/fhalab/evSeq) (**Figure 5-1c**). Extensive documentation of the software and its capabilities is available as a website (fhalab.github.io/evSeq). The software was designed to be accessible to users with varying degrees of computational experience and can be run through either a graphical user interface (GUI), a command line application, or in a Python environment (e.g., a Jupyter notebook). Outputs from the software range from high-level overviews of data (e.g., an interactive "Platemap" graphic that displays sequencing coverage and identified mutations in each well of each plate; see **Figure 5-1c** for an example) to low-level details about the population of reads assigned to each well (e.g., in a well identified as polyclonal, the percentage of reads mapping to each of the identified variants). Functional data can also be easily associated with identified variants using the evSeq software outputs to produce sequence-fitness datasets, and we provide Jupyter notebooks and web pages that walk users through the process.

5.2.3 evSeq facilitates library construction, validation, and sequence-fitness pairing

To highlight the utility of evSeq for engineering and biochemical experiments, we first examined how it could be used to construct high-confidence and informative sequence-fitness data. Specifically, we constructed and screened eight single-site-saturation libraries of the enzyme Tm9D8*—an engineered β-subunit of tryptophan synthase from *Thermotoga maritima* (*Tm*TrpB)—for tryptophan-forming activity at 30 °C (**Figure 5-2**).[34] In two of the screened libraries, we targeted two positions distant from the active site (A118 and S292) that have been seen to play a role in allosteric regulation of *Tm*TrpB enzymes; in the other six libraries, we targeted active-site residues known to modulate the activity of TrpB (E105, L162, I166, F184, S228, and Y301) (**Figure 5-2a**).[35–37] As we show below, this type of sequence-fitness data can be used to assess the quality of a protein engineering library, identify improved variants during a round of directed evolution, and give insight into the significance of a given residue in catalysis.

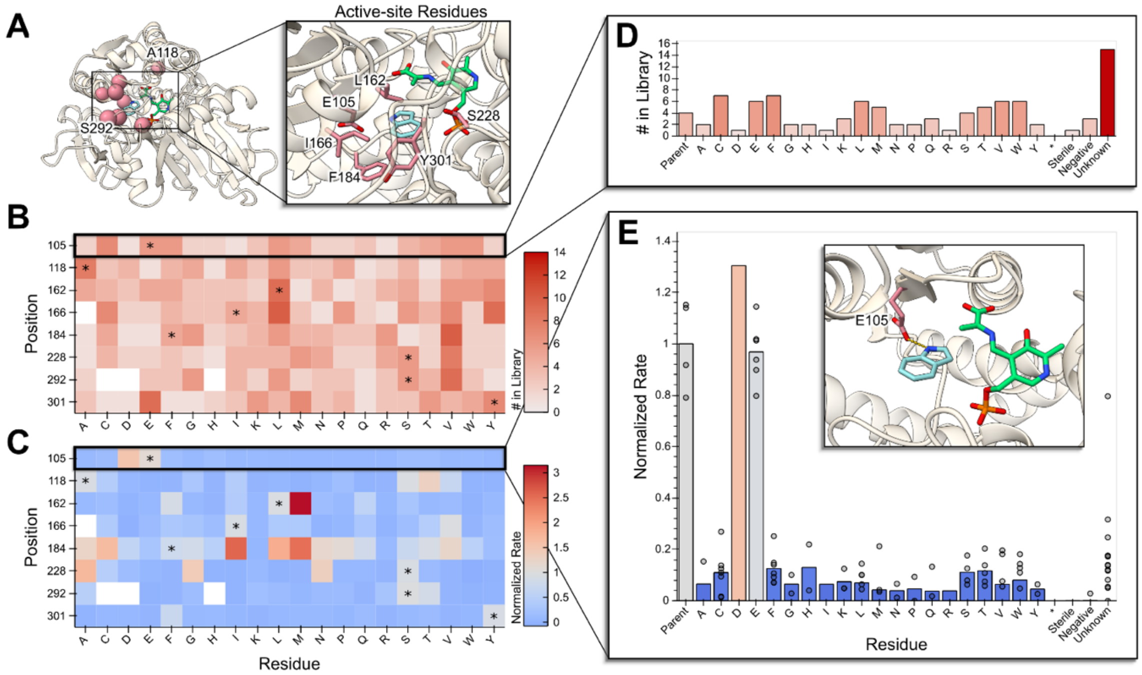


Figure 5-2. evSeq enables low-cost investigation of library quality and sequence-fitness pairing in site-saturation mutagenesis libraries. a. Eight residues (red) known to modulate the activity of Tm9D8* were independently targeted with site-saturation mutagenesis: A118 and S292 (distal residues), E105, L162, I166, F184, S228, and Y301 (active-site residues). An active form of the pyridoxal 5'-phosphate cofactor is represented in green, and the substrate indole is shown in light blue. **b.** Library quality can be investigated by plotting a heatmap of the number of times each variant/mutant was identified at each targeted position ("# in Library") from processed evSeq data. Parent amino acids are each marked with an asterisk. **c.** Likewise, the effect of mutations and mutational "hotspots" can be identified by plotting a heatmap of the average activity measurements for each variant/mutation in each library, normalized to the average parent activity for that library ("Normalized Rate"), when fitness data is combined with evSeq data. **d.** An example plot made possible by evSeq visualization functions shows the number of times each amino acid was found in a single TrpB library (position 105), also accounting for known controls and unidentified wells. **e.** Another example output of the evSeq software shows activity for a single library (position 105), showing biological replicates. The inset displays the role of the mutated residue in this library, which is to coordinate the nitrogen of the indole substrate. Note that the circles in this plot correspond to individual measurements while the bar plot represents the mean of these measurements. If no circles are present for a bar (e.g., E105D), then this is because only a single instance of this mutation was observed. Circles are not shown in this case to allow distinguishing between a single replicate and a tight distribution of multiple replicates.

Many factors can introduce bias into a site-saturation mutagenesis experiment, such as annealing bias for the native nucleotides during the PCR for library construction or contamination with the template plasmid during transformation. Without sequencing all of the variants, it is impossible to know that the library is representative of the experimental design. Since evSeq reports exactly which variants are contained in a library, researchers can leverage this to implement important quality control practices as part of the standard protein screening workflow. For instance, of all 153 possible unique variants in our eight single-site-saturation libraries, we observed 149 of them (**Figures 5-2b** and **5-2c**); only I166A, S292C, S292D, and S292H could not be assigned with confidence, where we define >80% abundance in a well with >10 reads as our confidence threshold. Of the variants identified, many were found in replicate (**Figure 5-2d**) due to oversampling during colony picking, which ensures that all protein variants have a chance to be found and screened (All libraries were constructed with the 22-codon trick[38] and 88 individual colonies were screened for each library, so we expected a 98% probability of seeing all variants assuming perfect construction of libraries). Conveniently, this oversampling also allows us to evaluate the noise in our functional screen (**Figure 5-2e**) which further improves the confidence in the quality of data gathered.

Given just the fitness data gathered in this experiment, a protein engineer would identify 50 wells that are at least 1.2-fold improved over the parent enzyme Tm9D8*. However, with the sequence-fitness pairs constructed via evSeq, we know that these 50 *wells* correspond to only 16 unique *variants*. Depending on how conservative the engineer was as to what should be sequenced, a decision to sequence hits with Sanger sequencing could result in anywhere from 12 (2-fold improvement) to 50 (1.2-fold improvement) wells sent off for sequencing for a total cost of $36 to $150 (using an estimate of $3 per sequence). It would cost ~$2000 to sequence all eight plates via Sanger. Using evSeq, however, we obtained the sequences of *all* 625 wells of variants for only $100, corresponding to $0.13 per non-control well. In other words, using evSeq, we can produce far more sequence-fitness information than sequencing just the top hits using Sanger all for a similar cost. Importantly, although the evSeq defaults currently allow only eight plates to be sequenced at once, the number of variants included in this experiment could likely have been increased as the median number of reads per well was 86 (mean: 98), which is above what is needed for reliable sequencing. Assuming that doubling the number of plates would halve the number of reads seen for each well, doubling the number of plates sequenced would cause only 14 non-control well sequences to drop below the confidence threshold.

The per-variant cost of evSeq may be reduced even further using different services and sequencing platforms. For instance, in both this section and the next, the reported number of reads and ~$100 total cost are from outsourced MiSeq runs, which returned hundreds of thousands of total reads per evSeq library. We report these numbers because outsourced multiplexed MiSeq is a standard service available to all research groups. As an alternative to outsourcing, however, our institution provides multiplexed sequencing (via the Caltech Millard and Muriel Jacobs Genetics and Genomics Laboratory) on an Illumina NextSeq platform, returning an average of ~10x more reads than the outsourced MiSeq run for a total cost of ~$10. At 10x more reads and 10x less the total cost, the per-variant evSeq cost could decrease 100-fold to <$0.01. Indeed, we were able to re-sequence the TrpB libraries at a per-variant cost of ~$0.01 with ~2.2 million total reads returned for an average of thousands of reads per variant, far higher than what is needed for reliable variant calling. It must be noted, however, that analysis of the millions of evSeq reads was no longer practical on a personal laptop, requiring a desktop workstation instead. Computational power beyond a laptop will be needed when processing more than hundreds of thousands of reads with the existing evSeq software.

Of final note, aside from providing valuable information for protein engineering experiments, evSeq can also facilitate investigation into the biochemical relevance of specific positions/mutations. Specifically, because all possible variants in a site-saturation library can be identified by evSeq, the sequence-fitness information generated can be used to explore the effects of mutations more fully than, for instance, an alanine scanning experiment.[39] Using an example from the TrpB data gathered here, an alanine scanning experiment would tell a biochemist that the mutation to the conserved catalytic residue E105A inactivates the enzyme, with no information about the effects of other amino acid changes at this position. Using site-saturation with evSeq, we instead find that all mutations to E105 except for E105D inactivate the enzyme. The fact that glutamate and aspartate are the only amino acids containing a carboxylic acid suggests that this functional group is critical for activity (**Figure 5-2e**, with inset).

5.2.4 evSeq can be used to generate data for machine learning-assisted protein engineering.

We next wanted to demonstrate the utility of evSeq for advancing and applying machine learning-assisted protein engineering (MLPE). In MLPE, models are trained to learn a function that relates protein sequence to protein fitness (i.e., they learn *f*(sequence) = fitness).[5,6,9–11] These models are then used for rapid, low-cost *in silico* prediction of protein fitness, avoiding or greatly reducing the need for often-costly laboratory screening of variants (**Figure 5-3**).

Sequence-fitness data is critical for effective MLPE. Indeed, even though strategies exist that *can* predict protein fitness from sequence alone (e.g., those that use evolutionary data to predict protein fitness), their effectiveness is improved with the inclusion of sequence-fitness information.[7,14,15,40] As a result, the most effective MLPE workflows require that both sequence *and* fitness data be collected, unlike a DE workflow, which requires only fitness data.

The need to collect sequence data in addition to fitness data is an often-overlooked additional cost of MLPE strategies compared to standard DE. For instance, we recently developed an ML strategy known as machine learning-assisted directed evolution (MLDE) for efficient navigation of epistatic combinatorial protein variant libraries.[41,42] Previously, we used MLDE to evolve *Rhodothermus marinus* nitric oxide dioxygenase (*Rma*NOD) for greater enantioselectivity in a carbon–silicon bond-forming reaction.[41] Over the course of the engineering campaign, we collected six 96-well plates of sequence-fitness data for training ML models. In total, sequencing the variants in these plates by Sanger sequencing cost ~$1700. High additional sequencing costs like these can make MLPE methods far less attractive, even if they are more effective than traditional DE at finding high-fitness protein variants.[42] However, given that evSeq enables sequencing all variants for a cost similar to standard DE methods, it enables use of MLPE without added cost. In essence, evSeq eliminates the sequencing burden of MLPE.

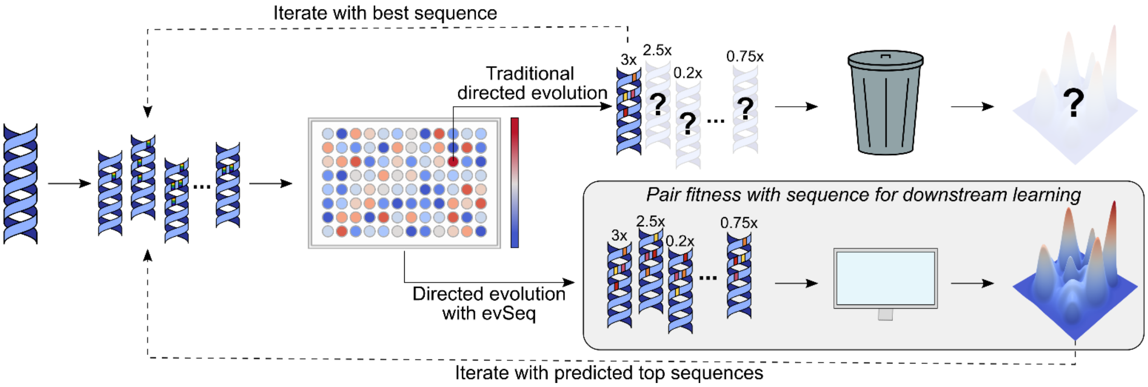

Iterate with best sequence
Traditional directed evolution
3x
2.5x
0.2x
0.75x
?
?
...
?
?
Pair fitness with sequence for downstream learning
3x
2.5x
0.2x
...
0.75x
Directed evolution with evSeq
Iterate with predicted top sequences

Figure 5-3. evSeq eliminates the sequencing burden of MLPE. Traditional DE only collects sequence information for top variants, essentially “throwing away” fitness data from inferior variants and learning nothing about the underlying fitness landscape. If, instead, evSeq is used to collect sequence information for all variants, MLPE methods, which require sequence-fitness pairs for supervised model training, can be implemented. Sampling from a fitness landscape, an ML model can be trained to predict the fitnesses of missing sequences and reconstruct the missing regions of this landscape.

To demonstrate the application of evSeq to MLPE, we used it to sequence five plates of *Rma*NOD variants from a four-site combinatorial library. Coupled with fitness data, the sequences resulting from this run could be used to drive a round of MLDE. Notably, sequencing these plates by Sanger sequencing would have cost ~$1400; in contrast, sequencing by evSeq using an outsourced multiplexed MiSeq run cost ~$100 for a per-variant cost of ~$0.21. The median read depth per variant in this run was 463 (mean: 506), much higher than is required for accurate sequencing, and so more plates—from either the same or a different library—could have reasonably been added to this evSeq run to decrease the per-variant sequencing cost even further (**Figure 5-3b**). Of course, as discussed in the previous section, in-house sequencing could have cut sequencing costs an additional ten-fold.

The cost of sequencing is most notably a barrier for MLPE strategies that focus on developing models for a single protein with a well-defined fitness (e.g., MLDE); however, the applicability of evSeq to MLPE is not limited solely to cost-reduction. For instance, ML strategies have been developed that, rather than focusing on a specific protein, train models on sequence-fitness information across multiple different protein scaffolds.[16,43] The goal is for these models to learn global determinants of protein fitness, then to use the models as general-purpose protein fitness predictors. By enabling the collection of sequence-fitness pairs across a wider array of proteins and fitness definitions, evSeq opens these approaches to new and more diverse data sources. Generally speaking, the more sequence-fitness data available to train and benchmark these strategies, the better we expect them to perform and the more rapidly we expect improvements to be developed.[16] It is no coincidence that large leaps forward in other ML disciplines have followed increased availability of large, diverse datasets, with the rapid advance in computer vision sparked by ImageNet being perhaps the most prominent example.[44] Widespread adoption of evSeq—and commitment to depositing sequence-fitness data in resources such as ProtaBank—would provide such a dataset for protein engineering.[8] This dataset would span the range of all engineered proteins and all target fitnesses, capture examples of sequences with both higher and lower/zero fitness relative to a parent (the latter of which is effectively never recorded with current DE sequencing practices), and overall enable rapid advancement in MLPE.

5.2.5 evSeq detects all variability in the sequenced amplicons

Although we focused here on demonstrating applications involving targeted mutagenesis strategies, evSeq is also applicable to other mutagenesis methods as the associated software can identify both user-specified and unspecified positions of variability (**Figure 5-4a**). This feature not only informs the user of potential unexpected mutations in the sequenced amplicon (Appendix D, **Table D-11**), but also allows it to work effectively with tile-based mutagenesis strategies and other semi-targeted mutagenesis strategies (e.g., error-prone PCR of specific regions or small genes). All that is required is that the amplicon length and read length be able to capture the full region containing mutations.

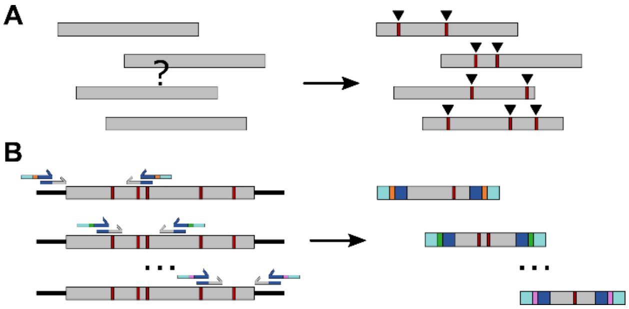


Figure 5-4. evSeq detects variability and can be expanded for random mutagenesis. a. evSeq does not require that the user specify which position in the amplicon was targeted. Instead, the software can identify variable regions by comparing to a reference **b.** evSeq can be used to sequence entire genes by designing a set of inner primer pairs which together capture the entire gene. Different evSeq barcodes can then be used for each region, and the user can reconstruct the entire sequence.

It should be noted that evSeq will not detect off-target mutations outside of the constructed amplicon as these regions are not sequenced, meaning that it is unable to identify other mutations in a larger DNA element that may be contributing to activity. Due to this fact, for exceedingly unexpected mutational effects that are not seen in replicate, we suggest sequencing the rest of the DNA element to confirm the presence or absence of any off-target mutations. However, this limitation is mitigated by the fact that off-target mutations are rare and, importantly, evSeq is agnostic to read length and will work with any length of paired-end sequencing.[45]

While the current software version is not yet suited for other, long-read sequencing technologies (e.g., PacBio or Oxford Nanopore), future versions could be updated and validated with these data formats and make full gene-length evSeq experiments more straightforward and cost effective. Given this, evSeq is currently best suited and most cost effective when all expected mutations exist in the sequenced amplicon, though sequencing of multiple overlapping amplicons can readily allow evSeq to be expanded to sequence entire genes of variants arrayed in microplates (**Figure 5-4b**). Care must be taken in such an application, however, to account for the fact that aggressive mutation rates could compromise the annealing efficiency of inner primers binding in the variable region, as could mutations to positions closer to the binding region of the 3' end of the inner primer. Such situations would lead to a higher proportion of wells failing sequencing.

5.3 Conclusion

Hundreds to thousands of protein variants (or more) are constructed and their fitnesses evaluated over the course of a standard protein engineering campaign. Without sequencing, these fitnesses are next to useless—the time, effort, and resources expended to produce them are largely wasted. Comparable in cost to existing protocols, accessible to scientists with no or minimal sequencing and computational experience, and easy to implement with existing technology, evSeq rescues these fitness data by making the collection of sequence data for every variant a practical and highly useful step of the protein engineering pipeline. Given the number of research groups working on DE and other protein engineering projects, widespread adoption of evSeq would lead to an explosion in the availability of sequence-fitness information. By sequencing every variant, no laboratory screening effort is wasted, and we open the door to advances in both our biochemical understanding of proteins and our ability to engineer them with data-driven methods.

5.4 Materials and Methods

5.4.1 Single-site-saturation library generation for TrpB.

Saturation mutagenesis libraries were prepared using a modification of the "22-codon trick" described by Kille *et al.*[38] We first designed primers using the templates given in Supplemental Table S12. For the forward primers, each sequence of "NNN" in these templates was replaced with "NDT", "VHG", and "TGG", resulting in a total of three degenerate primers which could then be mixed at a ratio of 12:9:1, respectively. The reverse primers were used without changes.

We also designed primers that bind within the ampicillin resistance (AmpR) gene in pET22b(+) with sequences as given in Appendix D, **Table D-13**. These primers were designed such that, when used in combination with the site-specific primers to run a PCR, two medium-length fragments would be created with a break in the AmpR gene. For the forward site-saturation primers, a PCR was performed using the reverse AmpR primer, resulting in a fragment from ~1500–2000 bp long. For the reverse site-saturation primers, a PCR was performed using the forward AmpR primer, resulting in a fragment ~4500–5000 bp long.

Once PCRs finished, 1 µL of DpnI (NEB R0176S) was added to each of the reactions, which were then incubated at 37 °C for 1 h to digest the unmutated template plasmid. The presence of correctly sized fragments was confirmed via gel electrophoresis and each fragment was then excised from the gel and purified with the Zymoclean Gel DNA Recovery Kit (Zymo Research D4002).

Purified fragments were then assembled following the standard Gibson assembly method.[46] After 1 h at 50 °C, the reaction mixtures were desalted with a DNA Clean & Concentrator-5 kit (Zymo Research D4013) and used to transform electrocompetent E. cloni® cells (Lucigen 60051-1). Libraries were spread onto solid agar selection medium consisting of Luria Broth (RPI L24040-5000.0) supplemented with 100 μg/mL carbenicillin (LB_{carb}) and incubated at 37 °C until single colonies were observed. Individual colonies were then transferred into the wells of 96-well 2-mL deep-well plates containing 300 μL of LB_{carb} to isolate monoclonal enzyme variants, with 8 wells being reserved for control conditions, giving 4-fold oversampling of the 22-codon library. These cultures were grown overnight at 37 °C, 220 rpm, and 80% humidity in an Infors Multitron HT until they reached stationary phase, at which point 100 μL from each well were mixed with an equal volume of 50% glycerol and stored at –80 °C for future use.

For protein expression, 20 μL of the remaining culture were used to inoculate 630 μL of Terrific Broth with 100 μg/mL carbenicillin (TB_{carb}). These were then grown at 37 °C, 220 rpm, and 80% humidity for 3 hours in an Infors Multitron HT, at which point they were placed on ice for 30 minutes. Following this, 50 μL of a 14 mM solution of isopropyl-β-D-thiogalactoside (IPTG; GoldBio #I2481C100) in TB_{carb} were added to each well to induce protein expression at a final concentration of 1 mM IPTG. Expression proceeded in the same Infors Multitron HT shaker as before at 22 °C, 220 rpm for roughly 18 hours. Cells were harvested via centrifugation at 4500*g* for 10 minutes, the supernatant was removed, and the plates (now containing pelleted, expressed cells) were placed at –20 °C until needed.

Once cells had been harvested, cultures for evSeq were prepared. These cultures were started from the 96-well plate glycerol stocks prepared prior to moving into the cell expression protocol; the cultures were grown overnight (~18hrs) in an Infors Multitron HT (220 rpm, 37 °C) to saturation in 96-well deep-well plates in 300 µL of LB_{carb}. These cultures were then frozen and stored at –20 °C to be used for sequencing with evSeq.

A GenBank file detailing the plasmid and primers used in this section is available on the evSeq GitHub at https://github.com/fhalab/evSeq/tree/master/genbank_files/tm9d8s.gb.

5.4.2 Sequencing TrpB libraries with evSeq.

Frozen overnight cultures (preparation detailed in the previous section) were thawed at room temperature. Libraries were then sequenced with the process described in Appendix D, **Section D.2.3**, *evSeq Library Preparation/Data Analysis Protocol*; the evSeq software was run using all default parameters (`average_q_cutoff = 25, bp_q_cutoff = 30, length_cutoff = 0.9, match_score = 1, mismatch_penalty = 0, gap_open_penalty = 3, gap_extension_penalty = 1, variable_thresh = 0.2, variable_count = 10`) with the "`return_alignments`" flag thrown. The inner primers used for library preparation are in Appendix D, **Table D-14**. The barcode plates (Appendix D, **Tables D-3–10**) were paired to positions as given in Appendix D, **Table D-15**.

5.4.3 Measuring the rate of tryptophan formation.

Rate of tryptophan formation data was collected with the same procedure described in Rix *et al.* for non-heat-treated lysate preparation (found in this thesis in Appendix C, **Section C.3.9.2,** *Indole rate measurements*) with a few modifications: lysis occurred in 300 µL KPi buffer with 100 μM pyridoxal 5’-phosphate (PLP) supplemented with 1 mg/mL lysozyme, 0.02 mg/mL bovine pancreas DNase I, and 0.1x BugBuster; lysis occurred at 37 °C for 1 h.[35]

5.4.4 Four-site-saturation library generation for *Rma*NOD.

Positions S28, M31, Q52, and L56 of a variant of *Rma*NOD (*Rma*NOD Y32G) were targeted for comprehensive site-saturation mutagenesis using a variant of the 22-codon trick originally described by Kille *et al.*[38] Due to the proximity of positions S28 and M31, it was easiest to use the same mutagenesis primers to target them; the same was done for positions Q52 and L56. Because the 22-codon trick requires three degenerate codons per position targeted, nine individual primers capturing all combinations (3 codons ^ 2 positions/per primer = 9 primers) of the degenerate codons had to be ordered for each of the two mutagenic primers. Sequences of these primers are given in Appendix D, **Table D-16**.

The primers from Appendix D, **Table D-16** were all ordered from IDT at 100 μM. Both a “forward” and a “reverse” primer mixture were prepared by combining individual forward and reverse primers in proportion to the number of individual codons they encoded. A 10 μM forward-reverse primer mixture was then prepared by adding 10 μL of both the forward and reverse primer mixtures to 80 μL ddH$_2$O. Once the forward-reverse primer mixture was prepared, it was used in a PCR to build a pool of DNA fragments containing the four-site combinatorial libraries. Two fragments that captured the remainder of the *Rma*NOD gene and host plasmid (pET22b(+)) were also produced by PCR. The primers used for these flanking fragments are given in Appendix D, **Table D-17**.

After PCR completed 1 μL DpnI (NEB R0176S) was added to each reaction. The reactions were then held at 37 °C in a thermalcycler for 1 h. The PCR fragments were then gel-extracted using a Zymoclean Gel DNA Recovery Kit (D4002).

Fragments were to eventually be assembled using Gibson assembly.[46] Because the efficiency of Gibson assembly increases with decreasing numbers of fragments, an assembly PCR was performed to combine flanking fragment 1 (see Appendix D, **Table D-17** for details) and the variant fragment. The resultant assembled fragment was then gel-extracted, again using a Zymoclean Gel DNA Recovery Kit (D4002).

To complete construction of the library of variant plasmids, a Gibson assembly was performed to combine the assembled PCR fragment and flanking fragment 0. After Gibson assembly, the Gibson reaction was cleaned using a Monarch PCR & DNA Cleanup Kit (NEB CAT T1030L). The cleaned Gibson product was next used to transform electrocompetent E. cloni® BL21 DE3. Transformed cells were spread onto solid agar selection medium consisting of Luria Broth (RPI L24040-5000.0) supplemented with 100 μg/mL ampicillin (LB_{amp}) and incubated at 37 °C until single colonies were observed.

To build the 96-well plates of *Rma*NOD variants used to demonstrate evSeq, 400 μL LB + 100 μg/mL ampicillin were first added to each well of 5x 96-well deep-well plates. Colonies from the agar plates grown overnight were then picked into the wells of the deep-well plates. The plates were placed in an Infors Multitron HT at 240 rpm, 37 °C for ~16 h. To glycerol stock the now-stationary-phase culture, 100 μL overnight culture were added to 100 μL 50% glycerol before being stored at –80 °C until its use in evSeq library preparation.

A GenBank file detailing the plasmid and primers used in this section is available on the evSeq GitHub at:

https://github.com/fhalab/evSeq/tree/master/genbank_files/rmanod_y32g.gb

5.4.5 Sequencing *Rma*NOD libraries with evSeq.

To begin preparation of culture for evSeq with the *Rma*NOD variants, cultures in 96-well deep-well plates (with 300 µL of LB_{carb}) were started from the 96-well plate glycerol stocks prepared in the previous section. The plates were placed in an Infors Multitron HT at 240 rpm; the cultures were grown overnight (~18hrs) before being frozen and stored at –20 °C.

To start the evSeq protocol, frozen overnight cultures were thawed in a room temperature water bath. Libraries were then sequenced with the process described in Appendix D, **Section D.2.3**, *evSeq Library Preparation/Data Analysis Protocol*; the evSeq software was run using the same parameters as for the TrpB data analysis (see **Section 5.4.2**, *Sequencing TrpB Libraries with evSeq*, above). The inner primers used for evSeq library preparation are given in Appendix D, **Table D-18**. The barcode plates (Appendix D, **Tables D-3–10**) were paired to positions as given in Appendix D, **Table D-19**.

Chapter V Bibliography

1. BCC Research Staff. Global markets for enzymes in industrial applications. *BCC Research LLC.* (2018) https://www.bccresearch.com/market-research/biotechnology/global-markets-for-enzymes-in-industrial-applications.html.
2. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
3. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, (2016).
4. Faure, A. J., Domingo, J., Schmiedel, J. M., Hidalgo-Carcedo, C., Diss, G. &

Lehner, B. Global mapping of the energetic and allosteric landscapes of protein binding domains. *bioRxiv* (2021).

5. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).

6. Li, G., Dong, Y. & Reetz, M. T. Can machine learning revolutionize directed evolution of selective enzymes? *Adv. Synth. Catal.* **361**, 2377–2386 (2019).

7. Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C. & Marks, D. S. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).

8. Wang, C. Y., Chang, P. M., Ary, M. L., Allen, B. D., Chica, R. A., Mayo, S. L. & Olafson, B. D. ProtaBank: A repository for protein design and engineering data. *Protein Sci.* **27**, 1113–1124 (2018).

9. Mazurenko, S. & Prokop, Z. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2020).

10. Siedhoff, N. E., Schwaneberg, U. & Davari, M. D. Machine learning-assisted enzyme engineering, in *Methods in Enzymology* 1st ed., pp 281–315 (2020).

11. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).

12. Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

13. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16** (2020).

14. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

15. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T. & Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* (2021).

16. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* **6**, 116–124 (2018).

17. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by

primed synthesis with DNA polymerase. *J. Mol. Bid* **94**, 441–448 (1975).

18. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

19. Smith, A. M., Heisler, L. E., St. Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., Harris, A. N., Perry, K. M., Giaever, G., Pourmand, N. & Nislow, C. Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* **38** (2010).

20. Appel, M. J., Longwell, S. A., Morri, M., Neff, N., Herschlag, D. & Fordyce, P. M. uPIC–M: Efficient and scalable preparation of clonal single mutant libraries for high-throughput protein biochemistry. *ACS Omega* **6**, 30542–30554 (2021).

21. Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D. & Meier, R. ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biol.* **19** (2021).

22. Glenn, T. C., *et al*. Adapterama I: Universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* **7**, e7755 (2019).

23. Wierbowski, S. D. *et al.* A massively parallel barcoded sequencing pipeline enables generation of the first ORFeome and interactome map for rice. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11836–11842 (2020).

24. Chubiz, L. M., Lee, M.-C., Delaney, N. F. & Marx, C. J. FREQ-Seq: A rapid, cost-effective, sequencing-based method to determine allele frequencies directly from mixed populations. *PLoS One 7*, e47959 (2012).

25. Weile, J. *et al*. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).

26. Campbell, N. R., Harmon, S. A. & Narum, S. R. Genotyping-in-thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* **15**, 855–867 (2015).

27. Wen, C., Wu, L., Qin, Y., Van Nostrand, J. D., Ning, D., Sun, B., Xue, K., Liu, F., Deng, Y., Liang, Y. & Zhou, J. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS One* **12**, e0176716 (2017).

28. Siloto, R. M. P. & Weselake, R. J. Site saturation mutagenesis: Methods and

applications in protein engineering. *Biocatal. Agric. Biotechnol.* **1**, 181–189 (2012).

29. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42**, e112 (2014).

30. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).

31. de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R. & Sundaram, A. Y. M. A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome* **5**, 68 (2017).

32. Tresnak, D. T. & Hackel, B. J. Mining and statistical modeling of natural and variant class IIa bacteriocins elucidate activity and selectivity profiles across species. *Appl. Environ. Microbiol.* **86**, e01646-20 (2020).

33. Illumina. Nextera XT DNA library prep reference guide. (2019) https://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-05.pdf

34. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *J. Org. Chem.* **83**, 7447–7452 (2018).

35. Rix, G., Watkins-Dulaney, E. J., Almhjell, P. J., Boville, C. E., Arnold, F. H. & Liu, C. C. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nat. Commun.* **11**, 5644 (2020).

36. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

37. Buller, A. R., Brinkmann-Chen, S., Romney, D. K., Herger, M., Murciano-Calles, J. & Arnold, F. H. Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

38. Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T. & Acevedo, J. P. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2**, 83–92 (2013).

39. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5**, 302–307 (2001).

40. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* (2022).

41. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8852–8858 (2019).

42. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026-1045.e7 (2021).

43. Alieva, A., Aceves, A., Song, J., Mayo, S., Yue, Y. & Chen, Y. Learning to make decisions via submodular regularization. *ICLR* (2021).

44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. ImageNet: Large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

45. McInerney, P., Adams, P. & Hadi, M. Z. Error rate comparison during polymerase chain reaction by DNA polymerase. *Mol. Biol. Int. 2014.*

46. Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A. & Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

Chapter VI

DIRECTED EVOLUTION OF TRYPTOPHAN SYNTHASE FOR NONCANONICAL L-TYROSINE SYNTHESIS

Material from this chapter appears in: "**Almhjell, P. J.**, Johnston, K. E., Porter, N. J., Kennemur, J. L., Ducharme, J. & Arnold, F. H. Directed evolution of tryptophan synthase for noncanonical L-tyrosine synthesis. *Manuscript submitted*".

P.J.A. conceived of the project, designed and executed research and wrote the manuscript with input from K.E.J., N.J.P., and J.L.K. Enzyme kinetics were performed by K.E.J., along with auxotroph complementation experiments. X-ray crystallography was performed by N.J.P, who also interpreted results. J.L.K. performed preparative-scale syntheses. J.D. assisted with rate measurements. All authors edited the manuscript.

ABSTRACT

The β-subunit of tryptophan synthase (TrpB) is responsible for the final step of all L-tryptophan biosynthesis, irreversibly coupling L-serine and indole *via* a Friedel-Crafts alkylation. Although this powerful synthetic paradigm could feasibly access other aromatic amino acids such as L-tyrosine and its derivatives, TrpB-mediated alkylation has only been demonstrated with indole-like substrates. Here, we describe directed evolution of TrpB for noncanonical L-tyrosine synthesis *via* the irreversible, regioselective alkylation of phenols. This new, nonnatural activity is unlocked by mutation of the catalytic glutamate conserved in all characterized TrpB enzymes and a phenol hydroxyl-coordinating water molecule that takes its place. Evolved TrpBs demonstrate a whole new biocatalytic route to the fundamental protein building block L-tyrosine and provide simple and efficient access to valuable L-tyrosine analogs at scale.

6.1 Introduction

The aromatic amino acids (aroAAs) L-phenylalanine (Phe), L-tyrosine (Tyr), and L-tryptophan (Trp) are required for all life as fundamental building blocks of proteins. These aroAAs also appear in secondary metabolites and pharmaceuticals,[1–3] where they are often derivatized to make noncanonical amino acids (ncAAs) that take on a variety of important functional roles. Derivatives of Tyr, for example, include the Parkinson's medication L-DOPA, the neurotransmitter dopamine, the structural elements of lignin, precursors to more complex natural products, and common biological probes, among many others (**Figure 6-1a**).[3–8]

The abundance of these ncAAs in natural and unnatural compounds belies the difficulty of their synthesis.[2,9] In all known organisms, *de novo* biosynthesis of the aroAAs occurs through a universally conserved set of chemistries that convert the common precursor chorismate into Phe, Tyr, or Trp (**Figure 6-1b**).[10,11] Their subsequent derivatization to ncAAs occurs by the action of a diverse array of enzymes. These pathways are often inefficient when used for other biosynthetic and biocatalytic purposes, as they have evolved to function well under specific biological conditions.[12] For example, aroAA biosynthesis is tightly regulated and uses complex substrates, which makes it generally unsuitable for fermentation-based ncAA production. Furthermore, although the enzymes that biosynthesize ncAAs from aroAA precursors can do so with high efficiencies and selectivities, they lack generality and can be difficult to express in heterologous organisms.[2] Chemical synthesis can provide a more general and modular framework for ncAA preparation, where the use of serine-derived electrophiles[13,14] or Negishi cross-couplings[15] are particularly effective (**Figure 6-1c**).

However, these strategies require strictly anaerobic and anhydrous conditions as well as multiple protection and deprotection steps.

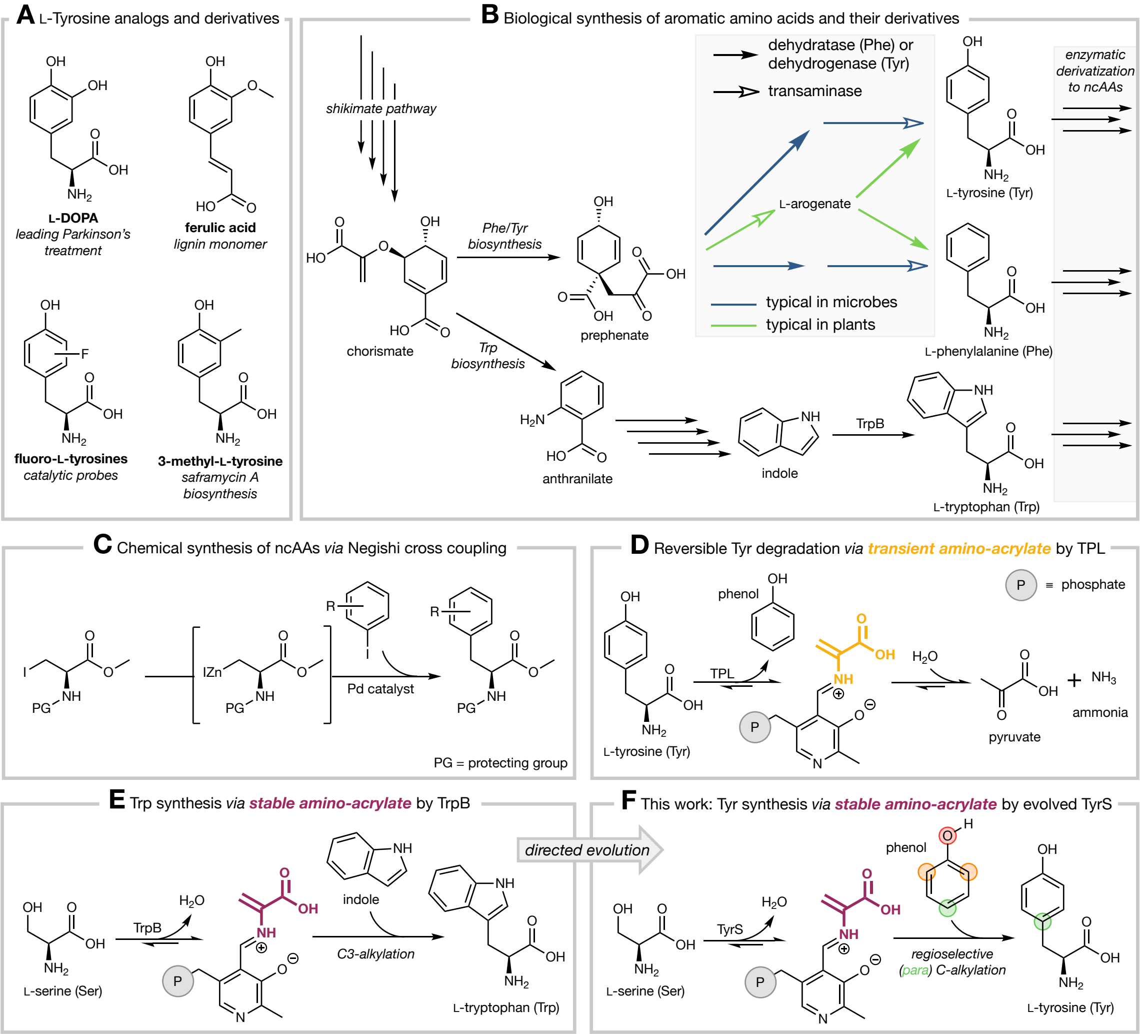


Figure 6-1. Biosynthesis, chemical synthesis, and biocatalytic synthesis of aromatic amino acids. a. Selected valuable L-tyrosine (Tyr) analogs and derivatives, both natural and unnatural. **b.** All known aromatic amino acid (aroAA) biosynthesis occurs by conversion of chorismate, the final product of the shikimate pathway, to the aroAAs using a limited and conserved set of chemistries. The aroAAs can then be used in protein synthesis, natural product biosynthesis, or further derivatized to noncanonical amino acids (ncAAs) by enzymes. **c**. An example of Negishi cross coupling as a general framework for chemical synthesis of aromatic ncAAs. **d.** Reversible degradation of Tyr by tyrosine phenol lyase (TPL), a pyridoxal 5'-phosphate (PLP)-dependent enzyme, can be used for biocatalytic Tyr analog synthesis. **e.** The final step of L-tryptophan (Trp) biosynthesis by TrpB, which irreversibly alkylates indole to form Trp, can be used for the biocatalytic synthesis of Trp analogs. **f.** An engineered "tyrosine synthase" (TyrS) uses the synthetic prowess of TrpB and the substrate preference of TPL to form Tyr from phenol and Ser.

Biocatalytic Friedel-Crafts alkylation marries the best features of enzymatic and chemical catalysis for ncAA preparation,[16] particularly *via* pyridoxal 5'-phosphate (PLP)-derived electrophilic amino-acrylate intermediates. For example, the PLP-dependent enzyme tyrosine phenol lyase (TPL) can accept phenols as nucleophiles to construct Tyr analogs.[5,17] Although this is one of the preferred methods to access valuable Tyr analogs, TPL is constrained by its native function, the degradation of Tyr to phenol, pyruvate, and ammonia, which occurs through a transient amino-acrylate intermediate (**Figure 6-1d**).[18] This activity reduces TPL efficiency *in vitro* and establishes strong equilibrium limitations *in vivo*.[19,20] Although a homolog of TPL, tryptophanase, exists for reversible Trp degradation and could be used to synthesize Trp analogs in a similar manner,[21] another PLP enzyme, the β-subunit of tryptophan synthase (TrpB), has proven far more useful.[22]

TrpB catalyzes the final step of all known *de novo* Trp biosynthesis (see **Figure 6-1b**).[22] Unlike TPL, TrpB generates a stable amino-acrylate intermediate *via* β-elimination of L-serine (Ser). This helps it couple Ser and indole in an effectively irreversible manner (**Figure 6-1e**). TrpB is used for synthesis of Trp analogs[23–25] and more diverse ncAAs[26–30] as well as *in vivo*.[31] Despite over half a century of study,[22,32–35] however, no TrpB has been shown to react with phenols to generate natural or unnatural Tyr analogs, nor has an equivalent "tyrosine synthase" (TyrS) been found in native Tyr synthesis. We anticipated, however, that directed evolution could be used to convert TrpB into a TyrS capable of noncanonical Tyr synthesis *via* the irreversible, regioselective Friedel-Crafts alkylation of phenols (**Figure 6-1f**).

6.2 Results

6.2.1 Identification of a starting point for evolution of tyrosine synthase (TyrS) activity.

Given that TPL uses an amino-acrylate intermediate to accomplish Tyr analog synthesis (and degradation), we suspected that TrpB could be engineered to react with phenols, but differences between the two enzyme reactions highlighted potential challenges. Phenol is a small, symmetric molecule with a single heteroatom and three nucleophilic positions: the carbon atoms *ortho* and *para* to the hydroxyl as well as the hydroxyl itself (**Figure 6-1f**). Tyr synthesis requires a highly *para*-selective catalyst. In TPL, multiple residues coordinate the phenolic hydroxyl group to facilitate *para* C–C bond breakage,[36] which also lowers the energy barrier for the reverse reaction (**Figure E-1**). A catalytic glutamate present in all characterized TrpB enzymes plays a similar role by coordinating indole for C–C bond formation during Trp synthesis.[28] This suggests that a TyrS might require an analogous coordination mechanism to favor *para* C–C bond formation over *ortho* or O-alkylation.

When representative engineered TrpB variants were challenged with phenol and Ser, however, none of the three possible amino-acid products was observed. This prompted us to look for a substrate that might serve as an evolutionary stepping stone from indole and phenol (**Figure 6-2a**). We chose 1-naphthol as an electron-rich phenol analog similar enough to indole to bind in the active site and also be aligned for *para* C–C bond formation (**Figure 6-2b**). To our delight, both TrpB variants tested—Tm9D8*, from *Thermotoga maritima* and engineered for 4-cyano-Trp formation at 37 °C,[37] and *Pf*2B9, from *Pyrococcus furiosus* and engineered for β-methyl-Trp production at 75 °C[24]—reacted with 1-naphthol and Ser to form an amino acid product (**Figure E-2**). Tm9D8* was selected for further evolution because it retains high thermostability while also displaying good activity at 37 °C, which reduces oxidation of electron-rich substrates and would be important for future *in vivo* applications.

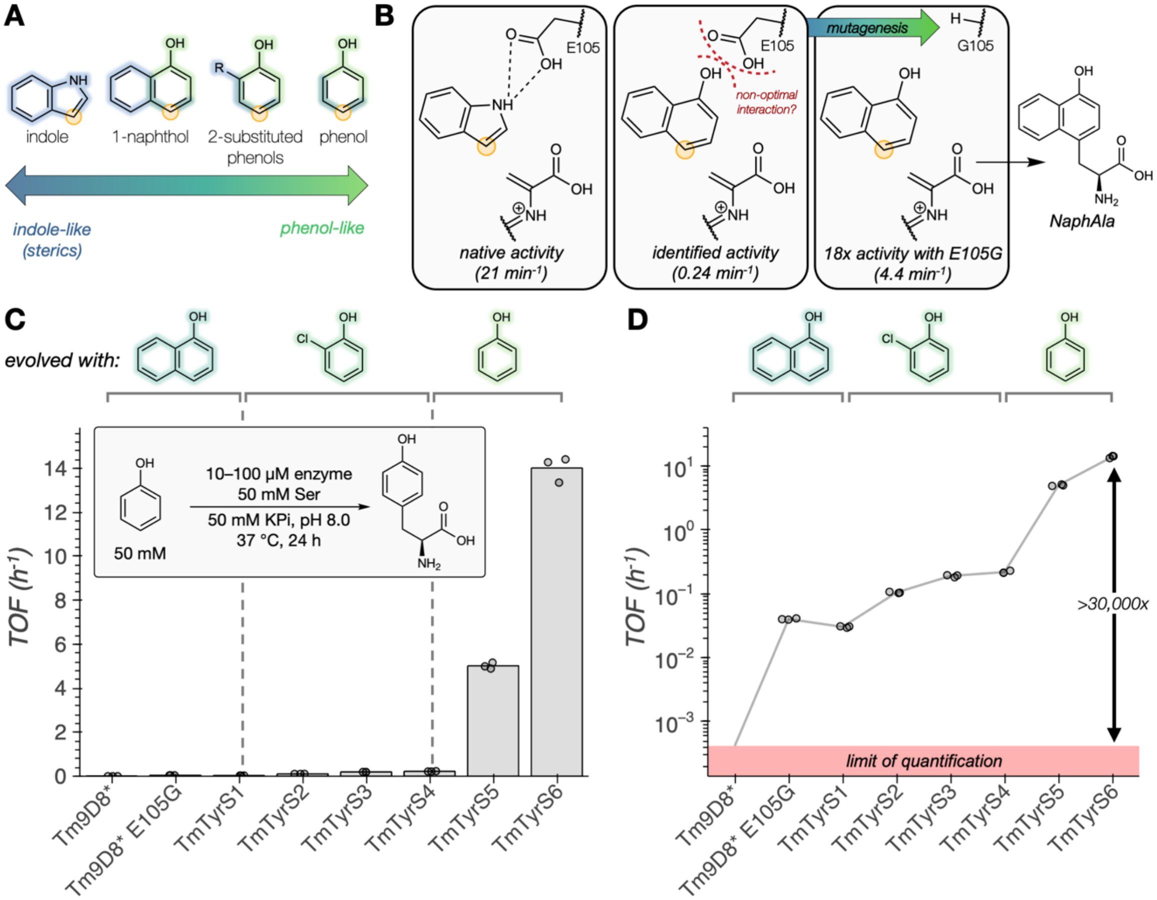


Figure 6-2. Directed evolution of a tyrosine synthase. a. Proposed 'substrate walk' from the native substrate of TrpB, indole, to phenol. **b.** The universally conserved catalytic glutamate (E105) side chain participates in interactions important for C3-alkylation of indole but which may not be optimal for para-alkylation of 1-naphthol. Mutating this glutamate to glycine (G105) enhanced activity 18-fold. The reaction exclusively forms the para-alkylation product β-(1-naphthol-4-yl)-L-alanine (NaphAla). **c.** Approximate turnover frequencies (TOFs, h^{-1}) for conversion of 50 mM phenol to Tyr by TyrS lineage. The screening substrate used during evolution is shown above the chart, with substrate change demarcated by vertical dashed lines. **d.** The TOFs presented on a log scale with a quantified limit for the turnover frequency required to detect Tyr under the presented experimental conditions.

6.2.2 Directed evolution of a tyrosine synthase.

Under the presumption that 1-naphthol was binding in a similar orientation to the natural indole substrate of TrpB, it appeared likely that the catalytic glutamate that aligns indole was not optimal for this non-native substrate (**Figures 6-2b** and **E-3**). Site-saturation mutagenesis and screening at this position identified three highly activating mutations: E105G, E105A, and E105S (all mutations to 'small' side chains). The E105G mutation provided the largest rate enhancement (18-fold over Tm9D8*). Preparative-scale synthesis allowed us to confirm that the *para*-alkylation product β-(1-naphthol-4-yl)-L-alanine (NaphAla) was the sole product of this enzymatic reaction (**Figure E-4**).

We reasoned that we could evolve Tm9D8* E105G for activity on phenol by first increasing activity on 1-naphthol and then moving to other substrates that are more similar to phenol as new activities are detected in a 'substrate walk' (**Figures 6-2a** and **6-2c**).[38] This would yield an enzyme for Tyr synthesis while simultaneously creating a panel of biocatalysts that could be tested for the synthesis of valuable Tyr analogs. Tm9D8* E105G was subjected to sequential rounds of mutagenesis and screening for NaphAla synthesis, eventually producing variant TmTyrS1 (*T. maritima*-derived tyrosine synthase enzyme 1; see Appendix E, **Section E.1.21**, *Enzyme sequences* for mutations/sequence of this and subsequent variants and **Section E.2.1**, *Evolutionary strategies* for specifics about enzyme evolution). Remarkably, this enzyme displays a k_{cat} of 11 min^{-1} for the conversion of 1-naphthol to NaphAla (**Figure E-5**), roughly half that of Tm9D8* for Trp formation (21 min^{-1}, **Figure 6-2b**). This approaches the rates of native TrpS enzymes, although the K_M value (2.8 mM, **Figure E-5e**) is far higher than that of native TrpS enzymes for indole (~1–100 μM).[23,39]

Continuation of this substrate walk required a more 'phenol-like' substrate for further screening. Fortunately, directed evolution for 1-naphthol activity also increased activity toward 2-chlorophenol to form the *para*-alkylated 3-chloro-Tyr product (**Figure E-6**). We subsequently evolved TmTyrS1 for activity on 2-chlorophenol to generate variants TmTyrS2–4 (**Figure 6-2c** and **Section E.2.1**, *Evolutionary strategies*). The rate of phenol conversion for these variants also increased over the course of evolution. However, activity toward phenol was routinely far lower than toward 2-chlorophenol, suggesting that the 2-chloro substituent has an activating effect (**Figure E-7**).

Comparison of 2-chloro- and 2-methylphenol as substrates revealed similar levels of activity, despite the electronic differences between these substituents (**Figure E-7**). This indicated that steric bulk at the 2-position likely plays an important role in the proper orientation of these substrates. Achieving a productive orientation with the unsubstituted phenol substrate is thus potentially challenging in the absence of sterically confining residues. By targeting the active site for further mutagenesis and screening for conversion of phenol to Tyr, we obtained substantially more active variants TmTyrS5 and 6 (**Figures 6-2c** and **6-2d**). The mutation of an active-site glycine to alanine (G229A) in TmTyrS5, which adds steric bulk (**Figure E-8**), proved particularly activating. The final variant, TmTyrS6, synthesizes Tyr at an apparent turnover frequency (TOF) of 14 h^{-1} (0.23 min^{-1}, 50 mM each substrate).

These evolved enzymes are at least 99.5% regio- and enantioselective for Tyr synthesis; we detected no D-Tyr, consistent with the TrpB mechanism (**Figure E-9**), and no *ortho*-alkylation was detected even when the concentration of enzymatically prepared Tyr was >1000-fold higher than the limit of detection (**Figure E-10**). The approximate turnover frequencies are presented on a log scale in **Figure 6-2d**, along with a conservative lower bound for the detectable amount of Tyr under the given conditions (see **Section E.2.2**, *Determination of limit for detectable turnover frequency*). Because Tm9D8* did not make

Tyr exceeding this threshold, the evolution presented here represents at least a 30,000-fold increase in activity from Tm9D8* to TmTyrS6.

6.2.3 TyrS enzymes are efficient biocatalysts for ncAA synthesis.

Given that different substrates were targeted over the course of TyrS evolution, variants in this lineage should serve as biocatalysts for efficient synthesis of a variety of noncanonical Tyr analogs. To assess the substrate preferences and efficiencies of these enzymes, a panel of phenolic substrates was tested against each TyrS variant, starting with Tm9D8* E105G. For consistent comparisons, each substrate was added at 10 mM. The reactions were performed with 1.1 equivalents of Ser relative to the phenolic substrate. Under these conditions, high yields can only be achieved with excellent conversion of both the phenol and Ser to the Tyr analog. This requires the exertion of kinetic control by the enzyme to avoid reversibility as well as minimal conversion of the amino-acrylate to pyruvate and ammonia, a known side reactivity of some TrpB enzymes.[24,25] A diverse profile of activities was observed, with high conversions achieved in many cases (**Figure 6-3a** and **Section E.2.3**, *Estimating product formation by HPLC peak area percentages* and **Figure E-11**). Selected products were isolated and always proved to be the *para*-alkylation product (Appendix E, **Supplementary Methods and Materials**). Importantly, at least one variant displays reasonable initial activity (>1%) for all but one substrate tested; further evolution could increase these activities.

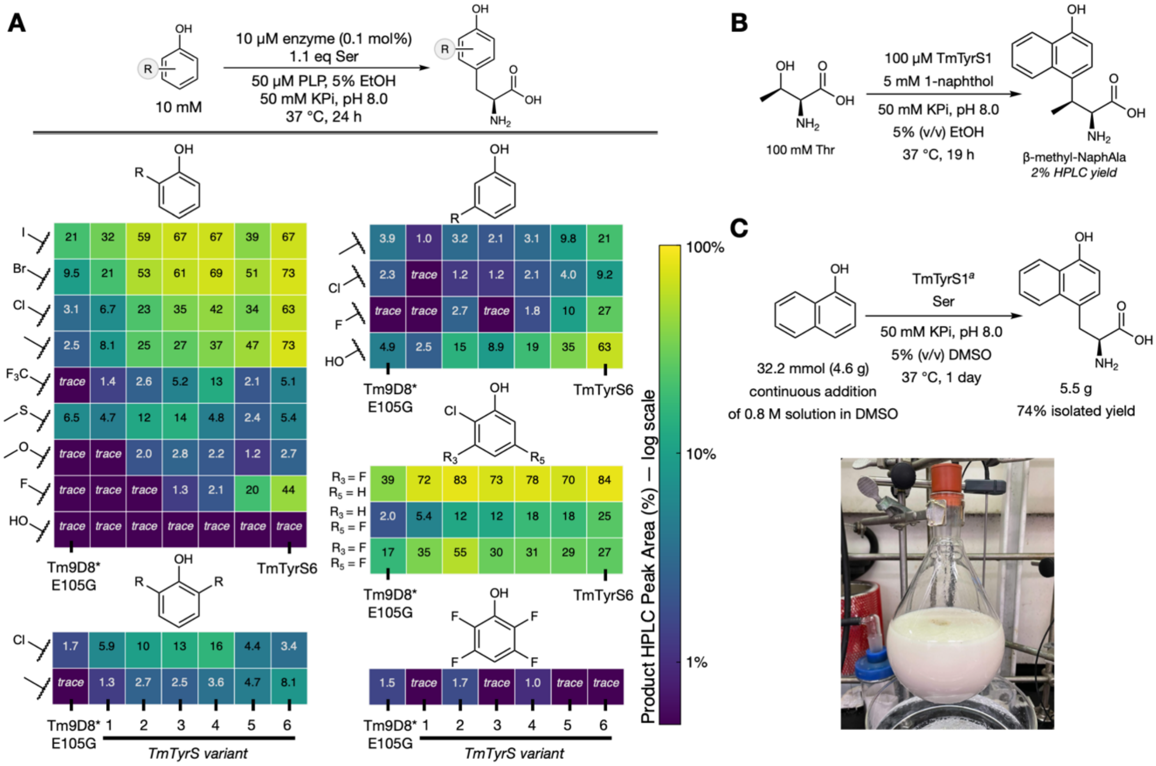


Figure 6-3. Utility of TyrS variants for Tyr analog synthesis. a. Substrate scope of enzymes in the evolutionary lineage. Substrates are grouped by position of substitution(s). Numbers represent the percent HPLC peak area of the product relative to both product and substrate, which serves as a reliable approximation of product formation (see **Section E.2.3**, *Estimating product formation by HPLC peak area percentages* for quantification details). Values are an average of two replicates. Values under 1% are designated as *trace*. **b.** TyrS variants can accept a β-branched Ser analog to synthesize a β-branched Tyr analog in a single step. HPLC yield is the average of three technical replicates. Stereochemistry assumed by analogy to TrpB. **c.** Chromatography-free preparation of NaphAla at multi-gram scale. [a]Catalyst prepared as a lyophilized powder from heat-treated lysate.

Consistent with the native activity of TrpB, these enzymes do not degrade their amino acid products: incubation of NaphAla, 3-chloro-Tyr, and 3-iodo-Tyr with 10 μM enzyme for 20 hours resulted in near-complete or complete retention of the amino acid (**Figure E-12**). This property aids in achieving high yields and identifying trace synthesis of other Tyr analogs. For example, formation of L-DOPA, the product of 2-hydroxyphenol, was low but clearly increased over the course of evolution, as detected by mass spectrometry, following the trends in **Figure 6-3a** and suggesting that evolution enhanced general TyrS-activity.

Furthermore, the rate of product formation for this and most other reactions can be improved by increasing the concentration of substrate above the 10 mM used here. Activity was also observed when L-threonine (Thr) was used in place of Ser to furnish β-methyl-NaphAla in a single step (**Figures 6-3b** and **E-13**), indicating that this platform could provide access to β-methyl-Tyr and other β-substituted Tyr analogs.[40]

TyrS variants can be used for gram-scale synthesis of valuable Tyr analogs in a manner similar to that described for TrpB in the production of other ncAAs at scale.[25,28,37] Although high concentrations of phenolic substrates destabilize the enzyme (e.g., above 50 mM 2-methylphenol, 25 mM 2-chlorophenol, or 10 mM 1-naphthol), this can be overcome by slow addition of the phenolic substrate. The preparation of NaphAla, a commercially unavailable blue-fluorescent ncAA whose applications have been limited by its challenging synthesis,[8] was examined first. Over the course of 24 hours, 1-naphthol ($0.14 / g, Millipore-Sigma) was slowly added to a solution of TmTyrS1 and Ser ($0.77 / g) to generate the NaphAla product, which precipitated from solution toward the end of the substrate addition (**Figure 6-3c** and Appendix E, **Section E.1.15**, *Multi-gram scale synthesis of NaphAla*). The resultant solid was collected over a filter, washed with ice-cold water and ethyl acetate to remove buffer salts and unreacted substrates, and subsequently dried *in vacuo*, affording 5.5 g of pure NaphAla (74% isolated yield relative to 1-naphthol, 91% weight purity, >99% enantiomeric excess (*ee*), **Figure E-14**) without significant reaction optimization. For comparison, the reported chemical synthesis requires three steps and an expensive rhodium catalyst to arrive at an enantioenriched, triply protected NaphAla product from 4-hydroxy-1-naphthaldehyde ($30 / g), with three more deprotection steps to yield NaphAla.[8]

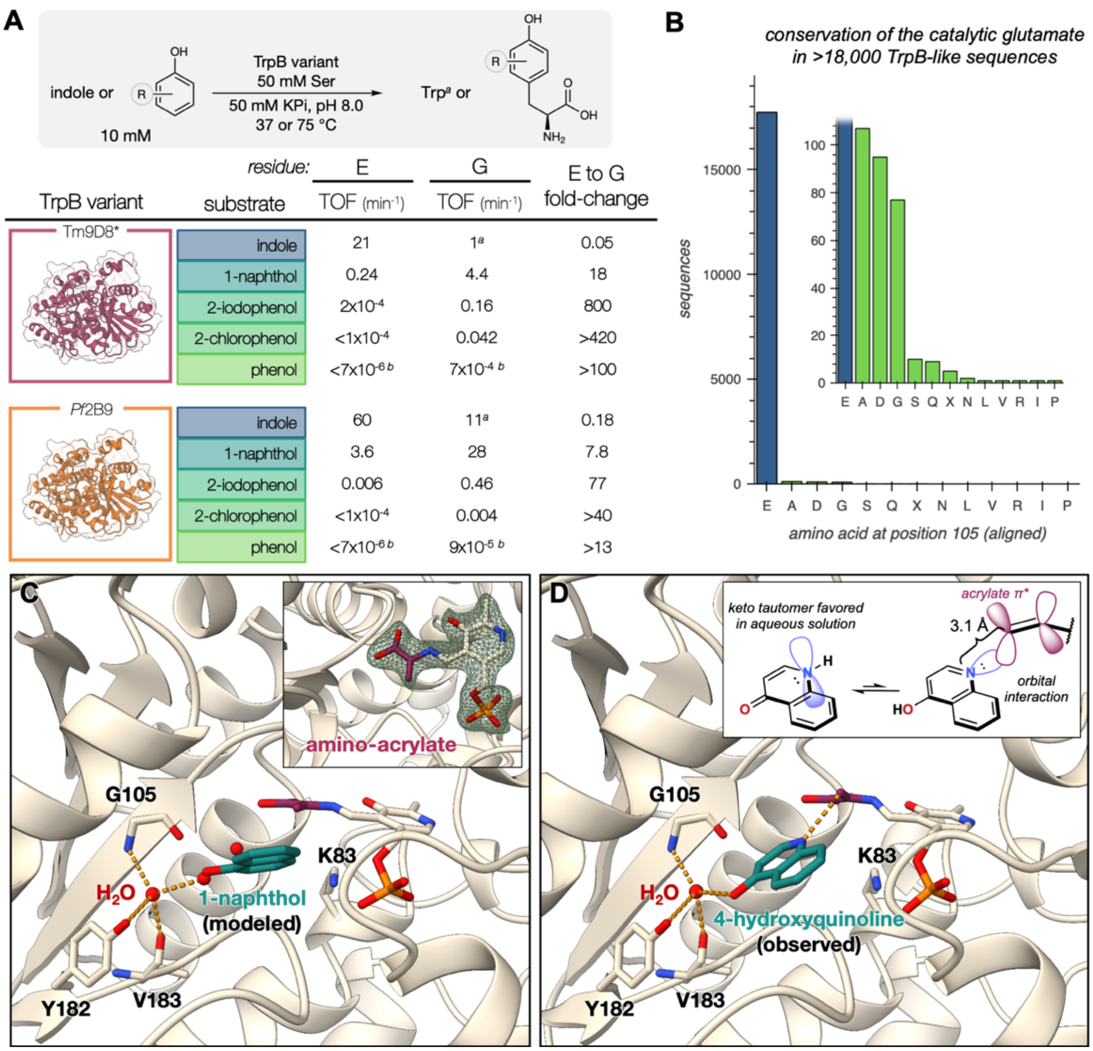
We used a similar approach to synthesize 3-methyl-Tyr, a Tyr analog made by radical *S*-adenosyl methionine (SAM)-catalyzed methylation of Tyr in the biosynthesis of saframycin A.[41] This simple ncAA is costly (~$1600 / g) and is prepared synthetically *via* cross coupling of tetramethyltin and 3-iodo-Tyr occurring over six days at 70 °C[42] or using TPL followed by chromatographic purification.[5,17] Using TmTyrS4, two sequential additions of 50 mM 2-methylphenol (<$0.1 / g) resulted in product precipitation at >90% conversion, allowing us to isolate 1.13 g of 3-methyl-Tyr without chromatographic purification (49% isolated yield relative to 2-methylphenol, 89% weight purity, >99% *ee*, **Figure E-15** and Appendix E, **Section E.1.16**, *Gram scale synthesis of 3-Me-Tyr*). As shown in **Figure 6-3a**, TmTyrS4 demonstrates moderate conversion of 2-methylphenol (42% product HPLC peak area), suggesting that this value can serve as a benchmark for the preparation of other compounds at scale.

These gram-scale syntheses are simple and effective. They take place at 37 °C in a ~100 mL volume per gram of product (roughly 10 g L^{-1} day^{-1} space-time yield) using inexpensive reagents and enzyme obtained from a 1-L bacterial culture. The high stability of the TyrS enzymes facilitates their preparation as bench-stable lyophilized powder from heat-treated lysate. The heat treatment removes nearly all mesophilic *E. coli* host proteins.

6.2.4 The E105G mutation converts TrpB to TyrS.

Because significant rate enhancement can be attributed to the single E105G mutation, we investigated whether this mutation is activating in other TrpBs. Installation of the equivalent E104G mutation in *Pf*2B9, which had previously demonstrated activity with 1-naphthol, increased this activity 7.8-fold at this enzyme's optimal temperature of 75 °C (**Figure 6-4a**). In both TrpBs, the mutation decreased activity toward indole to levels below those on 1-

naphthol and also decreased regioselectivity for indole alkylation (to <99.5%, **Figure E-16**), as previously observed.[25,28]



TrpB variant	substrate	E TOF (min⁻¹)	G TOF (min⁻¹)	E to G fold-change
Tm9D8*	indole	21	1[a]	0.05
	1-naphthol	0.24	4.4	18
	2-iodophenol	2×10^{-4}	0.16	800
	2-chlorophenol	$<1\times10^{-4}$	0.042	>420
	phenol	$<7\times10^{-6}$ [b]	7×10^{-4} [b]	>100
*Pf*2B9	indole	60	11[a]	0.18
	1-naphthol	3.6	28	7.8
	2-iodophenol	0.006	0.46	77
	2-chlorophenol	$<1\times10^{-4}$	0.004	>40
	phenol	$<7\times10^{-6}$ [b]	9×10^{-5} [b]	>13

Figure 6-4. Removal of the conserved catalytic glutamate in TrpB unlocks regioselective tyrosine synthase activity. a. Rates of conversion of different substrates in two TrpB variants with the catalytic glutamate sidechain (residue = E) or without it (residue = G). Tm9D8* tested at 37 °C; *Pf*2B9 tested at 75 °C. [a]Rate of indole conversion calculated using both Trp and isoTrp formation. [b]Performed with 50 mM substrate (**Figure E-17**). Reactions performed in duplicate. **b.** Conservation of the catalytic glutamate in 18,051 TrpB-like sequences. *Inset*: An axis-adjusted view for the sequences with a different residue. *X* = unidentified. **c.** Crystal structure of amino-acrylate bound TmTyrS1 with 1-naphthol naïvely modeled in a productive binding pose. *Inset*: amino-acrylate polder omit map contoured at 5σ. **d.** Crystal structure of amino-acrylate-bound TmTyrS1 in complex with the non-reactive 1-naphthol analog 4-hydroxylquinoline coordinated to an active-site water in place of E105. This interaction likely orients 1-naphthol and other phenolic nucleophiles for *para* C–C bond formation. *Inset*: The keto tautomer is favored in aqueous solution while the enol tautomer is observed in the enzyme interacting with the electrophilic amino-acrylate Cβ.

Removal of the glutamate side chain enhanced activity on simple phenol analogs under all tested conditions (**Figure 6-4a**). E104G effected a >40-fold increase in the activity of *Pf*2B9 toward 2-chlorophenol and a 77-fold increase toward 2-iodophenol. Tm9D8* E105G saw an even more impressive >420- and an 800-fold increase over Tm9D8* when provided 2-chlorophenol and 2-iodophenol, respectively. In both Tm9D8* and *Pf*2B9 variants, this sole glutamate-to-glycine substitution was sufficient to enable Tyr formation with increased enzyme and substrate concentration (**Figures 6-4a** and **E-17**).

The catalytic glutamate is strictly conserved in all characterized TrpB enzymes within the human-annotated SwissProt database (451 sequences). To examine the conservation of this residue in more detail, we analyzed 18,051 TrpB-like sequences with 14–93% aligned sequence identity to Tm9D8*. Of these, 98.28% (17,741) contained the catalytic glutamate (**Figure 6-4b**). Only three other amino acids occurred with a significant frequency: alanine (0.59%; 107), aspartate (0.53%; 95), and glycine (0.43%; 77). The residue corresponding to G229, where alanine was found to be activating in TmTyrS5, is equally conserved (98.20%; 17,726), but differences at this position are notably present in sequences in which the catalytic glutamate is absent (**Figure E-18**). It remains to be seen whether these are efficient Trp synthases or whether they have other biological roles.

6.2.5 Structural analysis of regioselective phenol alkylation.

The products of the TyrS reactions were always *para*-alkylated; *ortho*- and O-alkylation were never observed during evolution or in the substrate scope analysis. The regioselectivity of TyrS should, at a minimum, be achieved through active-site discrimination between the *para*-alkylating and *ortho*-alkylating binding modes, while the discrimination of C- and O-alkylation may be accomplished through other means.[43] Although steric factors can be used to justify the regioselectivity when alkylating a bulky substrate like 1-naphthol, the regioselective transformation of phenol to Tyr suggests that the active site interacts with the hydroxyl group. To probe this further we pursued structural characterization of TyrS enzymes in various catalytically relevant states, including in complex with substrate analogs. Whereas previous studies of *T. maritima* TrpB variants had to rely on homology models due to the absence of structural data,[28,37] we were able to obtain an experimental X-ray crystal structure of a TyrS for this study (**Figures E-19** and **E-20a** and **Table E-3**).

Previous reports showed that TrpB crystals formed in the resting state (known as the internal aldimine, E(A_{in}), state) can be used to form and observe the reactive amino-acrylate complex, E(A-A).[44] This is also the case for TyrS crystals, as soaking Ser into the E(A_{in}) crystals readily formed a stable E(A-A) complex in both subunits of the asymmetric unit (**Figure 6-4c**, inset, and **Figure E-20b**). The ability to observe this reactive intermediate highlights its stability within the TrpB scaffold. In contrast, the transient amino-acrylate of TPL could only be assumed to exist until spectroscopic evidence was obtained through kinetic trapping with an inhibitor.[18] In both the E(A_{in}) and E(A-A) structures of TmTyrS1, removal of the E105 sidechain made space for the coordination of a single water molecule in the active site. Furthermore, this coordinated water interacts with a second water molecule in the E(A-A) structure. Based on conserved indole interactions in TrpB, naïve modeling of 1-naphthol in a productive binding mode places the phenolic hydroxyl group in an orientation that would displace this second water molecule and interact with the water molecule coordinated in place of E105 (**Figure 6-4c**). Active-site water molecules have been shown to enable catalysis through electrostatic interactions with the functional groups of substrates,[45] providing a possible rationale for how and why the TyrS enzymes described here perform this non-native reaction with exquisite regioselectivity.

To obtain more conclusive evidence of the role of this water in coordinating the phenolic hydroxyl group, the structure of TmTyrS1 in the E(A-A) state was determined in complex with a non-reactive 1-naphthol analog, 4-hydroxyquinoline (**Figure E-21**). When soaked into crystals of the E(A-A) complex, the hydroxyl group of 4-hydroxyquinoline forms a hydrogen bond with the active-site water (O---O distance of 3.1 Å), supporting its putative role in directing regioselective bond formation by TmTyrS1 (**Figure 6-4d**). This molecule, which favors the keto tautomer (4-quinolone) in aqueous solution, is best modeled as the enol tautomer within the enzyme active site. This is based on the short interatomic distance between the quinoline nitrogen and the electrophilic β-carbon of the amino-acrylate (N---C distance of 3.1 Å) requiring a lone pair on the heterocyclic nitrogen (**Figure 6-4b**, inset). Such a binding mode for the structurally analogous 1-naphthol would thus direct C–C bond formation between the amino-acrylate and C4 of this and other phenolic substrates, providing a structural explanation for the regioselectivity of these reactions.

Soaking the TmTyrS1 crystals with 1-naphthol analog quinoline *N*-oxide showed this molecule could also bind within the active site of the E(A-A) complex and interact directly with the coordinated water. Interestingly, however, it binds in a non-productive mode, with C7 oriented toward the amino-acrylate (**Figure E-22**). Although the C7-alkylated product was never observed, this suggests that there may be an off-target binding mode for 1-naphthol that would increase the apparent K_M of the reaction in a competitive inhibition-like fashion. This may explain why 1-naphthol displays an elevated K_M of 2.8 mM (**Figure E-**

5e), which is roughly 10^2–10^3-fold higher than that of indole in native TrpB enzymes,[31,44] despite the similarity in steric bulk and the preservation of packing interactions around these two substrates. This observation provides a target for further improving the catalytic efficiency of these enzymes.

6.2.6 Kinetics of TyrS activity.

Consistent with the importance of substrate binding in these reactions, the classic reactivity patterns of electrophilic aromatic substitution are not observed, and—as noted during the evolution of TyrS and seen in the substrate scope—it appears that nucleophilicity of the arene ring is less important than other factors like sterics. Notably, however, while the increase in rate observed with the addition of the 3-fluoro substituent could be attributed to improved binding (**Figure 6-3a**), the rate could also be increased due to the higher acidity of C4 in the event that deprotonation after C–C bond formation is rate limiting (**Figure E-23**). Indeed, assays with two different deuterated phenols (2-chlorophenol-d_4 and 2-methylphenol-d_8) demonstrate clear primary kinetic isotope effects (KIEs) (**Table E-4**). Given that the only C–H bond that is broken is the one at C4, these KIEs indicate that deprotonation of C4—and thus rearomatization of the arene—is strongly rate limiting (**Figure E-23**).

Despite the 30,000-fold improvement achieved in novel tyrosine synthase activity, TmTyrS6 remains ~100-fold more sluggish than native TrpS enzymes with indole, even when the new substrates are provided at 50 mM (**Figure 6-2**). The apparent K_M for phenol is ~50 mM (**Figure E-24a**), compared to the low-micromolar K_M values for indole of TrpS.[31,39] The apparent K_M for Ser is also in the millimolar range (~3 mM, **Figure E-24b**) similar to the ~1 mM K_M values of native TrpS enzymes.[23,39] However, both concentrations far exceed what is achievable *in vivo* due to toxicity, particularly for phenol. Furthermore, Tyr is a more abundant metabolite than Trp, compounding the need for an efficient enzyme *in vivo*. Thus, perhaps unsurprisingly, initial efforts to apply TmTyrS6 within primary metabolism *via* the phenol-dependent complementation of Tyr-auxotrophic *E. coli* have not been successful.

6.3 Discussion.

The fact that TrpB is responsible for all known Trp biosynthesis but had never been reported to catalyze the formation of Tyr presented us with a compelling challenge: could we discover a new solution to Tyr synthesis that is fundamentally different from that used in nature? Through chemical intuition and substrate engineering, we quickly identified a single mutation at a highly conserved catalytic residue in TrpB that unlocked non-native activity on phenol analogs. However, activity for phenol itself remained low. Directed evolution using a 'substrate walk' approach[38] ultimately improved activity toward phenol by at least 30,000-fold, with the accumulation of roughly twenty mutations. Although the engineering efforts focused ultimately on the biocatalytic synthesis of Tyr—a relatively inexpensive and readily available product—the evolutionary engineering process delivered useful biocatalysts for the synthesis of valuable Tyr analogs at scale. We also identified activity for numerous other phenol analogs which can serve as starting points for future directed evolution campaigns.[2,6]

Rather than installing catalytic functionality directly through a new amino acid side chain, this new TyrS activity is greatly increased upon the effective removal of TrpB's catalytic glutamate—a residue conserved in all characterized TrpBs and 98.2% of TrpB-like sequences—by mutation to glycine. This small amino acid makes room for the serendipitous binding of a water molecule which is correctly positioned to bind and orient phenolic substrates for *para* C–C bond formation. These results highlight the often-observed fact that solutions discovered through evolution can be surprising and unintuitive. Evolutionary conservation suggests that mutating the catalytic glutamate would be strongly deleterious, and indeed it is for the native activity. Previous studies showed that mutation to alanine can impair formation of the amino-acrylate in *Salmonella typhimurium* TrpB.[35] Additionally, mutating this residue to glycine in native TrpS and engineered TrpB enzymes reduced the rate of Trp synthesis by 2–20 fold and effectively abrogated activity with azulene, a different non-indole nucleophile.[28] Yet the same mutation is highly activating for reactivity with phenols through a rare, water-mediated mechanism of substrate coordination. It is unlikely that this effective solution could be obtained using rational protein design methods, most of which do not explicitly include solvent. Although it is possible that future evolution might replace the coordinating water with a dedicated catalytic amino acid, the current solution is highly practical, given that the enzyme can catalyze the reaction at a rate of 11 min^{-1} (in the case of 1-naphthol) and produce valuable amino acids at scale.

It is interesting to speculate whether this biocatalyst represents a new possibility within nature's universally conserved pathways of aroAA biosynthesis. Throughout life, the only known route for the *de novo* biosynthesis of Tyr subjects chorismate—the precursor to all aroAAs—to the same three chemical steps: a Claisen rearrangement, dehydrogenation, and transamination (**Figure 6-1b**).[46] Given that phenol is accessible in just two enzymatic steps from chorismate (**Figure E-25**),[47] the regioselective and irreversible Friedel-Crafts alkylation of phenol by TyrS completes a new, similarly complex post-chorismate pathway for Tyr biosynthesis that is kinetically feasible, alleviating the thermodynamic limitations of TPL. Thus these new TyrS enzymes could open a new testing ground for innovation and discovery within primary metabolism.

Chapter VI Bibliography

1. Parmeggiani, F., Weise, N.J., Ahmed, S. T. & Turner, N. J. Synthetic and therapeutic applications of ammonia-lyases and aminomutases. *Chem. Rev.* **118**, 73–118 (2018).
2. Almhjell, P. J., Boville, C. E. & Arnold F. H., Engineering enzymes for noncanonical amino acid synthesis. *Chem. Soc. Rev.* **47**, 8980–8997 (2018).
3. Lütke-Eversloh, T., Santos, C. N. S. & Stephanopoulos, G. Perspectives of biotechnological production of L-tyrosine and its applications. *Appl. Microbiol. Biotechnol.* **77**, 751–762 (2007).
4. Cheng, Z., Kuru, E., Sachdeva, A. & Vendrell, M. Fluorescent amino acids as versatile building blocks for chemical biology. *Nat. Rev. Chem.* **4**, 275–290 (2020).
5. Kim, K., Parang, K., Lau, O. D. & Cole, P. A. Tyrosine analogues as alternative substrates for protein tyrosine kinase Csk: Insights into substrate selectivity and catalytic mechanism. *Bioorg. Med. Chem.* **8**, 1263–1268 (2000).
6. Rubini, R., Jansen, S. C., Beekhuis, H., Rozeboom, H. J. & Mayer, C. Selecting better biocatalysts by complementing recoded bacteria. *Angew. Chem. Int. Ed.* **62**, e202213942 (2023).
7. Seyedsayamdost, M. R., Reece, S. Y., Nocera, D. G. & Stubbe, J. A. Mono-, di-, tri-, and tetra-substituted fluorotyrosines: New probes for enzymes that use tyrosyl radicals in catalysis. *J. Am. Chem. Soc.* **128**, 1569–1579 (2006).
8. Knör, S., Laufer, B. & Kessler, H. Efficient enantioselective synthesis of condensed and aromatic-ring- substituted tyrosine derivatives. *J. Org. Chem.* **71**, 5625–5630 (2006).
9. Ager, D. J. Synthesis of unnatural/nonproteinogenic α-amino acids in *Amino Acids, Peptides, and Proteins in Organic Chemistry, Vol. 1 - Origins and Synthesis of Amino Acids*, A. B. Hughes, Ed. (WILEY-VCH, 2009), pp. 495–526.
10. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42** D459–D471 (2014).
11. Lynch, J. H. & Dudareva, N., Aromatic amino acids: A complex network ripe for future exploration. *Trends Plant Sci.* **25**, 670–681 (2020).
12. Rodriguez, A. *et al.* Engineering *Escherichia coli* to overproduce aromatic amino

acids and derived compounds. *Microb. Cell Fact*. **13** (2014).

13. Arnold, L. D., Kalantar, T. H. & Vederas, J. C. Conversion of serine to stereochemically pure β-substituted α-amino acids via β-lactones. *J. Am. Chem. Soc*. **107**, 7105–7109 (1985).

14. Tanner, D. Chiral aziridines—Their synthesis and use in stereoselective transformations. *Angew. Chem. Int. Ed*. **33**, 599–619 (1994).

15. Brittain, W. D. G. & Cobb, S. L. Negishi cross-couplings in the synthesis of amino acids. *Org Biomol. Chem*. **16**, 10–20 (2017).

16. Kumar, V., Turnbull, W. B. & Kumar, A. Review on recent developments in biocatalysts for Friedel–Crafts reactions. *ACS Catal*. **12**, 10742–10763 (2022).

17. Nagasawa, T. *et al.* Syntheses of L-tyrosine-related amino acids by tyrosine phenol-lyase of *Citrobacter intermedius*. *Eur. J. Biochem*. **117**, 33–40 (1981).

18. Phillips, R. S., Chen, H. Y. & Faleev, N. G. Aminoacrylate intermediates in the reaction of *Citrobacter freundii* tyrosine phenol-lyase. *Biochemistry*. **45**, 9575–9583 (2006).

19. Won, Y. *et al. In vivo* biosynthesis of tyrosine analogs and their concurrent incorporation into a residue-specific manner for enzyme engineering. *Chem. Commun*. **55**, 15133–15136 (2019).

20. Olson, N. M. *et al.* Development of a single culture *E. coli* expression system for the enzymatic synthesis of fluorinated tyrosine and its incorporation into proteins. *J. Fluor. Chem*. **261–262**, 110014 (2022).

21. Watanabe, T. & Snell, E. E. Reversibility of the tryptophanase reaction: Synthesis of tryptophan from indole, pyruvate, and ammonia. *Proc. Natl. Acad. Sci. U.S.A*. **69**, 1086–1090 (1972).

22. Watkins-Dulaney, E., Straathof, S. & Arnold, F. Tryptophan synthase: Biocatalyst extraordinaire. *ChemBioChem*. **22**, 5–16 (2021).

23. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).

24. Herger, M. *et al.* Synthesis of β-branched tryptophan analogues using an engineered subunit of tryptophan synthase. *J. Am. Chem. Soc*. **138**, 8388–8391 (2016).

25. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc*. **139**, 10769–10776 (2017).

26. Romney, D. K., Sarai, N. S. & Arnold, F. H. Nitroalkanes as versatile nucleophiles for enzymatic synthesis of noncanonical amino acids. *ACS Catal*. **9**, 8726–8730 (2019).

27. Dick, M., Sarai, N. S., Martynowycz, M. W., Gonen, T. & Arnold, F. H. Tailoring tryptophan synthase TrpB for selective quaternary carbon bond formation. *J. Am. Chem. Soc*. **141**, 19817–19822 (2019).

28. Watkins, E. J., Almhjell, P. J. & Arnold, F. H. Direct enzymatic synthesis of a deep-blue fluorescent noncanonical amino acid from azulene and serine. *ChemBioChem*. **21**, 80–83 (2019).

29. Goss, R. J. M. & Newill, P. L. A. A convenient enzymatic synthesis of L-halotryptophans. *Chem. Commun.*, 4924–4925 (2006).

30. Smith, D. R. M. *et al.* The first one-pot synthesis of L-7-iodotryptophan from 7-iodoindole and serine, and an improved synthesis of other L-7-halotryptophans. *Org. Lett*. **16**, 2622–2625 (2014).

31. Rix, G. *et al.* Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nat. Commun*. **11**, 5644 (2020).

32. Tatum, E. L. & Bonner, D. Indole and serine in the biosynthesis and breakdown of tryptophane. *Proc. Natl. Acad. Sci. U.S.A.* **30**, 30–37 (1944).

33. Hall, A. N., Lea, D. J. & Rhydon, H. N. The behaviour of the Bz-methylindoles as substrates and inhibitors for *Neurospora crassa* tryptophan synthase. *Biochem. J.* **84**, 12–16 (1962).

34. Brzovic, P. S., Kayastha, A. M., Miles, E. W. & Dunn, M. F. Substitution of glutamic acid 109 by aspartic acid alters the substrate specificity and catalytic activity of the β-subunit in the tryptophan synthase bienzyme complex from *Salmonella typhimurium*. *Biochemistry*. **31**, 1180–1190 (1992).

35. Ruvinov, S. B., Ahmed, S. A., McPhie, P. & Miles, E. W. Monovalent cations partially repair a conformational defect in a mutant tryptophan synthase α2β2 complex (β-E109A). *J. Biol. Chem*. **270**, 17333–17338 (1995).

36. Milić, D. *et al.* Structures of apo- and holo-tyrosine phenol-lyase reveal a catalytically critical closed conformation and suggest a mechanism for activation by K^+ ions. *Biochemistry*. **45**, 7544–7552 (2006).

37. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogs in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *J. Org. Chem*. **83**, 7447–7452 (2018).

38. Chen, Z., & Zhao, H. Rapid creation of a novel protein function by *in vitro* coevolution. *J. Mol. Biol*. **348**, 1273–1282 (2005).

39. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A panel of TrpB biocatalysts derived from tryptophan synthase through the transfer of mutations that mimic allosteric activation. *Angew. Chem. Int. Ed*. **55**, 11577–11581 (2016).

40. Boville, C. E. *et al.* Engineered biosynthesis of β-alkyl tryptophan analogues. *Angew. Chem. Int. Ed*. **57**, 14764–14768 (2018).

41. Tang, M. C., Fu, C. Y. & Tang, G. L. Characterization of SfmD as a heme peroxidase that catalyzes the regioselective hydroxylation of 3-methyltyrosine to 3-hydroxy-5-methyltyrosine in saframycin A biosynthesis. *J. Biol. Chem*. **287**, 5112–5121 (2012).

42. Schmidt, E. W., Nelson, J. T. & Fillmore, J. P. Synthesis of tyrosine derivatives for saframycin MX1 biosynthetic studies. *Tetrahedron Lett*. **45**, 3921–3924 (2004).

43. Smith, J. L., Harrison, I. M., Bingman, C. & Buller, A. R. Investigation of β-substitution activity of O-acetylserine sulfhydrolase from *Citrullus vulgaris*. *ChemBioChem*, e202200157 (2022).

44. A. R. Buller *et al.* Directed evolution mimics allosteric activation by stepwise tuning of the conformational ensemble. *J. Am. Chem. Soc*. **120**, 7256–7266 (2018).

45. Kraut, D. A., Sigala, P. A., Fenn, T. D. & Herschlag, D. Dissecting the paradoxical effects of hydrogen bond mutations in the ketosteroid isomerase oxyanion hole. *Proc. Natl. Acad. Sci. U.S.A*. **107**, 1960–1965 (2010).

46. Merino, E., Jensen, R. A. & Yanofsky, C. Evolution of bacterial trp operons and their regulation. *Curr. Opin. Microbiol*. **11**, 78–86 (2008).

47. Thompson, B., Machas, M. & Nielsen, D. R. Engineering and comparison of non-natural pathways for microbial phenol production. *Biotechnol. Bioeng*. **113**, 1745–1754 (2016).

APPENDICES

Supplementary Information for Chapters II–VI

Appendix A
SUPPLEMENTARY INFORMATION FOR CHAPTER II

A.1 Results of site-saturation mutagenesis libraries.

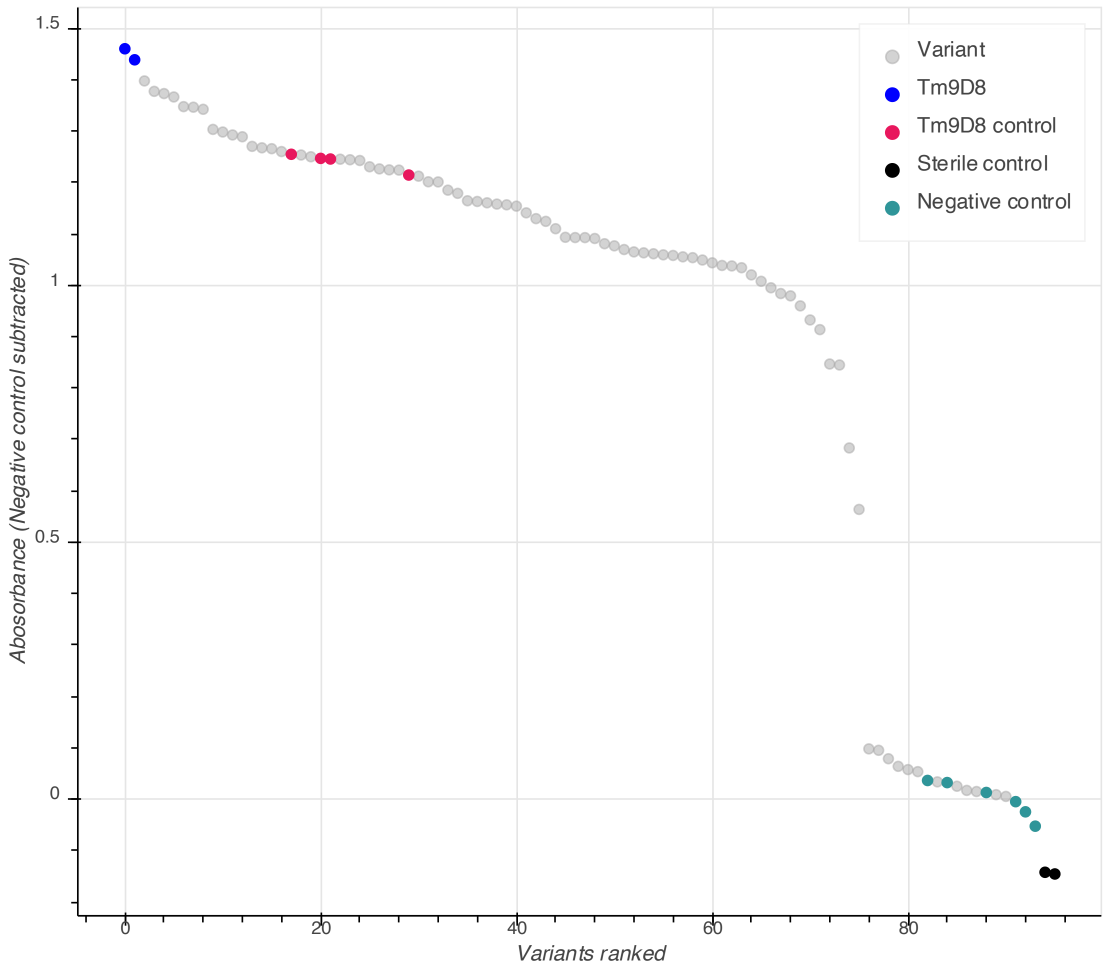


Figure A-1. Site-saturation mutagenesis of residue 30 in *Tm*9D8. Following the procedure in the Experimental Section, 80 colonies (4-fold oversampling) were assayed for 4-CN-Trp production at 50 °C. Although many variants showed parent-like activity, the top two variants were found to be parent (G30).

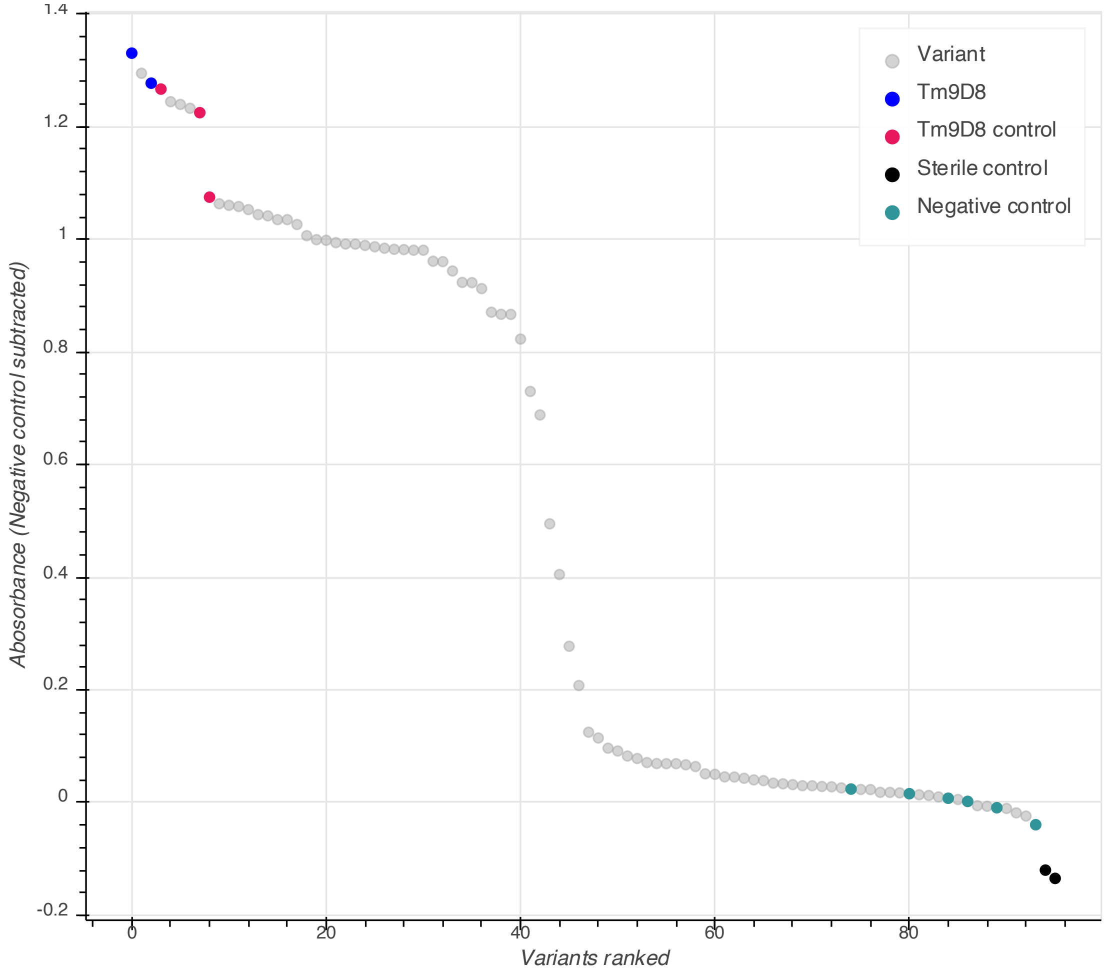


Figure A-2. Site-saturation mutagenesis of residue 228 in *Tm*9D8. Following the procedure in the Experimental Section, 80 colonies (4-fold oversampling) were assayed for 4-CN-Trp production at 50 °C. The top two variants were identified as parent (S228).

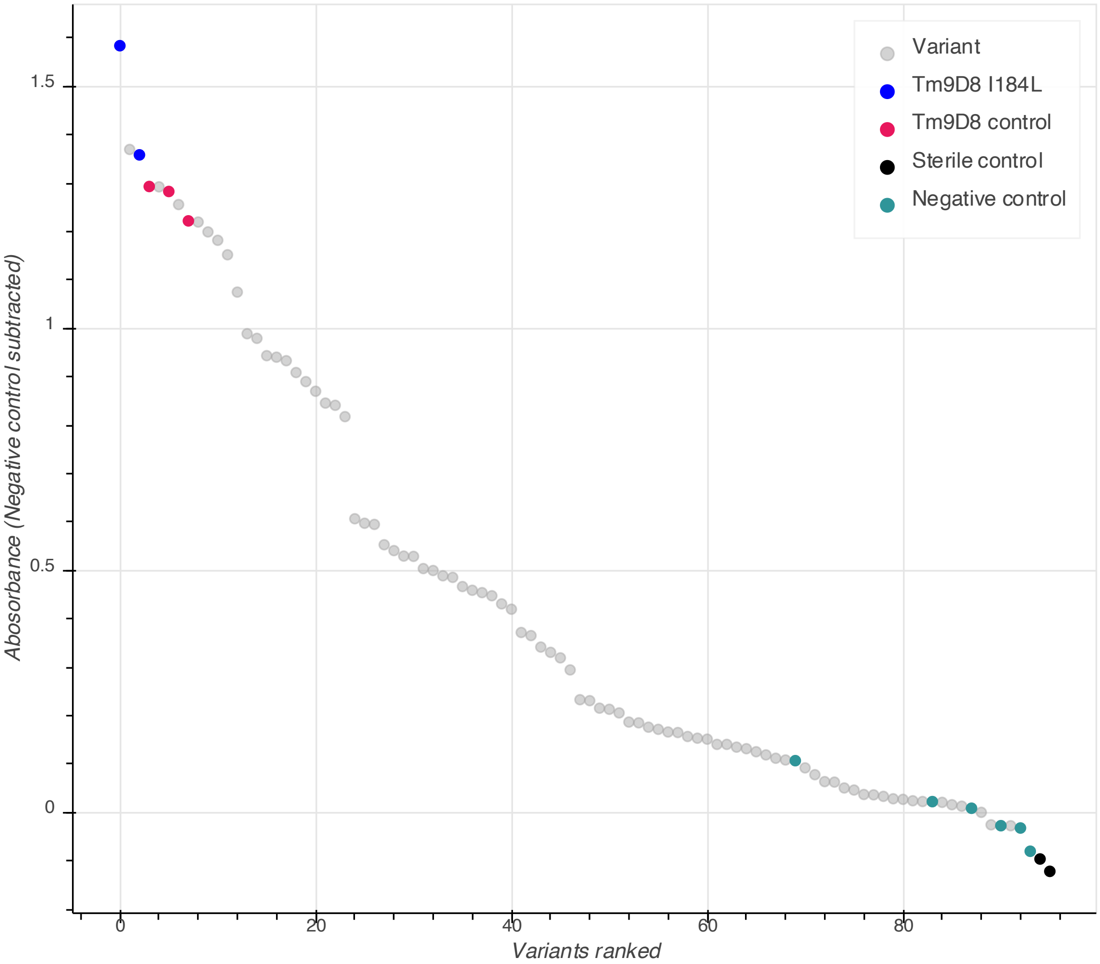
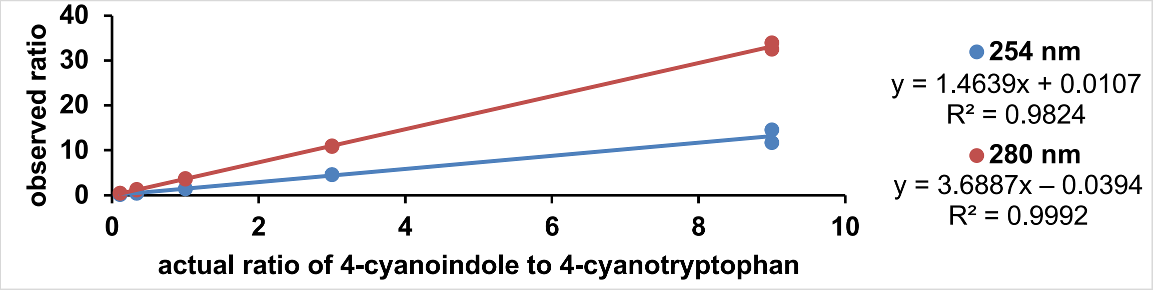


Figure A-3. Site-saturation mutagenesis of residue 184 in *Tm*9D8. Following the procedure in the Experimental Section, 80 colonies (4-fold oversampling) were assayed for 4-CN-Trp production at 50 °C. I184L was found in two of the top variants, but this mutation was not as beneficial as I184F in the rescreen.

A.2 LCMS calibration curves for Table 2-2.



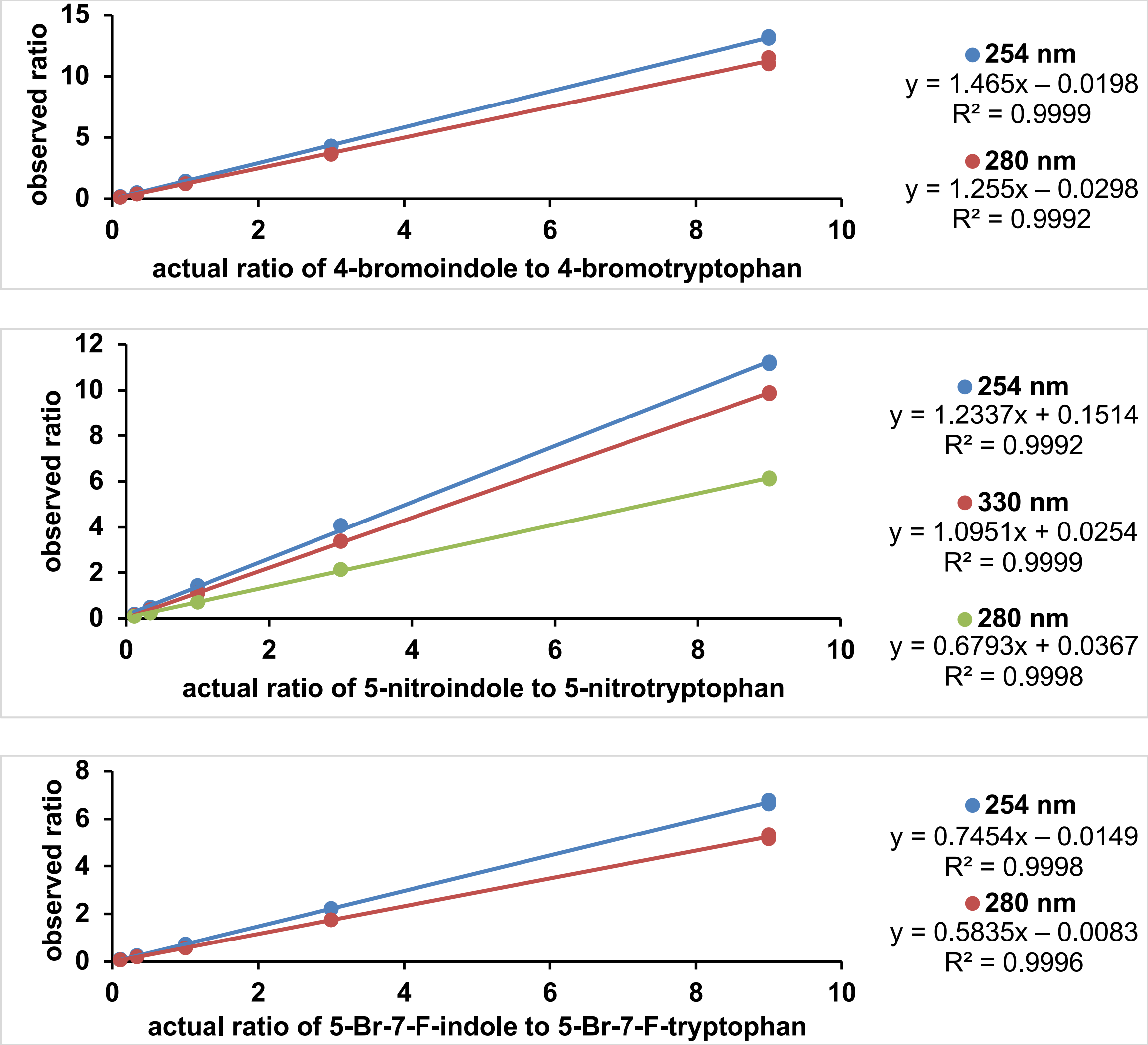


Figure A-4. Using an authentic product standard, mixtures of starting material and product at different ratios (9:1, 3:1, 1:1, 1:3, and 1:9) were used to generate a calibration curve. Each mixture was prepared in duplicate with a final concentration of 1 mM in 1:1 1 M aq. HCl/ CH_3CN. Mixtures were analyzed with LCMS at 254 nm and 280 nm (reference 360 nm, bandwidth 100 nm) and 330 nm (no reference wavelength) and were correlated to the actual ratios by a linear relationship. The correlation for 4-nitrotryptophan was published previously.[1] For tryptophan and 5-chloro-7-iodotryptophan, the HPLC yield was approximated by comparing absorption peaks of product and starting material at 280 nm.

A.3 Analytical reaction yield results.

Table A-1. HPLC data for Figure 2-2

	Temperature (°C)					
	37	37	50	50	75	75
***Tm*2F3**						
254 nm:	5.68%	5.76%	17.1%	17.3%	15.1%	18.7%
280 nm:	3.19%	3.00%	10.8%	10.0%	7.44%	9.78%
***Tm*9D8**						
254 nm:	30.5%	28.7%	38.0%	36.2%	18.0%	16.5%
280 nm:	16.6%	15.2%	21.9%	19.5%	11.7%	10.8%
Tm*9D8						
254 nm:	69.2%	66.5%	70.1%	71.1%	35.1%	35.9%
280 nm:	48.1%	47.6%	49.9%	48.7%	19.3%	20.5%

	Temperature (°C)					
	37	37	50	50	75	75
***Tm*2F3**			**Corrected**			
254 nm:	8.11%	8.22%	23.2%	23.5%	20.6%	25.3%
280 nm:	10.8%	10.2%	30.7%	29.0%	22.8%	28.5%
***Tm*9D8**			**Corrected**			
254 nm:	39.2%	37.2%	47.5%	45.5%	24.4%	22.5%
280 nm:	42.1%	39.6%	50.5%	46.9%	32.7%	30.7%
Tm*9D8			**Corrected**			
254 nm:	77.1%	74.8%	77.9%	78.7%	44.4%	45.1%
280 nm:	76.7%	76.4%	77.9%	77.1%	46.7%	48.6%

Table A-2. HPLC data for reactions after 1 hour[a]

	HPLC yield			**Corrected**				
Catalyst	**254 nm**	**254 nm**	**254 nm**	**#1**	**#2**	**#3**	**Average**	**Std. Dev.**
*Tm*2F3[b]	0.834%	0.800%	0.840%	1.22%	1.17%	1.22%	1.20%	0.03%
*Tm*9D8[b]	3.7%	3.7%	4.3%	5.4%	5.3%	6.1%	5.6%	0.4%
*Tm*9D8*[c]	3.9%	3.8%	4.3%	5.6%	5.5%	6.1%	5.7%	0.3%

[a]Maximum 1000 turnovers. [b]Reactions conducted at 50 °C. [c]Reactions conducted at 37 °C.

Table A-3. HPLC data for Table 2-2 and indole

4-bromotryptophan:

	HPLC yield at indicated wavelength							
Catalyst	**254 nm**	**254 nm**	**Average**	**Corrected**	**280 nm**	**280 nm**	**Average**	**Corrected**
*Tm*2F3	58.3%	58.1%	58.2%	66.5%	60.9%	59.7%	60.3%	64.6%
*Tm*9D8	63.2%	65.7%	64.4%	71.9%	67.1%	68.0%	67.6%	71.1%
*Tm*9D8*	69.5%	68.6%	69.1%	75.8%	72.7%	72.2%	72.4%	75.3%

4-nitrotryptophan:

	HPLC yield at indicated wavelength							
Catalyst	**254 nm**	**254 nm**	**Average**	**Corrected**	**330 nm**	**330 nm**	**Average**	**Corrected**
*Tm*2F3	3.45%	2.51%	2.98%	2.52%	1.09%	1.14%	1.11%	2.58%
*Tm*9D8	1.74%	2.70%	2.22%	1.87%	1.10%	1.08%	1.09%	2.53%
*Tm*9D8*	–	–	–	–	–	–	–	–

5-nitrotryptophan:

	HPLC yield at indicated wavelength							
Catalyst	**254 nm**	**254 nm**	**Average**	**Corrected**	**330 nm**	**330 nm**	**Average**	**Corrected**
*Tm*2F3	16.4%	17.9%	17.2%	20.9%	19.3%	19.4%	19.3%	20.9%
*Tm*9D8	4.28%	4.97%	4.62%	5.68%	4.79%	4.92%	4.86%	5.3%
*Tm*9D8*	5.57%	5.80%	5.68%	6.98%	6.77%	6.97%	6.87%	7.5%

5-bromo-7-fluorotryptophan:

	HPLC yield at indicated wavelength							
Catalyst	**254 nm**	**254 nm**	**Average**	**Corrected**	**280 nm**	**280 nm**	**Average**	**Corrected**
*Tm*2F3	40.8%	42.1%	41.4%	34.3%	46.8%	48.1%	47.4%	34.3%
*Tm*9D8	96.8%	97.1%	96.9%	94.1%	97.1%	97.7%	97.4%	94.4%
*Tm*9D8*	94.4%	95.4%	94.9%	91.5%	95.1%	96.4%	95.7%	91.7%

5-chloro-7-iodotryptophan:

		HPLC yield	
Catalyst	**280 nm**	**280 nm**	**Average**
*Tm*2F3	2.3%	2.2%	2.3%
*Tm*9D8	19.7%	20.6%	20.2%
*Tm*9D8*	18.9%	17.8%	18.4%

tryptophan:

		HPLC yield	
Catalyst	**280 nm**	**280 nm**	**Average**
*Tm*2F3	99.3%	97.3%	98.3%
*Tm*9D8	90.6%	88.3%	89.4%
*Tm*9D8*	79.0%	79.7%	79.3%

A.4 Primer sequences.

Table A-4. Primers for construction of recombination libraries

Fragment	Forward Primer	Reverse Primer
Fragment 1: *Nde*I to E30	GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG	CTCATCTTTCATGATTYCTTCGTACGCAGCTTC
Fragment 2: E30 to I184	GAAGCTGCGTACGAAGRAATCATGAAAGATGAG	ACCAACCACAGAGCCGTWCACGTAATAGGTGGT
Fragment 3: I184 to G228	ACCACCTATTACGTGWTCGGCTCTGTGGTTGGT	AGCGTTAGAACCACCGCYCACGCACGCAACGAT
Fragment 4: G228 to *Xho*I	ATCGTTGCGTGCGTGRGCGGTGGTTCTAACGCT	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAG
Gene Assembly	GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAG

Table A-5. Primers for construction of site-saturation libraries

target *Tm*9D8 residue	Forward primer	Reverse primer
E30	TGCGTACGAA**NNN**ATCATGAAAGATGAG	GCTTCCAGTTCTTCCAGAG
I184	CTATTACGTG**NNN**GGCTCTGTGGTTGG	GTGGTCTGCAGGTTGGTA
G228	TGCGTGCGTG**NNN**GGTGGTTCTA	ACGATGTAGTCCGGCAGAC

The codon indicated as NNN is replaced with NDT, VHG, or TGG.

A.5 NMR spectra.

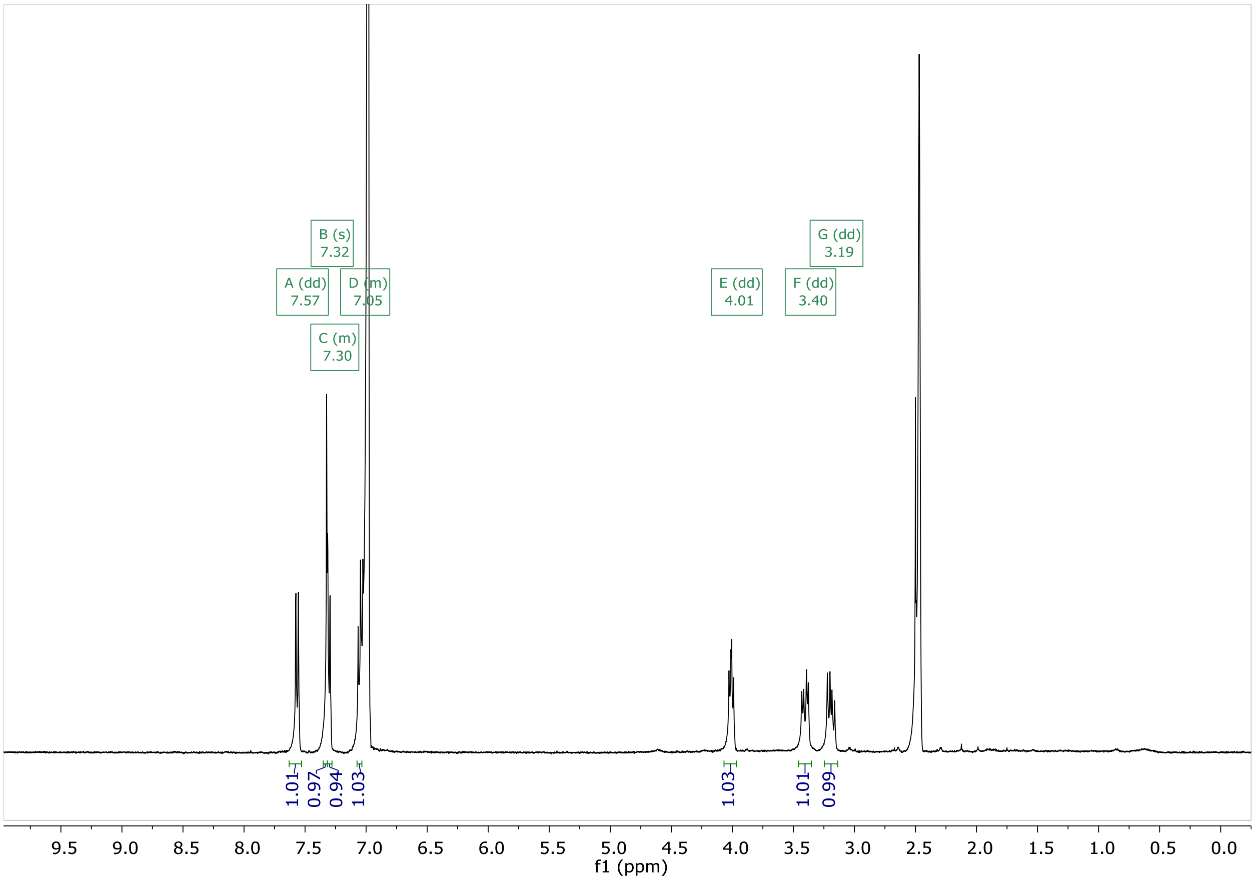


Figure A-5. 4-Cyanotryptophan NMR spectrum for Scheme 2-2.

Bibliography for Appendix A

1. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).

Appendix B

SUPPLEMENTARY INFORMATION FOR CHAPTER III

B.1 Figures and Tables

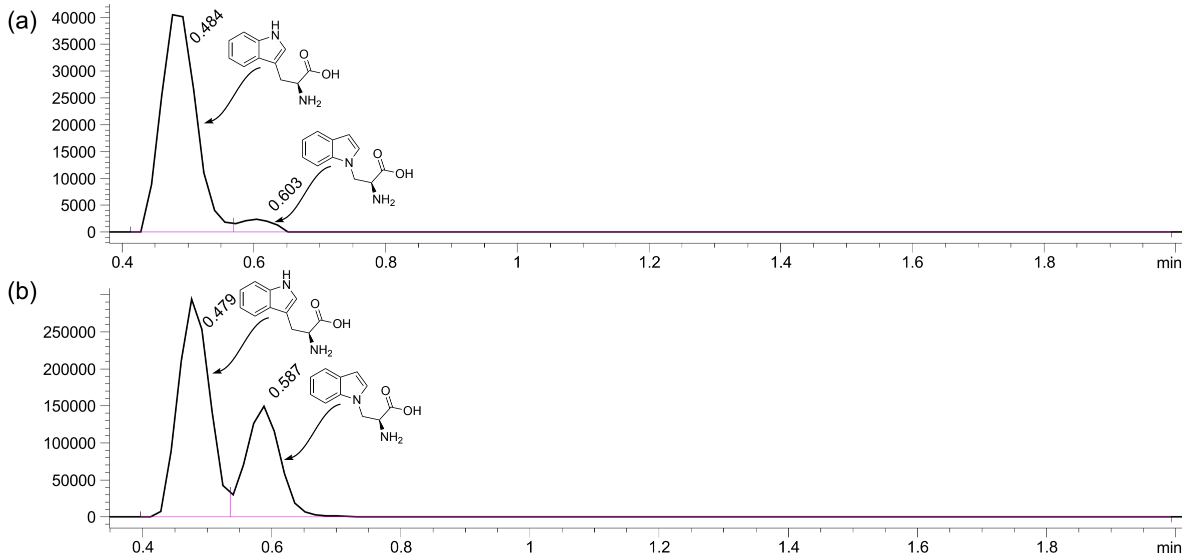


Figure B-1. Isotryptophan formation catalyzed by TrpB E104(105)G mutants. Traces shown as total ion count filtered by [M+H]+ for Trp (m/z = 205). **a.** A trace amount of isoTrp is formed by *Pf*5G8 E104G alongside Trp. **b.** *Tm*9D8* E105G also produces isoTrp as the minor product. Reactions prepared according to **Section B.2.6**, with conditions given in **Table B-1**.

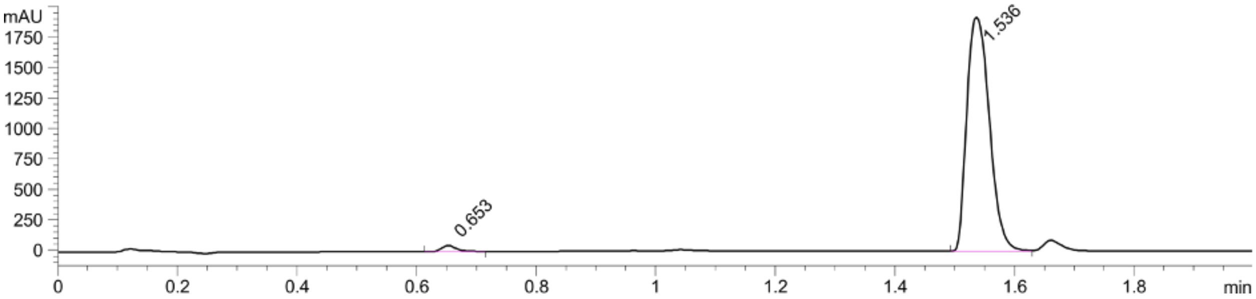


Figure B-2. Trace AzAla production by *Tm*9D8* E105G. After 16 hours at 37 °C with 0.1 mol % catalyst loading, only a small amount of AzAla (peak at 0.653 min) is formed by *Tm*9D8* E105G. The reaction was prepared according to **Section B.2.6**, but the reaction mixture was not diluted after centrifugation to increase the concentration of product in the sample. Trace AzAla yield is also seen for *Pf*5G8 E104G at 75 °C, but the sublimation of azulene interferes with analysis of the reaction at long timescales.

Table B-1. Rate comparisons among native and engineered TrpB enzymes[a]

	initial turnover frequency (min^{-1})			
	to Trp		to AzAla	
catalyst	E104(105)	G104(105)	E104(105)	G104(105)
*Tm*TrpS[b]	174 ± 1	—	12.0 ± 0.4	—
*Pf*5G8[b]	12.5 ± 0.3	6.8 ± 0.6	21.4 ± 1.4	n.d.
*Tm*9D8*[c]	19.0 ± 0.1	0.68 ± 0.1[d]	4.6 ± 0.6	n.d.
*Tm*Azul[c]	—	—	14.0 ± 0.4	—

[a]Reactions performed according to **Section B.2.6** with 0.02 mol % catalyst loading and allowed to react for 15 minutes (Trp production) or 1 hour (AzAla production). [b]Reactions run at 75 ◦C. [c]Reactions run at 37 ◦C. [d]Reaction time and catalyst loading increased to 2 hours and 0.1 mol % catalyst loading to improve quantification. n.d. = not detected by LC-MS.

B.2 Experimental Procedures

B.2.1 General experimental methods

Chemicals and reagents were purchased from commercial sources and used without further purification. Proton and carbon NMR were recorded on a Bruker 400 MHz (100 MHz) spectrometer equipped with a cryogenic probe. Proton chemical shifts are reported in ppm (δ) relative to tetramethylsilane and calibrated using the residual solvent resonance (D2O, δ 4.79 ppm). Data are reported as follows: chemical shift (multiplicity [singlet (s), doublet (d), doublet of doublets (dd), doublet of doublet of doublets (ddd), triplet (t), triplet of doublets (td), multiplet (m)], coupling constants [Hz], integration).

All cultures were grown in Terrific Broth supplemented with 100 μg/mL carbenicillin (TB$_{carb}$). Cultures were shaken in the New Brunswick Innova 4000 shaker (shaking diameter 19 mm), with the exception of 96-well plates, which were shaken in the Multitron INFORS HT shaker (shaking diameter 50 mm). Lysis buffer was composed of 50 mM potassium phosphate, pH 8.0 (KPi buffer), supplemented with 100 μM pyridoxal 5'-phosphate (PLP).

Lysis was performed in 75 °C water bath (Fisherbrand™ Isotemp™ Digital-Control Water Baths: Model 220) for >1 h. Reactions were performed in 50 mM KPi, pH 8.0, unless otherwise stated. High-resolution mass spectrometry (HRMS) was conducted on an Agilent 6200 TOF using electrospray ionization (ESI) to ionize the sample. Liquid chromatography/mass spectrometry (LCMS) was performed on an Agilent 1290 UPLC-LCMS equipped with a C-18 silica column (1.8 µm, 2.1 × 50 mm) using CH_3CN/H2O (0.1% acetic acid by volume): 5% to 95% CH_3CN over 2 min; 1 mL/min.

B.2.2 Cloning, expression, and purification of TrpB variants

The genes encoding *Tm*TrpA^{WT} (Uniprot P50908), *Tm*TrpB^{WT} (Uniprot G4FDT2), *Pf*5G8, *Pf*5G8 E104G, *Tm*9D8*, and *Tm*9D8* E105G were previously cloned into pET22b(+) with a C-terminal 6x His tag. Protein expression of the variants was carried out in *Escherichia coli* BL21(DE3) *E. cloni* Express® cells (Lucigen) by inoculating 5 mL TB_{carb} with a single colony and incubating this pre-culture overnight at 37 °C and 230 rpm. For expression, 2.5 mL culture were used to inoculate 500 mL Tb_{carb} in a 2-L flask and incubated at 37 °C and 130 rpm for 2.5 hours to reach OD_{600} 0.6–0.8. Cultures were chilled on ice for 20–30 minutes, and protein expression was induced with a final concentration of 1 mM isopropyl β-D-thiogalactopyranoside (IPTG). Expression proceeded at 20 °C and 130 rpm for approximately 24 h. Cells were harvested by centrifugation at 10,000*g* for 20 minutes at 4 °C and the supernatant was decanted. The pellet was stored at −20 °C until further use.

For protein purification, cells were thawed and were resuspended in 4 mL lysis buffer/g pellet. Cells were heat-treated at 75 °C for >1 h. The supernatant was collected from clarified lysate following centrifugation for 20 min at 14,000*g* and 4 °C. Purification was performed with a 1-mL Ni-NTA gravity flow column at room temperature. Buffer A: 20 mM imidazole, 25 mM KPi buffer, Buffer B: 500 mM imidazole, 25 mM KPi buffer. The column was equilibrated with 10 column volumes (CV) Buffer A. Subsequently, heat-treated lysate was loaded onto column and washed with 10 CV Buffer A. Next, 10 CV 1:1 Buffer A: Buffer B were added to elute non-target proteins from the column. Protein was eluted with 3 CV Buffer B. Proteins were dialyzed into 50 mM KPi buffer, flash frozen in liquid nitrogen, and stored at −80 °C.

To obtain the purified *Tm*TrpS complex, *Tm*TrpA^{WT} was co-purified with *Tm*TrpB^{WT}. Heat-treated lysate of both *Tm*TrpA^{WT} and *Tm*TrpB^{WT} were mixed together in a roughly 5:1 ratio (concentrations determined relative band intensities from SDS-PAGE analysis) and purification was followed as described above.

B.2.3 Construction of random mutagenesis libraries

Random mutagenesis libraries were generated with the *Tm*9D8* gene as template by the addition of 200–400 µM MnCl*2* to a *Taq* (New England Biolabs) PCR reaction as previously reported.[1] PCR fragments were treated with *Dpn*I for 1 h at 37 °C, purified by gel extraction, and then inserted into a pET22b(+) vector via Gibson assembly.[2] The Gibson assembly product was purified and concentrated using Zymo DNA Clean and Concentrate–5 kit (Catalog #: D4004). BL21(DE3). *E. cloni* Express® cells were transformed with the Gibson assembly product. Libraries generated with 200, 300, and 400 µM $MnCl_2$ were tested (one 96-well plate, each) to determine which library gave the optimal balance of high diversity and low rate of inactivation. The chosen library was then tested further (see below).

B.2.4 Library expression and screening

Individual colonies were grown in 300 μL TB_{carb} in deep-well 96-well polypropylene plates and grown overnight at 37 °C, 250 rpm, 80% humidity. The following day, 20 μL overnight culture were used to inoculate 630 μL TB_{carb} cultures in deep-well 96-well plates and grown at 37 °C, 250 rpm. After 2.5 h, cultures were chilled on ice for 20–30 minutes and protein expression was induced upon addition of 50 μL IPTG (final conc. 1 mM) diluted in TB_{carb}. Cultures were shaken at 20 °C, 250 rpm for 20–24 h, after which they were subjected to centrifugation at 4,000*g* for 10 min. The cell pellets could be frozen at −20 °C until further use or used immediately.

Pellets were lysed in 300 μL lysis buffer and heat-treated lysate clarified by centrifugation at 4,000*g* for 10 min. To UV-transparent 96-well assay plates (Caplugs, catalog #:290-8120-0AF) charged with 10 μL azulene dissolved in DMSO (final conc. 0.625 mM), 30 μL heat-treated lysate was transferred using Microlab NIMBUS96 liquid handler (Hamilton), followed by addition of 70 μL serine (final conc. 10 mM), and 90 μL 50 mM KPi buffer with a 12-channel pipet. Reactions were sealed with Microseal 'B' PCR plate sealing film (BioRad, catalog #: MSB1001) and incubated in a 37 °C water bath. Reaction progress was monitored by measuring absorption at 340 nm over the course of 24 h, in which more active variants retained AzAla in solution while the wells of inactive variants lost azulene due to absorption into the plastic assay plate and/or due to sublimation from solution. This can be seen below (**Figure B-3**), in a case where *Tm*9D8* is not given Ser, resulting in no activity, as opposed to the case where the addition of Ser allows for AzAla formation and retention of signal at 340 nm:

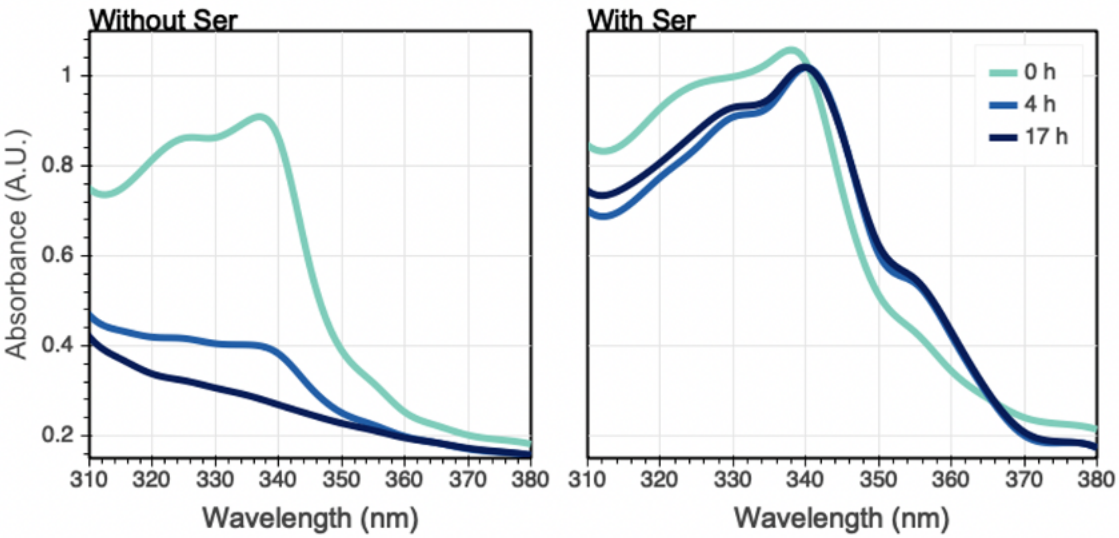


Figure B-3. Graphs of UV absorbance (A.U.) vs. wavelength for AzAla screen. Reactions (200 µL) were run in UV-transparent plates for 17 hours, incubated at 37 ºC. Lysate (30 µL) and 0.625 mM azulene were added to each well. No-serine control received 50 mM Kpi instead of 10 mM serine in 50 mM Kpi. Averaged data from two replicates of plate. Initially, samples have similar absorbance, but over time serine (-) control wells lose absorbance while serine (+) wells maintain signal.

B.2.5 Recombination of mutations

Site-directed mutagenesis was performed to combine mutations identified through screening to be beneficial for AzAla production. Gibson primers were designed to recombine the F184S and W286R mutations. Using *Tm*9D8* W286R as template, the F184S mutation was introduced. Two PCRs were performed using Phusion polymerase (New England Biolabs), each fragment originating from the F184S site and ending in the center of the ampicillin resistance gene. The PCR fragments were treated with *Dpn*I for 1 h at 37 °C, purified by gel extraction, and then combined via Gibson assembly.[2] The Gibson assembly product was purified using Zymo DNA Clean and Concentrate–5 kit (Catalog #: D4004). BL21(DE3) E. cloni Express® cells were transformed with the Gibson assembly product.

Table B-2. Gibson assembly primers

Fragment	Size (Kb)	Forward primer (5’ to 3’)	Reverse primer (5’ to 3’)
1	4.8	CCTGCAGACCACCTATTACGTGTCCGGCTCTGTGGTTGGTCCGCATCCATATCCG	CTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCG
2	1.7	CCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGC	GCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTG

B.2.6 Small-scale analytical reactions

All analytical reactions were performed in 2-mL glass HPLC vials (Agilent) charged with 10 μL azulene or indole (final conc. 10 mM) dissolved in DMSO (5% v/v), followed by addition of serine (final conc. 10 mM) and purified enzyme diluted in 50 mM KPi buffer to a final volume of 200 μL. Reactions were incubated at 75 °C or 37 °C for 24 h. The reaction was then diluted with 800 μL of 1:1 CH_3CN/1 M aq. HCl and subjected to centrifugation at 20,000*g*. For indole, this reaction mixture was analyzed directly by UHPLC-MS at 277 nm, representing the isosbestic point between indole and tryptophan and allowing quantification of yield by comparing the substrate and product peak areas.[3] For azulene, the reaction mixture was further diluted by adding 10 μL of this mixture into 190 μL of 1:1 CH_3CN/1 M aq. HCl and analyzed by UHPLC-MS. The yield was estimated comparing the integration of the product peak at 254 nm to a calibration curve (**Section A.3.8**).

B.2.7 Calibration for measuring HPLC yield of AzAla

Solutions of azulene and an authentic AzAla standard were made in 1:1 CH_3CN/1 M aq. HCl (total conc. 1 mM) and mixed in different ratios (9:1, 3:1, 1:1, 1:3, 1:9) in duplicate and analyzed by HPLC. The ratios of substrate to product peaks at 254 nm were correlated to the actual ratios by a linear relationship (**Figure B-4**).

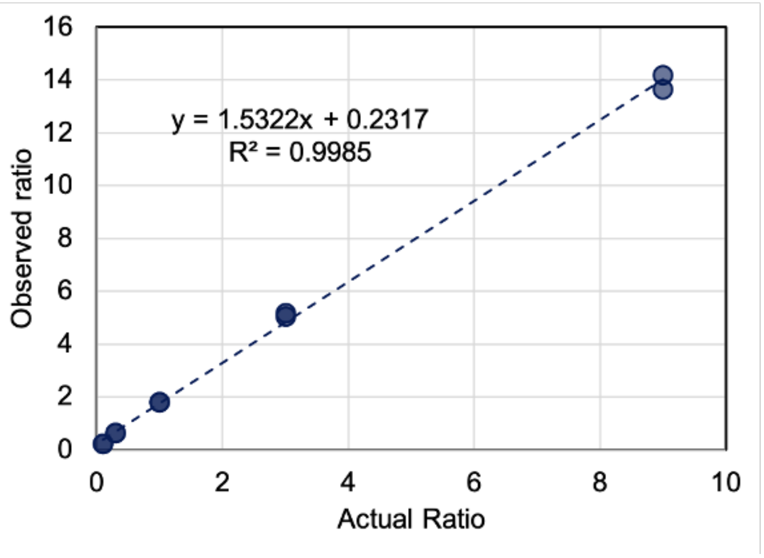


Figure B-4. AzAla calibration curve.

B.2.8 Large-scale preparation of AzAla

Two 500 mL cultures of BL21(DE3) cells expressing *Tm*Azul were grown according to expression conditions described in Section 3.2. The expression cultures were centrifuged for 20 min and 14,000*g*. The pellets (14 g) were resuspended in 50 mL lysis buffer (100 µM PLP, 50 mM KPi buffer) and lysed at 75 °C for >1 h. Heat-treated lysate was centrifuged 20 min, 14,000*g*, and the lysate decanted into a fresh container.

To a 500-mL Erlenmeyer flask azulene (7.8 mmol, 1.0 g) and serine (8.6 mmol, 0.90 g), 30 mL DMSO (20% v/v), 70 mL 50 mM KPi buffer, and 50-mL heat-treated lysate were added. The reaction was covered in aluminum foil, placed in 37 °C incubator, and shaken at 180 rpm. After 48 h, the reaction was washed with ethyl acetate to remove any remaining azulene, allowing crude starting material to be recovered from ethyl acetate by gently evaporating off the solvent under a constant stream of nitrogen. The aqueous layer was concentrated *in vacuo* and resulting in a dense paste comprised of DMSO, AzAla, and buffer salts. Ethyl acetate was added to remove DMSO and precipitate the amino acid product. The blue solid was filtered and washed with ethyl acetate. At this stage, AzAla is relatively pure by NMR and can be used for further applications. For highly pure AzAla, the blue solid was solubilized in 1 M HCl and immediately purified on a Biotage Isolera One purification system, using C-18 silica as the stationary phase, CH_3CN (0.1% acetic acid) as the strong solvent, and H_2O (0.1% acetic acid) as the weak solvent. AzAla should not spend long durations of time in strongly acidic conditions (such as 1 M HCl) as the product can decompose over time. The purified product fractions were combined and concentrated *in vacuo* to afford pure AzAla (blue solid, 0.965 g, 57% yield).

B.3 Variant Sequences

Variants identified through screening were DNA sequenced using Sanger Sequencing (Laragen) to determine their identities. The DNA sequences of all TrpB genes tested in this project are included here. All variants were cloned into a pET22b(+) vector as described above. T7 and T7-terminator primers were used to sequence all variants.

Table B-3. TrpB sequencing primers

Primer	Direction	Sequence
T7	Forward (5' to 3')	`taatacgactcactataggg`
T7-term	Reverse (3' to 5')	`ctagttattgctcagcggtg`

***Tm*TrpA$^{\mathrm{WT}}$:**

```
ATGAAAGGTTTTATCGCGTACATCCCGGCTGGTTTTCCGGATCTGGAAACCACCCGTAAAATTCTGATCGCACTGAACGAGCTGGGTATTA
CCGGTGTTGAAATTGGTGTCCCGTTCTCCGACCCGGTTGCGGATGGTCCGGTGATCCAACTGGCGCATAGCGTTGCTCTGCGTAACGGTGT
GACTATTAAAAAAATTCTGGAAATGCTGTCCGAGATTTCCGTAGATTACGACCTGTACCTGATGTCTTACCTGAACCCGATCGTTAATTAC
CCTGAAGGCAAAGAGAAACTGCTGGACGAACTGAAGAAGCTGGGCGTTAAAGGCCTGATTATCCCAGACCTGCCGCTGCGTGAAGTAAAAA
ACGTTGACATCGCTTACCCGATCGTTCCATTCGTTGCACCGAATACCAAAGACGAAGAGATCGACCTGATCAACTCCGTGCAGGCTCCGTT
CGTGTACTATATCTCTCGTTACGGTGTAACTGGTGAACGCGAAGACCTGCCGTTTGCAGATCACATCAAACGCGTGAAAGAACGTATCAAA
CTGCCACTGTTCGTCGGTTTCGGTATCTCCCGTCACGAACAAGTTAAAAAAGTTTGGGAAATCGCTGATGGTGTTATTGTTGGCAGCGCAC
TGGTCCGCATCATGGAAGAAAACCCGAAAGATGAGATCCCACGTAAAGTTGTTGAAAAAGTTAAAGAGCTGCTGGGCAAAtga
```

***Tm*TrpB$^{\mathrm{WT}}$:**

```
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGCCAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGAAA
TCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCT
GTCCGAAAAATACGGTGCTCGCATCTATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTT
CTGCTGGCAAAAAAAATGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGT
TCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACCGAACGTTGAACGTATGAAACTGCTGGGTGCTAAAGT
TGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTAACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTAC
```

```
GTGATCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTC
TGGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGGGTGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTC
TGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTAC
CTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACTGGGGTCAAGTTCAGGTGACGCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTG
TCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACT
GTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTG
GTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCtcgagcaccaccatc
accatcactga
```

***Pf*5G8:**

```
ATGTGGTTCGGTGAATTTGGTGGTCAGTACGTGCCAGAAACGCTGGTTGGACCCCTGAAAGAGCTGGAAAAAGCTTACAAACGTTTCAAAG
ATGACGAAGAATTCAATCGTCAGCTGAATTACTACCTGAAAACCTGGGCAGGTCGTCCAACCCCACTGTACTACGCAAAACGCCTGACTGA
AAAAATCGGTGGTGCTAAAGTCTACCTGAAACGTGAAGACCTGGTTCACGGTGGTGCACACAAGACCAACAACGCCATCGGTCAGGCACTG
CTGGCAAAGCTCATGGGTAAAACTCGTCTGATCGCTGAGACCGGTGCTGGTCAGCACGGCGTAGCGACTGCAATGGCTGGTGCACTGCTGG
GCATGAAAGTGGACATTTACATGGGTGCTGAGGACGTAGAACGTCAGAAATTGAACGTATTCCGTATGAAGCTGCTGGGTGCAAACGTAAT
TCCAGTTAACTCCGGTTCTCGCACCCTGAAAGACGCAATCGACGAGGCTCTGCGTGATTGGGTGGCTACTTTTGAATACACCCACTACCTA
ATCGGTTCCGTGGTCGGTCCACATCCGTATCCGACCATCGTTCGTGATTTTCAGTCTGTTATCGGTCGTGAGGCTAAAGCGCAGATCCCGG
AGGCTGAAGGTCAGCTGCCAGATGTAATCGTTGCTTGTGTTGGTGGTGGCTCTAACGCGATGGGTATCTTTTACCCGTTCGTGAACGACAA
AAAAGTTAAGCTGGTTGGCGTTGAGGCTGGTGGTAAAGGCCTGGAATCTGGTAAGCATTCCGCTAGCCTGAACGCAGGTCAGGTTGGTGTG
TCCCATGGCATGCTGTCCTACTTTCTGCAGGACGAAGAAGGTCAGATCAAACCAAGCCACTCCATCGCACCAGGTCTGGATTATCCAGGTG
TTGGTCCAGAACACGCTTACCTGAAAAAAATTCAGCGTGCTGAATACGTGGCTGTAACCGATGAAGAAGCACTGAAAGCGTTCCATGAACT
GAGCCGTACCGAAGGTATCATCCCAGCTCTGGAATCTGCGCATGCTGTGGCTTACGCTATGAAACTGGCTAAGGAAATGTCTCGTGATGAG
ATCATCATCGTAAACCTGTCTGGTCGTGGTGACAAAGACCTGGATATTGTCCTGAAAGCGTCTGGCAACGTGCtcgagcaccaccaccacc
accactga
```

***Pf*5G8 E104G**

```
ATGTGGTTCGGTGAATTTGGTGGTCAGTACGTGCCAGAAACGCTGGTTGGACCCCTGAAAGAGCTGGAAAAAGCTTACAAACGTTTCAAAG
ATGACGAAGAATTCAATCGTCAGCTGAATTACTACCTGAAAACCTGGGCAGGTCGTCCAACCCCACTGTACTACGCAAAACGCCTGACTGA
AAAAATCGGTGGTGCTAAAGTCTACCTGAAACGTGAAGACCTGGTTCACGGTGGTGCACACAAGACCAACAACGCCATCGGTCAGGCACTG
CTGGCAAAGCTCATGGGTAAAACTCGTCTGATCGCTGGGACCGGTGCTGGTCAGCACGGCGTAGCGACTGCAATGGCTGGTGCACTGCTGG
GCATGAAAGTGGACATTTACATGGGTGCTGAGGACGTAGAACGTCAGAAATTGAACGTATTCCGTATGAAGCTGCTGGGTGCAAACGTAAT
TCCAGTTAACTCCGGTTCTCGCACCCTGAAAGACGCAATCGACGAGGCTCTGCGTGATTGGGTGGCTACTTTTGAATACACCCACTACCTA
ATCGGTTCCGTGGTCGGTCCACATCCGTATCCGACCATCGTTCGTGATTTTCAGTCTGTTATCGGTCGTGAGGCTAAAGCGCAGATCCCGG
AGGCTGAAGGTCAGCTGCCAGATGTAATCGTTGCTTGTGTTGGTGGTGGCTCTAACGCGATGGGTATCTTTTACCCGTTCGTGAACGACAA
AAAAGTTAAGCTGGTTGGCGTTGAGGCTGGTGGTAAAGGCCTGGAATCTGGTAAGCATTCCGCTAGCCTGAACGCAGGTCAGGTTGGTGTG
TCCCATGGCATGCTGTCCTACTTTCTGCAGGACGAAGAAGGTCAGATCAAACCAAGCCACTCCATCGCACCAGGTCTGGATTATCCAGGTG
TTGGTCCAGAACACGCTTACCTGAAAAAAATTCAGCGTGCTGAATACGTGGCTGTAACCGATGAAGAAGCACTGAAAGCGTTCCATGAACT
GAGCCGTACCGAAGGTATCATCCCAGCTCTGGAATCTGCGCATGCTGTGGCTTACGCTATGAAACTGGCTAAGGAAATGTCTCGTGATGAG
ATCATCATCGTAAACCTGTCTGGTCGTGGTGACAAAGACCTGGATATTGTCCTGAAAGCGTCTGGCAACGTGCtcgagcaccaccaccacc
accactga
```

***Tm*9D8*:**

```
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGGAA
TCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCT
GTCCGAAAAATACGGTGCTCGCGTATATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTT
CTGCTGGCAAAACTAATGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGT
TCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTTGAACGTATGAAACTGCTGGGTGCTAAAGT
TGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTAC
GTGTTCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTC
CAGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTC
TGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTAC
CTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTG
TCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACT
GTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTG
GTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCtcgagcaccaccacc
accaccactga
```

***Tm*9D8* E105G:**

```
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGGAA
TCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCT
GTCCGAAAAATACGGTGCTCGCGTATATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTT
CTGCTGGCAAAACTAATGGGCAAAACCCGTATCATTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGT
TCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTTGAACGTATGAAACTGCTGGGTGCTAAAGT
TGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTAC
```

```
GTGTTCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTC
CAGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTC
TGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTAC
CTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTG
TCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACT
GTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTG
GTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCtcgagcaccaccacc
accaccactga
```

***Tm*Azul:**

```
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACTGGAAGCTGCGTACGAAGGAA
TCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATTATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCT
GTCCGAAAAATACGGTGCTCGCGTATATCTGAAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTT
CTGCTGGCAAAACTAATGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAGCAGCGCTGT
TCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTTGAACGTATGAAACTGCTGGGTGCTAAAGT
TGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAATTGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTAC
GTGTCCGGCTCTGTGGTTGGTCCGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTC
CAGAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTTCTATCCGTTTATCGATTC
TGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTAAACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTAC
CTGCACGGTTCTAAGACGTTCGTTCTGCAGGATGACAGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTG
TCGGTCCGGAACACGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCATCGAACT
GTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAGATCAACATCAAGGGTAAAGTTGTGGTG
GTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACTGAACCACCCGTATGTTCGCGAACGCATCCGCCtcgagcaccaccacc
accaccactga
```

B.4 Structural Modeling

A crystal structure of a homologous TrpS complex (*Salmonella typhimurium*, *St*TrpS) has been reported with the amino-acrylate in the β-subunit stabilized in the active site by benzimidazole, a competitive inhibitor (PDB: 4HPX).[4] This TrpB structure served as the template for construction of a homology model of *Tm*9D8* (57% sequence identity) using the SWISS-MODEL program.[5] The alignment of *Tm*9D8* homology model and the *St*TrpB structure allowed benzimidazole and the amino-acrylate to be placed within the *Tm*9D8* structure in PyMOL, and the five-membered ring of azulene was aligned directly to the five-membered ring of benzimidazole to simulate a productive binding pose.

B.5 Characterization of AzAla

The enzymatic AzAla product was characterized by NMR, HRMS, and chiral derivatization. The spectra were taken in a mixture of D_2O/5% DCl to increase solubility, resulting in deuterium exchange of one of the protons on the five-membered rings. ^1H NMR (400 MHz, D_2O) δ 7.85 (dd, *J* = 10.9, 9.6 Hz, 2H), 7.29 (s, 1H), 7.16 (t, *J* = 9.9 Hz, 1H), 6.72 (dt, *J* =

11.7, 9.8 Hz, 2H), 3.89 (dd, *J* = 7.1, 5.8 Hz, 1H), 3.23 (dd, *J* = 15.3, 5.9 Hz, 1H), 3.15 (dd, *J* = 15.3, 7.1 Hz, 1H). ^{13}C NMR (100 MHz, D_2O) δ 172.35, 140.79, 138.75, 137.59, 137.54, 137.48, 136.66, 133.71, 123.74, 123.09, 120.83, 54.63, 27.64. HRMS (*m/z*) for $[M+H]^+$ $C_{13}H_{14}NO_2$ requires 216.1019, observed 216.1019.

Enantiopurity was determined by derivatization with enantiopure and racemic FDNP-alanamide. In a 2-mL vial, AzAla (0.5 μmol) was dissolved in 1 M aq. $NaHCO_3$ (100 μL), to which 10 μL of a 33-mM FDNP-alanamide solution in acetone (0.33 μmol) was added. The vial was shaken for 2 h at 230 rpm, 37 °C. The reaction was allowed to cool to room temperature, then diluted with 1:1 CH_3CN/1 M aq. HCl (600 μL). The solution was analyzed via LC-MS (40% to 95% CH_3CN, monitored by using total ion count filtered for the expected mass of 468). Absolute stereochemistry for AzAla was inferred by analogy to L-tryptophan and determined to have >99% enantiomeric excess.

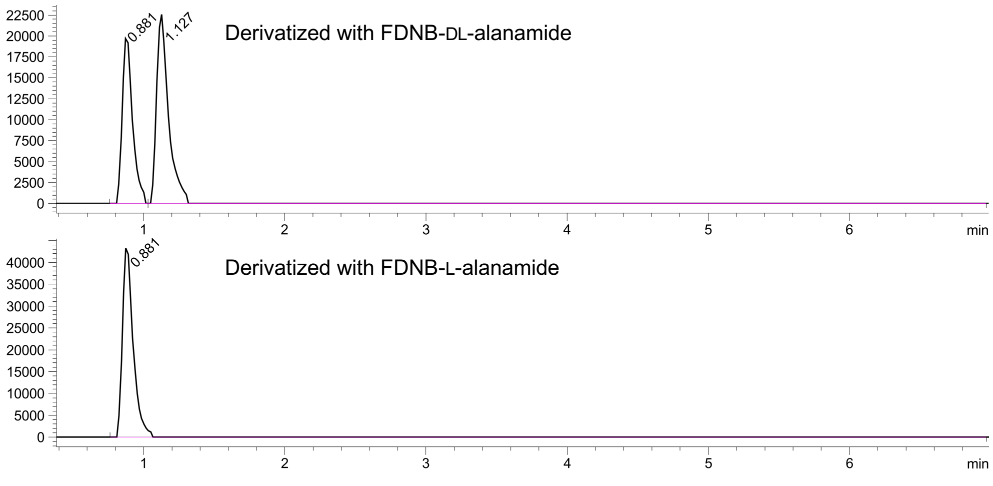


Figure B-5. HPLC traces of AzAla enantiopurity experiment.

B.6 NMR Spectra

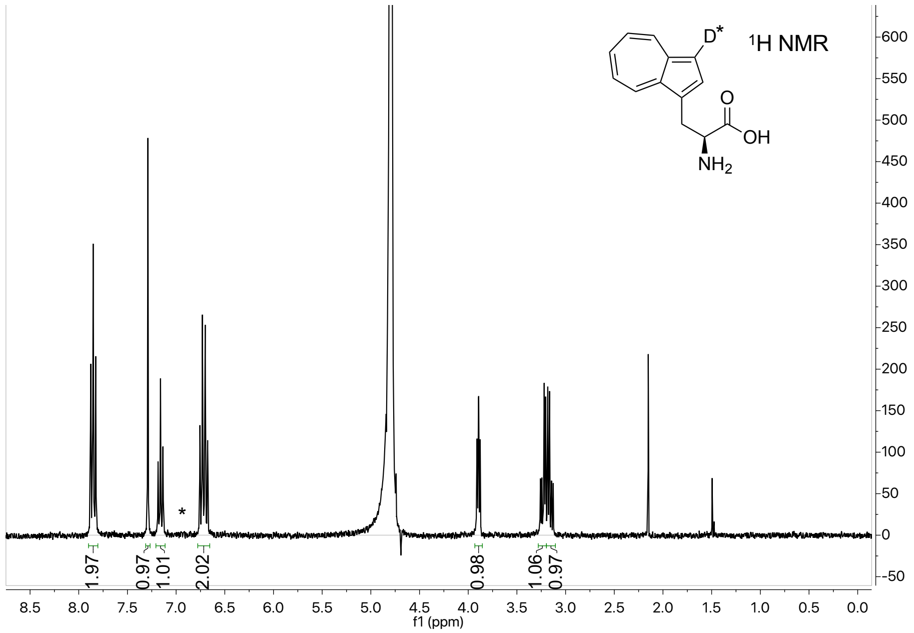


Figure B-6. ^{1}H NMR of pure AzAla. *Acidic conditions in deuterated solvents result in the exchange of one of the Cp^{-} protons (6.9 ppm) for deuterium.

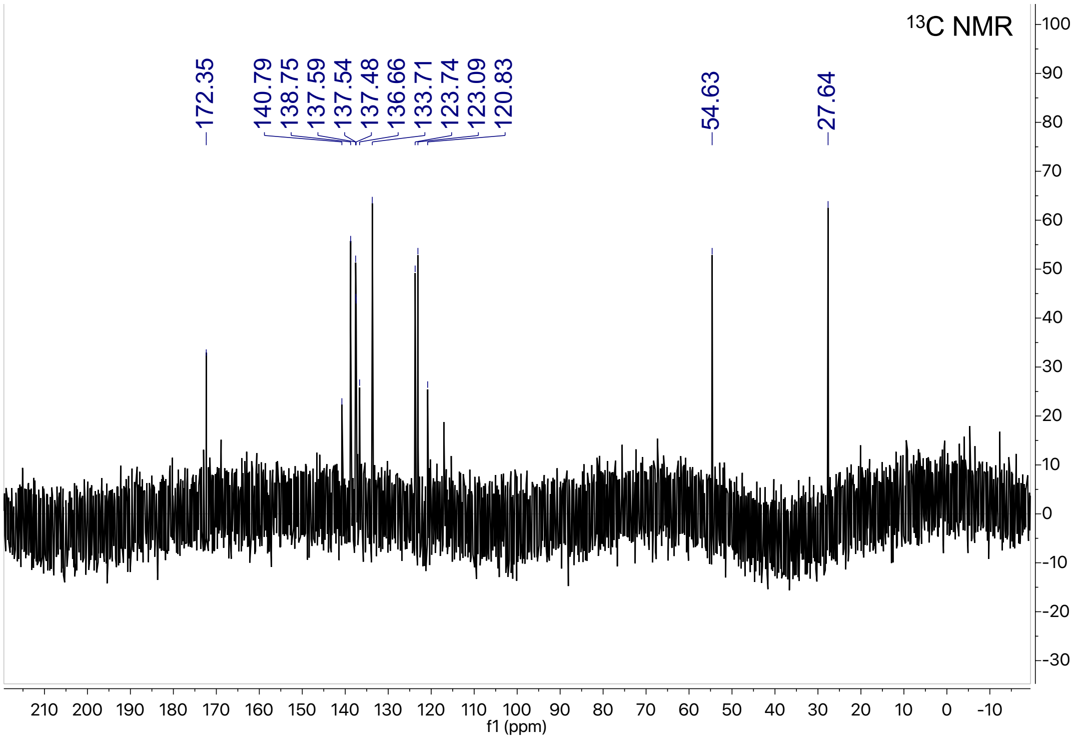


Figure B-7. ^{13}C NMR of pure AzAla.

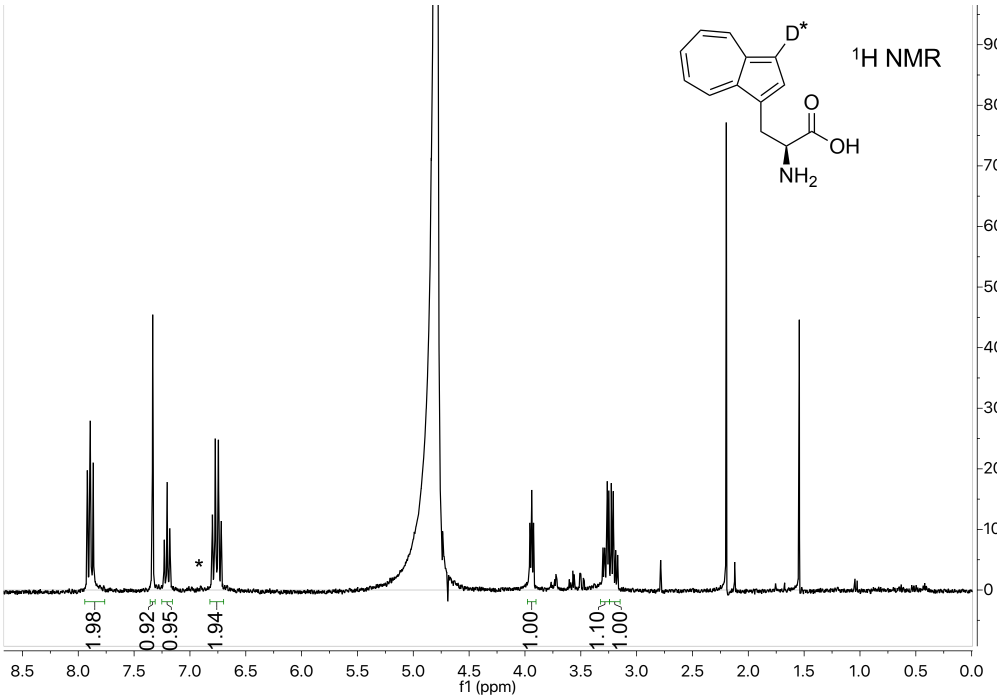


Figure B-8. ^{1}H NMR of crude AzAla. *Acidic conditions in deuterated solvents result in the exchange of one of the Cp$^-$ protons (6.9 ppm) for deuterium.

Bibliography for Appendix B

1. Boville, C. E. *et al*. Engineered biosynthesis of β-alkyl tryptophan analogs. *Angew. Chem. Int. Ed.* (2018).
2. Gibson, D. G. *et al*. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
3. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc* **139**, 10769–10776 (2017).
4. Niks, D. *et al*. Allostery and substrate channeling in the tryptophan synthase bienzyme complex: Evidence for two subunit conformations and four quaternary states. *Biochemistry* **52**, 6396–6411 (2013).
5. Waterhouse, A. *et al*. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018)

Appendix C
SUPPLEMENTARY INFORMATION FOR CHAPTER IV

C.1 Supplementary Tables

Table C-1. Summary of all cultures passaged for evolution of *Tm*TrpB variants.

TrpB variant (strain)	**culture volume (mL)**	**total number passaged**	**total number successful**
wt *Tm*TrpB (GR-Y053)	3	4	1
	100	2	2
***Tm*Triple (GR-Y055)**	3	8	2
	100	4	4
***Tm*TripleQ90* (GR-Y057)**	3	8	1
	100	0	0

Table C-2. Mutation summary statistics for OrthoRep-evolved TrpB populations.

		non-synonymous mutations			**synonymous mutations**	
variant set	total number of sequences	total unique mutations	mean	standard deviation	mean	standard deviation
consensus	10	43	5.6	2.3	3.0	2.3
1	16	85	8.7	2.1	6.3	4.4
2	60	194	9.3	2.8	6.5	3.0

Table C-3. Mutations and identification information for all individual *Tm*TrpB sequences.

variant set	**variant name**	**number of non-synonymous mutations**	**number of synonymous mutations**	**starting sequence**
1	WT-100-1-A	13	3	*Tm*TrpB
1	WT-100-2-A	9	7	*Tm*TrpB
1	WT-003-1-A	7	1	*Tm*TrpB
1	Q90*-003-1-A	6	11	*Tm*TripleQ90*
1	Q90*-003-1-B	11	15	*Tm*TripleQ90*
1	Tri-003-1-A	9	3	*Tm*Triple
1	Tri-003-1-B	11	4	*Tm*Triple
1	Tri-003-1-C	9	3	*Tm*Triple
1	Tri-003-2-A	9	6	*Tm*Triple
1	Tri-100-1-A	10	5	*Tm*Triple
1	Tri-100-2-A	9	7	*Tm*Triple
1	Tri-100-2-B	8	6	*Tm*Triple
1	Tri-100-3-A	5	3	*Tm*Triple

1	Tri-100-4-A	10	1	*Tm*Triple
1	Tri-100-4-B	6	14	*Tm*Triple
1	Tri-100-4-C	7	11	*Tm*Triple
2	WT-100-1-B	14	9	WT *Tm*TrpB
2	WT-100-1-C	12	7	WT *Tm*TrpB
2	WT-100-1-D	12	9	WT *Tm*TrpB
2	WT-100-1-E	14	10	WT *Tm*TrpB
2	WT-100-1-F	16	11	WT *Tm*TrpB
2	WT-100-2-B	7	9	WT *Tm*TrpB
2	WT-100-2-C	7	5	WT *Tm*TrpB
2	WT-100-2-D	9	8	WT *Tm*TrpB
2	WT-100-2-E	8	9	WT *Tm*TrpB
2	WT-100-2-F	9	10	WT *Tm*TrpB
2	WT-100-2-G	7	6	WT *Tm*TrpB
2	WT-100-2-H	9	9	WT *Tm*TrpB
2	WT-100-2-I	10	5	WT *Tm*TrpB
2	WT-003-1-B	6	2	WT *Tm*TrpB
2	WT-003-1-C	7	5	WT *Tm*TrpB
2	WT-003-1-D	7	7	WT *Tm*TrpB
2	WT-003-1-E	10	8	WT *Tm*TrpB
2	WT-003-1-F	8	4	WT *Tm*TrpB
2	WT-003-1-G	9	4	WT *Tm*TrpB
2	WT-003-1-H	8	4	WT *Tm*TrpB
2	WT-003-1-I	8	6	WT *Tm*TrpB
2	Q90*-003-1-C	10	6	*Tm*TripleQ90*
2	Q90*-003-1-D	11	9	*Tm*TripleQ90*
2	Q90*-003-1-E	14	8	*Tm*TripleQ90*
2	Q90*-003-1-F	14	8	*Tm*TripleQ90*
2	Q90*-003-1-G	16	6	*Tm*TripleQ90*
2	Q90*-003-1-H	12	9	*Tm*TripleQ90*
2	Tri-003-1-D	9	4	*Tm*Triple
2	Tri-003-1-E	7	5	*Tm*Triple
2	Tri-003-1-F	12	3	*Tm*Triple
2	Tri-003-1-G	10	5	*Tm*Triple
2	Tri-003-1-H	15	6	*Tm*Triple
2	Tri-003-1-I	6	4	*Tm*Triple
2	Tri-003-2-B	8	11	*Tm*Triple

2	Tri-003-2-C	13	4	*Tm*Triple
2	Tri-003-2-D	12	8	*Tm*Triple
2	Tri-003-2-E	10	3	*Tm*Triple
2	Tri-100-1-B	5	5	*Tm*Triple
2	Tri-100-1-C	6	5	*Tm*Triple
2	Tri-100-1-D	9	6	*Tm*Triple
2	Tri-100-1-E	10	8	*Tm*Triple
2	Tri-100-1-F	7	2	*Tm*Triple
2	Tri-100-1-G	6	5	*Tm*Triple
2	Tri-100-1-H	8	4	*Tm*Triple
2	Tri-100-2-C	8	5	*Tm*Triple
2	Tri-100-2-D	10	15	*Tm*Triple
2	Tri-100-2-E	13	10	*Tm*Triple
2	Tri-100-2-F	9	10	*Tm*Triple
2	Tri-100-2-G	9	15	*Tm*Triple
2	Tri-100-2-H	10	10	*Tm*Triple
2	Tri-100-2-I	13	4	*Tm*Triple
2	Tri-100-3-B	7	3	*Tm*Triple
2	Tri-100-3-C	6	3	*Tm*Triple
2	Tri-100-3-D	8	4	*Tm*Triple
2	Tri-100-3-E	6	2	*Tm*Triple
2	Tri-100-3-F	5	3	*Tm*Triple
2	Tri-100-4-D	8	8	*Tm*Triple
2	Tri-100-4-E	7	7	*Tm*Triple
2	Tri-100-4-F	6	6	*Tm*Triple
2	Tri-100-4-G	6	5	*Tm*Triple

Note that variant names are abbreviations of the evolution experiment from which the variant was harvested. For example, the variant name WT-100-1-A refers to the arbitrarily designated unique clone A that came from replicate 1 of the 100 mL evolution experiment that started from wt *Tm*TrpB. Likewise, Q90*-003-1-B refers to the arbitrarily designated unique clone B that came from replicate 1 of the 3 mL evolution experiment that started from *Tm*TripleQ90*. Likewise, Tri-003-2-A refers to the arbitrarily designated unique clone A that came from replicate 2 of the 3 mL evolution experiment that started from *Tm*Triple.

Table C-4. Kinetic parameters of selected *Tm*TrpB variants at 30 °C.

variant	k_{cat} [95% credible region] (s^{-1})	K_M [95% credible region] (µM)	k_{cat}/K_M [95% credible region] (mM^{-1} s^{-1})
***Tm*Triple**	0.2 [0.16, 0.31]	41.23 [14.32, 192.66]	4.89 [1.54, 12.12]
WT-003-1-A	0.53 [0.49, 0.58]	3.89 [1.85, 7.99]	137.22 [70.29, 276.04]
Q90*-003-1-A	0.77 [0.72, 0.83]	5.79 [3.82, 8.8]	133.38 [91.24, 193.71]
Tri-100-2-A	0.62 [0.59, 0.66]	5.58 [3.99, 7.91]	111.89 [81.52, 152.25]

C.2 Supplementary Figures

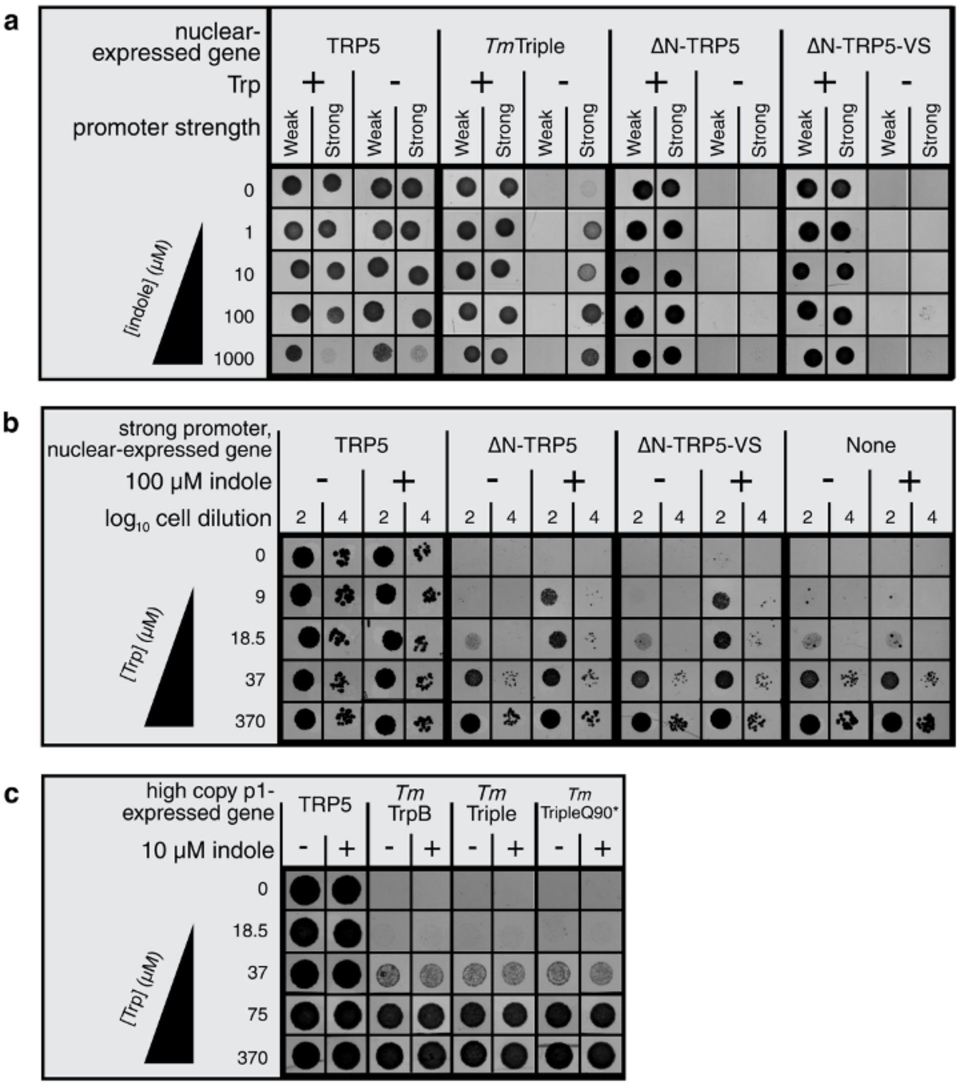


Figure C-1. Evaluation of indole-dependent TRP5 complementation of TrpB variants. a-c, Spot plating assays for Δ*trp5* yeast strains expressing TrpB variants from a nuclear plasmid under two different promoter strengths from **a.** a nuclear plasmid under a strong promoter or **b.** p1 at a high copy number (wt TP-DNAP1 expressed *in trans*) or **c.** grown on indicated growth medium. ΔN-TRP5, N-terminally truncated yeast TRP5 constituting only the region of TRP5 homologous to TmTrpB. ΔN-TRP5-VS, ΔN-TRP5 with two of the three *Tm*Triple mutations relative to wt TmTrpB. The markedly reduced growth at 1000 μM indole only when TRP5 is expressed by a strong promoter may be explained by indole toxicity induced by additional indole production by TRP5.

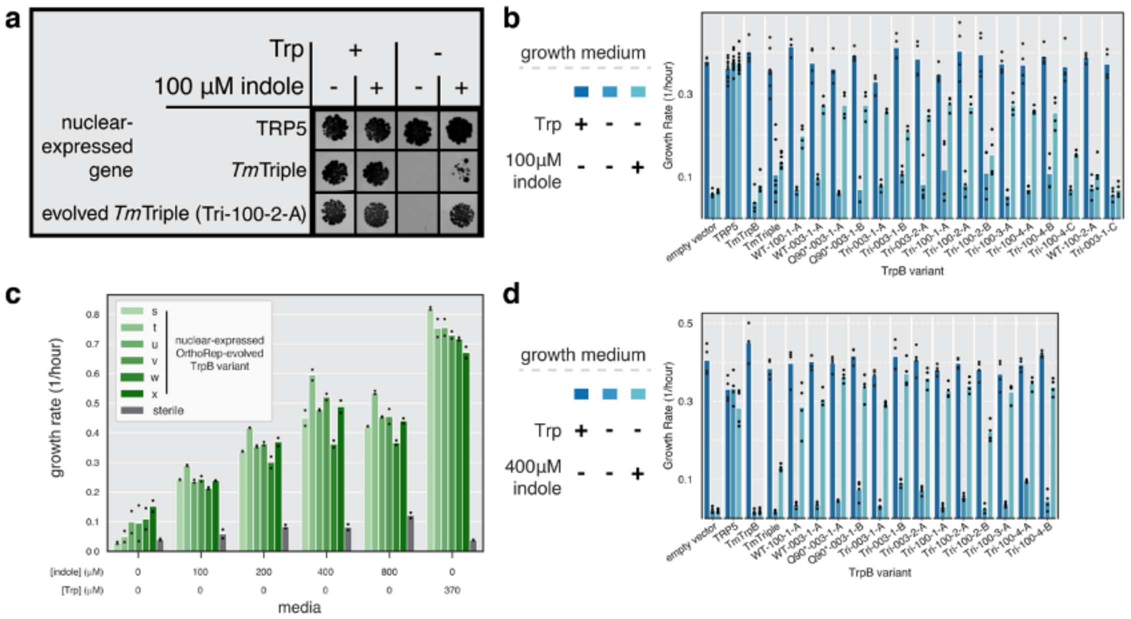


Figure C-2. *In vivo* Trp production by evolved TrpBs. a. Spot plating assay for TRP5-deleted yeast with a nuclear plasmid expressing TRP5, TmTriple, or an individual OrthoRep-evolved *Tm*TrpB variant driven by a promoter (pRNR2) that approximates expression of *Tm*TrpB variants from OrthoRep's p1 plasmid, grown on indicated media. **b-d.** Evaluation of TRP5 complementation by evolved variants through a growth rate assay. Maximum growth rates during exponential growth phase (when rate is above ~0.15 per hour) over a 24-hour period for Δtrp5 yeast strains transformed with a nuclear plasmid expressing the indicated TmTrpB variant, grown in the indicated growth medium. Points represent growth rate for individual replicates, bars represent the mean growth rate for all replicates. Note that growth rates below ~0.15 per hour correspond to cultures that did not enter exponential phase; in these cases, the reported growth rate is not meaningful and instead can be interpreted as no quantifiable growth. **b.** Low indole growth rate test. TRP5 tested in n=13 biologically independent replicates over 5 independent experiments. *Tm*Triple tested in n=8 biologically independent replicates over 4 independent experiments. All other variants tested in n=4 biologically independent replicates in a single experiment. **c.** Optimization of indole concentration. All conditions tested in technical duplicate, although exact sequences of TmTrpB variants were not determined, as the sole purpose of these variants was to evaluate the effect of indole concentration. **d.** High indole growth rate test. A subset of this data is shown in **Figure 4-2a**. All variants tested in n=4 biologically independent replicates in a single experiment.

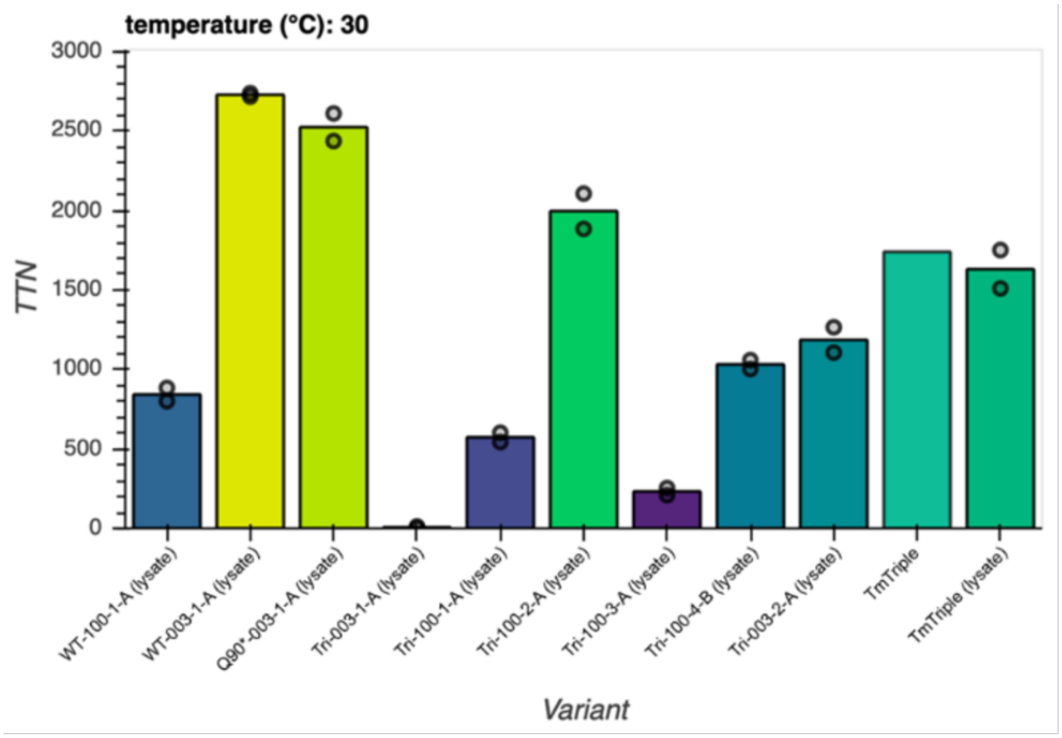


Figure C-3. ***In vitro*** **Trp production by evolved TrpBs with heat-treated lysate.** Trp production at 30 °C by indicated TmTrpB variants. Reactions with *Tm*Triple were performed with both heat treated lysate (1 hour incubation at 75 °C) and with purified protein, while all other reactions were performed only with heat treated lysate. TTN, total turnover number. Maximum TTN is 5,000. Points represent TTN for individual biological replicates, bars represent mean TTN for reactions with replicates, or TTN for a single replicate otherwise. Reaction with purified *Tm*Triple performed in a single replicate, and all lysate reactions performed in n=2 biologically independent replicates.

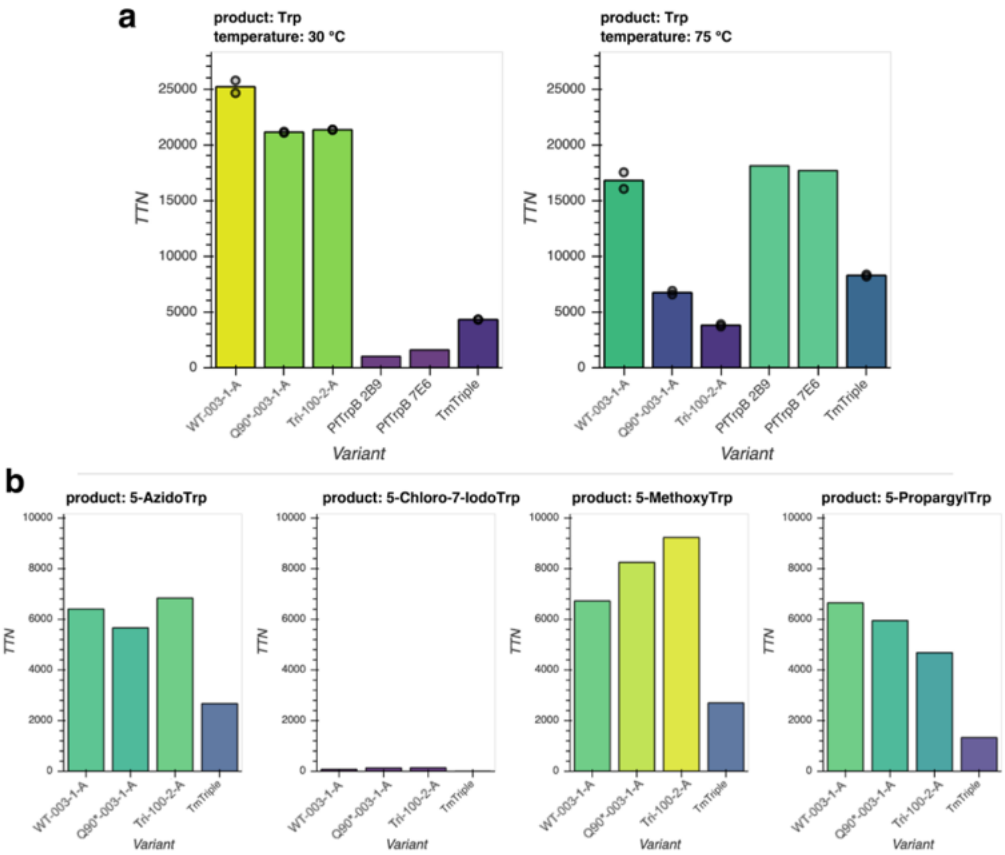


Figure C-4. *In vitro* Trp and Trp analog production with purified enzyme. Production of: **a.** Trp at 30 °C or 75 °C with 40,000 maximum TTN, or **b.** indicated Trp analogs at 30 °C by column-purified TmTrpB variants. TTN, total turnover number with 10,000 as maximum TTN. Points represent TTN for individual replicates, bars represent mean TTN for reactions with replicates, or TTN for a single replicate otherwise.

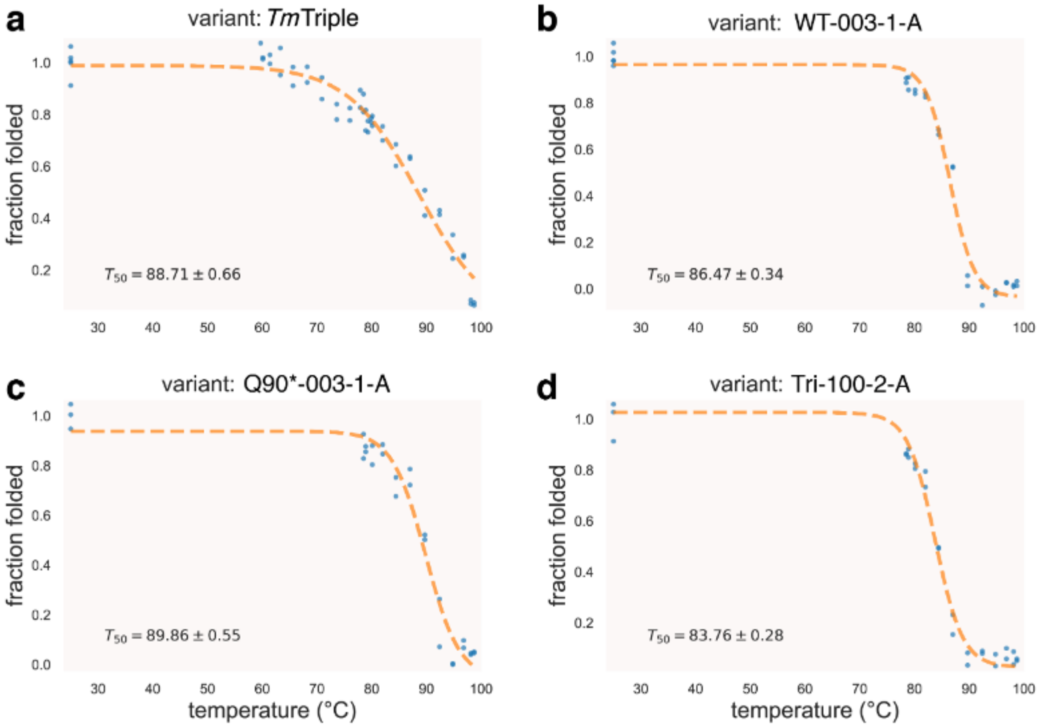


Figure C-5. Thermal shift assay on various *Tm*TrpBs. Proportion of *Tm*TrpB variants *Tm*Triple (**a**), WT-003-1-A (**b**), Q90*-003-1-A (**c**), or Tri-100-2-A (**d**) that remain folded after incubation at indicated temperature for 1 hour, as measured by the fraction of Trp production relative to incubation at 25 °C. T_{50}, temperature at which 50% of enzyme is irreversibly inactivated, as estimated by best fit logistic model (dotted line). Each temperature was tested in technical duplicate.

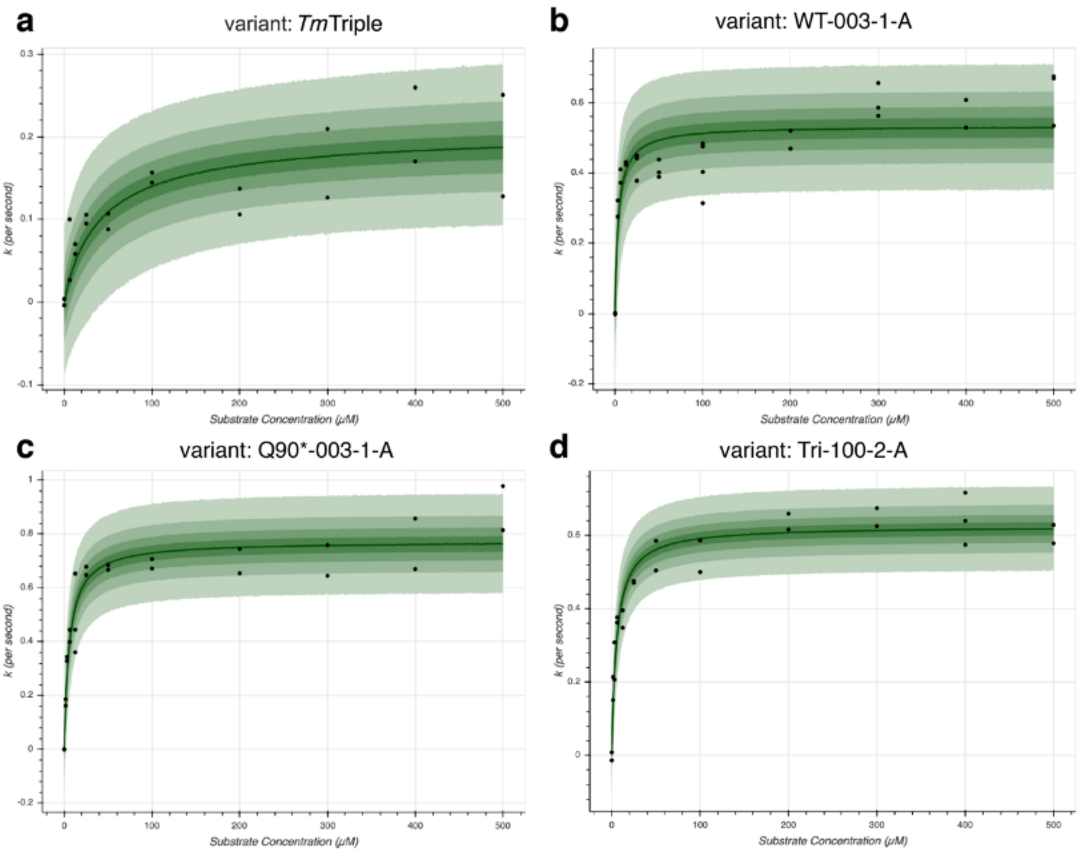


Figure C-6. Michaelis-Menten plots for rate of Trp production at saturating serine for evolved *Tm*TrpB variants. Initial rate of Trp formation (*k*, per second) with TrpB variants TmTriple (**a**), WT-003-1-A (**b**), Q90*-003-1-A (**c**), or Tri-100-2-A (**d**) at saturating L-serine concentration (40 mM) vs. indole concentration. Points, median estimates for initial rate based on absorbance change over time (see Methods). For all four variants, the median estimated Michaelis-Menten curve is shown as a dark green line, with the 25, 50, 75, and 95% credible regions displayed from dark to light green, respectively. Measurements performed in n=3 technical replicates for all substrate concentrations for WT-003-1-A and 400 μM for Tri-100-2-A, and n=2 technical replicates for all other substrate concentrations and samples.

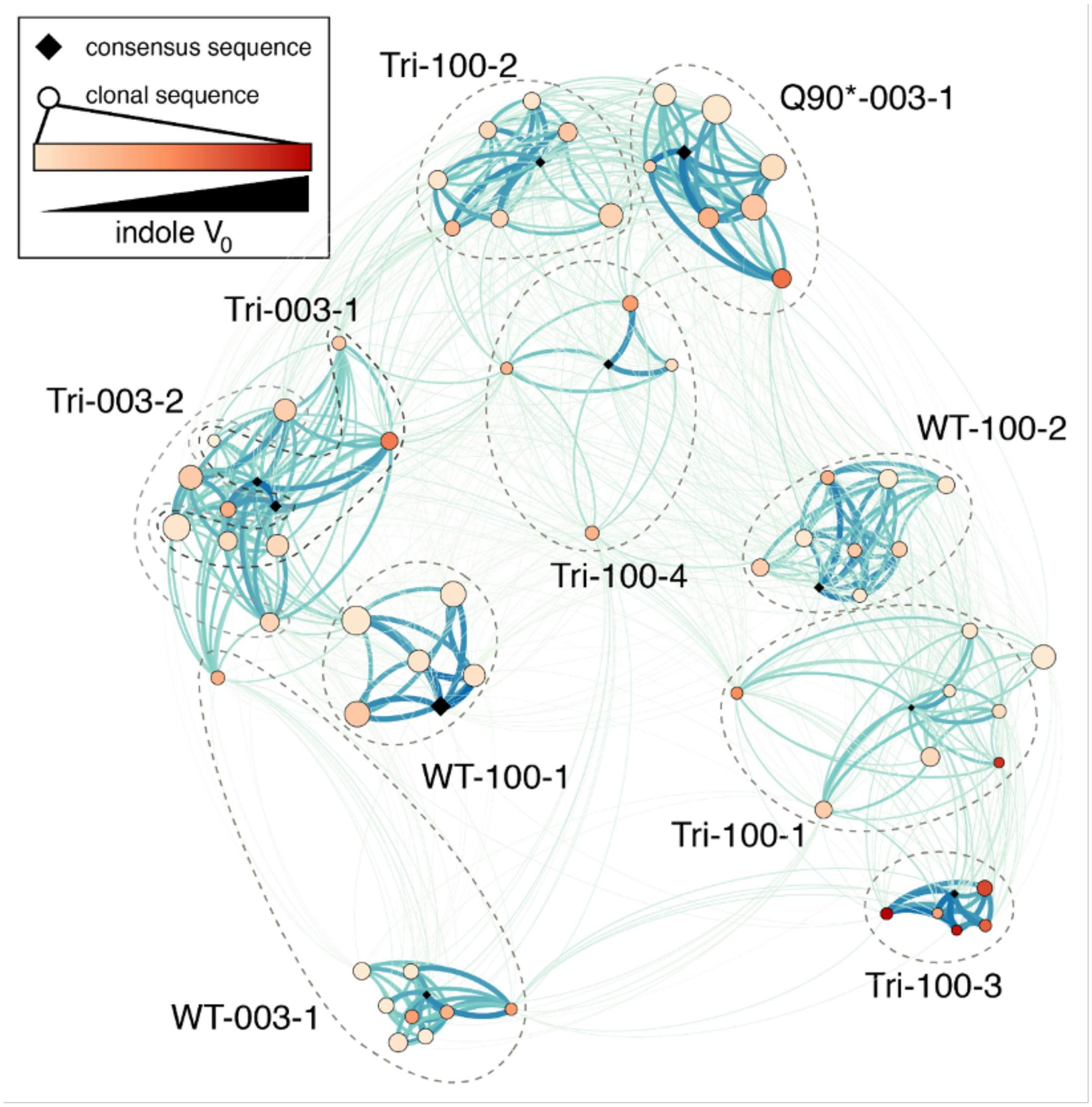


Figure C-7. Relatedness of TrpB panel sequences generated by OrthoRep evolution. Force directed graph where each node represents an individual sequence (all variants from set 2, and variants WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A) or consensus sequence for one of the ten evolved populations. Edge weights are proportional to the number of shared mutations between two nodes. Higher edge weight yields a stronger attractive force between two nodes and is visualized as a darker color and a thicker line. Nodes for individual sequences are colored according to initial rate of Trp formation, similar to **Figure 4-5b**. Dotted lines are drawn around consensus sequences and individual sequences that are derived from the same evolved culture, if nodes are sufficiently clustered to allow it.

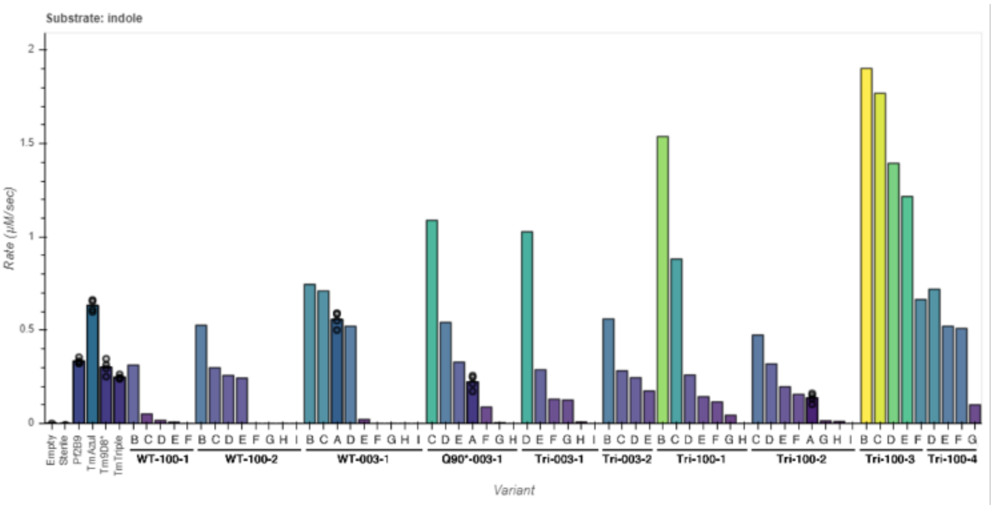
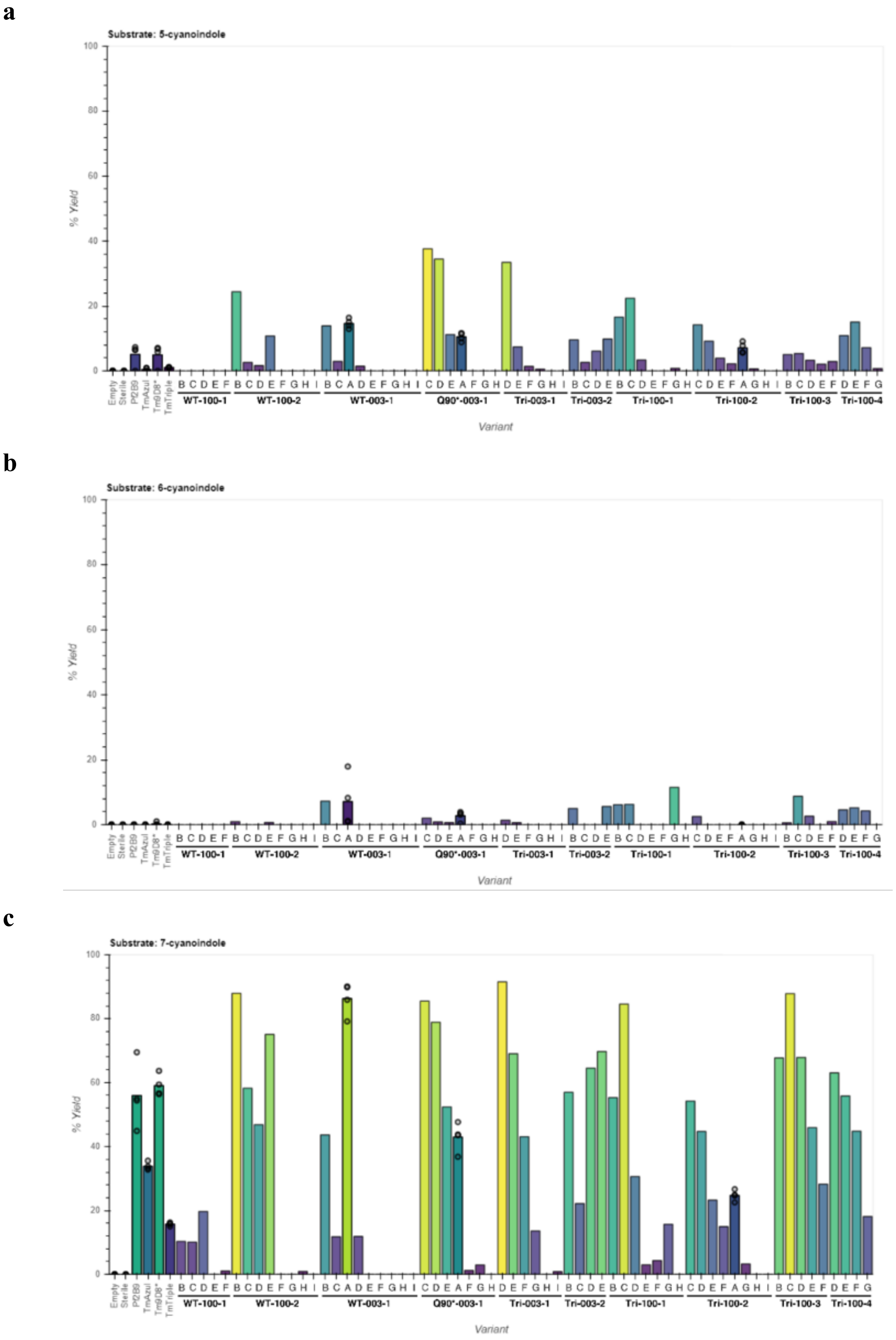
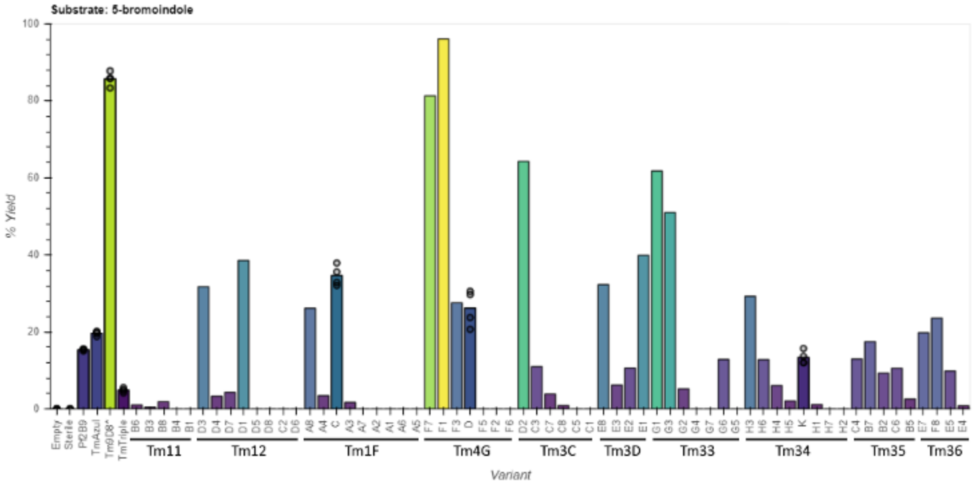


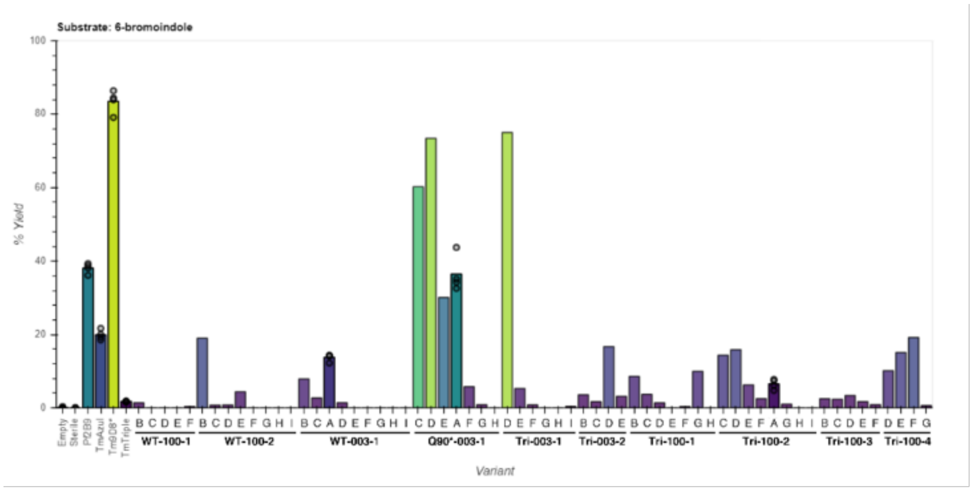
Figure C-8. TrpB panel indole activity by initial rate of Trp formation. Initial rate of Trp formation at saturating L-serine by UV-vis spectrophotometry. Points represent rate for individual replicates, bars represent mean rate for reactions with multiple replicates, or rate for a single replicate otherwise. OrthoRep-evolved variants are ordered first by the population from which they were derived, then by indole activity. Empty, expression vector without TrpB variant. Sterile, reaction master mix without heat-treated lysate added. Empty, *Pf*2B9, *Tm*Azul, *Tm*9D8*, *Tm*Triple, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A all performed in n=4; empty performed in n=3 biological replicate; sterile performed in n=2 technical replicates; all other reactions performed in a single replicate. (Data are identical to those shown in **Figure 4-5b**.)


a
Substrate: 5-cyanoindole
% Yield
100
80
60
40
20
0
Empty
Sterile
P2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant
b
Substrate: 6-cyanoindole
% Yield
100
80
60
40
20
0
Empty
Sterile
P2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant
c
Substrate: 7-cyanoindole
% Yield
100
80
60
40
20
0
Empty
Sterile
P2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant

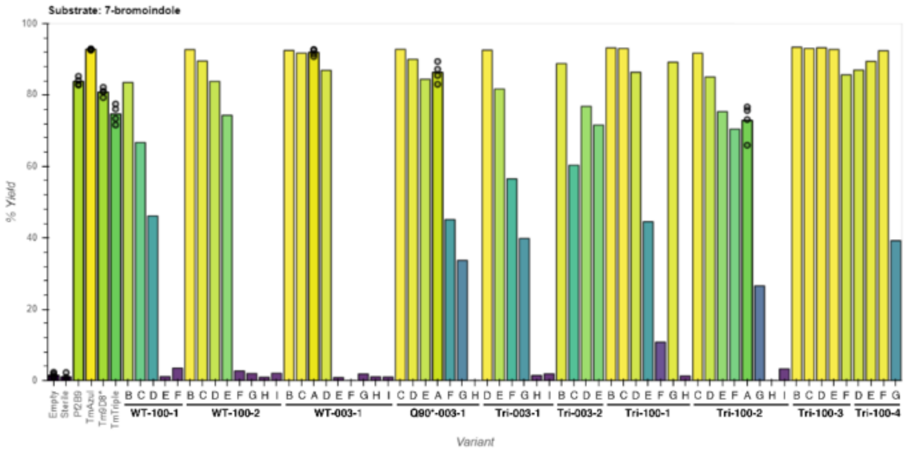
d


Substrate: 5-bromoindole
100
80
60
40
20
0
% Yield
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
B6 B3 B8 B4 B1
Tm11
D3 D4 D7 D1 D5 D8 C2 D6
Tm12
A8 A4 C A3 A7 A2 A1 A6 A5
Tm1F
F7 F1 F3 D F5 F2 F6
Tm4G
D2 C3 C7 C8 C5 C1
Tm3C
E8 E3 E2 E1
Tm3D
G1 G3 G2 G4 G7 G6 G5
Tm33
H3 H6 H4 H5 K H1 H7 H2
Tm34
C4 B7 B2 C6 B5
Tm35
E7 F8 E5 E4
Tm36
Variant

e


Substrate: 6-bromoindole
100
80
60
40
20
0
% Yield
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant

f


Substrate: 7-bromoindole
100
80
60
40
20
0
% Yield
Empty
Sterile
Pf2B9
TmAzul
Tm9D8*
TmTriple
B C D E F
WT-100-1
B C D E F G H I
WT-100-2
B C A D E F G H I
WT-003-1
C D E A F G H
Q90*-003-1
D E F G H I
Tri-003-1
B C D E
Tri-003-2
B C D E F G H
Tri-100-1
C D E F A G H I
Tri-100-2
B C D E F
Tri-100-3
D E F G
Tri-100-4
Variant

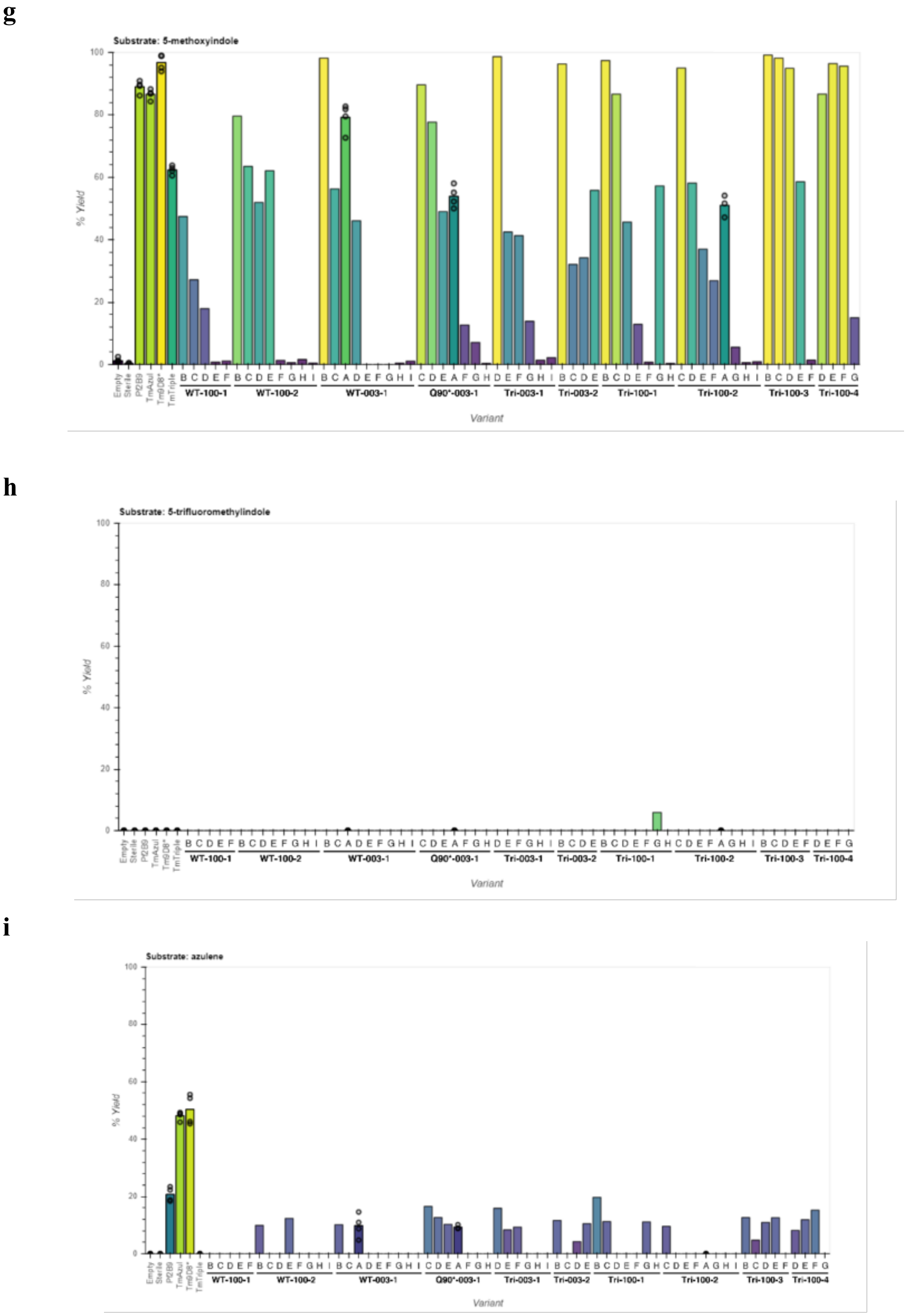


Figure C-9. TrpB panel activity with indole analogs by HPLC yield. a-i. HPLC yield of (**a**) 5-cyanoTrp, (**b**) 6-cyanoTrp, (**c**) 7-cyanoTrp, (**d**) 5-bromoTrp, (**e**) 6-bromoTrp, (**f**) 7-bromoTrp, (**g**)

5-methoxyTrp, (**h**) 5-trifluoromethylTrp, and (**i**) β-(1-azulenyl)-L-alanine for indicated variants supplied with L-serine and each indole substrate. Points represent % yield for individual replicates, bars represent mean % yield for reactions with replicates, or % yield for a single replicate otherwise. Replicates and variant order are as in Figure C-8, and populations from which OrthoRep-evolved variants were derived are annotated. Empty, *Pf*2B9, *Tm*Azul, *Tm*9D8*, *Tm*Triple, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A performed in n=4 biologically independent replicates; sterile performed in n=2 technical replicates; all other reactions performed a single replicate. (Data are identical to those shown in **Figure 4-5b**.)

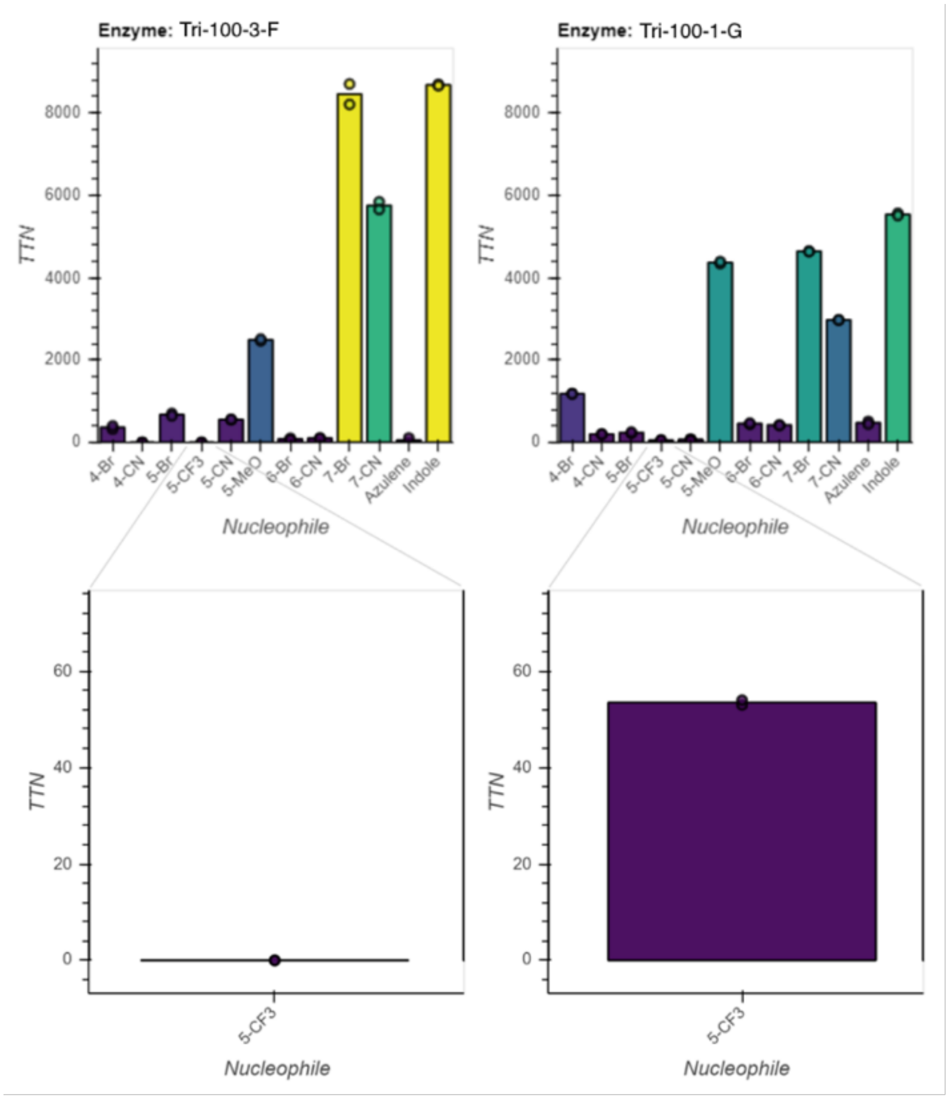


Figure C-10. Substrate activity profiles for large scale purification of variants Tri-100-3-F and Tri-100-1-G. Total turnover number (TTN) for *Tm*TrpB variants Tri-100-3-F and Tri-100-1-G purified at large scale (see **Section B.3**, *Methods*) and supplied with L-serine and the indicated indole analog, azulene, or indole (nucleophile), with a maximum TTN of 10,000. All reactions were performed in duplicate. Points represent TTN for individual replicates, while bars represent mean for two replicates. Insets, TTN for 5-trifluoromethylindole with y-axis scale adjusted for clarity.

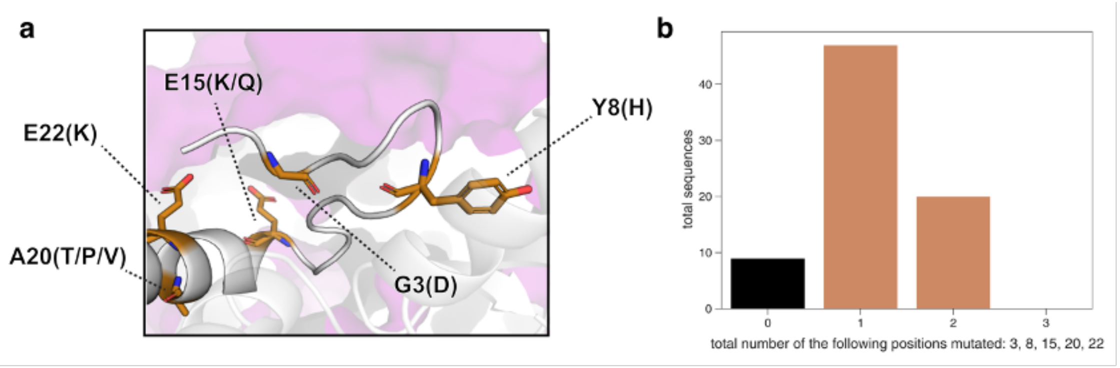


Figure C-11. Commonly observed mutations at the α-subunit interaction interface. a. Homology-predicted *Tm*TrpB structure (based on engineered stand-alone *Pf*TrpB, PDB 6AM8), with commonly mutated TmTrpB residues located near the TrpA interaction interface highlighted. Solvent-exposed regions of TrpA (purple) (PDB 1WDW) are shown as a surface. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. **b.** Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panel **a**.

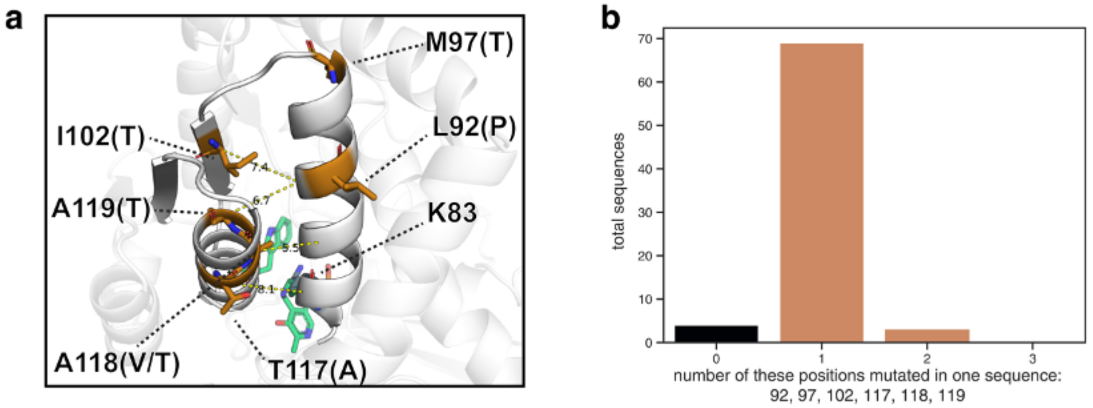


Figure C-12. Commonly observed mutations to residues near a catalytic α-helix. a. Homology-predicted *Tm*TrpB structure (aligned to engineered stand-alone *Pf*TrpB, PDB 6AM8) with wt residues on or near the α-helix housing K83, which are commonly mutated in OrthoRep-evolved populations (orange). PLP (green) and Trp (green) are shown as sticks, and the catalytic lysine K83 (teal) is shown as spheres. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. Dotted lines connect the α-carbon of residues not located on the K83 α-helix with the α-carbon of the nearest residue on the K83 α-helix, with the distance noted in Ångstroms. **b.** Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panel **a**.

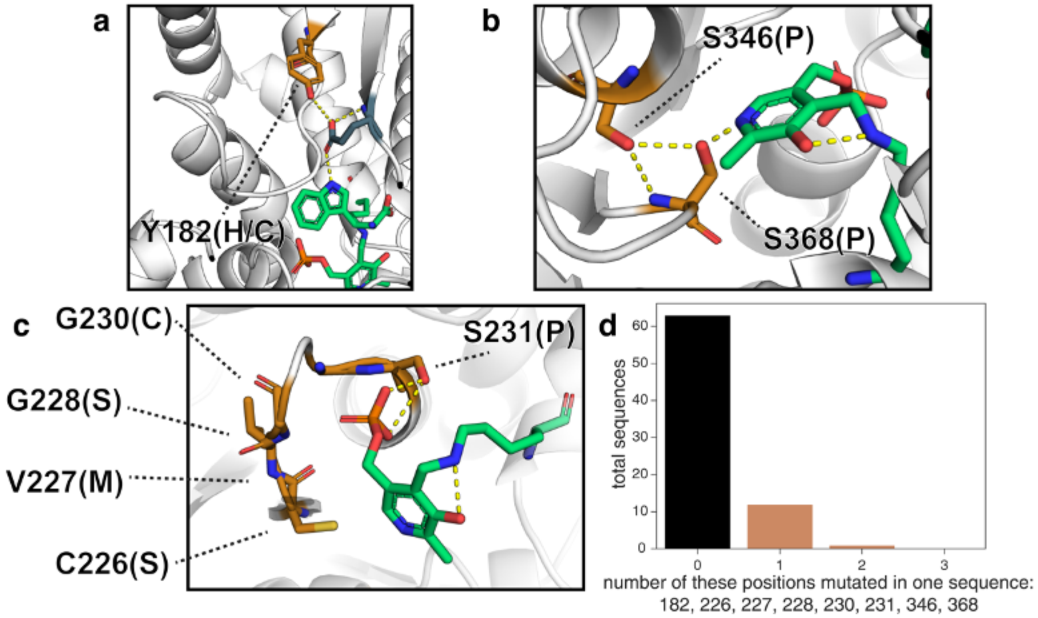


Figure C-13. First- and second-shell active site mutations. a-c. Homology model of *Tm*TrpB (aligned to *Pf*TrpB, PDB: 6AM8) highlighting residues mutated in OrthoRep-evolved variants (orange) that may influence (**a**) indole charge, (**b**) PLP six-member ring binding, and (**c**) PLP-phosphate binding. Mutations are indicated by the wt residue and position, followed by any residues to which this wt residue is mutated in OrthoRep-evolved TrpB sequences. **d.** Total number of sequences in both variant sets 1 and 2 that contain the indicated number of mutations to any of the residues highlighted in panels **a**, **b**, or **c**.

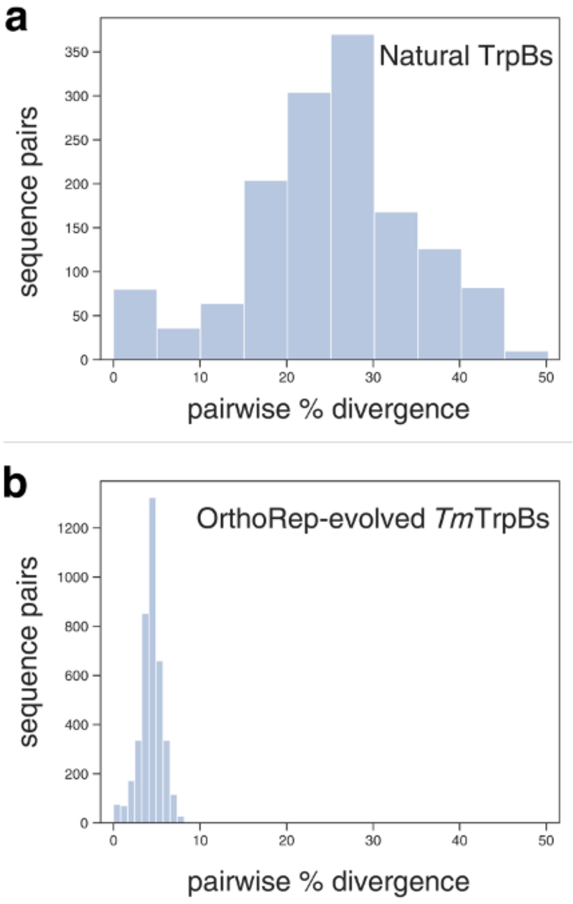


Figure C-14. Sequence divergence for natural and OrthoRep-evolved TrpBs. a-b. Distributions of pairwise % amino acid sequence divergence for a diverse group of 38 naturally occurring mesophilic TrpB variants (**a**) and OrthoRep-evolved variant sets 1 and 2 (**b**).

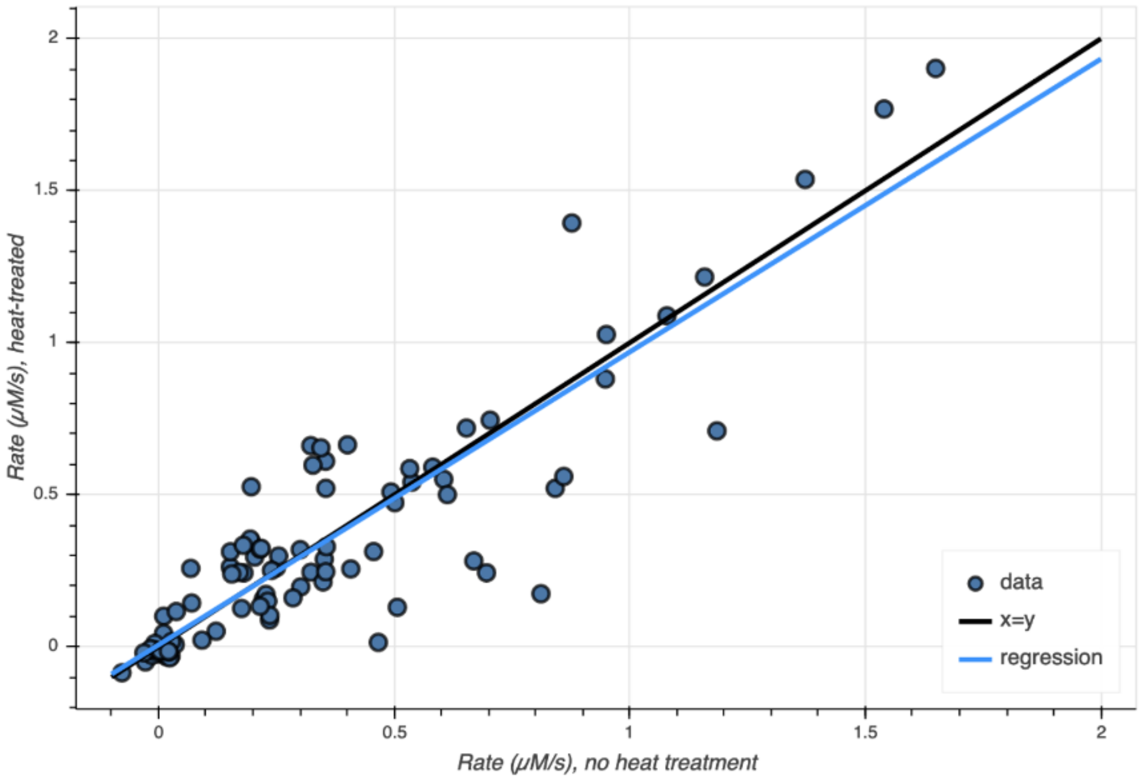


Figure C-15. Correlation between Trp formation by lysates with and without heat treatment. Trp formation by each OrthoRep-evolved variant from variant set 2, WT-003-1-A, Q90*-003-1-A, and Tri-100-2-A with or without heat treatment at 625 µM indole and 25 mM serine, evaluated by UV-vis spectrophotometry. Each point represents the initial rate of Trp formation by cell lysate generated via both heat treatment (1 hour at 75 °C) (y-axis) and a more mild method (x-axis) (see Methods). Linear regression on these data demonstrates a slope of 0.96, suggesting a negligible systematic decrease in activity with heat treatment across variants.

C.3 Methods

C.3.1 DNA plasmid construction

All plasmids that were not generated in a previous study were constructed via Gibson assembly[1] from parts derived from the Yeast Toolkit,[2] from previously described OrthoRep integration cassette plasmids,[3] from *E. coli* expression vectors for previously described TrpB variants,[4,5] from synthesized oligonucleotides, from yeast genomic DNA, or from the standard *E. coli* expression vector, pET-22b(+). All DNA cloning steps and *E. coli* protein expression steps were performed in *E. coli* strains TOP10 and BL21(DE3), respectively. All oligonucleotides used for PCR were purchased from IDT, and all enzymes and reagents used for cloning were purchased from NEB.

Parts used to generate yeast nuclear expression plasmids for testing the selection and p1 integration plasmids were PCR amplified from DNA sources listed above, Gibson assembled, used to transform *E. coli*, and plated onto selective LB agar plates. Individual clones were picked, grown to saturation in selective LB liquid media, miniprepped, and sequence confirmed. Following evolution of TrpB, individual variants were assembled into new yeast or *E. coli* expression vectors through PCR amplification of purified DNA from evolved yeast cultures, bulk cloning into the appropriate expression vector, picking individual colonies, and confirming absence of any frameshift mutations by Sanger sequencing.

C.3.2 Yeast strains and media

Yeast were incubated at 30 °C, with shaking at 200 rpm for liquid cultures, and were typically grown in synthetic complete (SC) growth medium (20 g/L dextrose, 6.7 g/L yeast nitrogen base w/o amino acids (US Biological), 2 g/L SC dropout (US Biological) minus nutrients required for appropriate auxotrophy selection(s)), or were grown in YPD growth medium (10 g/L bacto yeast extract, 20 g/L bacto peptone, 20 g/L dextrose) with or without antibiotics, if no auxotrophic markers were being selected for. Media agar plates were made by combining 2X concentrate of molten agar and 2X concentrate of desired media formulation. Prior to all experiments, cells were grown to saturation in media selecting for maintenance of any plasmids present.

C.3.3 Yeast transformation

All yeast transformations were performed as described in Gietz and Shiestl.[6] After all transformations, transformed cells were streaked onto selective media agar plates, and resulting single colonies were picked for all further uses. Transformations for integration onto p1 were performed as described previously:[7] 2–4 μg of plasmid DNA with *Sca*I restriction sites adjacent to integration flanks was cut with *Sca*I-HF (NEB) and used to transform yeast harboring the wt p1 and p2 plasmids. Proper integration was validated by miniprepping resulting clonal strains as previously described,[7] visualizing the recombinant p1 band of the desired size by gel electrophoresis, and PCR and Sanger sequencing of the gene of interest integrated onto p1. Resulting strains were then transformed with either of two plasmids for nuclear expression of an OrthoRep terminal protein DNA polymerase 1 (TP-DNAP1) variant: wt TP-DNAP1 (pAR-Ec318) for evaluating trp5 complementation of TrpB variants without mutagenesis, or error-prone TP-DNAP1 (pAR-Ec633) for generating strains ready for TrpB evolution. These strains were passaged for ~40 generations in order to stabilize copy number of the recombinant p1 species, prior to any use in experiments.

Genomic deletions were made through co-transformation of a CRISPR/Cas9 plasmid targeting the region of interest and a linear DNA fragment comprised of two concatenated 50 bp homology flanks to the region of interest.[8] Transformations were then plated on selective media agar, colonies were re-streaked onto nonselective media agar, and resulting colonies were grown to saturation in liquid media. The region of interest was PCR amplified and Sanger sequenced to confirm presence of desired modification.

C.3.4 Plating assays

Yeast strains expressing a TrpB variant either from a nuclear plasmid, or from p1, with wt OrthoRep polymerase (TP-DNAP1) expressed from a nuclear plasmid, were grown to saturation in SC –L or SC –LH, spun down, washed once with 0.9% NaCl, then spun down again, and the resulting pellet was resuspended in 0.9% NaCl. Washed cells were then diluted 1:100 (or 1:10,000 where indicated) in 0.9% NaCl, and 10 µL of each diluted cell suspension were plated onto media agar plates in pre-marked positions. After three days of growth, cell spots were imaged (Bio-Rad ChemiDoc™). Resulting images were adjusted uniformly ('High' set to 40,000) to improve visibility of growth (Bio-Rad Image Lab™ Software). Figures utilizing these images (**Figures C-1** and **C-2**) were made by manually combining images of different plates, but all images of the same media condition within each figure panel were derived from the same image of a single plate.

C.3.5 *Tm*TrpB evolution

Yeast strains with a nuclear plasmid expressing error-prone TP-DNAP1 and with wt *Tm*TrpB, *Tm*Triple, or *Tm*TripleQ90* encoded on p1 (GR-Y053, GR-Y055, and GR-Y057) were grown to saturation in SC –LH, prior to passaging for evolution. All cultures passaged for evolution of *Tm*TrpB regardless of success are described in **Table C-1**. To provide enough indole substrate for sufficient Trp production, but not enough to induce toxicity, all growth media used for evolution of TrpB activity were supplemented with 100 µM indole, as informed by results shown in **Figure C-1**. All passages for evolution were carried out as 1:100 dilutions. In order to induce a growth defect but still allow for some growth, the first passage for each evolution culture was carried out in SC –LH media with 37 µM Trp (7.6 mg/L). After two or three days of shaking incubation, if OD_{600} > 1.0 (Bio-Rad SmartSpec™

3000) for 100-mL cultures, or if most wells in a 24-well block of 3-mL cultures were saturated to a similar degree by eye, cultures were passaged into fresh growth medium with a slightly reduced Trp concentration. If the level of growth was beneath this threshold, the culture was passaged into growth medium with the same Trp concentration. This process was continued until cultures were capable of growth in a Trp concentration of 3.7 μM (or, in the sole case of WT-100-1, 4.7 μM), at which point a passage into media lacking Trp was attempted, which typically resulted in successful growth. Resulting cultures were then passaged six additional times into growth medium lacking Trp.

C.3.6 Growth rate assays

Yeast strains containing nuclear plasmids encoding one of several OrthoRep-evolved TrpB variants, wt *Tm*TrpB, *Tm*Triple, or none of these (denoted 'empty') were grown to saturation in SC –L, washed as described above, then inoculated 1:100 into multiple media conditions in 96-well clear bottom plates, with four biological replicates per media/strain combination. Plates were then sealed with a porous membrane and allowed to incubate with shaking at 30 °C for 24 hours, with OD_{600} measurements taken automatically every 30 minutes (Tecan Infinite M200 Pro), according to a previously described protocol.[9] Multiple 24 hour periods were required for each experiment, but empty controls were included in each individual 96-well plate to ensure validity of growth in other cultures. Raw OD_{600} measurements were fed into a custom MATLAB script,[10] which carries out a logarithmic transformation to linearize the exponential growth phase, identifies this growth phase, and uses this to calculate the doubling time (*T*). Doubling time was then converted to growth rate by the equation *ln(2)/T*.

C.3.7 Enzyme characterization—general experimental methods

Chemicals and reagents were purchased from commercial sources and used without further purification. All cultures were grown in Terrific Broth supplemented with 100 μg/mL carbenicillin (TB_{carb}). Cultures were shaken in a New Brunswick Innova 4000 shaker (shaking diameter 19 mm), with the exception of the 96-well deep-well plates (USA Scientific), which were shaken in a Multitron INFORS HT shaker (shaking diameter 50 mm). Lysis buffer was composed of 50 mM potassium phosphate, pH 8.0 (KPi buffer), supplemented with 100 or 200 μM pyridoxal 5'-phosphate (PLP). Heat lysis was performed in a 75 °C water bath (Fisher) for >1 hour. Protein concentrations were determined using a Pierce™ BCA Protein Assay Kit (Thermo Scientific). Reactions were performed in KPi buffer. Liquid chromatography/mass spectrometry (LCMS) was performed on an Agilent 1290 UPLC-LCMS equipped with a C-18 silica column (1.8 μm, 2.1 × 50 mm) using CH_3CN/H_2O (0.1% acetic acid by volume): 5% to 95% CH_3CN over 2 min; 1 mL/min. Liquid chromatography/mass spectrometry (LCMS) was also performed on an Agilent 1260 HPLC-MS equipped with Agilent InfinityLab Poroshell 120 EC-C18 column (2.7 μm, 4.6×50 mm): hold 5% CH_3CN for 0.5 min, 5-95% CH_3CN over 2 min; 1 mL/min.

TrpB variants selected for further characterization were cloned into a pET-22b(+) vector with a C-terminal 6X His-tag and used to transform *E. coli* BL21(DE3) cells (Lucigen).

C.3.8 Expression and characterization of variants from set 1

C.3.8.1 Large scale expression and lysis

A single colony containing the appropriate TrpB gene was used to inoculate 5 mL TB_{carb} and incubated overnight at 37 °C and 230 rpm. For expression, 0.5 mL of overnight culture were used to inoculate 50 mL TB_{carb} in a 250-mL flask and incubated at 37 °C and 250 rpm for 3 hours to reach an OD_{600} of 0.6–0.8. Cultures were chilled on ice for 20 min and expression was induced with a final concentration of 1 mM isopropyl β-D-thiogalactopyranoside (IPTG). Expression proceeded at 25 °C and 250 rpm for approximately 20 hours. Cells were harvested by centrifugation at 5,000*g* for 5 min at 4 °C, and then the supernatant was decanted. The pellet was stored at −20 °C until further use or used immediately for whole cell transformations.

Pellets were lysed in 5 mL of lysis KPi buffer with 200 μM PLP, supplemented with 1 mg/mL lysozyme (HEWL, Sigma Aldrich), 0.02 mg/mL bovine pancreas DNase I, and 0.1X BugBuster (Novagen) and incubated at 37 °C for 30 minutes. Lysate was clarified by centrifugation at 5,000*g* for 10 min, divided into 1-mL aliquots, and stored at −20 °C until further use.

C.3.8.2 Expression and characterization of variants from set 1 — lysate and whole cell small-scale reactions

Protein concentration in lysate was quantified by BCA. Lysate reactions were performed in 2-mL glass HPLC vials (Agilent) charged with indole (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of lysate (final enzyme conc. 4 μM), and serine (final conc. 20 mM) to achieve a final volume of 200 μL. Whole-cell reactions were performed in 2-mL glass HPLC vials (Agilent) charged with indole (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of cells diluted in KPi buffer (final OD_{600}=6), and serine (final conc. 20 mM) to achieve a final volume of 200 μL. Reactions were incubated at 30 °C for 24 hours, diluted with 800 μL 1:1 CH_3CN/1 M aq. HCl, and analyzed via UHPLC-MS.

C.3.8.3 Expression and characterization of variants from set 1 — thermostability determination

Enzyme T_{50} measurements (the temperature at which 50% of the enzyme is irreversibly inactivated after a 1-hour incubation) were used to report on the thermostability of the enzyme. In a total volume of 100 μL, samples were prepared in KPi buffer with 1 μM enzyme in PCR tubes and either set aside (25 °C) or heated in a thermal cycler on a gradient from 79–99 °C (OrthoRep-generated variants), or 59–99 °C (TmTriple), for 1 hour, with each temperature performed in duplicate. Precipitated protein was pelleted via centrifugation, and 75 μL of each sample were carefully removed and added to the wells of a 96-well UV-transparent assay plate containing 0.5 mM indole and 0.5 mM serine. Relative product formation was observed by measuring the change in absorbance at 290 nm to determine the temperature at which the sample had 50% residual activity compared to the 25 °C samples (modeled as a logistic function).

C.3.8.4 Enzyme kinetics

Enzymatic parameters, k_{cat} and K_{M}, for the conversion of indole to Trp were estimated via Bayesian inference assuming Michaelis-Menten behavior under saturating serine (40 mM) in KPi buffer. Briefly, initial velocities (v) were determined by monitoring Trp formation in a Shimadzu UV-1800 spectrophotometer at 30 °C for 1 min over a range of indole concentrations at 290 nm using the reported indole-Trp difference in absorbance coefficient ($\Delta\varepsilon_{290}$ = 1.89 mM^{-1} cm^{-1}).[11] These velocities were modeled using the equation:

$$v = \frac{V_{\text{max}}[\text{indole}]}{K_{\text{M}} + [\text{indole}]}$$

and estimates for *v* and $V_{\max}$ were converted to *k* and k_{cat} by normalizing for enzyme concentration. Parameter estimates are obtained as Hamiltonian Markov chain Monte Carlo (MCMC) posterior samples and reported as the median with their 95% credible regions (CR). The MCMC software used for sampling was Stan (pystan version 2.19.0.0). The sampling was performed with four separate chains, each starting with 2000 warm-up (disregarded) steps followed by 12000 posterior sampling steps. The priors chosen for k_{cat} and K_{M} were lognormal distributions with means log(150) and log(500), standard deviations 2.5 and 1.5, with units of sec^{-1} and µM, respectively. This provided non-negative probability density that covered low-to-moderate values of the parameters for many known enzymes, but still had significant density out to very high values for each parameter. (In all cases, the data was shown to significantly inform the prior.) The code used to generate these estimates (along with example data) can be found at http://github.com/palmhjell/bayesian_kinetics.

C.3.9 Expression and characterization of variants from set 2

C.3.9.1 Small scale expression and lysis

Variants were arrayed into a 96-well deep-well plate along with *Tm*Triple, *Tm*9D8*, *Tm*Azul, and *Pf*2B9. Individual colonies were grown in 600 µL TB_{carb} in 96-well polypropylene plates overnight at 37 °C, 250 rpm, 80% humidity. The following day, 20 µL of overnight culture were used to inoculate 630 µL TB_{carb} in deep-well 96-well plates and grown at 37 °C, 250 rpm. After 4 hours, cultures were chilled on ice for 20–30 min and protein expression was induced with 50 µL IPTG (final conc. 1 mM) diluted in TB_{carb}. Cultures were shaken at 20 °C, 250 rpm for 20–24 hours, after which they were subjected to centrifugation at 5,000*g* for 10 min. The cell pellets were frozen at −20 °C until further use or used immediately.

C.3.9.2 Indole rate measurements

Pellets were lysed in either 600 µL of KPi buffer with 100 µM PLP and heat treated at 75 °C for 1 hour, or in 600 µL of this buffer supplemented with 1 mg/mL lysozyme, 0.02 mg/mL bovine pancreas DNase I, and 0.1X BugBuster and incubated at 37 °C for 1 hour. Lysate from both conditions was clarified by centrifugation at 4,500*g* for 10 min and stored at 4 °C until further use.

Reaction master mix composed of 625 µM indole and 25 mM serine in KPi buffer was prepared and, before reactions, plates and master mix were incubated in a 30 °C water bath for 30 min. The microplate reader (Tecan Spark) was also pre-heated to 30 °C.

To UV-transparent 96-well assay plates (Caplugs, catalog # 290-8120-0AF), 160 µL pre-heated reaction master mix was added by 12-channel pipet followed by 40 µL of lysate from the pre-heated plate using a Microlab NIMBUS96 liquid handler (Hamilton). Plates were immediately transferred into the plate reader, shaken for 10 sec to mix and the absorbance of each well at 290 nm was recorded as rapidly as possible (~20 s between measurements) for 120 cycles. The rate of product formation was determined by finding the rate of absorbance change over time and converting to units of concentration using $\Delta\varepsilon_{290}$ = 1.89 mM^{-1} cm^{-1} (see above) and a determined path length of 0.56 cm. We observed no systematic difference in activity between the two lysate preparations (**Figure C-15**), suggesting that most enzyme variants retained sufficient thermostability for purification via heat treatment, and this method was used in subsequent experiments.

C.3.9.3 Substrate scope screen

Pellets were lysed in 300 µL KPi buffer with 200 µM PLP and clarified by centrifugation at 4,000*g* for 10 min. To a 96-well deep-well plate charged with 10 µL nucleophile dissolved in DMSO (final conc. denoted in the table directly below), 40 µL of the heat treated lysate were transferred using a Microlab NIMBUS96 liquid handler (Hamilton), followed by addition of 150 µL serine (final conc. 20 mM) with a 12-channel pipet. Reactions were sealed with 96-well ArctiSeal™ Silicone/PTFE Coating (Arctic White) and incubated in a 30 °C water bath for ~24 hours. Reactions were diluted with 600 µL 2:1 CH_3CN/1 M aq. HCl, subjected to centrifugation at 5,000*g*, and 400 µL were transferred to 2-mL glass HPLC vials (Agilent). Samples were analyzed by HPLC-MS. Azulene samples were further diluted 20X to avoid oversaturation of the UV-detector and analyzed via UHPLC-MS.

Table C-5. **Nucleophiles tested in substrate scope screen**

Nucleophile	Source	Catalog #	CAS	Final conc. (mM)
5-Cyanoindole	Chem Impex	21849	15861-24-2	20
6-Cyanoindole	Chem Impex	21181	15861-36-6	20
7-Cyanoindole	Sigma	CDS008484		20
5-Bromoindole	Sigma	B68607	10075-50-0	20
6-Bromoindole	Sigma	524344	52415-29-9	20
7-Bromoindole	Sigma	473723	51417-51-7	20
5-Methoxyindole	Combi-Blocks	IN0049	1006-94-6	20
5-Trifluoromethylindole	Sigma	701068	100846-24-0	10
Azulene	Alfa Aesar	L08271	275-51-4	20

All samples except those containing azulene were analyzed at 277 nm, representing the isosbestic point between indole and Trp and allowing estimation of yield by comparing the substrate and product peak areas for indole analogs.[12] Azulene yield was estimated as described previously.[13] Nucleophile retention times were determined though injection of authentic standards and product retention times were identified by extracting their expected mass from the mass spectrum.

C.3.10 Characterization of Tri-100-3-F and Tri-100-1-G

C.3.10.1 Large-scale expression and purification

A single colony containing the appropriate TrpB gene was used to inoculate 5 mL TB_{carb} and incubated overnight at 37 °C and 230 rpm. For expression, 2.5 mL of overnight culture were used to inoculate 250 mL TB_{carb} in a 1-L flask and incubated at 37 °C and 250 rpm for 3 hours to reach OD_{600} 0.6–0.8. Cultures were chilled on ice for 20 min and expression was induced with a final concentration of 1 mM IPTG. Expression proceeded at 25 °C and 250 rpm for approximately 20 hours. Cells were harvested by centrifugation at 5,000*g* for 5 min at 4 °C, and then the supernatant was decanted. The pellet was stored at −20 °C until further use.

Pellets were lysed in 25 mL KPi buffer with 200 μM PLP for >1 hour at 75 °C. Lysate was clarified by spinning 14,000*g* for 20 min at 4 °C (New Brunswick Avanti J-30I). Protein was purified over hand-packed HisPur™ Ni-NTA Resin (Thermo Scientific, catalog # 88221), dialyzed into KPi buffer and quantified by BCA.

C.3.10.2 Tri-100-3-F PLP-binding assay

Variant Tri-100-3-F did not exhibit the characteristic yellow color of PLP-bound TrpB variants after purification, however BCA indicated protein concentrations comparable to the Tri-100-1-G variant. We have previously observed that some TrpB variants lose binding affinity for PLP resulting in non-functional apoenzyme. We evaluated Trp formation of Tri-100-3-F supplemented with 0, 0.1, 0.25, 0.5, 1, 2, 5, and 100 μM PLP via UV-Vis spectrophotometry. Serine (final conc. 25 mM) + PLP master mixes of the eight concentrations were prepared and dispensed into a 96-well UV-transparent plate. Enzyme dilutions (final conc. 1 μM) with or without indole master mixes were prepared, and 100 μL thereof were dispensed into the 96-well plate. The plate was immediately transferred into plate reader, shaken for 10 s to mix, and product formation was measured ~20 s for 120 cycles at 290 nm.

Only the 100 μM PLP condition restored activity, supporting our hypothesis that the purified enzyme was apoprotein and binds PLP poorly, requiring supplementation of PLP to re-form a functional holoenzyme. Thus, we chose to supplement PLP in the subsequent purified protein reactions.

C.3.10.3 Small-scale analytical reactions

Reactions were performed in 2-mL glass HPLC vials (Agilent) charged with nucleophile (final conc. 20 mM) dissolved in DMSO (5% w/v), followed by the addition of purified protein (final enzyme concentrations were either 2 μM or 40 μM), PLP (final conc. 100 μM) and serine (final conc. 20 mM) to achieve a final volume of 200 μL. Reactions were incubated at 30 °C for ~24 hours. Reactions were diluted with 600 μL 2:1 CH_3CN/1 M aq. HCl, subjected to centrifugation at 5,000*g*, and 400 μL transferred to 2-mL glass HPLC vials (Agilent). Samples were analyzed by HPLC-MS. Azulene samples were further diluted 20X to avoid oversaturation of the UV-detector and analyzed via UHPLC-MS.

Appendix C Bibliography

1. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
2. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A highly characterized

yeast toolkit for modular, multipart assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).

3. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, Continuous evolution of genes at mutation rates above genomic error thresholds. *Cell* **175**, 1946–1957 (2018).
4. Boville, C. E., Romney, D. K., Almhjell, P. J., Sieben, M. & Arnold, F. H. Improved synthesis of 4-cyanotryptophan and other tryptophan analogues in aqueous solvent using variants of TrpB from *Thermotoga maritima*. *J. Org. Chem.* **83**, 7447–7452 (2018).
5. Murciano-Calles, J., Romney, D. K., Brinkmann-Chen, S., Buller, A. R. & Arnold, F. H. A panel of TrpB biocatalysts derived from tryptophan synthase through the transfer of mutations that mimic allosteric activation. *Angew. Chem. Int. Ed.* **55**, 11577–11581 (2016).
6. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
7. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).
8. Ryan, O. W. & Cate, J. H. D. Multiplex engineering of industrial yeast genomes using CRISPRm. *Methods in Enzymology* **546**, 473–489 (2014).
9. Jung, P. P., Christian, N., Kay, D. P., Skupin, A. & Linster, C. L. Protocols and programs for high-throughput growth and aging phenotyping in yeast. *PLoS One* **10**, e0119807 (2015).
10. Zhong, Z. *et al.* Automated continuous evolution of proteins *in vivo*. *ACS Synth. Biol.* (2020). doi:10.1021/acssynbio.0c00135
11. Lane, A. N. & Kirschner, K. The catalytic mechanism of tryptophan synthase from *Escherichia coli*. *Eur. J. Biochem.* **129**, 571–582 (1983).
12. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: A general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).
13. Watkins, E. J., Almhjell, P. J. & Arnold, F. H. Direct enzymatic synthesis of a deep-blue fluorescent noncanonical amino acid from azulene and serine. *ChemBioChem* **21**, 80–83 (2020).

Appendix D

SUPPLEMENTARY INFORMATION FOR CHAPTER V

D.1 Oligo Design

D.1.1 Inner primer design

The inner primers of evSeq are specific to the region of interest. Each region of interest is captured by both a forward and reverse primer. These primers have the below general layout:

F: 5′ - CACCCAAGACCACTCTCCGG**XXXXXXX**… - 3′
R: 5′ - CGGTGTGCGAAGTAGGTGC**XXXXXXXX**… - 3′

The 5' region is a universal adapter to which outer primers bind (see **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes*, below) while the 3' region (denoted by "**X**" in the primers above) is specific to the region of interest. Note that the length of the variable 3' region will vary depending on the target gene (this is indicated by the ellipses at the end of the poly-**X** region). Note that there is no need for the two primers in the pair to be equal length—we show them as such to highlight the fact that the forward universal adapter is one base longer than the reverse universal adapter. Detailed instructions for effective primer construction are provided on the evSeq wiki (https://fhalab.github.io/evSeq/1-lib_prep.html#inner-primer-design).

D.1.2 Outer primer design

The barcode (outer) primers used in evSeq all follow the below layout:

F: 5′ - TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**XXXXXXX**CACCCAAGACCACTCTCCGG - 3′
R: 5′ - GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**XXXXXXX**CGGTGTGCGAAGTAGGTGC - 3′

Each of these primers consists of (1) a 5’ sequence matching the Illumina Nextera transposase adapters, (2) a central unique 7-nucleotide barcode (**Table D-1**), and (3) a 3’ universal seed that matches the 5’ adapter of the inner primers (see **Section D.1.1**, *Inner Primer Design*, above). Note that only Illumina indices compatible with the Nextera transposase adapters can be used with the provided outer primer designs; other indexing systems would require different adapters. The full set of outer primers used in this study can be found in **Table D-2**; they can be ordered from IDT by following the instructions provided in **Section D.2.1**, *Ordering Barcode Primers from IDT*, below.

D.1.3 Barcode design

evSeq uses 192 unique 7-nucleotide barcodes (**Table D-1**). The barcodes were designed to satisfy the below criteria:

1. All barcodes must have GC-content of 40–60%.
2. All barcodes must be at least 3 substitutions apart. This is to prevent misassignment of reads due to sequencing errors of the barcodes.
3. No barcode can have 3 of the same bases in a row. This is to reduce sequencing errors.
4. No barcode can be a sub-sequence of the Nextera transposase adapters or their reverse complements (see below). This is to avoid interference with downstream Illumina chemistry.
5. No barcode can be a sub-sequence of the Illumina p5 and p7 flow cell-binding sequences or their reverse complements (see below for sequences). Again, this is to avoid interference with downstream Illumina chemistry.

The Nextera transposase adapter sequences are below:

```
5′ - TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG - 3'
5′ - GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG - 3'
```

The p5 and p7 flow cell-binding sequences are below:

```
p5: 5′ - AATGATACGGCGACCACCGAGATCTACAC - 3'
p7: 5′ - CAAGCAGAAGACGGCATACGAGAT - 3′
```

D.2 Supplemental Protocols

D.2.1 Ordering barcode primers from IDT

We provide a pre-filled IDT order form for all evSeq primers on the evSeq GitHub repository (https://github.com/fhalab/evSeq/tree/master/lib_prep_tools/IdtOrderForm.xlsx). This order form can be used to order evSeq primers in the 96-well plate layout needed to prepare the evSeq barcode primer mixes (see **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes*, below). To order evSeq primers:

1. Navigate to the IDT DNA oligo ordering page: https://www.idtdna.com/pages/products/custom-dna-rna/dna-oligos/custom-dna-oligos.
2. Under “Ordering”, select “Plates”.
3. From the “Single-stranded DNA” table, select the amount (in nanomoles) of oligo you wish to order (denoted in the “Product” column) by clicking “Order” under the “96 Well” column. For the work described in this paper, 25 nmol oligos were ordered.
4. On the next page, click “UPLOAD PLATE(S)”. Using the pop-up that results, upload the “IdtOrderForm.xls” provided on the evSeq GitHub repository. The pop-up should recognize two plates—one called “FBC” and the other called “RBC”—each consisting of 96 wells. Click “ADD PLATES” followed by “CLOSE THIS WINDOW” to close the window.
5. For the “FBC” plate, click “Plate Specifications”. Confirm that the below specifications are set as follows:
    a. Purification: Standard Desalting
    b. Plate Type: Deep Well

c. **Ship Option: Wet**
d. Buffer: IDTE 8.0 pH
e. Normalization Type: Full Yield
f. **Concentration: 100 µM**

Note that the bolded specifications are different from default. **While not strictly required, it is strongly recommended that primers be ordered *wet* at 100 µM; reconstituting plates of dry primers to 100 µM can be very time-consuming without robotic support.**

6. Once specifications are correctly set for the “FBC” plate, click “APPLY SETTINGS TO ALL PLATES” at the bottom of the specifications pop-up, followed by “YES” on the window that follows. Quickly check to make sure that the same settings as recommended in step 5 were applied to “RBC” by clicking on the “RBC” “Plate Specifications” option.
7. Add the primers to your order by clicking “ADD TO ORDER”, then follow standard IDT procedures for purchasing.

D.2.2 Preparation of evSeq barcode primer mixes

There are 96 unique forward and 96 unique reverse outer primers (**Table D-2**), corresponding to 96 unique forward and 96 unique reverse barcodes (**Table D-1**). The forward and reverse outer primers were ordered following the procedure given above in **Section D.2.1**, *Ordering Barcode Primers from IDT*.

Each well sequenced in evSeq is encoded by a different combination of forward and reverse barcode. Different primers from the forward and reverse outer primer plates can be mixed together to associate a barcode combination with a specific well in a specific plate. Because the same outer primers can be used regardless of inner primer, it is convenient to keep plates of barcode combinations on hand. Plates of outer primer combinations (hereafter also referred to as “barcode plates”) can be stored for long periods of time.

Throughout this work, we used the same 8 barcode plates (consisting of 768 different combinations of forward and reverse outer primers) to encode plate and well locations. Barcode plates are named DI01–DI08, where "DI" stands for "dual-indexed". The exact barcode combinations used by evSeq are given in **Tables D-3–10**; these combinations can also be found in the "index_map.csv" file on the evSeq GitHub (https://github.com/fhalab/evSeq/tree/master/evSeq/util/index_map.csv). By default, the evSeq software assumes the barcode plates used for library preparation are laid out in the order given in the "index_map.csv" file. To build the barcode plates depicted in **Tables D-3–10**, we followed the below procedure:

1. 10-fold dilutions of each of the forward and reverse outer primer plates ordered from IDT were prepared by adding 10 µL of each primer stock to 90 µL ddH2O, keeping the well layout constant. Dilutions were performed in fully-skirted PCR plates (Bio-Rad HSP9601). The plates from IDT had a starting concentration of 100 µM, so the final concentration of these two diluted plates was 10 µM.

2. To 8 fully-skirted PCR plates, 80 µL ddH2O was added, followed by 10 µL diluted (10 µM) forward barcode plate. The well layout was kept constant for the forward barcode primers.

3. To the each of the 8 plates, 10 µL of diluted (10 µM) reverse barcode plate was added, shifting the well layout down by 1 row per plate. For instance, row A of the reverse plate went into row A of the first barcode plate, row B of the second barcode plate, row C of the third barcode plate, and so on; row H of the reverse plate went into row H of the first barcode plate, row A of the second barcode plate, row B of the third barcode plate, and so on.

4. When not in use, the 10-fold dilutions prepared in step 1 were stored at –20 °C, while the barcode plates (each well of which had a combination of a specific

forward and reverse primer at a final concentration of 1 µM) were stored at 4 °C. Both the 10 µM stock plates and 1 µM barcode plates can be stored for long periods of time—we have noticed no drop in effectiveness even after years of storage.

D.2.3 evSeq library preparation/data analysis protocol

The evSeq library preparation protocol was designed to be as cost-effective as possible. The quantities used in the below protocol were chosen to fit within the constraints of the resources available to our research group (these are the quantities used for all evSeq experiments performed in this paper). However, with automation support (e.g., liquid handling robots) and higher-capacity molecular biology equipment, the entire protocol could be scaled down to lower quantities, further improving cost-effectiveness.

The list of steps below can be followed to prepare an evSeq library for sequencing using the outer primers described in **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes*, above. Note that when first using a new set of inner primers, it is recommended to complete the below protocol for a few wells as a test before deploying them for plate-scale reactions.

The library preparation protocol can be completed with the below steps. Note that provided part numbers are for the materials/reagents we used while developing this protocol—the same components from other providers will almost certainly work as well. This protocol is also provided on the evSeq wiki (https://fhalab.github.io/evSeq/1-lib_prep.html#pcr-protocol).

1. Prepare a PCR master mix for the number of wells to be sequenced according to the below table. Note that we provide an excel calculator on the evSeq GitHub repository for easy calculation of master mix volumes based on the number of plates to be sequenced

(https://github.com/fhalab/evSeq/tree/master/lib_prep_tools/MastermixCalculator.xlsx).

Component	Amount per 10 µL rxn (µL)
Thermopol Buffer (NEB B9004S)	1.00
10 mM dNTPs (NEB N0447)	0.20
Taq Polymerase (NEB M0267)	0.05
ddH$_2$O	5.33
Mol-Bio Grade DMSO (MP 194819)	0.40
Inner Primer Mix (10 µM)	0.02

a. Note that the above table assumes that each evSeq PCR reaction will be 10 µL—if scaling down, adjust volumes accordingly.

b. Note that the above table also assumes the same set of inner primers is used to prepare all plates. If this is not the case, a separate master mix will need to be prepared for each set of inner primers.

c. The Inner Primer Mix (10 µM) is a combination of forward and reverse inner primers at a final concentration of 10 µM each in diH2O (this can be prepared, e.g., by adding 10 µL of 100 µM forward inner primer and 10 µL of 100 µM reverse inner primer to 80 µL diH2O).

2. Add 7 µL of master mix to each well of as many half-skirted PCR plates (USA Scientific 1402-9700) as will be sequenced. These are referred to as "PCR plates" in the remainder of this protocol.

3. Stamp 1 µL of overnight culture from each plate to be sequenced into the PCR plates.

   a. "Stamp" means "apply to all wells, keeping the plate layout consistent". For example, 1 µL of culture from library 01 F02 is moved to PCR plate 01 F02, 1 µL of culture from library 02 C07 is moved to PCR plate 02 C07, etc.

   b. Note that both fresh culture and previously frozen culture (thawed before use as template) will work here. No modifications need to be made to the protocol.

4. Complete stage 1 PCR using the below thermal cycler conditions. This PCR amplifies the fragment of interest from the template DNA contained in the cell culture.

Step	Temperature (°C)	Time
1	95	5 min

2	95	20 s
3	TD 63-> 54	20 s
4	68	30 s
5	Return to 2, 9 x	
6	4	Hold

   a. "TD" above stands for "touchdown". A touchdown step decrements the temperature by 1 °C each cycle. The touchdown in the above PCR starts at 63 °C and drops to 54 °C by the end.
   b. Note that the extension step (step 4) is long enough to amplify a 500 bp fragment. Longer fragments will need a longer extension time. Note, however, that you may see reduced sequencing efficiency with fragments that are too large.
   c. While developing this protocol, we used the below thermal cycler models:
      i. Eppendorf Mastercycler ep Gradient S Thermal Cycler, Model 5345 with 96-well universal block
      ii. Eppendorf Mastercycler pro S vapo.protect
      iii. Eppendorf Mastercycler X50s 96-well silver block thermal cycler

5. Once PCR has completed, stamp 2 µL of 1 µM barcode primer mix from the barcode plates into the PCR plates (see **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes*, above, for details on preparation of barcode plates). **Record which barcode plate was stamped into which PCR plate.**
6. Perform the second step PCR using the below conditions:

Step	Temperature (°C)	Time
1	95	20s
2	68	50 s
3	Return to 1, 24 x	
4	68	5 min
5	4	Hold

   a. Again, longer fragments may need a longer extension time.

7. While the second PCR runs, prepare a 2% agarose gel with SYBR gold added (Thermo Fisher Scientific, S11494).
8. Once the second PCR has completed, for each plate, pool 5 µL of each reaction into 100 mM EDTA to a final concentration of 20 mM EDTA—this step quenches the reactions. Pooling will leave you with as many tubes as you have plates, each

containing ~600 μL [96 rxns/plate × (5 μL per rxn + 1.25 μL 100mM EDTA per reaction)].

   a. Note: The most efficient way to do the pooling varies depending on the equipment available. Our group relies on 12-channel multichannel pipets for this step, and so will accomplish pooling by (1) adding 10 μL 100 mM EDTA to each well in a single row of a fresh PCR plate, (2) transferring 5 μL reaction from each row in the plate-to-be-pooled into the single row of EDTA, and (3) transferring 40 μL from each well in the single row of pooled reactions using a single-channel pipet (leaving 10 μL dead volume in each well) to a microcentrifuge tube. An alternate strategy might be, for instance, adding 120 μL 100 mM EDTA to a trough, then pipetting 5 μL of all reactions from a plate into this trough. **Whatever strategy is taken, what is important in pooling is that the ratios of the reactions in the pool remain equal—sacrificing some reaction as dead volume is perfectly acceptable to achieve equal mixing in this step.**

9. For each tube made in step 8, take 100 μL of pooled reaction and add it to 20 μL 6x loading dye (NEB B7025S) in a microcentrifuge tube. **It is critical that the loading dye does not contain SDS.** At this point, the remaining pooled reaction from step 8 can be stored at –20 °C for future use (i.e., if the later steps of this protocol ever need to be redone).
   a. Note that most of the pooled reaction is not moved into later steps with this protocol. Again, if relevant automation and molecular biology equipment is available, reactions can be scaled down below 10 μL, reducing wasted reaction. Current reaction sizes are set to minimize pipetting error.
10. Load the contents of each tube made in step 9 into the agarose gel prepared in step 7. The contents of each tube should be kept separate (i.e., loaded into different lanes in the gel). Load a ladder (we typically use 100 bp ladder from NEB, N3231S) in the flanking lanes.
11. Run the agarose gel at 130 V until the bands have sufficiently migrated. Often, you will see two bands: the lower band is usually primer dimer and the upper is the target. Reference the ladder to identify your product, remembering that the two-step PCR

adds 120 bp of additional length (from the universal adapter, barcode, and transposase adapters) onto the gene fragment of interest.

12. Gel-extract the target bands from the agarose gel, again keeping bands from different plates separate. We typically use Zymoclean Gel DNA Recovery Kit (Zymo Research, D4001) for this step. Elution should be performed at a low volume—we typically elute in 10 μL of ddH2O.
13. After gel extraction, combine the gel-extracted pools from each plate in equimolar concentrations. We provide a calculator on the evSeq GitHub repository that can be used to normalize *equal-length* fragments to a pre-specified concentration (https://github.com/fhalab/evSeq/tree/master/lib_prep_tools/LibDilCalculator.xlsx).
    a. Note that the quantification here need not be extremely robust. For all results presented in this work, we performed this step using DNA concentrations output by a GE NanoVue Plus.
    b. Tip: It is generally not advised to pool amplicons drastically different in length. Shorter fragments are preferentially sequenced in NGS, and so the shorter amplicon will dominate the number of reads. Separate submissions should be made for libraries with very different lengths.
14. After the previous step, you should have a single tube of cleaned, normalized DNA consisting of all amplicons from all plates to be pooled. This DNA will be submitted to your sequencing provider for inclusion in a multiplexed sequencing run. You should work with your sequencing provider to ensure that all requirements are met to slot into their pipeline. For instance, this protocol assumes that the sequencing provider can add Nextera-compatible Illumina indices and flow-cell-binding sequences via PCR—it should be confirmed that your sequencing provider can do this before submitting your sample.
    a. Note: Throughout this work, we used the "Customized PCR Amplicon Sequencing" services of Laragen Inc., available at http://www.laragen.com/laragen_nextgen.php.
    b. Also note that, depending on your sequencing provider, it may be possible (or even necessary) to add the Illumina indices yourself. Again, you should work with your provider to determine the best course of action for submitting

evSeq libraries. Adding indices simply requires one final PCR on the pooled evSeq library.

15. Once sequencing is complete, your sequencing provider should return two fastq (or fastq.gz) files to you. One will contain the forward reads for your pooled samples and the other will contain the reverse reads—both files are needed by the evSeq software for processing.
16. Using the files returned in step 15, run the evSeq software to process results and assign variants to their original wells. Detailed instructions on how to use the evSeq software and interpret its outputs are provided on the evSeq Wiki https://fhalab.github.io/evSeq/4-usage.html

D.3 Supplemental Figures

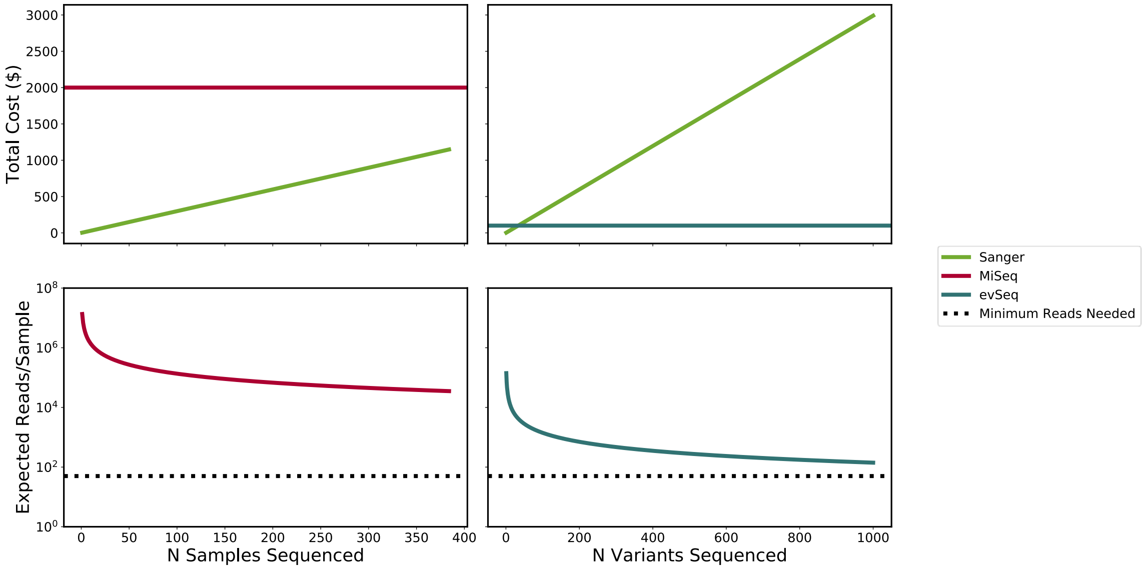


Figure D-1. Comparison of the tradeoff between sequencing depth and cost for Sanger sequencing (green), a multiplexed MiSeq run (red), and an evSeq library (blue). The top row gives the total cost for sequencing a given number of variants; the bottom row gives the expected number of reads per variant for sequencing a given number of variants. Note that the x-axes for the left and right columns are different. The limit on the x-axis for the left column is set to reflect what is typically the maximum level of multiplexed NGS available (384 samples) when outsourcing sequencing. To be consistent with the language used throughout the main text, the x-axis labels refer to elements run in a multiplexed NGS run as “samples” and elements contained in an evSeq library as “variants”. We assume that the elements sequenced in these examples are derived from protein mutant libraries amenable to sequencing by evSeq (i.e., the sequenced elements are targeted amplicons). **Top Row:** We see that both multiplexed NGS on a commercial MiSeq run and evSeq have constant cost with an increasing number of elements sequenced; Sanger, in contrast, scales linearly with the number of elements sequenced. Many elements (669 with the cost estimates used to make this figure) need to be added to a multiplexed MiSeq run before it becomes more cost-effective than Sanger. Even though research groups may frequently meet or exceed 669 variants in a standard protein engineering experiment, the flat cost of $2000 is far too high to justify regular sequencing of every variant. Many fewer variants (34) need to be added to an evSeq run before it becomes cost-effective over Sanger. A flat cost of ~$100 is justifiable for regularly sequencing all variants. **Bottom Row:** NGS technologies trade off sequencing depth for cost effectiveness. Notably, the per-sample sequencing depth achieved by commercially available multiplexed runs is much higher than what is needed for reliable sequencing. evSeq, in contrast, more efficiently spreads reads, keeping the expected number of reads closer to, yet still above the minimum needed for effective sequencing. **Notes on Figure Generation:** Cost of a single MiSeq run ($2000) is based on an estimate provided by Laragen Inc. Cost of a single Sanger sequencing run ($2.99) is based on a quote from MCLAB for sequencing a single 96-well plate. The number of expected reads from a MiSeq run (13.5 million) is based on estimates provided by Illumina for a MiSeq Reagent Kit v2 (note that almost double the number of reads can be achieved using a v3 kit—we used v2 here to be conservative with our estimates for NGS/evSeq). The number of expected reads for a variant sampled with evSeq assumes the evSeq library was sequenced as 1 of 96 samples on a multiplexed sequencing run using a MiSeq Reagent Kit v2. The cost of a single evSeq run is based on an estimate provided by Laragen for a single sample in a multiplexed sequencing run using a PE150 kit.

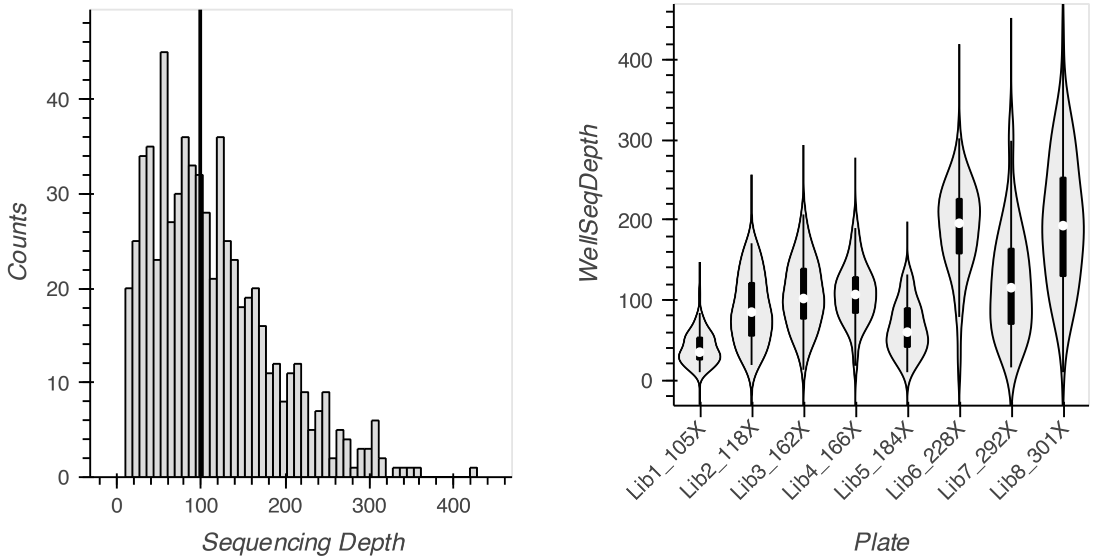


Figure D-2. Sequencing depths for the Tm9D8* evSeq libraries. Left: A histogram of sequencing depths for each Tm9D8* variant contained in the full evSeq library. The vertical black line gives the median. **Right:** Violin plots showing the distribution of read depths over the wells in each sequenced plate. Variability between plates likely indicates inaccurate quantification of pooled plates prior to final assembly of the evSeq library. Notable, libraries 1-5 use different evSeq primers than libraries 6-8.

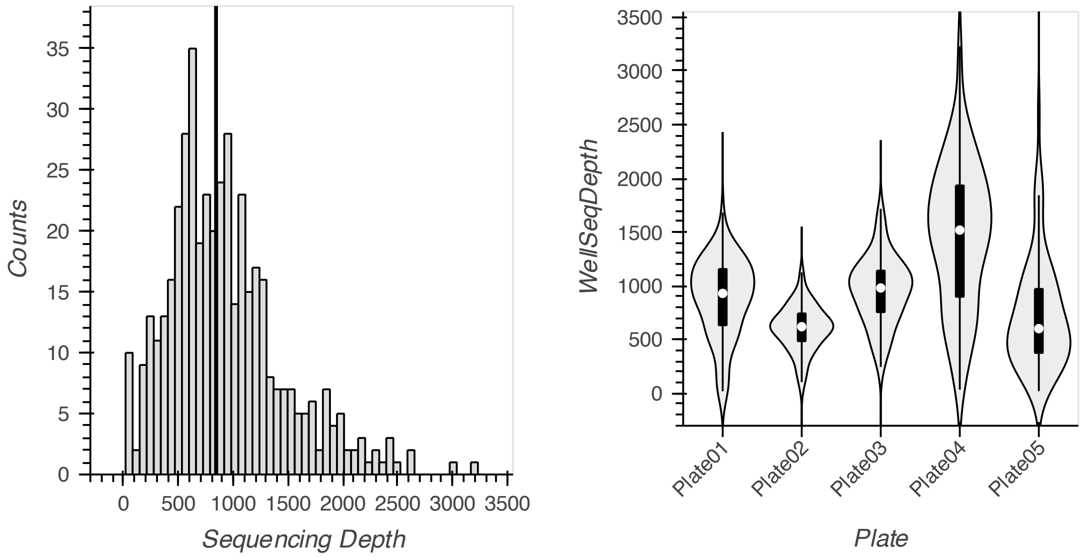


Figure D-3. Sequencing depths for the *Rma*NOD evSeq libraries. Left: A histogram of sequencing depths for each *Rma*NOD variant contained in the full evSeq library. The vertical black line gives the median. **Right:** Violin plots showing the distribution of read depths over the wells in each sequenced plate. Variability between plates likely indicates inaccurate quantification of pooled plates prior to final assembly of the evSeq library.

D.4 Barcode and Outer Primer Sequences

Table D-1. evSeq barcode sequences used in this work. The "Plate" and "Well" columns give the location of these sequences in the IDT order form provided on the evSeq GitHub repository (see **Section D.2.1**, *Ordering Barcode Primers from IDT* and **Section D.1.3**, *Barcode Design*, above). Note that barcode sequences can also be found in the "index_map.csv" file found on the evSeq GitHub repository (https://github.com/fhalab/evSeq/tree/master/evSeq/util/index_map.csv); this csv file also gives the combinations of barcodes used to define the dual indexing (DI) plates.

Plate	Well	Barcode
FBC	A01	GATCATG
FBC	A02	TACATGG
FBC	A03	AAGCACC
FBC	A04	TGGCTCA
FBC	A05	CTTGCTC
FBC	A06	GAAGCGT
FBC	A07	TCTCCAT
FBC	A08	TTGAAGG
FBC	A09	GAATGTC
FBC	A10	ATCTCCA
FBC	A11	GCGTTAT
FBC	A12	TGCACCA
FBC	B01	TGCCTAT
FBC	B02	AGGAATC
FBC	B03	TCCACTG
FBC	B04	TTGTACC
FBC	B05	TTCGAGT
FBC	B06	CTTCAGC
FBC	B07	CAGTGCA
FBC	B08	TGCTGTC
FBC	B09	CGCCATT
FBC	B10	GCCATGA
FBC	B11	CACAACG
FBC	B12	CTTCGCT
FBC	C01	TCGTGAA
FBC	C02	TTATCGG
FBC	C03	AGACCAT
FBC	C04	ACATGAG
FBC	C05	ACGTACT
FBC	C06	CACCTCA
FBC	C07	GTTGGAG
FBC	C08	TGTTCTG
FBC	C09	CTTACGT
FBC	C10	GAGGTTG
FBC	C11	ATGGACA

FBC	C12	ACACTGA
FBC	D01	ATCTGTG
FBC	D02	AATGTGC
FBC	D03	GAGTTGA
FBC	D04	TTCTCAC
FBC	D05	TGAAGCG
FBC	D06	GCTACAA
FBC	D07	AGAGAAC
FBC	D08	CAGAGTG
FBC	D09	TTCCGAA
FBC	D10	GTACGAC
FBC	D11	ACTCTTG
FBC	D12	CCAACCA
FBC	E01	CTCTAGA
FBC	E02	AATCGGA
FBC	E03	CGTCCTA
FBC	E04	GGAATGT
FBC	E05	TCCAAGC
FBC	E06	GCACCTA
FBC	E07	TTGCGTT
FBC	E08	CAGGATT
FBC	E09	CTGCATA
FBC	E10	CGTTGAG
FBC	E11	TGCTACT
FBC	E12	GTGATCC
FBC	F01	GCATGGT
FBC	F02	GTCGTTA
FBC	F03	CCTGACA
FBC	F04	AGTGTAG
FBC	F05	CGAGCAA
FBC	F06	CTACTCC
FBC	F07	GATGCCA
FBC	F08	GACCGAT
FBC	F09	ACGTTGG
FBC	F10	ATGAGCG
FBC	F11	TACTCCG
FBC	F12	GATTCAC
FBC	G01	ATGACGC
FBC	G02	GGTTGTT
FBC	G03	GTACTTG
FBC	G04	TAGCAAG
FBC	G05	CTGCCAT
FBC	G06	GAGAACA

FBC	G07	GTATAGC
FBC	G08	TGATGGA
FBC	G09	GGCAGTA
FBC	G10	GAAGAAG
FBC	G11	AGCGGTT
FBC	G12	TAAGGCC
FBC	H01	AACCTGT
FBC	H02	AGTACAC
FBC	H03	CTCGTAG
FBC	H04	CTAGGTG
FBC	H05	CGATACC
FBC	H06	TCGGCTA
FBC	H07	CGGTTGT
FBC	H08	ATTGCCT
FBC	H09	CATTCGA
FBC	H10	GCACAAT
FBC	H11	GCAGTAA
FBC	H12	CCTAATC
RBC	A01	GAACTGC
RBC	A02	ACCAGGT
RBC	A03	TCTAGAG
RBC	A04	CACACAA
RBC	A05	GTGGAAC
RBC	A06	ATATGCC
RBC	A07	GGTCTGA
RBC	A08	GTGAGAT
RBC	A09	TTGGCAG
RBC	A10	ATGCCTG
RBC	A11	TCCGAAG
RBC	A12	GGCTTAC
RBC	B01	AGTTGGC
RBC	B02	AACGATG
RBC	B03	ACTACCG
RBC	B04	GGTGTCT
RBC	B05	CCAGCTT
RBC	B06	TTAGACG
RBC	B07	ACCATAC
RBC	B08	GACGACT
RBC	B09	GTCACCT
RBC	B10	CGTGATG
RBC	B11	GCTTCCT
RBC	B12	TAGACGT
RBC	C01	CGGACTT

RBC	C02	ACCGGAA
RBC	C03	CCGAAGT
RBC	C04	TCACGCA
RBC	C05	ATCCTCG
RBC	C06	CGAATAG
RBC	C07	TATCCGG
RBC	C08	AGCAAGA
RBC	C09	TGTCGAC
RBC	C10	TTCCATG
RBC	C11	GCAATCG
RBC	C12	TGAGTGG
RBC	D01	TAGGAGA
RBC	D02	AGTCAGT
RBC	D03	GTGCTGT
RBC	D04	CAACAAC
RBC	D05	AATAGCC
RBC	D06	TCTGTGA
RBC	D07	TGTGGTA
RBC	D08	GCGTATG
RBC	D09	AGTTACG
RBC	D10	TTCCTGC
RBC	D11	TATGTCG
RBC	D12	GGAGAGA
RBC	E01	CCTTAGG
RBC	E02	TGTATCC
RBC	E03	CAACCTG
RBC	E04	CTGATGA
RBC	E05	AAGACAG
RBC	E06	AGCTCGT
RBC	E07	GATTGCG
RBC	E08	TCCTTCA
RBC	E09	TCACAGG
RBC	E10	AGAGCTG
RBC	E11	CCTCTGT
RBC	E12	CCTCGAA
RBC	F01	GTGTCTC
RBC	F02	ATTGAGG
RBC	F03	GACAATC
RBC	F04	CACTTGC
RBC	F05	TGAACGC
RBC	F06	CGTAGCA
RBC	F07	AGGTTCC
RBC	F08	GTACACA

RBC	F09	GATAGGT
RBC	F10	TAGCCTC
RBC	F11	TTCAGCC
RBC	F12	GGATTCA
RBC	G01	TGAGCCT
RBC	G02	AACGCGA
RBC	G03	TCATTGC
RBC	G04	AGCATCT
RBC	G05	TTGGTCT
RBC	G06	CAAGGAT
RBC	G07	AGACGTC
RBC	G08	AGGTCAA
RBC	G09	ATGCTAC
RBC	G10	CTCTGAT
RBC	G11	TCAAGTC
RBC	G12	TCGAGCT
RBC	H01	ACAGTCT
RBC	H02	CAGATAC
RBC	H03	TACGTTC
RBC	H04	ACGGTTC
RBC	H05	CATCGTC
RBC	H06	TACGCAT
RBC	H07	CTTAGAC
RBC	H08	AACTGAC
RBC	H09	ACTTGCA
RBC	H10	ACGCGAT
RBC	H11	TCGACAC
RBC	H12	ACTCAAC

Table D-2. Full-length evSeq barcode (outer) primer sequences used in this work. The "Plate" and "Well" columns give the location of these sequences in the IDT order form provided on the evSeq GitHub repository (see **Section D.2.1**, *Ordering Barcode Primers from IDT* and **Section D.2.2**, *Preparation of evSeq Barcode Primer Mixes*, above).

Plate	Well	Sequence
FBC	A01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATCATGCACCCAAGACCACTCTCCGG
FBC	A02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACATGGCACCCAAGACCACTCTCCGG
FBC	A03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGCACCCACCCAAGACCACTCTCCGG
FBC	A04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGCTCACACCCAAGACCACTCTCCGG
FBC	A05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGCTCCACCCAAGACCACTCTCCGG
FBC	A06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGCGTCACCCAAGACCACTCTCCGG
FBC	A07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTCCATCACCCAAGACCACTCTCCGG
FBC	A08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGAAGGCACCCAAGACCACTCTCCGG
FBC	A09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAATGTCCACCCAAGACCACTCTCCGG
FBC	A10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTCCACACCCAAGACCACTCTCCGG
FBC	A11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCGTTATCACCCAAGACCACTCTCCGG
FBC	A12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCACCACACCCAAGACCACTCTCCGG
FBC	B01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCCTATCACCCAAGACCACTCTCCGG
FBC	B02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGAATCCACCCAAGACCACTCTCCGG
FBC	B03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCACTGCACCCAAGACCACTCTCCGG
FBC	B04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGTACCCACCCAAGACCACTCTCCGG
FBC	B05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCGAGTCACCCAAGACCACTCTCCGG
FBC	B06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCAGCCACCCAAGACCACTCTCCGG
FBC	B07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGCACACCCAAGACCACTCTCCGG
FBC	B08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCTGTCCACCCAAGACCACTCTCCGG
FBC	B09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCCATTCACCCAAGACCACTCTCCGG
FBC	B10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCATGACACCCAAGACCACTCTCCGG
FBC	B11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACAACGCACCCAAGACCACTCTCCGG
FBC	B12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCGCTCACCCAAGACCACTCTCCGG
FBC	C01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGTGAACACCCAAGACCACTCTCCGG
FBC	C02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTATCGGCACCCAAGACCACTCTCCGG
FBC	C03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGACCATCACCCAAGACCACTCTCCGG
FBC	C04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACATGAGCACCCAAGACCACTCTCCGG
FBC	C05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACGTACTCACCCAAGACCACTCTCCGG
FBC	C06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACCTCACACCCAAGACCACTCTCCGG
FBC	C07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGGAGCACCCAAGACCACTCTCCGG
FBC	C08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTTCTGCACCCAAGACCACTCTCCGG
FBC	C09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTACGTCACCCAAGACCACTCTCCGG
FBC	C10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGGTTGCACCCAAGACCACTCTCCGG
FBC	C11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGACACACCCAAGACCACTCTCCGG
FBC	C12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACACTGACACCCAAGACCACTCTCCGG
FBC	D01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTGTGCACCCAAGACCACTCTCCGG
FBC	D02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATGTGCCACCCAAGACCACTCTCCGG

FBC	D03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGTTGACACCCAAGACCACTCTCCGG
FBC	D04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCACCACCCAAGACCACTCTCCGG
FBC	D05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGAAGCGCACCCAAGACCACTCTCCGG
FBC	D06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTACAACACCCAAGACCACTCTCCGG
FBC	D07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGAACCACCCAAGACCACTCTCCGG
FBC	D08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAGTGCACCCAAGACCACTCTCCGG
FBC	D09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCCGAACACCCAAGACCACTCTCCGG
FBC	D10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACGACCACCCAAGACCACTCTCCGG
FBC	D11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTCTTGCACCCAAGACCACTCTCCGG
FBC	D12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAACCACACCCAAGACCACTCTCCGG
FBC	E01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCTAGACACCCAAGACCACTCTCCGG
FBC	E02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATCGGACACCCAAGACCACTCTCCGG
FBC	E03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTCCTACACCCAAGACCACTCTCCGG
FBC	E04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAATGTCACCCAAGACCACTCTCCGG
FBC	E05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCAAGCCACCCAAGACCACTCTCCGG
FBC	E06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCACCTACACCCAAGACCACTCTCCGG
FBC	E07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGCGTTCACCCAAGACCACTCTCCGG
FBC	E08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGGATTCACCCAAGACCACTCTCCGG
FBC	E09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCATACACCCAAGACCACTCTCCGG
FBC	E10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTTGAGCACCCAAGACCACTCTCCGG
FBC	E11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCTACTCACCCAAGACCACTCTCCGG
FBC	E12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGATCCCACCCAAGACCACTCTCCGG
FBC	F01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCATGGTCACCCAAGACCACTCTCCGG
FBC	F02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCGTTACACCCAAGACCACTCTCCGG
FBC	F03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTGACACACCCAAGACCACTCTCCGG
FBC	F04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGTAGCACCCAAGACCACTCTCCGG
FBC	F05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGAGCAACACCCAAGACCACTCTCCGG
FBC	F06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACTCCCACCCAAGACCACTCTCCGG
FBC	F07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGCCACACCCAAGACCACTCTCCGG
FBC	F08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACCGATCACCCAAGACCACTCTCCGG
FBC	F09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACGTTGGCACCCAAGACCACTCTCCGG
FBC	F10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGAGCGCACCCAAGACCACTCTCCGG
FBC	F11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACTCCGCACCCAAGACCACTCTCCGG
FBC	F12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATTCACCACCCAAGACCACTCTCCGG
FBC	G01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGACGCCACCCAAGACCACTCTCCGG
FBC	G02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTGTTCACCCAAGACCACTCTCCGG
FBC	G03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACTTGCACCCAAGACCACTCTCCGG
FBC	G04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAGCAAGCACCCAAGACCACTCTCCGG
FBC	G05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCCATCACCCAAGACCACTCTCCGG
FBC	G06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGAACACACCCAAGACCACTCTCCGG
FBC	G07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTATAGCCACCCAAGACCACTCTCCGG
FBC	G08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGATGGACACCCAAGACCACTCTCCGG
FBC	G09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCAGTACACCCAAGACCACTCTCCGG

FBC	G10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGAAGCACCCAAGACCACTCTCCGG
FBC	G11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCGGTTCACCCAAGACCACTCTCCGG
FBC	G12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAAGGCCCACCCAAGACCACTCTCCGG
FBC	H01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACCTGTCACCCAAGACCACTCTCCGG
FBC	H02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTACACCACCCAAGACCACTCTCCGG
FBC	H03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCGTAGCACCCAAGACCACTCTCCGG
FBC	H04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTAGGTGCACCCAAGACCACTCTCCGG
FBC	H05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATACCCACCCAAGACCACTCTCCGG
FBC	H06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGGCTACACCCAAGACCACTCTCCGG
FBC	H07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGGTTGTCACCCAAGACCACTCTCCGG
FBC	H08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCCTCACCCAAGACCACTCTCCGG
FBC	H09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCATTCGACACCCAAGACCACTCTCCGG
FBC	H10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCACAATCACCCAAGACCACTCTCCGG
FBC	H11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCAGTAACACCCAAGACCACTCTCCGG
FBC	H12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTAATCCACCCAAGACCACTCTCCGG
RBC	A01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAACTGCCGGTGTGCGAAGTAGGTGC
RBC	A02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCAGGTCGGTGTGCGAAGTAGGTGC
RBC	A03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTAGAGCGGTGTGCGAAGTAGGTGC
RBC	A04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCACACAACGGTGTGCGAAGTAGGTGC
RBC	A05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGGAACCGGTGTGCGAAGTAGGTGC
RBC	A06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATATGCCCGGTGTGCGAAGTAGGTGC
RBC	A07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTCTGACGGTGTGCGAAGTAGGTGC
RBC	A08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGAGATCGGTGTGCGAAGTAGGTGC
RBC	A09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGGCAGCGGTGTGCGAAGTAGGTGC
RBC	A10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGCCTGCGGTGTGCGAAGTAGGTGC
RBC	A11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCGAAGCGGTGTGCGAAGTAGGTGC
RBC	A12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCTTACCGGTGTGCGAAGTAGGTGC
RBC	B01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTTGGCCGGTGTGCGAAGTAGGTGC
RBC	B02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACGATGCGGTGTGCGAAGTAGGTGC
RBC	B03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTACCGCGGTGTGCGAAGTAGGTGC
RBC	B04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTGTCTCGGTGTGCGAAGTAGGTGC
RBC	B05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCAGCTTCGGTGTGCGAAGTAGGTGC
RBC	B06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTAGACGCGGTGTGCGAAGTAGGTGC
RBC	B07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCATACCGGTGTGCGAAGTAGGTGC
RBC	B08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACGACTCGGTGTGCGAAGTAGGTGC
RBC	B09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCACCTCGGTGTGCGAAGTAGGTGC
RBC	B10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTGATGCGGTGTGCGAAGTAGGTGC
RBC	B11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTTCCTCGGTGTGCGAAGTAGGTGC
RBC	B12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACGTCGGTGTGCGAAGTAGGTGC
RBC	C01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGGACTTCGGTGTGCGAAGTAGGTGC
RBC	C02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCGGAACGGTGTGCGAAGTAGGTGC
RBC	C03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGAAGTCGGTGTGCGAAGTAGGTGC
RBC	C04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCACGCACGGTGTGCGAAGTAGGTGC

RBC	C05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATCCTCGCGGTGTGCGAAGTAGGTGC
RBC	C06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGAATAGCGGTGTGCGAAGTAGGTGC
RBC	C07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTATCCGGCGGTGTGCGAAGTAGGTGC
RBC	C08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCAAGACGGTGTGCGAAGTAGGTGC
RBC	C09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTCGACCGGTGTGCGAAGTAGGTGC
RBC	C10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCATGCGGTGTGCGAAGTAGGTGC
RBC	C11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCAATCGCGGTGTGCGAAGTAGGTGC
RBC	C12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAGTGGCGGTGTGCGAAGTAGGTGC
RBC	D01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGGAGACGGTGTGCGAAGTAGGTGC
RBC	D02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTCAGTCGGTGTGCGAAGTAGGTGC
RBC	D03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGCTGTCGGTGTGCGAAGTAGGTGC
RBC	D04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACAACCGGTGTGCGAAGTAGGTGC
RBC	D05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAATAGCCCGGTGTGCGAAGTAGGTGC
RBC	D06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTGTGACGGTGTGCGAAGTAGGTGC
RBC	D07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTGGTACGGTGTGCGAAGTAGGTGC
RBC	D08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGTATGCGGTGTGCGAAGTAGGTGC
RBC	D09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTTACGCGGTGTGCGAAGTAGGTGC
RBC	D10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCTGCCGGTGTGCGAAGTAGGTGC
RBC	D11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTATGTCGCGGTGTGCGAAGTAGGTGC
RBC	D12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAGAGACGGTGTGCGAAGTAGGTGC
RBC	E01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTTAGGCGGTGTGCGAAGTAGGTGC
RBC	E02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTATCCCGGTGTGCGAAGTAGGTGC
RBC	E03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACCTGCGGTGTGCGAAGTAGGTGC
RBC	E04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGATGACGGTGTGCGAAGTAGGTGC
RBC	E05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGACAGCGGTGTGCGAAGTAGGTGC
RBC	E06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCTCGTCGGTGTGCGAAGTAGGTGC
RBC	E07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATTGCGCGGTGTGCGAAGTAGGTGC
RBC	E08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCTTCACGGTGTGCGAAGTAGGTGC
RBC	E09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCACAGGCGGTGTGCGAAGTAGGTGC
RBC	E10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGAGCTGCGGTGTGCGAAGTAGGTGC
RBC	E11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCTGTCGGTGTGCGAAGTAGGTGC
RBC	E12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCGAACGGTGTGCGAAGTAGGTGC
RBC	F01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGTCTCCGGTGTGCGAAGTAGGTGC
RBC	F02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTGAGGCGGTGTGCGAAGTAGGTGC
RBC	F03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACAATCCGGTGTGCGAAGTAGGTGC
RBC	F04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCACTTGCCGGTGTGCGAAGTAGGTGC
RBC	F05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAACGCCGGTGTGCGAAGTAGGTGC
RBC	F06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTAGCACGGTGTGCGAAGTAGGTGC
RBC	F07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGTTCCCGGTGTGCGAAGTAGGTGC
RBC	F08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTACACACGGTGTGCGAAGTAGGTGC
RBC	F09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATAGGTCGGTGTGCGAAGTAGGTGC
RBC	F10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGCCTCCGGTGTGCGAAGTAGGTGC
RBC	F11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCAGCCCGGTGTGCGAAGTAGGTGC

RBC	F12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGATTCACGGTGTGCGAAGTAGGTGC
RBC	G01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAGCCTCGGTGTGCGAAGTAGGTGC
RBC	G02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACGCGACGGTGTGCGAAGTAGGTGC
RBC	G03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCATTGCCGGTGTGCGAAGTAGGTGC
RBC	G04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCATCTCGGTGTGCGAAGTAGGTGC
RBC	G05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGGTCTCGGTGTGCGAAGTAGGTGC
RBC	G06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAAGGATCGGTGTGCGAAGTAGGTGC
RBC	G07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGACGTCCGGTGTGCGAAGTAGGTGC
RBC	G08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGTCAACGGTGTGCGAAGTAGGTGC
RBC	G09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGCTACCGGTGTGCGAAGTAGGTGC
RBC	G10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCTGATCGGTGTGCGAAGTAGGTGC
RBC	G11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCAAGTCCGGTGTGCGAAGTAGGTGC
RBC	G12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGAGCTCGGTGTGCGAAGTAGGTGC
RBC	H01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAGTCTCGGTGTGCGAAGTAGGTGC
RBC	H02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGATACCGGTGTGCGAAGTAGGTGC
RBC	H03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACGTTCCGGTGTGCGAAGTAGGTGC
RBC	H04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGGTTCCGGTGTGCGAAGTAGGTGC
RBC	H05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATCGTCCGGTGTGCGAAGTAGGTGC
RBC	H06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACGCATCGGTGTGCGAAGTAGGTGC
RBC	H07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTAGACCGGTGTGCGAAGTAGGTGC
RBC	H08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACTGACCGGTGTGCGAAGTAGGTGC
RBC	H09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTTGCACGGTGTGCGAAGTAGGTGC
RBC	H10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGCGATCGGTGTGCGAAGTAGGTGC
RBC	H11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGACACCGGTGTGCGAAGTAGGTGC
RBC	H12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTCAACCGGTGTGCGAAGTAGGTGC

D.5 Dual-Indexing Platemaps

This section contains all platemaps for the dual indexing plates (DI plates) used in this study. The tables that follow show how the primers from the forward and reverse barcode plates (**Table D-2**) were arrayed to produce the barcode plates. Each entry in the below platemaps follows the format “Well-Barcode Plate”, where the “-” delimits the plate and well. An “F” after the delimiter indicates that the well preceding the delimiter was from the forward barcode plate (“FBC” in **Table D-2**) and an “R” indicates that the well was from the reverse barcode plate (“RBC”). A detailed protocol for how the dual index plates were produced is given in **Section 2.2.2**, *Preparation of evSeq Barcode Primer Mixes*, above.

Table D-3. Platemap for DI01 used in this study.

DI01	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, A01-R	A02-F, A02-R	A03-F, A03-R	A04-F, A04-R	A05-F, A05-R	A06-F, A06-R	A07-F, A07-R	A08-F, A08-R	A09-F, A09-R	A10-F, A10-R	A11-F, A11-R	A12-F, A12-R
B	B01-F, B01-R	B02-F, B02-R	B03-F, B03-R	B04-F, B04-R	B05-F, B05-R	B06-F, B06-R	B07-F, B07-R	B08-F, B08-R	B09-F, B09-R	B10-F, B10-R	B11-F, B11-R	B12-F, B12-R
C	C01-F, C01-R	C02-F, C02-R	C03-F, C03-R	C04-F, C04-R	C05-F, C05-R	C06-F, C06-R	C07-F, C07-R	C08-F, C08-R	C09-F, C09-R	C10-F, C10-R	C11-F, C11-R	C12-F, C12-R
D	D01-F, D01-R	D02-F, D02-R	D03-F, D03-R	D04-F, D04-R	D05-F, D05-R	D06-F, D06-R	D07-F, D07-R	D08-F, D08-R	D09-F, D09-R	D10-F, D10-R	D11-F, D11-R	D12-F, D12-R
E	E01-F, E01-R	E02-F, E02-R	E03-F, E03-R	E04-F, E04-R	E05-F, E05-R	E06-F, E06-R	E07-F, E07-R	E08-F, E08-R	E09-F, E09-R	E10-F, E10-R	E11-F, E11-R	E12-F, E12-R
F	F01-F, F01-R	F02-F, F02-R	F03-F, F03-R	F04-F, F04-R	F05-F, F05-R	F06-F, F06-R	F07-F, F07-R	F08-F, F08-R	F09-F, F09-R	F10-F, F10-R	F11-F, F11-R	F12-F, F12-R
G	G01-F, G01-R	G02-F, G02-R	G03-F, G03-R	G04-F, G04-R	G05-F, G05-R	G06-F, G06-R	G07-F, G07-R	G08-F, G08-R	G09-F, G09-R	G10-F, G10-R	G11-F, G11-R	G12-F, G12-R
H	H01-F, H01-R	H02-F, H02-R	H03-F, H03-R	H04-F, H04-R	H05-F, H05-R	H06-F, H06-R	H07-F, H07-R	H08-F, H08-R	H09-F, H09-R	H10-F, H10-R	H11-F, H11-R	H12-F, H12-R

Table D-4. Platemap for DI02 used in this study.

DI02	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, H01-R	A02-F, H02-R	A03-F, H03-R	A04-F, H04-R	A05-F, H05-R	A06-F, H06-R	A07-F, H07-R	A08-F, H08-R	A09-F, H09-R	A10-F, H10-R	A11-F, H11-R	A12-F, H12-R
B	B01-F, A01-R	B02-F, A02-R	B03-F, A03-R	B04-F, A04-R	B05-F, A05-R	B06-F, A06-R	B07-F, A07-R	B08-F, A08-R	B09-F, A09-R	B10-F, A10-R	B11-F, A11-R	B12-F, A12-R
C	C01-F, B01-R	C02-F, B02-R	C03-F, B03-R	C04-F, B04-R	C05-F, B05-R	C06-F, B06-R	C07-F, B07-R	C08-F, B08-R	C09-F, B09-R	C10-F, B10-R	C11-F, B11-R	C12-F, B12-R
D	D01-F, C01-R	D02-F, C02-R	D03-F, C03-R	D04-F, C04-R	D05-F, C05-R	D06-F, C06-R	D07-F, C07-R	D08-F, C08-R	D09-F, C09-R	D10-F, C10-R	D11-F, C11-R	D12-F, C12-R
E	E01-F, D01-R	E02-F, D02-R	E03-F, D03-R	E04-F, D04-R	E05-F, D05-R	E06-F, D06-R	E07-F, D07-R	E08-F, D08-R	E09-F, D09-R	E10-F, D10-R	E11-F, D11-R	E12-F, D12-R
F	F01-F, E01-R	F02-F, E02-R	F03-F, E03-R	F04-F, E04-R	F05-F, E05-R	F06-F, E06-R	F07-F, E07-R	F08-F, E08-R	F09-F, E09-R	F10-F, E10-R	F11-F, E11-R	F12-F, E12-R
G	G01-F, F01-R	G02-F, F02-R	G03-F, F03-R	G04-F, F04-R	G05-F, F05-R	G06-F, F06-R	G07-F, F07-R	G08-F, F08-R	G09-F, F09-R	G10-F, F10-R	G11-F, F11-R	G12-F, F12-R
H	H01-F, G01-R	H02-F, G02-R	H03-F, G03-R	H04-F, G04-R	H05-F, G05-R	H06-F, G06-R	H07-F, G07-R	H08-F, G08-R	H09-F, G09-R	H10-F, G10-R	H11-F, G11-R	H12-F, G12-R

Table D-5. Platemap for DI03 used in this study.

DI03	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, G01-R	A02-F, G02-R	A03-F, G03-R	A04-F, G04-R	A05-F, G05-R	A06-F, G06-R	A07-F, G07-R	A08-F, G08-R	A09-F, G09-R	A10-F, G10-R	A11-F, G11-R	A12-F, G12-R
B	B01-F, H01-R	B02-F, H02-R	B03-F, H03-R	B04-F, H04-R	B05-F, H05-R	B06-F, H06-R	B07-F, H07-R	B08-F, H08-R	B09-F, H09-R	B10-F, H10-R	B11-F, H11-R	B12-F, H12-R
C	C01-F, A01-R	C02-F, A02-R	C03-F, A03-R	C04-F, A04-R	C05-F, A05-R	C06-F, A06-R	C07-F, A07-R	C08-F, A08-R	C09-F, A09-R	C10-F, A10-R	C11-F, A11-R	C12-F, A12-R
D	D01-F, B01-R	D02-F, B02-R	D03-F, B03-R	D04-F, B04-R	D05-F, B05-R	D06-F, B06-R	D07-F, B07-R	D08-F, B08-R	D09-F, B09-R	D10-F, B10-R	D11-F, B11-R	D12-F, B12-R
E	E01-F, C01-R	E02-F, C02-R	E03-F, C03-R	E04-F, C04-R	E05-F, C05-R	E06-F, C06-R	E07-F, C07-R	E08-F, C08-R	E09-F, C09-R	E10-F, C10-R	E11-F, C11-R	E12-F, C12-R
F	F01-F, D01-R	F02-F, D02-R	F03-F, D03-R	F04-F, D04-R	F05-F, D05-R	F06-F, D06-R	F07-F, D07-R	F08-F, D08-R	F09-F, D09-R	F10-F, D10-R	F11-F, D11-R	F12-F, D12-R
G	G01-F, E01-R	G02-F, E02-R	G03-F, E03-R	G04-F, E04-R	G05-F, E05-R	G06-F, E06-R	G07-F, E07-R	G08-F, E08-R	G09-F, E09-R	G10-F, E10-R	G11-F, E11-R	G12-F, E12-R
H	H01-F, F01-R	H02-F, F02-R	H03-F, F03-R	H04-F, F04-R	H05-F, F05-R	H06-F, F06-R	H07-F, F07-R	H08-F, F08-R	H09-F, F09-R	H10-F, F10-R	H11-F, F11-R	H12-F, F12-R

Table D-6. Platemap for DI04 used in this study.

DI04	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, F01-R	A02-F, F02-R	A03-F, F03-R	A04-F, F04-R	A05-F, F05-R	A06-F, F06-R	A07-F, F07-R	A08-F, F08-R	A09-F, F09-R	A10-F, F10-R	A11-F, F11-R	A12-F, F12-R
B	B01-F, G01-R	B02-F, G02-R	B03-F, G03-R	B04-F, G04-R	B05-F, G05-R	B06-F, G06-R	B07-F, G07-R	B08-F, G08-R	B09-F, G09-R	B10-F, G10-R	B11-F, G11-R	B12-F, G12-R
C	C01-F, H01-R	C02-F, H02-R	C03-F, H03-R	C04-F, H04-R	C05-F, H05-R	C06-F, H06-R	C07-F, H07-R	C08-F, H08-R	C09-F, H09-R	C10-F, H10-R	C11-F, H11-R	C12-F, H12-R
D	D01-F, A01-R	D02-F, A02-R	D03-F, A03-R	D04-F, A04-R	D05-F, A05-R	D06-F, A06-R	D07-F, A07-R	D08-F, A08-R	D09-F, A09-R	D10-F, A10-R	D11-F, A11-R	D12-F, A12-R
E	E01-F, B01-R	E02-F, B02-R	E03-F, B03-R	E04-F, B04-R	E05-F, B05-R	E06-F, B06-R	E07-F, B07-R	E08-F, B08-R	E09-F, B09-R	E10-F, B10-R	E11-F, B11-R	E12-F, B12-R
F	F01-F, C01-R	F02-F, C02-R	F03-F, C03-R	F04-F, C04-R	F05-F, C05-R	F06-F, C06-R	F07-F, C07-R	F08-F, C08-R	F09-F, C09-R	F10-F, C10-R	F11-F, C11-R	F12-F, C12-R
G	G01-F, D01-R	G02-F, D02-R	G03-F, D03-R	G04-F, D04-R	G05-F, D05-R	G06-F, D06-R	G07-F, D07-R	G08-F, D08-R	G09-F, D09-R	G10-F, D10-R	G11-F, D11-R	G12-F, D12-R
H	H01-F, E01-R	H02-F, E02-R	H03-F, E03-R	H04-F, E04-R	H05-F, E05-R	H06-F, E06-R	H07-F, E07-R	H08-F, E08-R	H09-F, E09-R	H10-F, E10-R	H11-F, E11-R	H12-F, E12-R

Table D-7. Platemap for DI05 used in this study.

DI05	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, E01-R	A02-F, E02-R	A03-F, E03-R	A04-F, E04-R	A05-F, E05-R	A06-F, E06-R	A07-F, E07-R	A08-F, E08-R	A09-F, E09-R	A10-F, E10-R	A11-F, E11-R	A12-F, E12-R
B	B01-F, F01-R	B02-F, F02-R	B03-F, F03-R	B04-F, F04-R	B05-F, F05-R	B06-F, F06-R	B07-F, F07-R	B08-F, F08-R	B09-F, F09-R	B10-F, F10-R	B11-F, F11-R	B12-F, F12-R
C	C01-F, G01-R	C02-F, G02-R	C03-F, G03-R	C04-F, G04-R	C05-F, G05-R	C06-F, G06-R	C07-F, G07-R	C08-F, G08-R	C09-F, G09-R	C10-F, G10-R	C11-F, G11-R	C12-F, G12-R
D	D01-F, H01-R	D02-F, H02-R	D03-F, H03-R	D04-F, H04-R	D05-F, H05-R	D06-F, H06-R	D07-F, H07-R	D08-F, H08-R	D09-F, H09-R	D10-F, H10-R	D11-F, H11-R	D12-F, H12-R
E	E01-F, A01-R	E02-F, A02-R	E03-F, A03-R	E04-F, A04-R	E05-F, A05-R	E06-F, A06-R	E07-F, A07-R	E08-F, A08-R	E09-F, A09-R	E10-F, A10-R	E11-F, A11-R	E12-F, A12-R
F	F01-F, B01-R	F02-F, B02-R	F03-F, B03-R	F04-F, B04-R	F05-F, B05-R	F06-F, B06-R	F07-F, B07-R	F08-F, B08-R	F09-F, B09-R	F10-F, B10-R	F11-F, B11-R	F12-F, B12-R
G	G01-F, C01-R	G02-F, C02-R	G03-F, C03-R	G04-F, C04-R	G05-F, C05-R	G06-F, C06-R	G07-F, C07-R	G08-F, C08-R	G09-F, C09-R	G10-F, C10-R	G11-F, C11-R	G12-F, C12-R
H	H01-F, D01-R	H02-F, D02-R	H03-F, D03-R	H04-F, D04-R	H05-F, D05-R	H06-F, D06-R	H07-F, D07-R	H08-F, D08-R	H09-F, D09-R	H10-F, D10-R	H11-F, D11-R	H12-F, D12-R

Table D-8. Platemap for DI06 used in this study.

DI06	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, D01-R	A02-F, D02-R	A03-F, D03-R	A04-F, D04-R	A05-F, D05-R	A06-F, D06-R	A07-F, D07-R	A08-F, D08-R	A09-F, D09-R	A10-F, D10-R	A11-F, D11-R	A12-F, D12-R
B	B01-F, E01-R	B02-F, E02-R	B03-F, E03-R	B04-F, E04-R	B05-F, E05-R	B06-F, E06-R	B07-F, E07-R	B08-F, E08-R	B09-F, E09-R	B10-F, E10-R	B11-F, E11-R	B12-F, E12-R
C	C01-F, F01-R	C02-F, F02-R	C03-F, F03-R	C04-F, F04-R	C05-F, F05-R	C06-F, F06-R	C07-F, F07-R	C08-F, F08-R	C09-F, F09-R	C10-F, F10-R	C11-F, F11-R	C12-F, F12-R
D	D01-F, G01-R	D02-F, G02-R	D03-F, G03-R	D04-F, G04-R	D05-F, G05-R	D06-F, G06-R	D07-F, G07-R	D08-F, G08-R	D09-F, G09-R	D10-F, G10-R	D11-F, G11-R	D12-F, G12-R
E	E01-F, H01-R	E02-F, H02-R	E03-F, H03-R	E04-F, H04-R	E05-F, H05-R	E06-F, H06-R	E07-F, H07-R	E08-F, H08-R	E09-F, H09-R	E10-F, H10-R	E11-F, H11-R	E12-F, H12-R
F	F01-F, A01-R	F02-F, A02-R	F03-F, A03-R	F04-F, A04-R	F05-F, A05-R	F06-F, A06-R	F07-F, A07-R	F08-F, A08-R	F09-F, A09-R	F10-F, A10-R	F11-F, A11-R	F12-F, A12-R
G	G01-F, B01-R	G02-F, B02-R	G03-F, B03-R	G04-F, B04-R	G05-F, B05-R	G06-F, B06-R	G07-F, B07-R	G08-F, B08-R	G09-F, B09-R	G10-F, B10-R	G11-F, B11-R	G12-F, B12-R
H	H01-F, C01-R	H02-F, C02-R	H03-F, C03-R	H04-F, C04-R	H05-F, C05-R	H06-F, C06-R	H07-F, C07-R	H08-F, C08-R	H09-F, C09-R	H10-F, C10-R	H11-F, C11-R	H12-F, C12-R

Table D-9. Platemap for DI07 used in this study.

DI07	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, C01-R	A02-F, C02-R	A03-F, C03-R	A04-F, C04-R	A05-F, C05-R	A06-F, C06-R	A07-F, C07-R	A08-F, C08-R	A09-F, C09-R	A10-F, C10-R	A11-F, C11-R	A12-F, C12-R
B	B01-F, D01-R	B02-F, D02-R	B03-F, D03-R	B04-F, D04-R	B05-F, D05-R	B06-F, D06-R	B07-F, D07-R	B08-F, D08-R	B09-F, D09-R	B10-F, D10-R	B11-F, D11-R	B12-F, D12-R
C	C01-F, E01-R	C02-F, E02-R	C03-F, E03-R	C04-F, E04-R	C05-F, E05-R	C06-F, E06-R	C07-F, E07-R	C08-F, E08-R	C09-F, E09-R	C10-F, E10-R	C11-F, E11-R	C12-F, E12-R
D	D01-F, F01-R	D02-F, F02-R	D03-F, F03-R	D04-F, F04-R	D05-F, F05-R	D06-F, F06-R	D07-F, F07-R	D08-F, F08-R	D09-F, F09-R	D10-F, F10-R	D11-F, F11-R	D12-F, F12-R
E	E01-F, G01-R	E02-F, G02-R	E03-F, G03-R	E04-F, G04-R	E05-F, G05-R	E06-F, G06-R	E07-F, G07-R	E08-F, G08-R	E09-F, G09-R	E10-F, G10-R	E11-F, G11-R	E12-F, G12-R
F	F01-F, H01-R	F02-F, H02-R	F03-F, H03-R	F04-F, H04-R	F05-F, H05-R	F06-F, H06-R	F07-F, H07-R	F08-F, H08-R	F09-F, H09-R	F10-F, H10-R	F11-F, H11-R	F12-F, H12-R
G	G01-F, A01-R	G02-F, A02-R	G03-F, A03-R	G04-F, A04-R	G05-F, A05-R	G06-F, A06-R	G07-F, A07-R	G08-F, A08-R	G09-F, A09-R	G10-F, A10-R	G11-F, A11-R	G12-F, A12-R
H	H01-F, B01-R	H02-F, B02-R	H03-F, B03-R	H04-F, B04-R	H05-F, B05-R	H06-F, B06-R	H07-F, B07-R	H08-F, B08-R	H09-F, B09-R	H10-F, B10-R	H11-F, B11-R	H12-F, B12-R

Table D-10. Platemap for DI08 used in this study.

DI08	01	02	03	04	05	06	07	08	09	10	11	12
A	A01-F, B01-R	A02-F, B02-R	A03-F, B03-R	A04-F, B04-R	A05-F, B05-R	A06-F, B06-R	A07-F, B07-R	A08-F, B08-R	A09-F, B09-R	A10-F, B10-R	A11-F, B11-R	A12-F, B12-R
B	B01-F, C01-R	B02-F, C02-R	B03-F, C03-R	B04-F, C04-R	B05-F, C05-R	B06-F, C06-R	B07-F, C07-R	B08-F, C08-R	B09-F, C09-R	B10-F, C10-R	B11-F, C11-R	B12-F, C12-R
C	C01-F, D01-R	C02-F, D02-R	C03-F, D03-R	C04-F, D04-R	C05-F, D05-R	C06-F, D06-R	C07-F, D07-R	C08-F, D08-R	C09-F, D09-R	C10-F, D10-R	C11-F, D11-R	C12-F, D12-R
D	D01-F, E01-R	D02-F, E02-R	D03-F, E03-R	D04-F, E04-R	D05-F, E05-R	D06-F, E06-R	D07-F, E07-R	D08-F, E08-R	D09-F, E09-R	D10-F, E10-R	D11-F, E11-R	D12-F, E12-R
E	E01-F, F01-R	E02-F, F02-R	E03-F, F03-R	E04-F, F04-R	E05-F, F05-R	E06-F, F06-R	E07-F, F07-R	E08-F, F08-R	E09-F, F09-R	E10-F, F10-R	E11-F, F11-R	E12-F, F12-R
F	F01-F, G01-R	F02-F, G02-R	F03-F, G03-R	F04-F, G04-R	F05-F, G05-R	F06-F, G06-R	F07-F, G07-R	F08-F, G08-R	F09-F, G09-R	F10-F, G10-R	F11-F, G11-R	F12-F, G12-R
G	G01-F, H01-R	G02-F, H02-R	G03-F, H03-R	G04-F, H04-R	G05-F, H05-R	G06-F, H06-R	G07-F, H07-R	G08-F, H08-R	G09-F, H09-R	G10-F, H10-R	G11-F, H11-R	G12-F, H12-R
H	H01-F, A01-R	H02-F, A02-R	H03-F, A03-R	H04-F, A04-R	H05-F, A05-R	H06-F, A06-R	H07-F, A07-R	H08-F, A08-R	H09-F, A09-R	H10-F, A10-R	H11-F, A11-R	H12-F, A12-R

D.6 Supplemental Tables

Table D-11. evSeq captures off-target mutations. This table is derived from the "AminoAcids_Coupled_Max.csv" output file from evSeq for the TrpB run, and shows all confident (defined as ≥0.80 alignment frequency and ≥10 total reads) unexpected mutations captured by evSeq; some columns have been removed. Note in the "VariantCombo" column that the amino acid at the expected mutagenized position has a "?" as the original amino acid—this is because the evSeq run generating this data was told the variable positions with the "NNN" convention. For unexpected variable positions, both the original amino acid and the new amino acid are shown.

IndexPlate	Plate	Well	VariantCombo	AlignmentFrequency	WellSeqDepth
DI02	Lib2_118X	E03	?118V_D164G	0.964286	28
DI04	Lib4_166X	B02	P154S_?166Q	0.977011	87
DI08	Lib8_301X	H11	G250D_?301L	0.99537	216

Table D-12. Primer sequences for TrpB saturation mutagenesis library construction.

Site	Direction	Sequence
105	Forward	GGCAAAACCCGTATCATTGCTNNNACGGGTGCTGGTCAGCAC
105	Reverse	AGCAATGATACGGGTTTTGCCCATTAGTTTTGCCAGCAGAACCTGGC
118	Forward	GGCGTAGCAACTGCTACCNNNGCAGCGCTGTTCGGTATGGAATGTGTAATCTATATGG
118	Reverse	GGTAGCAGTTGCTACGCCGTGCTGACCAGC
162	Forward	GTAAAATCCGGTAGCCGTACCNNNAAAGACGCAATTGACGAAGCTCTG
162	Reverse	GGTACGGCTACCGGATTTTACCGGTACAACTTTAGCACCCAGCAG
166	Forward	CGTACCCTGAAAGACGCANNNGACGAAGCTCTGCGTGACTGGATTACCAACC
166	Reverse	TGCGTCTTTCAGGGTACGGCTACCGGATTTTACCGG
184	Forward	CTGCAGACCACCTATTACGTGNNNGGCTCTGTGGTTGGTCC
184	Reverse	CACGTAATAGGTGGTCTGCAGGTTGGTAATCCAGTCACGCAGAGCT
228	Forward	TACATCGTTGCGTGCGTGNNNGGTGGTTCTAACGCTGCC
228	Reverse	CACGCACGCAACGATGTAGTCCGGCAGACGGCCTTCT
292	Forward	GATGACTGGGGTCAAGTTCAGGTGNNNCACTCCGTCTCCGCTG
292	Reverse	CACCTGAACTTGACCCCAGTCATCCTGCAGAACGAACGTCTTAGAACCG
301	Forward	TCCGCTGGCCTGGACNNNTCCGGTGTCGGTCCGGA
301	Reverse	GTCCAGGCCAGCGGAGACGGAGTGGCTCACCTGAACT

Table D-13. Primers specific to the ampicillin resistance gene of pET22b(+) used in TrpB library construction.

Site	Direction	Sequence
AmpR	Forward	CCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGC
AmpR	Reverse	CGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAG

Table D-14. Inner primers used for evSeq library preparation from the TrpB site-saturation mutagenesis libraries.

Name	Direction	Sites	Sequence
evSeq_102_f	Forward	105, 118, 162, 166, 184	CACCCAAGACCACTCTCCGGGCAAAACTAATGGGCAAAACCCG
evSeq_184_r	Reverse	105, 118, 162, 166, 184	CGGTGTGCGAAGTAGGTGCGATGCGGACCAACCACAGAG
evSeq_226_f	Forward	228, 292, 301	CACCCAAGACCACTCTCCGGGCCGGACTACATCGTTGCG
evSeq_304_r	Reverse	228, 292, 301	CGGTGTGCGAAGTAGGTGCCAATAGGCGTGTTCCGGACC

Table D-15. The evSeq barcode plates used for sequencing each position of the TrpB site-saturation mutagenesis libraries.

Position targeted	Barcode plate
105	DI01
118	DI02
162	DI03
166	DI04
184	DI05
228	DI06
292	DI07
301	DI08

Table D-16. Mutagenic primers used for the construction of the *Rma*NOD four-site-saturation library. Note that the names of the primers are delimited by "-" and that the delimited sections reflect the mutagenized positions, the degenerate codons at those positions, and the direction of the primer on the template DNA ([Positions]-[Codon1]-[Codon2]-[Direction]).

Name	Sequence
S28M31-NDT-NDT-F	AAACACTCAGTCGCTATTNDTGCCACGNDTGGTCGGCTGCTTTTCG
S28M31-NDT-VHG-F	AAACACTCAGTCGCTATTNDTGCCACGVHGGGTCGGCTGCTTTTCG
S28M31-NDT-TGG-F	AAACACTCAGTCGCTATTNDTGCCACGTGGGGTCGGCTGCTTTTCG
S28M31-VHG-NDT-F	AAACACTCAGTCGCTATTVHGGCCACGNDTGGTCGGCTGCTTTTCG
S28M31-VHG-VHG-F	AAACACTCAGTCGCTATTVHGGCCACGVHGGGTCGGCTGCTTTTCG
S28M31-VHG-TGG-F	AAACACTCAGTCGCTATTVHGGCCACGTGGGGTCGGCTGCTTTTCG
S28M31-TGG-NDT-F	AAACACTCAGTCGCTATTTGGGCCACGNDTGGTCGGCTGCTTTTCG
S28M31-TGG-VHG-F	AAACACTCAGTCGCTATTTGGGCCACGVHGGGTCGGCTGCTTTTCG
S28M31-TGG-TGG-F	AAACACTCAGTCGCTATTTGGGCCACGTGGGGTCGGCTGCTTTTCG
Q52L56-AHN-AHN-R	GGCCAACAGGGCCGACGCAHNCTTGTGTATAHNTCTCTCAGGAAGTTCAAACAAG
Q52L56-AHN-CDB-R	GGCCAACAGGGCCGACGCAHNCTTGTGTATCDBTCTCTCAGGAAGTTCAAACAAG
Q52L56-AHN-CCA-R	GGCCAACAGGGCCGACGCAHNCTTGTGTATCCATCTCTCAGGAAGTTCAAACAAG
Q52L56-CDB-AHN-R	GGCCAACAGGGCCGACGCCDBCTTGTGTATAHNTCTCTCAGGAAGTTCAAACAAG
Q52L56-CDB-CDB-R	GGCCAACAGGGCCGACGCCDBCTTGTGTATCDBTCTCTCAGGAAGTTCAAACAAG
Q52L56-CDB-CCA-R	GGCCAACAGGGCCGACGCCDBCTTGTGTATCCATCTCTCAGGAAGTTCAAACAAG
Q52L56-CCA-AHN-R	GGCCAACAGGGCCGACGCCCACTTGTGTATAHNTCTCTCAGGAAGTTCAAACAAG
Q52L56-CCA-CDB-R	GGCCAACAGGGCCGACGCCCACTTGTGTATCDBTCTCTCAGGAAGTTCAAACAAG
Q52L56-CCA-CCA-R	GGCCAACAGGGCCGACGCCCACTTGTGTATCCATCTCTCAGGAAGTTCAAACAAG

Table D-17. Additional primers used to build flanking fragments during construction of the four-site-saturation *Rma*NOD library.

Flanking Fragment	Primer Type	Primer Name	Sequence
0	Forward	Universal-F	CCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGC
0	Reverse	S28M31_Const-R	AATAGCGACTGAGTGTTTCTGCAGTGCAGGCAC
1	Forward	L56_Const-F	GCGTCGGCCCTGTTGGCCTACGCCCGTAGTATCGACAACCC
1	Reverse	Universal-R	CGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAG

Table D-18. Inner primers used for evSeq library preparation from the *Rma*NOD four-site-saturation mutagenesis library.

Plates	Forward primer	Reverse Primer
All plates	CACCCAAGACCACTCTCCGGCACTGCAGAAACACTCAGTCG	CGGTGTGCGAAGTAGGTGCACTACGGGCGTAGGCCAAC

Table D-19. The evSeq barcode plates used for sequencing each position of the *Rma*NOD four-site-saturation mutagenesis library.

Position targeted	Barcode plate
Plate #1	DI01
Plate #2	DI02
Plate #3	DI03
Plate #4	DI04
Plate #5	DI05

Appendix E

SUPPLEMENTARY INFORMATION FOR CHAPTER VI

E.1 Material and Methods

E.1.1 General experimental methods.

Chemicals were purchased from commercial sources and used without additional purification. Analytical LCMS was performed on an Agilent 1260 Infinity II LC/MSD iQ equipped with a reversed-phase Poroshell 120 EC-C18, 4.6x50 mm, 2.7 µm column using a gradient of H_2O/MeCN with 0.1% acetic acid by volume. Unless otherwise stated, the gradient applied was 1–95% MeCN over 3 minutes, then held for 0.5 minutes, then immediately dropped to 1% MeCN for 0.5 minutes. NMR spectra were collected on a Bruker 400 MHz (100 MHz) spectrometer equipped with a cryogenic probe. Proton chemical shifts are reported in parts per million (ppm) relative to tetramethylsilane and calibrated using the residual solvent resonance (H_2O/HDO in D_2O, 4.79 ppm). Regioselectivities were determined by LCMS comparison to known standards when they could be chromatographically separated and by performing NMR experiments when sufficient product could be isolated. Preparative reversed-phase chromatography was used to isolate Tyr analog products on a Biotage Isolera One purification system equipped with a C-18 column, using acidified H_2O (0.01% of either HCl by weight when HCl salt was to be isolated, or acetic acid by volume when the pure product was to be isolated) as the weak solvent and MeCN as the strong solvent.

E.1.2 Cloning, expression, and preparation of enzyme catalysts.

Genes coding for all enzyme variants were cloned into pET22b(+) between the *Nde*I and *Xho*I restriction sites, in frame with the Lac-inducible T7 promotor and C-terminal 6xHis tag for expression in transformed BL21(DE3) *Escherichia coli* cells. Single colonies of *E. coli* harboring gene variants were isolated on Lysogeny Broth (LB) agar medium supplemented with 100 µg/mL carbenicillin. For large-scale expression, a single colony was transferred to 5 mL of LB with 100 µg/mL carbenicillin (LB_{carb}) and grown to stationary phase at 37 °C and 230 rpm. The culture was then diluted 1:250 into 250 mL Terrific Broth (TB) supplemented with 100 µg/mL carbenicillin (TB_{carb}) and grown for 6 hours at 37 °C at 250 rpm, yielding a dense, uninduced culture. Protein expression was then induced with 1 mM isopropyl β-d-thiogalactopyranoside (IPTG) and proceeded at 30 °C for 24 hours. Cells were harvested via centrifugation at >5,000*g* for 10 minutes, the supernatant discarded, and then the cells were stored at –20 °C until needed.

To purify, thawed pellets were resuspended to 10 mL with a lysis buffer containing 25 mM potassium phosphate, 100 mM NaCl, and 20 mM imidazole, pH 8.0 (Buffer A), then supplemented with 100 µM pyridoxal 5'-phosphate (PLP), 0.02 mg/mL DNase I, and BugBuster® at 1/10th the manufacturer's recommendation. Cell lysis proceeded at 37 °C for 1 hour with shaking at 220 rpm, at which point the lysate was heat treated at 75 °C for 30–60 minutes. The lysate was clarified by centrifugation for 20 minutes at 14,000*g*, and the supernatant was collected. The lysate was run over a gravity column prepared with 1–2.5 mL Ni-NTA Agarose (Qiagen) pre-equilibrated with Buffer A. (*Note*: in many cases the enzymes used in this study could not be purified on an FPLC system, as they appeared to form a strongly bound complex at the top of the column that over-pressurized FPLC systems.) The bound protein was then washed with 10 column volumes (CVs) of Buffer A, and protein was eluted with a mixture of 50% Buffer A and 50% 25 mM potassium phosphate, 100 mM NaCl, and 500 mM imidazole, pH 8.0 (Buffer B) and collected. An additional 1 mM PLP was added to the collected protein solution to ensure full cofactor incorporation, and the solution was then buffer exchanged into 50 mM potassium phosphate, pH 8.0 (KPi; *Note:* "KPi" used in this text always refers to 50 mM, pH 8.0) by dialysis. The purified protein was then flash frozen in ~20-μL drops in liquid nitrogen and stored at –80 °C. Protein concentrations were determined using the Pierce™ BCA Protein Assay Kit (ThermoFisher) according to the manufacturer's recommendations.

Alternatively, protein catalyst could be prepared as a heat-treated lysate and used directly for preparative-scale reactions, obviating the need for any chromatography. In these instances, thawed cell pellets were resuspended in a volume of KPi containing 100 μM PLP that was appropriate for the given reaction, usually 50–100 mL. The dilute resuspension was then heat treated at 75 °C for >1 hour to efficiently lyse the cells and denature the *E. coli* proteins, and then clarified by centrifugation at 14,000*g* for 20 minutes. (*Note*: more concentrated resuspensions result in lower efficiency lysis by heat treatment alone and should be supplemented with lysozyme or BugBuster® performed similarly to the lysis stages for purifying protein, withholding the further chromatography and, unless extra purity is desired, buffer exchange steps.)

For plate-based expression, single colonies of a desired variant or library were transferred into the wells of a 96-well plate containing 300 μL LB_{carb}. The cultures were covered with a sterile, breathable film and grown to stationary phase at 37 °C and 220 rpm. From these plates, 20 µL of each culture were transferred to new 96-well plates containing 630 µL of TB_{carb} and grown for 6 hours at 37 °C and 220 rpm. Protein expression was induced with the addition of 50 µL of 14 mM IPTG in TB_{carb} for a final concentration of 1 mM IPTG in a total volume of 700 µL of TB_{carb}. Expression proceeded for 24 hours at 30 °C, 250 rpm, at which point cells were harvested by centrifugation at 4,500*g* for 10 minutes, discarding the supernatant, and, unless preparing lysate immediately, covered with a non-breathable film and frozen at –20 °C until needed. Heat-treated lysate was prepared by resuspending thawed cells in 300 µL of KPi supplemented with 100 µM PLP, heat treating for 15–90 minutes (depending on the desired thermostability challenge) at 75 °C, and clarifying by centrifugation at 4,500*g* for 10 minutes. Lysate was then added directly to pre-prepared reaction plates as described below.

E.1.3 Error-prone PCR mutagenesis.

Error-prone PCR (epPCR) was performed by a modified *Taq* PCR to amplify the gene between the *Nde*I and *Xho*I restriction sites using the following primers:

Name	**Direction**	**Sequence**
*Nde*I_f	Forward	GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG
*Xho*I_r	Reverse	GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAG

And the following thermal cycle:

Step	**Temperature (°C)**	**Time**
1	95	5 min
2	95	30 s
3	55	30 s
4	72	90 s
5	Return to 2, 29 x	
6	72	5 min
7	10	Hold

Different concentrations of $MnCl_2$ (typically 200, 300, and 400 µM) were added to increase the error rate of the polymerase, resulting in libraries with different error rates. The PCR products were treated with *Dpn*I at 37 °C for 1 hour and then isolated via gel extraction, assembled into the pET22b(+) plasmid vector via Gibson Assembly®,[1] and used to transform chemically competent *E. coli* as described above.

To determine which library to screen in more depth, a single plate (88 variants) was screened (see below for screening details), looking for which gave the best balance of retention of enzyme function and sufficient genetic diversity. Once selected, additional variants of this library were screened until one or more variants were identified with improved activity, which were then used directly in a subsequent round of mutagenesis and screening or recombined (see below) to identify additive mutations.

E.1.4 Site-saturation mutagenesis.

Site-saturation mutagenesis (SSM) was performed via the "22-codon trick" as described previously[2] with a few modifications. Briefly, a forward primer template was designed at the selected site comprising three parts: an assembly region, the mutated site, and an annealing region. The assembly region is immediately upstream of the mutated site with a T_m of ~55 °C. The annealing region is immediately downstream of the mutated site with a T_m of ~68 °C, ideally ending on one or more G or C bases. From this template, three primers were ordered with the codons NDT, VHG, and TGG in place of the native codon at the mutated site, comprising 22 codons that cover all 20 amino acids (leucine and valine are sampled twice). A reverse primer was designed completely overlapping the assembly region of the forward primer (immediately adjacent to the mutated site) and extending to a final T_m of ~68 °C, again ending on one or more G or C bases. The secondary structures were examined to ensure that no strong monomeric or dimeric primer-primer interactions would interfere with the primer-template interactions and adjusted as necessary. Once ordered, the primers were used in a QuikChange^{TM}-like PCR using Phusion® polymerase as described previously,[3] isolated via gel extraction, and assembled via Gibson Assembly®. Alternatively, to reduce the chances for non-specifically assembled constructs, two fragments were generated with the SSM primers, splitting the plasmid template at the resistance cassette (ampicillin (AmpR) in pET22b(+), as used here) using the following primers:

Name	**Direction**	**Sequence**
AmpR_f	Forward	`CCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGC`
AmpR_r	Reverse	`CGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAG`

with the forward SSM primer paired with the reverse AmpR primer and the reverse SSM primer paired with the forward AmpR primer in pET22b(+). The fragments were generated via PCR with Phusion® polymerase, treated with *Dpn*I at 37 °C for 1 hour, and purified by gel extraction, then assembled again via Gibson Assembly®. Once assembled, the DNA was used to transform chemically competent *E. coli*. For each site targeted, a single plate of colonies (88 variants, providing 4-fold oversampling of the 22 codons and a 95% chance of complete library coverage in an unbiased library)[2] was screened (see below) and the identities of improved variants were confirmed by Sanger sequencing.

E.1.5 Recombination via Staggered Extension Process.

When improved variants were identified containing mutations at different sites, recombination via a modification of the Staggered Extension Process (StEP) PCR[4] was performed. In all cases, >250 ng of total plasmid DNA (usually 500 ng) was used per 20 µL

PCR in Standard Taq Buffer, with a final concentration of 50 nM forward and reverse primers flanking the gene for amplification. For a large number of variants (>20), cultures were grown to saturation in LB_{carb} and combined in equal volumes before isolating all plasmid DNA simultaneously. For a lower number of variants, plasmid DNA was isolated individually and then normalized. To recombine the variant genes, six identical reactions were created and placed along a temperature gradient in an Eppendorf Mastercycler X50 (96-well silver block) and run on the following thermal cycle:

Step	**Temperature (°C)**	**Time**
1	95	20 s
2	95	5 s
3	55	2 s
4	Gradient, 50–72	1 s
5	Return to 2, 120 x	
6	68	5 min
7	4	Hold

The template plasmid was digested with the addition of 1 μL *Dpn*I at 37 °C for at least 1 hour and analyzed by gel electrophoresis. The reaction with the lowest temperature along the gradient that gave a discrete PCR product (usually, but not always, 50 °C) was then used as template for a subsequent amplification PCR with Phusion® polymerase. The PCR product containing recombined variants was isolated by gel extraction, assembled into pET22b(+), and used for expression as described above for error-prone PCR mutagenesis, then screened as described below to identify variants with improved activity.

E.1.6 Absorbance-based screening.

Enzyme variants were tested for activity by combining heat-treated lysate (prepared as described above) with L-serine (Ser) and an appropriate nucleophile (e.g., indole or a phenol analog), along with 5% EtOH by volume as cosolvent, directly into a UV-transparent assay plate and measuring the change in absorbance over time at a given wavelength that has been validated to increase over the course of the reaction. For example, the absorbance of 1-naphthol increases as it is converted to the ncAA product at wavelengths between 284 nm and ~350 nm (**Figure SE-5, a** and **b**). For reactions containing 5 mM 1-naphthol, the change in absorbance at 335 nm increases over the course of the reaction and is directly proportional to product formation. Changes in activity were therefore quantified by looking at differences in absorbance at a given timepoint in the linear range of the reaction (steady state) if the reaction is continuously monitored, or at the endpoint if it occurs before the reaction has slowed. For continuously monitored reactions, initial rates could also be determined for each variant and used to determine the change in activity.

E.1.7 LCMS screening.

Enzyme variants could also be tested for activity via LCMS when conditions could not be optimized for absorbance-based screening, such as for low levels activity or poorly absorbing nucleophiles. In this case, heat-treated lysate was combined with Ser and the phenol analog, along with 5% EtOH by volume as cosolvent, and allowed to react overnight (typically ~18 hours) at 37 °C. The reactions were worked up with 300 µL of 1:1 1 M aq. HCl/MeCN and then filtered through a 0.2-µm 96-well filter plate (Pall AcroPrep #8019) via centrifugation at 5,000*g* until the soluble reaction components were collected in the wells of a 96-well assay plate. This plate was then sealed and run on a suitable LCMS method and column that can separate reaction components sufficiently for quantification by UV and/or MS. See **Section E.2.1**, *Evolutionary Strategies* for more details.

E.1.8 Analytical-scale vial reactions.

Analytical reactions were performed in 2-mL glass vials in a total reaction volume of 200 μL. Vials were first charged with 10 μL of 20X stock of the nucleophile in EtOH (final concentration of 5% EtOH by volume), to which 190 μL of a mixture of Ser and purified enzyme in KPi were added. The reactions were generally protected from light (primarily 1-naphthol, which is a photoacid) and allowed to react at 37 °C. At the end of the reaction time, the 200-μL reactions were worked up with 800 μL of 1:1 1 M aq. HCl/MeCN, transferred to a microcentrifuge tube, and clarified by centrifugation at 14,000*g*. A 200-μL aliquot of this mixture was then collected and analyzed via LCMS as described in the general experimental methods.

E.1.9 Analytical-scale plate reactions.

Analytical-scale reactions were also carried out in plates using a specified volume of heat-treated lysate in place of a known final concentration of purified enzyme. Because screening used heat-treated lysate where the concentration of enzyme was not known or measured, improvements in expression, stability, or other factors were allowed to manifest as improvements in the catalyst. These reactions were performed identically to the vial reaction specified above, without controlling for enzyme concentration. At the end of the reaction time, the 200-μL reactions were worked up with either 300 or 800 μL of 1:1 1 M aq. HCl/MeCN (depending on the expected yield of the reaction) and then filtered through a 0.2 μm 96-well filter plate (Pall AcroPrep #8019) via centrifugation at 5,000*g* until the soluble reaction components are collected in the wells of a 96-well LCMS assay plate. This plate was sealed and analyzed via LCMS as described in the general experimental methods.

E.1.10 Preparative-scale reactions.

Preparation of Tyr analogs for characterization and further experiments was performed using either purified enzyme or a large volume of heat-treated lysate. First, a 1.1 molar equivalent of Ser was weighed into a flask followed by the phenol analog as the limiting reagent. The phenol analog was dissolved in EtOH (5% by volume final concentration), which was then mixed with an appropriate amount of KPi for the reaction volume. Solutions were incubated at the reaction temperature (typically 37 °C) in a water bath, followed by the addition of enzyme, reaching the desired final concentrations of all reaction components. The reactions were protected from light and allowed to react for up to three days, taking small samples as timepoints for reaction progress analysis by LCMS. Reactions were then concentrated *in vacuo* and the Tyr analog products isolated by reversed-phase chromatography. Collected fractions containing the Tyr analog were pooled and again concentrated *in vacuo* to afford a the final product.

E.1.11 Michaelis-Menten kinetics.

Enzyme kinetic parameters (k_{cat} and K_{M}) for the conversion of 1-naphthol to NaphAla by TmTyrS1 were inferred from initial rate measurements of a continuous colorimetric screen (**Figure E-5**). Briefly, a reaction containing 400 μM 1-naphthol and 10 μM 1-naphthol was monitored for 40 minutes, scanning infrequently (once every 2 minutes) to reduce photo-induced oxidation of 1-naphthol (**Figure E-5a**). An exponential function was used to model the reaction time course at 335 nm (**Figure E-5b**), and these parameters were used to obtain the total absorptivity change (ΔA) for the conversion of 400 μM 1-naphthol to 400 μM NaphAla (**Figure E-5c**). This in turn can be used to determine the molar absorptivity change at 335 nm ($\Delta\varepsilon_{335}$) for the conversion of 1-naphthol to NaphAla at pH 8.0. Using this conversion ratio, short time courses (0.2 sec intervals for 300 seconds) could be obtained at varying concentrations of 1-naphthol (400–5,000 μM) and 20 mM Ser (**Figure E-5d**) that correspond to rates of NaphAla formation. Correcting for enzyme concentration and modeling these rates to the Michaelis-Menten equation of enzyme kinetics provided estimations of k_{cat} and K_{M} (**Figure E-5e**).

For conversion of phenol to Tyr by TmTyrS6, parameters were inferred from LCMS-based rate measurements, due to the slow rate and small extinction coefficient provided by phenol. Small-scale analytical reactions were performed as described above with 10 μM TmTyrS6 and either varying phenol at 100 mM Ser or varying Ser at 50 mM phenol. Reactions were worked up at 4, 8 or 20 hours. For the 0.5, 1, and 2 mM concentrations of Ser, reactions were instead worked up at 0.5, 1, and 2 hours to better estimate these initial rates. Time courses were fit to either exponential or linear equations and the initial rates were obtained from these fits. For phenol, the rates for each concentration were modeled to the standard Michaelis-Menten equation of enzyme kinetics and provided estimations of k_{cat} and K_{M} (**Figure E-24a**). A moderate inhibitory effect was observed for Ser, and instead the substrate-inhibition model of enzyme kinetics was used to provide estimations of k_{cat}, K_{M}, and K_{i} (**Figure E-24b**). It should be noted that estimates of k_{cat} are likely underestimated due to necessary sub-saturating concentrations of phenol.

E.1.12 Determination of enantiopurity.

Enantiopurity was assessed using Nα-(5-fluoro-2,4-dinitrophenyl)-L-alaninamide (Marfey's reagent), a common chiral derivatization agent.[5] To assess the enantiopurity of an enzymatic reaction without purification, a small-scale analytical reaction was first carried out as described above. After the reaction time, 100 μL of 1 M aq. $NaHCO_3$ were added. Alternatively, for purified product, a ~10 mM solution in KPi was prepared which was then mixed 2:1 with 1 M aq. $NaHCO_3$. From these solutions, 125 μL were transferred into a new 2-mL vial along with 33 μL of a 33-mM solution of Marfey's reagent in acetone. This mixture was incubated at 37 °C and 220 rpm for 2 hours, and then the mixture was diluted with 600 μL 1:1 1 M aq HCl/MeCN.

The products of the reaction were analyzed by LCMS using the following gradient of MeCN: 25–45% over 7 minutes. Products were monitored by MS in single-ion mode selected for the expected molecular ion of the S_NAr product (e.g., 434 m/z for the Tyr product). This provided baseline separation of DL-tyrosine peaks (**Figure E-11**). Only one of these peaks (with the shorter retention time) was seen for L-tyrosine, along with the enzymatic product of TmTyrS5 and phenol (**Figure E-11**). Putative O-alkylation product of the tyrosine and Marfey's reagent, with the same mass as the desired N-alkylation product, was seen as an early single peak. A similar O-alkylation product of unreacted phenol and Marfey's reagent was also seen in enzymatic reactions with leftover phenolic substrate. In all cases, enzymatic products derivatized with Marfey's reagent were identified to have only a single peak.

To confirm enantiopurity in the absence of accessible D enantiomers of these compounds, the D enantiomer of Marfey's reagent was used to prepare a racemic mixture of this reagent. Enzymatic Tyr products were derivatized with racemic and enantiopure Marfey's reagent mixtures which resulted in two or one peaks, respectively, and confirmed the L configuration as the only observable enzymatic product.

E.1.13 Measuring kinetic isotope effects.

Kinetic isotope effects (KIEs) were measured from reactions as described in **Section E.1.8**, *Analytical-scale vial reactions* with a few minor changes. The cosolvent used was DMSO rather than EtOH, because stocks of the deuterated phenols were previously prepared in commercial DMSO-d_6 for long-term storage. KIEs were measured in direct competition with equimolar concentrations of the standard and deuterated substrate at a total concentration of 5 mM phenolic substrate and 50 mM Ser. Reactions were analyzed by LCMS, extracting the ions corresponding to the appropriate product masses and comparing their ratios. Ser KIEs were measured with Ser-d_3 using 2-chlorophenol as the phenolic substrate. (*Note*: the +2 isotope of chlorine yielded a 30% relative abundance of a 218 m/z product for 3-chloro-Tyr, which also corresponded with the di-deuterated product of Ser-d_3 and 2-chlorophenol. Reported ratios of 216/218 ion counts are therefore not exactly the 3-chloro-Tyr 216 m/z ion over the deuterated 3-chloro-Tyr 218 m/z ion, which would give a value of slightly lower than 1 for a true KIE of 1.)

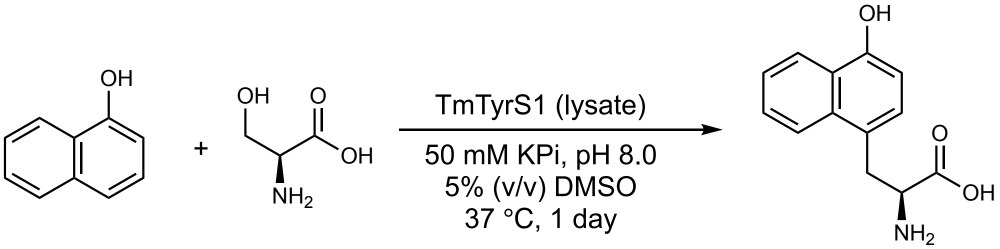
E.1.14 Preparation of lyophilized enzyme catalyst for large-scale reactions.

Heat-treated lysate was prepared as described in **Section E.1.2**, *Cloning, expression, and purification of enzyme catalysts*. For the large-scale reactions reported, dialysis of the lysate into KPi was performed to remove small-molecule impurities from the *E. coli* host cell. However, in many instances this step was not explicitly needed, as these impurities were removed during the washing steps when isolating the ncAA product (see below). Once prepared, the lysate was transferred into a tared 50-mL Falcon® conical tube and flash-frozen in liquid nitrogen while agitating. Once completely frozen, the tube was topped with a

Kimtech Science™ Kimwipe™ and lyophilized to dryness, resulting in a benchtop-stable powder. The mass of the powder was recorded (typically ~1 gram).

To quantify the activity of the powder, a small portion was removed and resuspended to a final concentration of 2 mg/mL in deionized water. The rate of conversion (mM/mg) of the desired substrate at concentrations similar to those to be used for the large-scale reactions was obtained at 1, 0.5, and 0.25 mg/mL powder. The expected amount of product produced per gram of powder was determined, and reaction conditions were scaled accordingly. The specific activities in these analytical reactions were uniformly comparable to the large-scale reactions.

E.1.15 Multi-gram-scale synthesis of NaphAla.



In an oven-dried 1-L 1-N round-bottom flask (RBF) equipped with a magnetic stir bar, Ser (8.02 g, 76.3 mmol) was dissolved in 510 mL of KPi (50 mM, pH = 8.0) and 30 mL of DMSO. The solution was sparged with argon for 30 minutes and a balloon of argon was placed over the reaction mixture. In a separate 50-mL Falcon® conical tube, lysate derived from TmTyrS1 (1.2 g powder from 1-L culture, which had a previously quantified activity per mg of powder) was gently dissolved in 60.0 mL of deionized water and added to the reaction while stirring. (*Note*: during addition, a liquid funnel was used as an argon dispersion funnel and the lysate solution was added by pouring directly into the flask). The

RBF was then dropped into a pre-warmed water bath at 37 °C and allowed to stir at 500 rpm, starting as a clear yellow-green solution (**Figure E-14a**). Meanwhile, 1-naphthol (5.00 g, 34.7 mmol) was dissolved in 45.0 mL of DMSO (resulting in a 50.0-mL solution) and added dropwise to the reaction mixture using a syringe pump (rate = 2 mL/hour). Following the addition of ~46.3 mL (~23 hours of dropwise addition) of the 1-naphthol solution, a white solid had crashed out of the solution (**Figure 6-3c** and **Figure E-14b**). The reaction mixture was then filtered through filter paper using a Buchner funnel (**Figure E-14c**) and washed with 500 mL of cold, deionized water followed by 500 mL of EtOAc. A white solid was collected from the filter pad, transferred to a 125-mL RBF, and dried under reduced pressure (on high vacuum) for 48 hours with mild heating (40 °C) to yield 5.5 g (74% yield) of NaphAla as a white solid (>99% HPLC purity over 1-naphthol, **Figure E-14d**). *Note:* The theoretical yield was adjusted to 32.2 mmol to take into account the amount of 1-naphthol added to the reaction mixture before the product crashed out and the reaction was ceased.

Note: An aliquot for ^{1}H NMR spectroscopy was prepared by dissolving ~5 mg of **NaphAla** in ~600 μL of D_2O and adding 3–5 drops of DCl. To obtain an aliquot for ^{13}C NMR spectroscopy, 6 mg of **NaphAla** were stirred in an ethereal solution of HCl (0.5 mL, 4 M in Et_2O) for 30 minutes. The Et_2O was subsequently removed under reduced pressure and the resultant white solid was dissolved in 600 μL of D_2O for NMR analysis.

^{1}H NMR (400 MHz, D_2O) δ 8.07 (dd, J = 8.8, 1.4 Hz, 1H), 7.88 – 7.81 (m, 1H), 7.50 (ddd, J = 8.5, 6.8, 1.5 Hz, 1H), 7.44 (ddd, J = 8.2, 6.8, 1.2 Hz, 1H), 7.13 (d, J = 7.8 Hz, 1H), 6.76

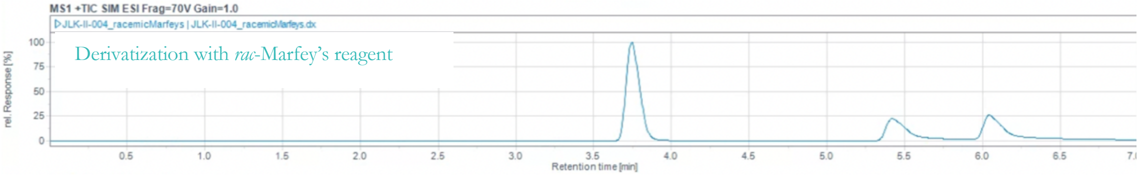
(d, J = 7.8 Hz, 1H), 4.24 (dd, J = 9.1, 5.7 Hz, 1H), 3.64 (dd, J = 14.8, 5.7 Hz, 1H), 3.26 (dd, J = 14.9, 9.2 Hz, 1H).

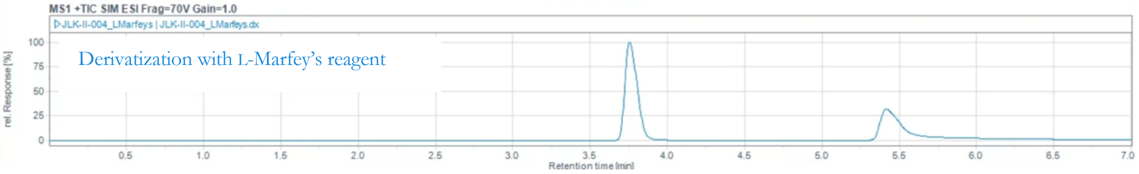
^{13}C NMR (101 MHz, D_2O) δ 172.00, 151.78, 132.25, 128.78, 127.33, 125.59, 124.96, 123.06, 122.45, 121.88, 108.22, 53.70, 32.96.

Full NMR provided in **Section E.5**, *NMR Spectra*.

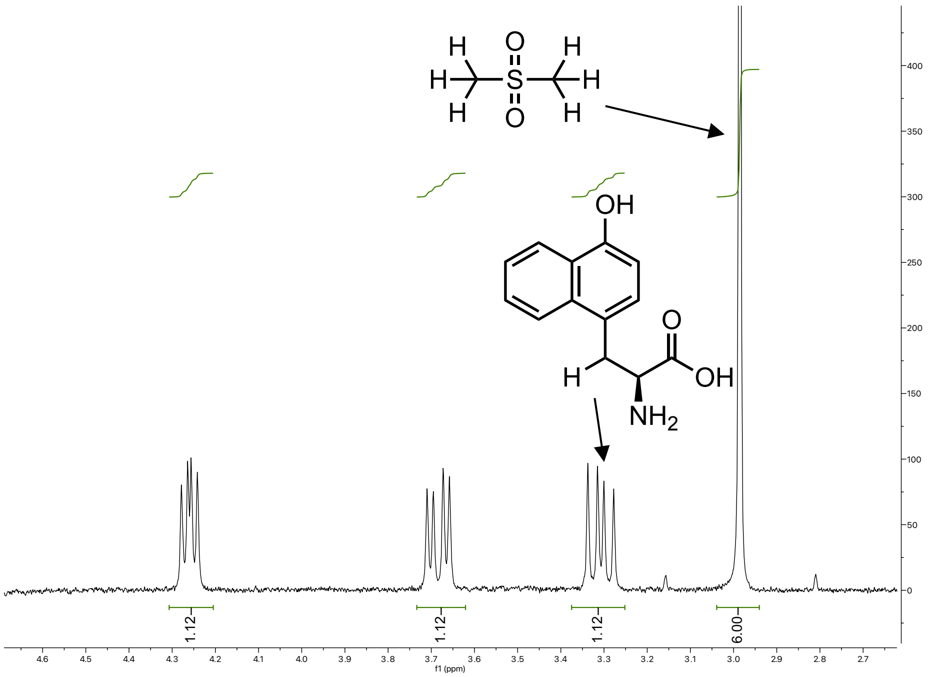
HRMS (FD+) Calculated for $C_{13}H_{13}NO_3$ (M^+): 231.08954; Found: 231.08980.

Determination of enantiopurity by chemical derivatization with Marfey's reagent. Enantiopurity was determined by derivatization with enantiopure (L) and racemic Marfey's reagent, as described in **Section E.1.12**, *Determination of enantiopurity*. Specifically, in a 1.5-mL Eppendorf tube, NaphAla (0.5 µmol) was dissolved in 1 M aq. $NaHCO_3$ (100 µL), to which 10 µL of a 33-mM solution of Marfey's reagent in acetone (0.33 µmol) were added. The vial was shaken for 2 h at 500 rpm, 37 °C. The reaction was allowed to cool to room temperature, then diluted with 1:1 1 M aq. HCl/MeCN (600 µL). The solution was analyzed via LCMS (25% to 45% MeCN, monitored by using single-ion mode for the molecular ion of the S_NAr product of 483 m/z). Absolute stereochemistry for NaphAla was inferred by analogy to L-tyrosine and determined to have >99.5% enantiomeric excess.

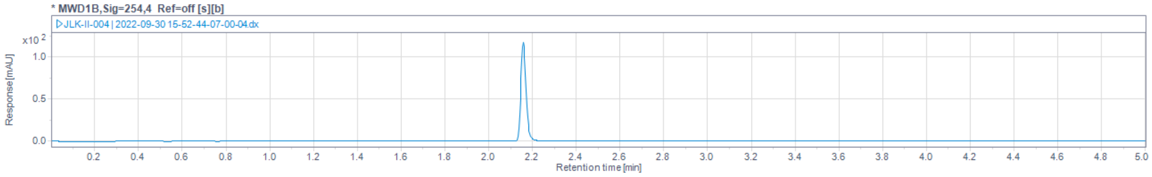




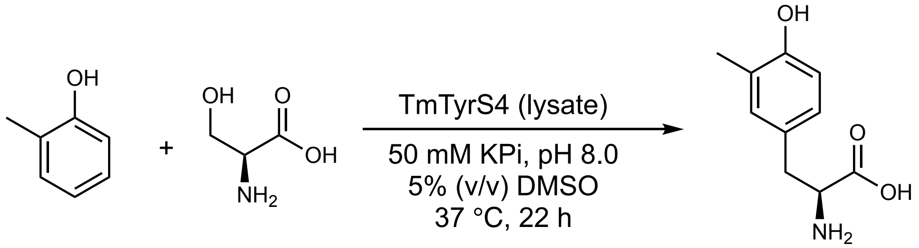
Determination of chemical purity using quantitative ^{1}H NMR spectroscopy. NaphAla (4.82 mg, 20.85 µmol) and dimethyl sulfone (1.63 mg, 17.32 µmol) were weighed into a 2-mL Eppendorf tube and dissolved in 1.5 mL of D_2O and 200 µL of DCl. An aliquot was removed and a ^{1}H NMR was obtained with a relaxation delay of 30 s. The chemical purity was determined to be 93% (relevant portion of ^{1}H NMR shown below). This procedure was performed in duplicate, and the reported chemical purity (91%) is an average of duplicate procedures. We suspect that weight impurities potentially include water and/or salts from the buffer/lysate. **NaphAla** is >99% pure of UV-absorbing chemical species (e.g., 1-naphthol) by HPLC analysis (see trace below).



HPLC-MS trace of isolated NaphAla at 254 nm.



E.1.16 Gram-scale synthesis of 3-methyl-Tyr.



To an oven-dried 250-mL 1-N RBF equipped with a magnetic stir bar, Ser (2.08 g, 19.8 mmol) was added. Subsequently, *o*-cresol (2-methylphenol, 0.659 g, 6.00 mmol, solution in 6.00 mL DMSO) was added, followed by 100 mL of KPi (50 mM, pH = 8.0). The RBF was then dropped into a pre-warmed oil bath at 37 °C. In a separate 50-mL Falcon® conical tube, lysate derived from TmTyrS4 (1.2 g powder from a 1-L culture, which had a previously quantified activity per mg of powder) was gently dissolved in 14.0 mL of deionized water and added to the reaction mixture using a Pasteur pipette. The reaction mixture was allowed to stir at 37 °C. After a 12-hour reaction time, an aliquot was removed and analyzed by LCMS, which indicated ~90% conversion to 3-methyl-Tyr relative to *o*-cresol (**Figure E-15a**). A second batch of *o*-cresol (0.6456 g, 5.878 mmol, solution in 6.00 mL DMSO) was then added and the reaction was allowed to continue stirring at 37 °C. After approximately 10 hours, a white solid crashed out of solution (**Figure E-15b**). The reaction mixture was filtered through filter paper using a Buchner funnel and washed with 75.0 mL of cold, deionized water followed by 75.0 mL of EtOAc. A white solid was collected from the filter pad, transferred to a 20-mL scintillation vial, and dried under reduced pressure (on high vacuum) for 24 hours with mild heating (40 °C) to yield 1.13 g (48.7% yield) of 3-methyl-Tyr as a white solid (**Figure E-15c**) with >99% HPLC purity over *o*-cresol (**Figure E-15d**).

Note: An aliquot for ^{1}H NMR spectroscopy was prepared by dissolving ~5 mg of **3-Me-Tyr** in ~600 µL of D_2O and adding 3–5 drops of DCl. To obtain an aliquot for ^{13}C NMR spectroscopy, 6 mg of **3-Me-Tyr** were stirred in an ethereal solution of HCl (0.5 mL, 4 M in Et_2O) for 30 minutes. The Et_2O was subsequently removed under reduced pressure and the resultant white solid was dissolved in 600 µL of D_2O for NMR analysis.

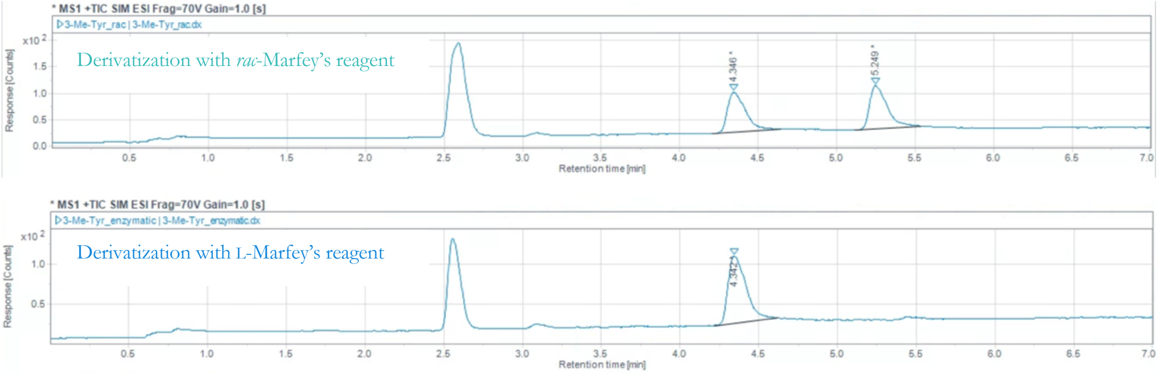
^{1}H NMR (400 MHz, D_2O) δ 6.97 – 6.92 (m, 1H), 6.87 (dd, *J* = 8.2, 2.4 Hz, 1H), 6.71 (d, *J* = 8.1 Hz, 1H), 4.17 (dd, *J* = 7.5, 5.6 Hz, 1H), 3.10 (dd, *J* = 14.7, 5.6 Hz, 1H), 2.98 (dd, *J* = 14.7, 7.5 Hz, 1H), 2.03 (s, 3H).

^{13}C NMR (101 MHz, D_2O) δ 171.78, 153.21, 131.99, 127.92, 125.78, 125.64, 115.49, 54.38, 34.79, 15.17.

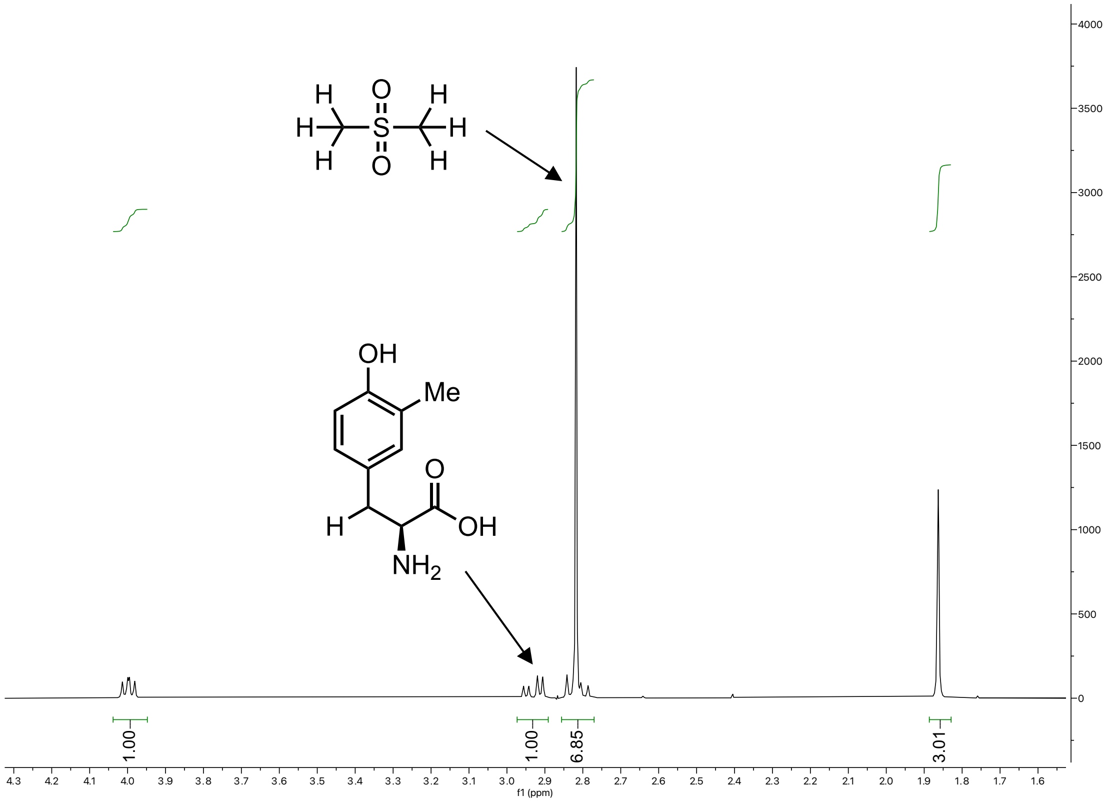
Full NMR provided in **Section E.5**, *NMR Spectra*.

HRMS (FD+) Calculated for $C_{10}H_{13}NO_3$ (M^+): 195.08954; Found: 195.08901.

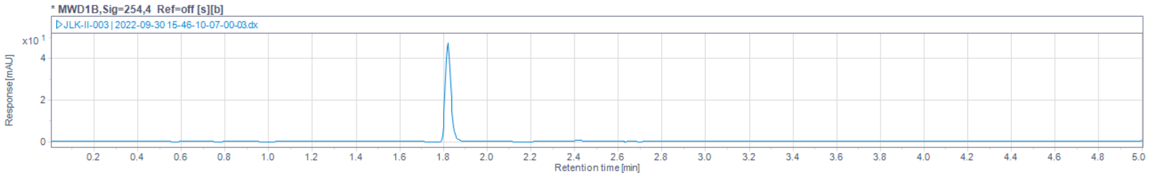
Determination of enantiopurity by chemical derivatization with Marfey's reagent. Enantiopurity was determined by derivatization with enantiopure (L) and racemic Marfey's reagent as described above. Specifically, in a 1.5-mL Eppendorf tube, 3-Me-Tyr (0.5 µmol) was dissolved in 1 M aq. $NaHCO_3$ (100 µL), to which 10 µL of a 33-mM solution of Marfey's reagent in acetone (0.33 µmol) was added. The vial was shaken for 2 h at 500 rpm, 37 °C. The reaction was allowed to cool to room temperature, then diluted with 1:1 1 M aq. HCl/MeCN (600 µL). The solution was analyzed via LCMS (25% to 45% MeCN, monitored by using single-ion mode for the molecular ion of the S_NAr product of 447 m/z). Absolute stereochemistry for 3-Me-Tyr was inferred by analogy to L-tyrosine and determined to have >99.5% enantiomeric excess.



Determination of chemical purity using quantitative ^{1}H NMR spectroscopy. 3-Me-Tyr (4.31 mg, 22.09 µmol) and dimethyl sulfone (2.02 mg, 21.46 µmol) were weighed into a 2-mL Eppendorf tube and dissolved in 1.5 mL of D_2O and 200 µL of DCl. An aliquot was removed and a ^{1}H NMR was obtained with a relaxation delay of 30 s. The chemical purity was determined to be 92% (relevant portion of ^{1}H NMR shown below). The dimethylsulfone peak overlaps with one of the benzylic protons; therefore, 1.0 was subtracted from the integration of the standard. This procedure was performed in duplicate, and the reported chemical purity (89%) is an average of duplicate procedures. We suspect that weight impurities potentially include water and/or salts from the buffer/lysate. **3-Me-Tyr** is >99% pure of UV-absorbing chemical species (*e.g.*, *o*-cresol) by HPLC analysis (see trace below).



LCMS trace of isolated 3-Me-Tyr at 254 nm.



E.1.17 Crystallization of TmTyrS1.

For the crystallization of TmTyrS1, protein was purified as described above. For initial screening, protein was thawed from –80 °C to room temperature and diluted to 10 and 20 mg/mL in storage buffer (KPi). Using a Crystal Gryphon robot (Art Robbins Instruments), sparse matrix screening was performed using the Wizard HT 1 & 2 (Rigaku), JCSG+ (Molecular Dimensions), Index and PEGRx (Hampton Research) crystallization screens in Intelli-Plate 96-2 drop crystallization plates (Art Robbins Instruments) using 0.2 μL drops of precipitant followed by 0.2 μL of protein solution. Plates were sealed with transparent adhesive covers and incubated at room temperature. After 2 days, crystals were observed in well C3 of the Wizard Screen (1.2 M NaH_2PO_4/0.8 M K_2HPO_4, 0.1 M *N*-cyclohexyl-3-aminopropanesulfonic acid (CAPS), 0.2 M Li_2SO_4), which served as the precipitant for all crystallization presented here.

These crystals were then optimized by drop ratio variation in 24-well CrysChem M Plates (Hampton Research) using 1–6 μL protein drops and 2–5 μL precipitant drops. Yellow crystals with an atypical morphology (**Figure E-19**) appeared in all wells after 1–3 days, with larger crystals generally observed at higher protein and lower precipitant concentrations.

E.1.18 Crystal soaking and cryoprotection.

For crystals of the TmTyrS1 holoenzyme, a cryoprotectant solution was prepared by mixing 80 μL of equilibrated reservoir solution with 20 μL of ethylene glycol. This solution was then added to the crystal drop, sequentially adding and removing equivalent volumes until no schlieren was observed.

To trap the amino-acrylate intermediate E(A-A) state of TmTyrS1, a solution was prepared that consisted of the precipitant supplemented with 100 mM Ser. This was serially added and removed from the crystallization drop in 2-μL aliquots until no schlieren was observed. Crystals were incubated for 30 minutes, during which they turned from yellow to colorless, indicating that the amino-acrylate had formed. At this point, the serine-containing precipitant was further supplemented with 20% ethylene glycol and used as a cryoprotectant as stated above.

To obtain structures containing 1-naphthol mimics 4-hydroxyquinoline (QOH) and quinoline *N*-oxide (QOX), the amino-acrylate-containing crystals were first prepared as stated above. The cryoprotectant solution (20% ethylene glycol) was then further supplemented with 20 mM QOH or QOX, then applied to the crystals. The only addition to the cryoprotection procedure described above is the incubation of crystals in the cryoprotectant for 2–10 minutes. Following cryoprotection, all crystals were mounted in nylon loops, cooled in liquid nitrogen, and stored in liquid nitrogen prior to data collection.

E.1.19 Crystal structure determination.

Diffraction data were collected at the Stanford Synchrotron Radiation Laboratory (SSRL) beamline 12-2. Data reduction and integration were carried out using XDS[6] and scaled using Aimless in the CCP4 suite of programs.[7] For the structure of holo TmTyrS1, molecular replacement (MR) was performed using the structure of a holo TrpB from *Pyrococcus furiosus* (*Pf*TrpB; PDB 5DVZ)[8] as a search model in Phaser.[9] For all other structures, the protein chain of holo TmTyrS1 was used for MR. Model building and modification in the electron density was performed using Coot[10] and structure refinement was performed using Phenix.[11] Other ligands, specifically QOX and QOH, as well as water molecules and ethylene glycol were added during later stages of refinement. Occasionally, spurious electron density peaks were present in the active site, dimer interface, and COMM domain that could not be unambiguously modeled by alternative protein conformations, solvent, or other additives applied during the procedure, so these were left uninterpreted. The quality of the final models was evaluated with MolProbity[12] and PROCHECK[13]. Data collection and refinement statistics are presented in **Table E-3**.

E.1.20 Conservation of the catalytic glutamate in TrpB-like sequences.

Human-annotated TrpB sequences from the SwissProt database were obtained and aligned to obtain a multiple sequence alignment (MSA) referenced to the sequence of Tm9D8*. The catalytic glutamate was conserved at position 105 in all 451 sequences.

To probe this deeper, we obtained a multiple sequence alignment (MSA) of 18,719 TrpB-like sequences from the EVcouplings software.[14] We discarded 693 variants that did not contain an appropriately positioned catalytic lysine (K83) or had an insertion at position 105 in the MSA (i.e., from an improperly aligned and/or non-TrpB-like sequence), leaving 18,051 sequences. These sequences were analyzed based on their amino acids at position 105 (**Figure 6-3a** and **Figure E-18a**) and 229 (**Figure E-18b**). Correlations were observed by examining the identities at these positions within a given sequence (**Figure E-18c**). Notably, most sequences with the E105A+G229S/T pair were derived from plants, while those with the E105G+G229A pair were found in *Streptomyces* and related soil bacteria from the phylum *Actinobacteria*.

E.1.21 Enzyme sequences.

Amino acid substitutions accrued during evolution from Tm9D8* are presented in **Table E-1**. DNA sequences of individual variants used in this study are presented below.

```
>Tm9D8*
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCATTGCTGAAACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAACTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGTTCGGCTCTGTGGTTGGTC
CGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCA
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
```

```
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCACCACCACCACCACCACTGA

>Tm9D8* E105G
ATGAAAGGCTACTTCGGTCCGTACGGTGGCCAGTACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATTCAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCATTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAACTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGTTCGGCTCTGTGGTTGGTC
CGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCA
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGTGAGCCACTCCGTCTCCGCTGGCCTGGACTACTCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACCGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCGCCTCGAGCACCACCACCACCACCACTGA

>TmTyrS1
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACGCTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAATCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAACTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGCCGGGCTCTGTGGTTGGTC
CGCATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCA
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>TmTyrS2
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAGTCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAGCTGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGCCGGGCTCTGTGGTTGGTC
```

```
CGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCA
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCGTGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>TmTyrS3
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAGTCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAGCAGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGATTACCAACCTGCAGACCACCTATTACGTGCCGGGCTCTGTGGTTGGTC
CGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCT
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCATGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>TmTyrS4
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAGTCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAGCAGCTGGGTGCTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGACTACCAACCTGCAGACCACCTATTACGTGGCGGGCTCTGTGGTTGGTC
CGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCT
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCATGAGCGGTGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTCCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>TmTyrS5
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
```

CAGCGCTGTTCGGTATGGAATGTGTAGTCTATATGGGCGAAGAAGACACGATCCGCCAGAAACTAAACGTT
GAACGTATGAAGCAGCTGGGTGTTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTCTGCGTGACTGGACTACCAACCTGCAGACCACCCATTACGTGGCGGGCTCTGTGGTTGGTC
CGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCT
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCATGAGCGCGGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACTGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>TmTyrS6
ATGAAAGGCAACTTCGGTCCGTACGGTGGCCAGAACGTGCCGGAAATCCTGATGGGAGCTCTGGAAGAACT
GGAAGCTGCGTACGAAGGAATCATGAAAGATGAGTCTTTCTGGAAAGAATACAATGACCTGCTGCGCGATT
ATGCGGGTCGTCCGACTCCGCTGTACTTCGCACGTCGTCTGTCCGAAAAATACGGTGCTCGCGTATATCTG
AAACGTGAAGACCTGCTGCATACTGGTGCGCATAAAATCAATAACACTATCGGCCAGGTTCTGCTGGCAAA
ACTAATGGGCAAAACCCGTATCACTGCTGGTACGGGTGCTGGTCAGCACGGCGTAGCAACTGCTACCGCAG
CAGCGCTGTTCGGTATGGAATGTGTAGTCTATATGGGCGAAGAAGACACGATCCGCCAGAGACTAAACGTT
GAACGTATGAAGCAGCTGGGTGTTAAAGTTGTACCGGTAAAATCCGGTAGCCGTACCCTGAAAGACGCAAT
TGACGAAGCTTTTCGTGACTGGACTACCAACCTGCAGACCACCCATTACGTGGCGGGCTCTGTGGTTGGTC
CGTATCCATATCCGATTATCGTACGTAACTTCCAAAAGGTTATCGGCGAAGAGACCAAAAAACAGATTCCT
GAAAAAGAAGGCCGTCTGCCGGACTACATCGTTGCGTGCATGAGCGCGGGTTCTAACGCTGCCGGTATCTT
CTATCCGTTTATCGATTCTGGTGTGAAGCTGATCGGCGTAGAAGCCGGTGGCGAAGGTCTGGAAACCGGTA
AACATGCGGCTTCTCTGCTGAAAGGTAAAATCGGCTACCTGCACGGTTCTAAGACGTTCGTTCTGCAGGAT
GACGGGGGTCAAGTTCAGGCGAGCCACTCCGTCTCCGCTGGCCTGGACTACCCCGGTGTCGGTCCGGAACA
CGCCTATTGGCGTGAGACCGGTAAAGTGCTGTACGATGCTGTGACTGATGAAGAAGCTCTGGACGCATTCA
TCGAACTGTCTCGCCTGGAAGGCATCATCCCAGCCCTGGAGTCTTCTCACGCACTGGCTTATCTGAAGAAG
ATCAACATCAAGGGTAAAGTTGTGGTGGTTAATCTGTCTGGTCGTGGTGACAAGGATCTGGAATCTGTACT
GAACCACCCGTATGTTCGCGAACGCATCCACCTCGAGCACCACCACCACCACCACTGA

>Pf2B9
ATGTGGTTCGGTGAATTTGGTGGTCAGTACGTGCCAGAAACGCTGGTTGGACCCCTGAAAGAGCTGGAAAA
AGCTTACAAACGTTTCAAAGATGACGAAGAATTCAATCGTCAgCTGAATTACTACCTGAAAACCTGGGCAG
GTCGTCCAACCCCACTGTACTACGCAAAACGCCTGACTGAAAAAATCGGTGGTGCTAAAGTCTACCTGAAA
CGTGAAGACCTGGTTCACGGTGGTGCACACAAGACCAACAACGCCATCGGTCAGGCACTGCTGGCAAAGCT
CATGGGTAAAACTCGTCTGATCGCTGAGACCGGTGCTGGTCAGCACGGCGTAGCGACTGCAATGGCTGGTG
CACTGCTGGGCATGAAAGTGGACATTTACATGGGTGCTGAGGACGTAGAACGTCAGAAAATGAACGTATTC
CGTATGAAGCTGCTGGGTGCAAACGTAATTCCAGTTAACTCCGGTTCTCGCACCCTGAAAGACGCAATCAA
CGAGGCTCTGCGTGATTGGGTGGCTACTTTTGAATACACCCACTACCTAATCGGTTCCGTGGTCGGTCCAC
ATCCGTATCCGACCATCGTTCGTGATTTTCAGTCTGTTATCGGTCGTGAGGCTAAAGCGCAGATCCTGGAG
GCTGAAGGTCAGCTGCCAGATGTAATCGTTGCTTGTGTTGGTGGTGGCTCTAACGCGATGGGTATCTTTTA
CCCGTTCGTGAACGACAAAAAAGTTAAGCTGGTTGGCGTTGAGGCTGGTGGTAAAGGCCTGGAATCTGGTA
AGCATTCCGCTAGCCTGAACGCAGGTCAGGTTGGTGTGTCCCATGGCATGCTGTCCTACTTTCTGCAGGAC
GAAGAAGGTCAGATCAAACCAAGCCACTCCATCGCACCAGGTCTGGATTATCCAGGTGTTGGTCCAGAACA
CGCTTACCTGAAAAAAATTCAGCGTGCTGAATACGTGGCTGTAACCGATGAAGAAGCACTGAAAGCGTTCC
ATGAACTGAGCCGTACCGAAGGTATCATCCCAGCTCTGGAATCTGCGCATGCTGTGGCTTACGCTATGAAA
CTGGCTAAGGAAATGTCTCGTGATGAGATCATCATCGTAAACCTGTCTGGTCGTGGTGACAAAGACCTGGA
TATTGTCCTGAAAGCGTCTGGCAACGTGCTCGAGCACCACCACCACCACCACTGA

>Pf2B9 E104G
ATGTGGTTCGGTGAATTTGGTGGTCAGTACGTGCCAGAAACGCTGGTTGGACCCCTGAAAGAGCTGGAAAA
AGCTTACAAACGTTTCAAAGATGACGAAGAATTCAATCGTCAgCTGAATTACTACCTGAAAACCTGGGCAG

```
GTCGTCCAACCCCACTGTACTACGCAAAACGCCTGACTGAAAAAATCGGTGGTGCTAAAGTCTACCTGAAA
CGTGAAGACCTGGTTCACGGTGGTGCACACAAGACCAACAACGCCATCGGTCAGGCACTGCTGGCAAAGCT
CATGGGTAAAACTCGTCTGATCGCTGGGACCGGTGCTGGTCAGCACGGCGTAGCGACTGCAATGGCTGGTG
CACTGCTGGGCATGAAAGTGGACATTTACATGGGTGCTGAGGACGTAGAACGTCAGAAAATGAACGTATTC
CGTATGAAGCTGCTGGGTGCAAACGTAATTCCAGTTAACTCCGGTTCTCGCACCCTGAAAGACGCAATCAA
CGAGGCTCTGCGTGATTGGGTGGCTACTTTTGAATACACCCACTACCTAATCGGTTCCGTGGTCGGTCCAC
ATCCGTATCCGACCATCGTTCGTGATTTTCAGTCTGTTATCGGTCGTGAGGCTAAAGCGCAGATCCTGGAG
GCTGAAGGTCAGCTGCCAGATGTAATCGTTGCTTGTGTTGGTGGTGGCTCTAACGCGATGGGTATCTTTTA
CCCGTTCGTGAACGACAAAAAAGTTAAGCTGGTTGGCGTTGAGGCTGGTGGTAAAGGCCTGGAATCTGGTA
AGCATTCCGCTAGCCTGAACGCAGGTCAGGTTGGTGTGTCCCATGGCATGCTGTCCTACTTTCTGCAGGAC
GAAGAAGGTCAGATCAAACCAAGCCACTCCATCGCACCAGGTCTGGATTATCCAGGTGTTGGTCCAGAACA
CGCTTACCTGAAAAAAATTCAGCGTGCTGAATACGTGGCTGTAACCGATGAAGAAGCACTGAAAGCGTTCC
ATGAACTGAGCCGTACCGAAGGTATCATCCCAGCTCTGGAATCTGCGCATGCTGTGGCTTACGCTATGAAA
CTGGCTAAGGAAATGTCTCGTGATGAGATCATCATCGTAAACCTGTCTGGTCGTGGTGACAAAGACCTGGA
TATTGTCCTGAAAGCGTCTGGCAACGTGCTCGAGCACCACCACCACCACCACTGA
```

E.2 Supplementary Text

E.2.1 Evolutionary strategies.

Evolution for this study used many different strategies to achieve the rate enhancement in TmTyrS6. If not specified, experimental details for each type of mutagenesis (error-prone PCR mutagenesis, StEP recombination, and site-saturation mutagenesis (SSM)) and screening (colorimetric, LCMS) approach can be found in the methods above. This section provides general details about what was performed to obtain each variant in the lineage.

Tm9D8* E105G was identified by generating a saturation mutagenesis library at position 105 and screening enzyme variants against 0.5 mM 1-naphthol for 4 and 18 hours using a colorimetric assay (see **Figure E-5a**), and wells demonstrating high activity were analyzed by LCMS to confirm product formation. Following this saturation mutagenesis at position 184—which has been shown to often have a beneficial effect for new substrates—was performed and screened similarly, identifying F184P as a beneficial mutation (~2-fold boost for NaphAla formation).

TmTyrS1 was identified by subjecting Tm9D8* E105G F184P to highly error-prone PCR mutagenesis (>600 µM $MnCl_2$), which generated a library of highly mutated enzyme variants (~8 mutations per variant). Enzymes were prepared as heat-treated lysates (3-hour heat-treatment). A total of eight plates (704 variants) were screened. Activity determination took place via a continuous colorimetric screen against 5 mM 1-naphthol at room temperature using a Tecan® Spark® Spark-Stack™ in kinetics mode (wavelength = 335 nm). Variants that retained >50% parent activity (~40 variants) were subjected to StEP recombination. From this recombination library, four plates were screened in a similar way. The most-improved variants were once again subjected to recombination and four plates were again screened. This resulted in a panel of improved variants with groups of common mutations. Variants were compiled in biological replicate into a new plate and screened against 5 mM 1-naphthol (in the same way as previously) as well as 25 mM 2-chlorophenol and 50 mM 2-fluorophenol via LCMS. Mutations were recombined that were general for all substrates, yielding TmTyrS1. (Incidentally, F184P was identified to be neutral for 2-chlorophenol and deleterious for 2-fluorophenol, despite being highly activating for 1-naphthol.)

Following TmTyrS1, evolution proceeded in a more routine manner, using standard error-prone PCR mutagenesis techniques (see **Section E.1.3**, *Error-prone PCR mutagenesis*) and StEP recombination (**Section E.1.5**, *Recombination via Staggered Extension Process*). TmTyrS2 was identified by screening against 25 mM 2-chlorophenol using a 0.65-minute LCMS method screening on a C-18 guard column for sufficient separation of substrate and product with a total of 1.2 minutes between injections (2 mL/min flow rate; 0.00 min: 1%

MeCN; 0.01 min: 95% MeCN; 0.26 min: 1% MeCN; hold to 0.65 min; post-time: 0.25 min). TmTyrS3 was identified by screening against 10 mM 2-chlorophenol in the same way.

TmTyrS3 was sufficiently active that it could be used to detect Tyr formation in enzyme lysate, which harbors background Tyr from the cells that previously made such screening impossible. However, this required 50 mM phenol loading, long reaction times, and a 4-minute LCMS method to reliably detect the Tyr. During this time, a final round of screening against 5 mM 2-chlorophenol was performed, and enzyme variant sequences were determined using the evSeq method.[15] TmTyrS3 was subjected to SSM within the active site to identify the mutation P184A as highly activating. TmTyrS3 P184A was subjected to additional rounds of error-prone PCR mutagenesis and recombination and screened via both LCMS and a colorimetric screen (wavelength = 310 nm). This resulted in TmTyrS4.

When the TmTyrS3 SSM libraries were screened against phenol, the mutation G229A was observed to be highly activating (~3-fold improvement). This mutation was added to TmTyrS4 but was not observed to have the same effect. TmTyrS4 and TmTyrS3 P184A were recombined via StEP, and G229A was found to be activating only in the absence of the S265P mutation, which was kept reverted. This variant was subjected to additional error-prone PCR mutagenesis and recombination, screening against 10 mM phenol on the 0.65-min LCMS method (single-ion mode for 182 m/z) to identify TmTyrS5. Screening against 5 mM phenol conversion, a final round of SSM (to identify L170F) and error-prone PCR mutagenesis and recombination led to TmTyrS6.

E.2.2 Determination of limit for detectable turnover frequency.

The activity of Tm9D8* for Tyr formation was sufficiently low that it could not be reliably quantified. As other phenol analogs reacted at lower but more reliable levels, it became important to assign a limit of detection for Tyr to understand the rate enhancement achieved by directed evolution, specifically the E105G mutation.

A series of Tyr solutions in KPi ranging from 100 nM to 5 mM were prepared. These were combined 1:4 with 1:1 1 M aq. HCl/MeCN in the same way as analytical vial reactions and analyzed by LCMS. These data are presented in **Figure E-10c**, showing the integrated peak area for Tyr vs. the known concentration of Tyr from the original solution. On a log-log plot, these data are linear above 1 μM, below which there is a basal level of peak that can be integrated. This likely arises from cross-contamination between samples, with Tyr not being completely washed off the column in all cases. Exhaustive washing failed to provide significant improvements, and thus 1 μM was designated as the lower limit of Tyr detectability.

Given this limit, and the maximum concentration of enzyme used in the reactions (100 μM), the minimum enzymatic turnovers required to detect Tyr is:

$$\frac{1\ \mu\text{M Tyr}}{100\ \mu\text{M enzyme}} = 0.01\ \text{turnovers}$$

Reactions were carried out for exactly 24 hours, which converts this to a turnover frequency (TOF, h^{-1}) of:

$$\frac{0.01\ \text{turnovers}}{24\ \text{hours}} = 0.00042\ \text{h}^{-1}$$

This limit is represented in Chapter IV, **Figure 6-2d**. As 100 µM Tm9D8* did not exceed 1 µM Tyr in 24 hours (**Figure E-17**) this value provides a confident upper bound on its activity. TmTyrS6 reacts at a TOF of 14 h^{-1}, thus:

$$\frac{14\ \mathrm{h}^{-1}}{0.00042\ \mathrm{h}^{-1}} = 33{,}300\text{x difference.}$$

E.2.3 Estimating product formation by HPLC peak area percentages.

The data presented in **Figure 6-3a** are shown to give an idea of the activity of each enzyme for each substrate. The percentages listed within the boxes are estimates of HPLC yields/conversions, but are not adjusted with a calibration curve, as many of the standards are commercially inaccessible. Instead, product formation is estimated by assuming that the HPLC peak areas are equally proportional to compound concentration. Formally this is only true at the isosbestic point between the substrate and the red-shifted product (see **Figure E-5a**). However, red shifts are generally slight, and can still provide a reliable means of estimating product formation among different reactions. Moreover, the relative trends are exactly true within a given substrate row. In other words, in **Figure 6-3a** the effect of evolution on a given substrate can be seen exactly, while the effect of evolution across different substrates can be reasonably approximated.

These values will be the least accurate where there is an obvious difference in the total absorbance of the substrate and product peaks compared to high- and low-yielding reactions. Thus, if the formation of product significantly changes the *total* absorbance in the system, then there is an obvious difference in absorbance between the substrate and product peaks at the given wavelength. These data are shown in **Figure E-11**. Generally, the change in total absorbance is not dependent on the area of the product peak, making these reliable estimates of conversion. All product absorbance peaks are correlated with their respective MS peaks, allowing confident identification of the product peak in even low-conversion reactions.

All reactions were run according to the conditions presented in **Figure 6-3a** (10 µM enzyme, 10 mM phenolic substrate, 11 mM Ser, in KPi at 37 °C for 24 hours) using the methods exactly as written in **Section E.1.8**, *Analytical-scale vial reactions*. LCMS traces were processed by integrating the absorbance peak of the product and the known substrate absorbance peak. In cases of particularly low activity where a reliable mass peak could not be observed by extracted ion counts (EIC) obtained from the total ion count (TIC) signal, single-ion mode (SIM) was used to observe the product MS peak. If no reliable absorbance peak was observed above background but an MS peak was observed in the EIC or SIM signals, activity was labeled as “trace”.

For simplicity, and because the enzymatic reactions are generally highly specific (e.g., no side products), only the substrate and product peaks were integrated and processed. Only two substrates showed some deviation. Traces for the (2-MeS-phenol) reactions indicated the presence of an oxidized substrate, presumably to the sulfoxide (2-MeSO-phenol) based on the mass and absorbance difference. However, the peak was relatively minor compared to the substrate peak, had significant differences in its absorbance, and was omitted from the conversion analysis for consistency. Additionally, many peaks were observed for the 2-CF_3-phenol substrate, suggesting that this substrate was not of high purity. While this does not affect the comparison of yields across the different enzymes (e.g., TmTyrS4 does form more product than TmTyrS3), care should be taken to compare these values to other substrates.

E.3 Supplemental Figures

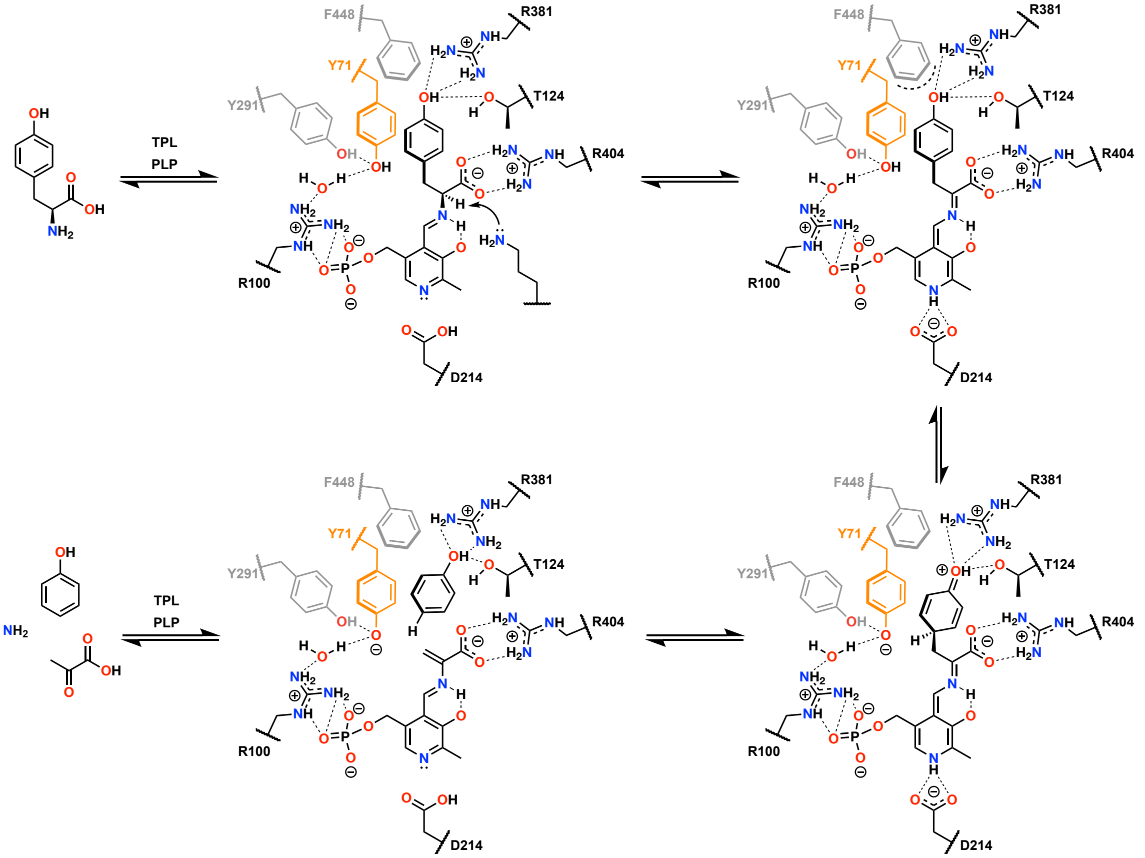


Figure E-1. Regioselective bond breakage and formation in the active site of tyrosine phenol lyase (TPL). Numbering and proposed mechanism according to studies with *Citrobacter freundii* TPL (*Cf*TPL).[16] A catalytic Tyr (Y71, orange) carries out protonation and deprotonation at C4 of the phenolic group, which is coordinated by T124 and R381.

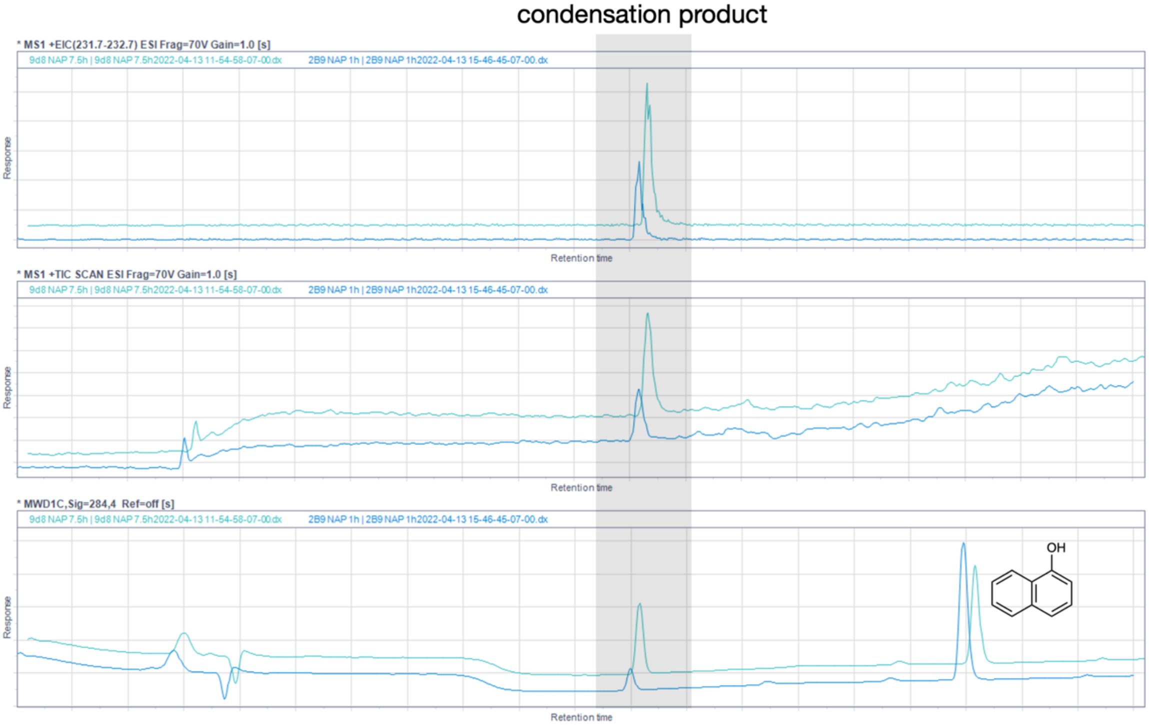


Figure E-2. Detection of amino acid product of 1-naphthol and TrpB by LCMS. LCMS traces are shown for additions of 10 mM 1-naphthol to 10 µM Tm9D8* for 7.5 hours at 37 °C (teal) or 10 µM *Pf*2B9 for 1 hour at 75 °C (blue). Traces are slightly offset in both axes for clarity. *Upper*: Extracted ion counts for 232 m/z, the expected molecular ion of the condensation product of Ser and 1-naphthol. *Middle*: Total ion counts. *Lower*: Absorbance at 284 nm.

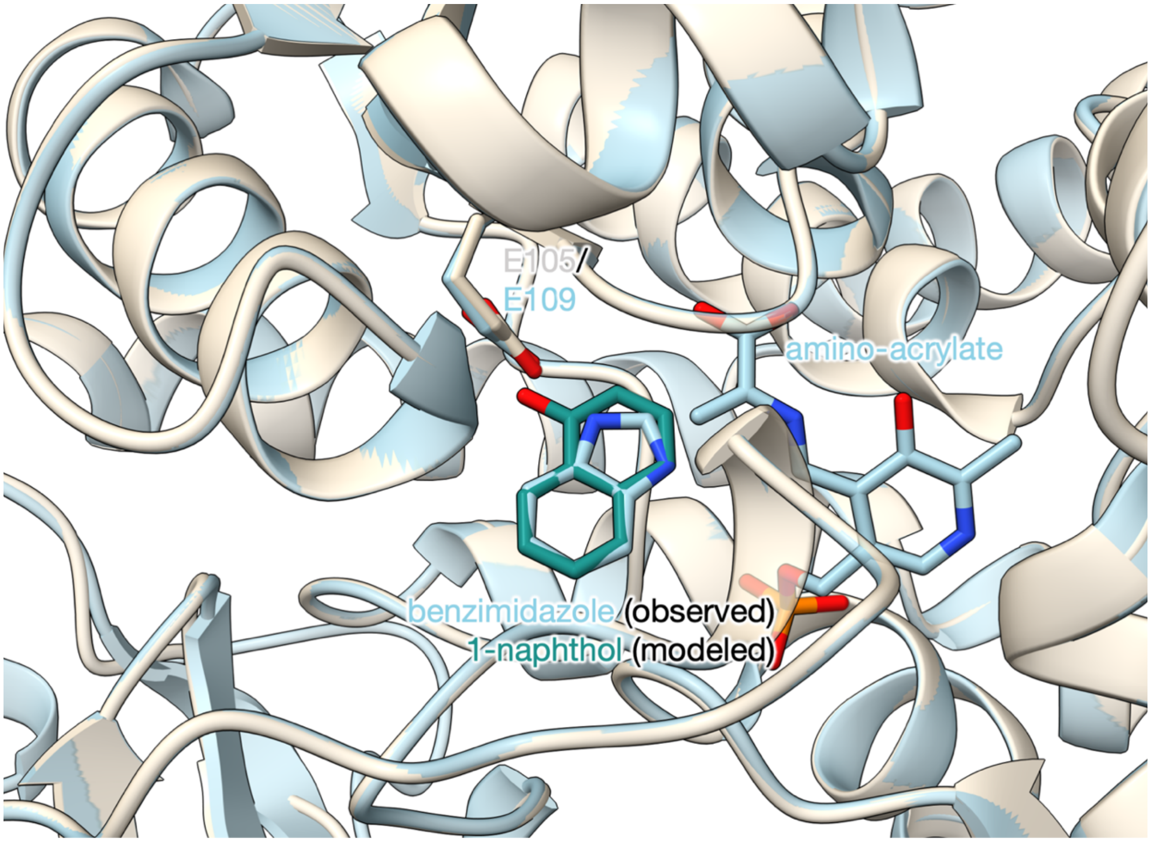


Figure E-3. Analysis of 1-naphthol in the active site of Tm9D8*. A homology model of Tm9D8* (beige) was prepared using *Salmonella typhimurium* TrpB (*St*TrpB) as a reference structure (blue). In this structure, the amino-acrylate E(A-A) state was captured binding a non-reactive indole analog, benzimidazole.[17] The homology model of Tm9D8* was aligned to *St*TrpB, superimposing the amino-acrylate and benzimidazole into the active site of Tm9D8*. 1-Naphthol (teal) was aligned to benzimidazole to orient it for *para* C–C bond formation, which revealed a likely sub-optimal interaction between the conserved catalytic glutamate (E105 in Tm9D8*, E109 in *St*TrpB) and 1-naphthol.

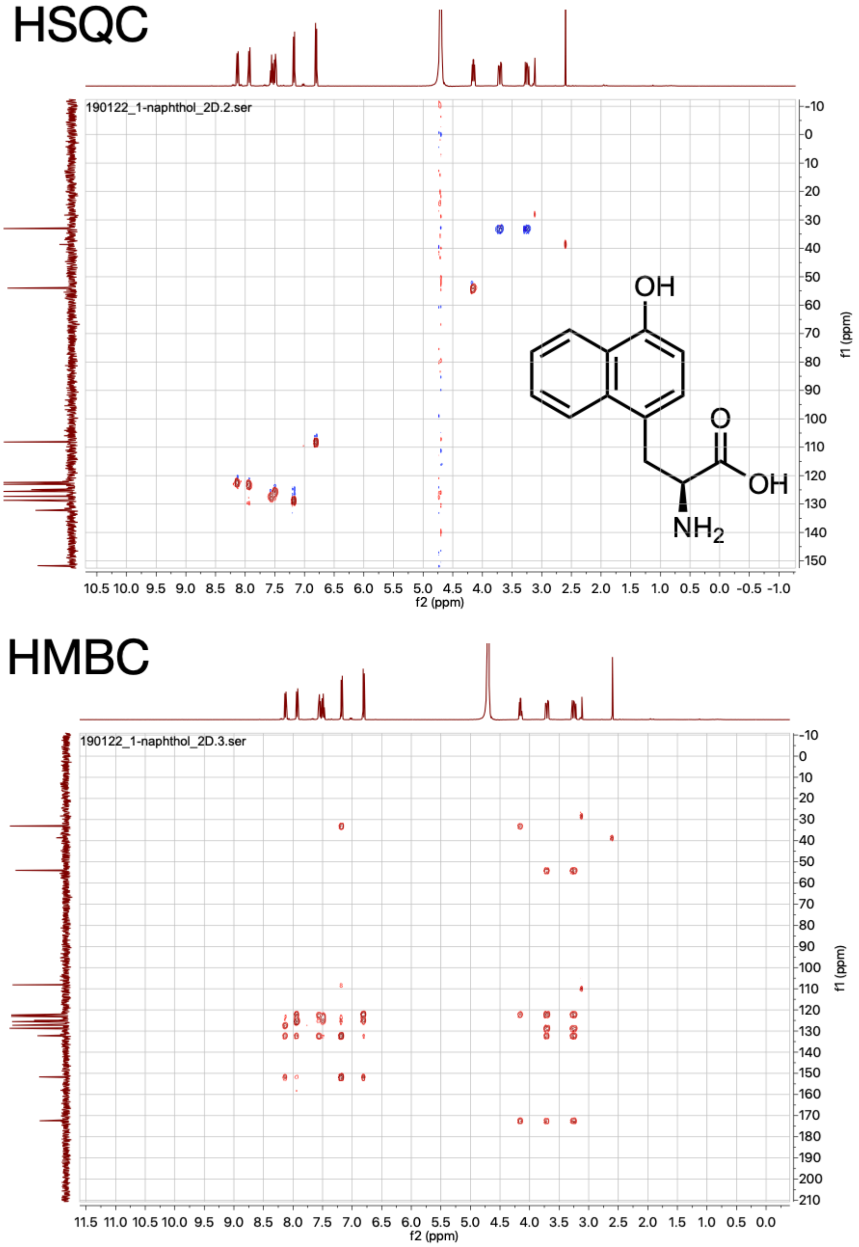


Figure E-4. 2D NMR characterization of the 1-naphthol product of Tm9D8* E105G. *Upper:* Heteronuclear single-quantum correlation (HSQC) spectrum. *Lower:* Heteronuclear multiple bond correlation (HMBC). Assignment of peaks leads to the *para* alkylation product NaphAla as the sole compound.

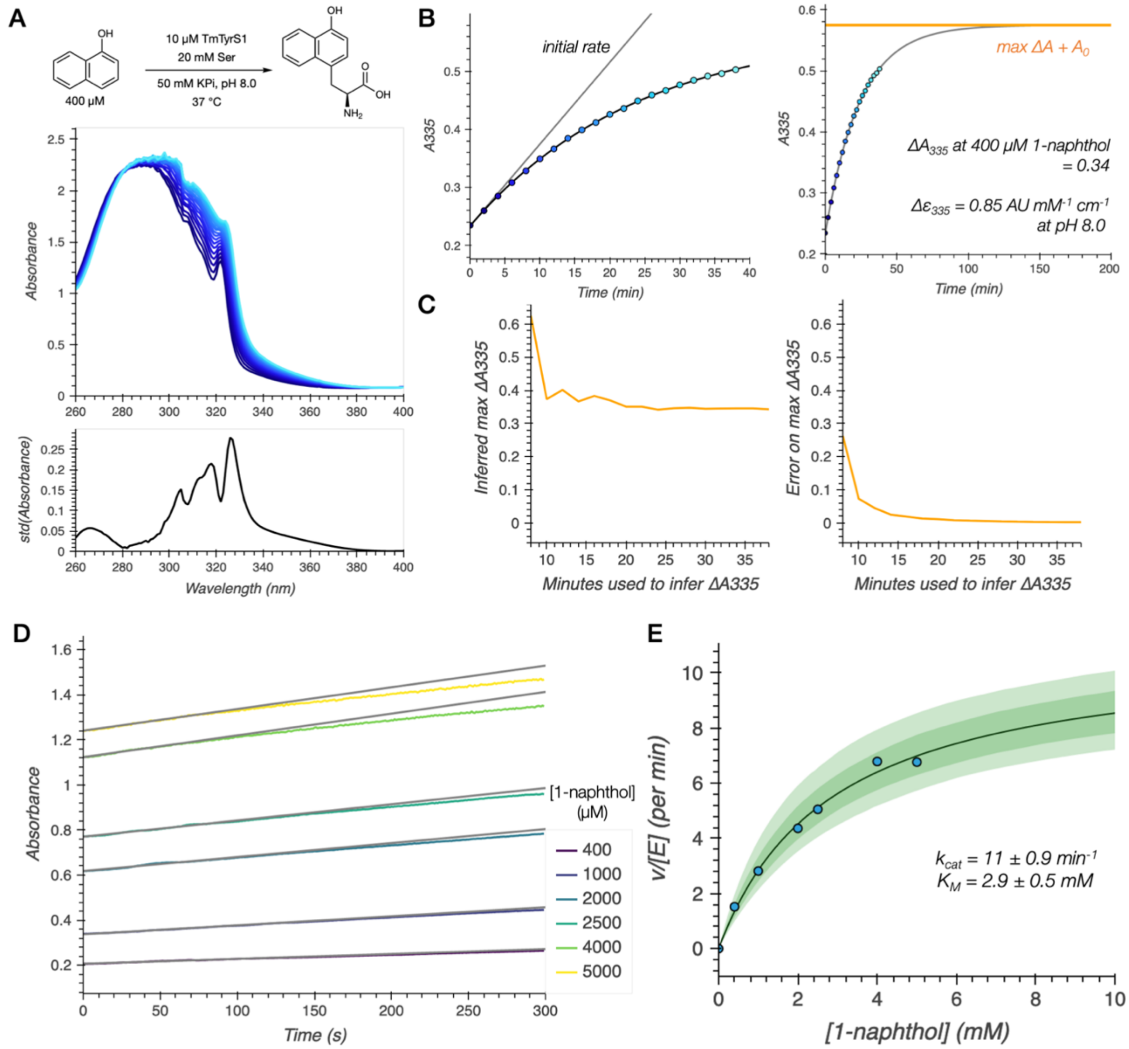


Figure E-5. Development of a continuous colorimetric assay for NaphAla production. a. The conversion of 1-naphthol to NaphAla results in a red-shift in the absorbance spectrum of 1-naphthol, with an isosbestic point between 280–284 nm (284 nm is used for this study). **b.** Reaction progress curve for 400 µM 1-naphthol at 335 nm, fit to an exponential function. The fit and estimated initial rate are plotted as lines. The orange horizontal line represents the predicted maximum change in absorbance (max ΔA) over the starting absorbance (A_0). As this represents the full colorimetric change at 335 nm for the conversion of 400 µM 1-naphthol, the molar change in absorptivity ($\Delta\varepsilon_{335}$) for the conversion of 1-naphthol to NaphAla can be inferred to be 0.85 AU mM^{-1} cm^{-1} at pH 8.0 to convert absorbance units to real concentrations. **c.** Estimates and errors of max ΔA with increasing data used, confirming that the estimate was made with sufficient data. **d.** Full absorbance time courses at 335 nm at varying 1-naphthol concentrations, with initial rates as gray lines. **e.** Michaelis-Menten model of initial rates obtained from time course data (blue points). 70% and 95% confidence intervals are shown in dark and light green, respectively.

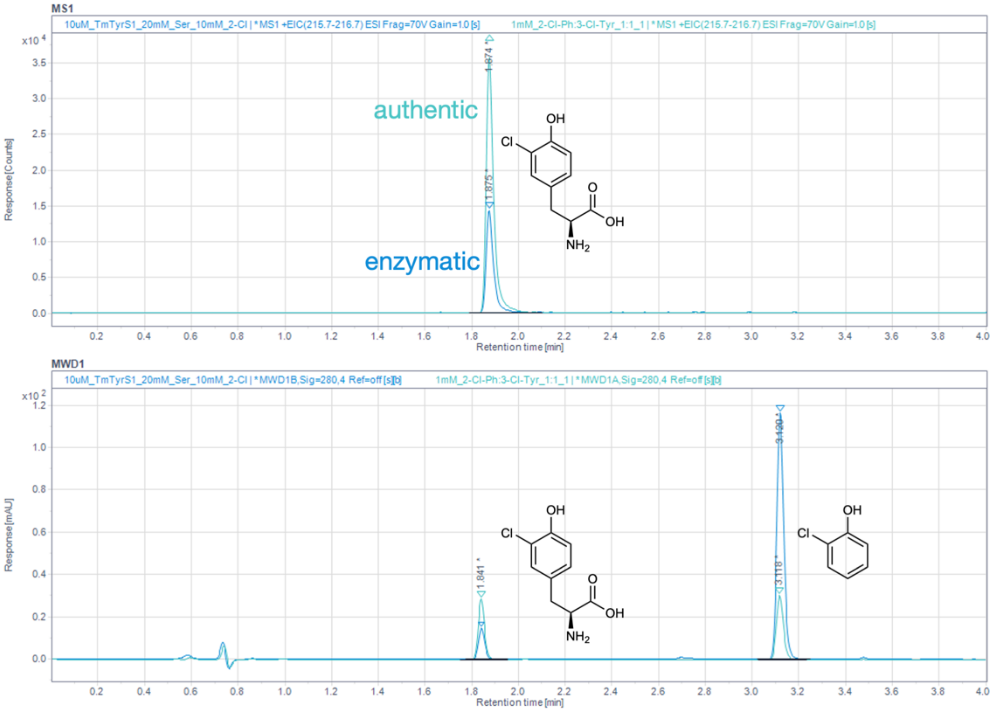


Figure E-6. Comparison of enzymatic 2-chlorophenol reactions (blue) and a 1:1 mix of 2-chlorophenol and authentic 3-chloro-Tyr (teal). 2-Chlorophenol elutes at 3.12 min, while 3-chloro-Tyr elutes at 1.84 min. *Upper*: Extract ion counts for 216 m/z, the molecular ion of chloro-Tyr. *Lower*: Absorbance at 280 nm.

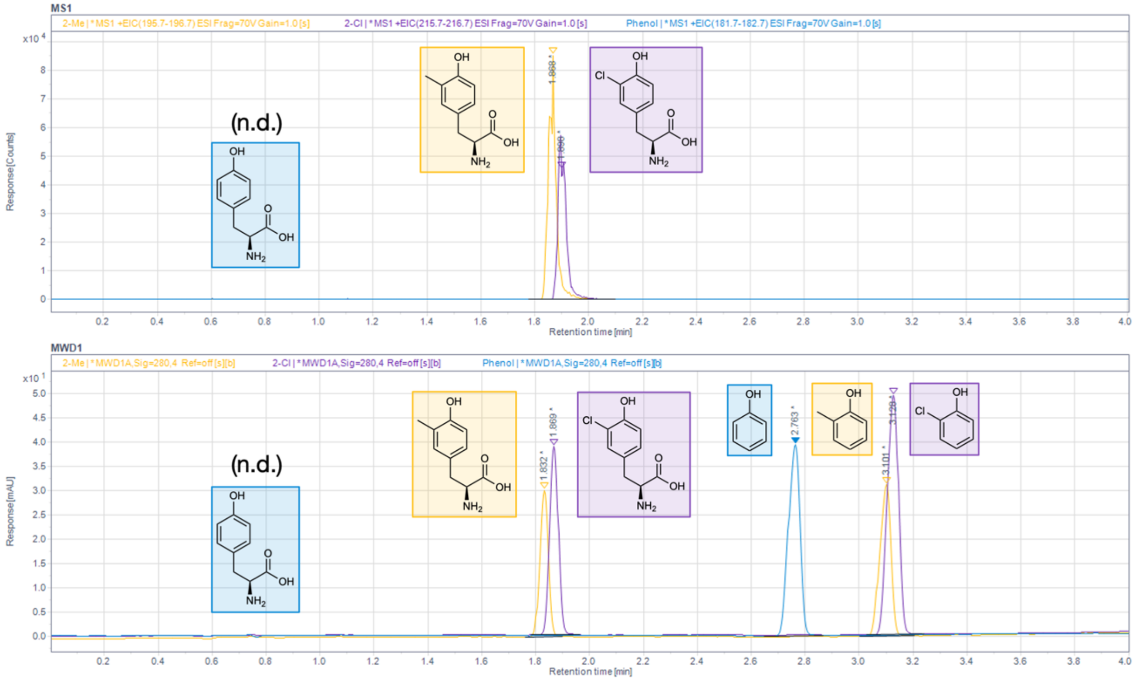


Figure E-7. Comparison of 10 µM TmTyrS4 activity toward 10 mM phenol (blue), 2-chlorophenol (purple), and 2-methylphenol (yellow) in 24-hour reactions at 37 °C. *Upper*: Extracted ion counts for the molecular ion of each respective Tyr analog. *Lower*: Absorbance at 280 nm. Both 2-chlorophenol and 2-mehtylphenol show clear and roughly equivalent yield for 3-chloro-Tyr (1.87 min) and 3-methyl-Tyr (1.83 min), while no product is detected under these conditions for phenol (a single ion mode channel is required for sufficient signal-to-noise). Only the phenol absorbance peak at 2.76 min is observed. (n.d. = not detected.)

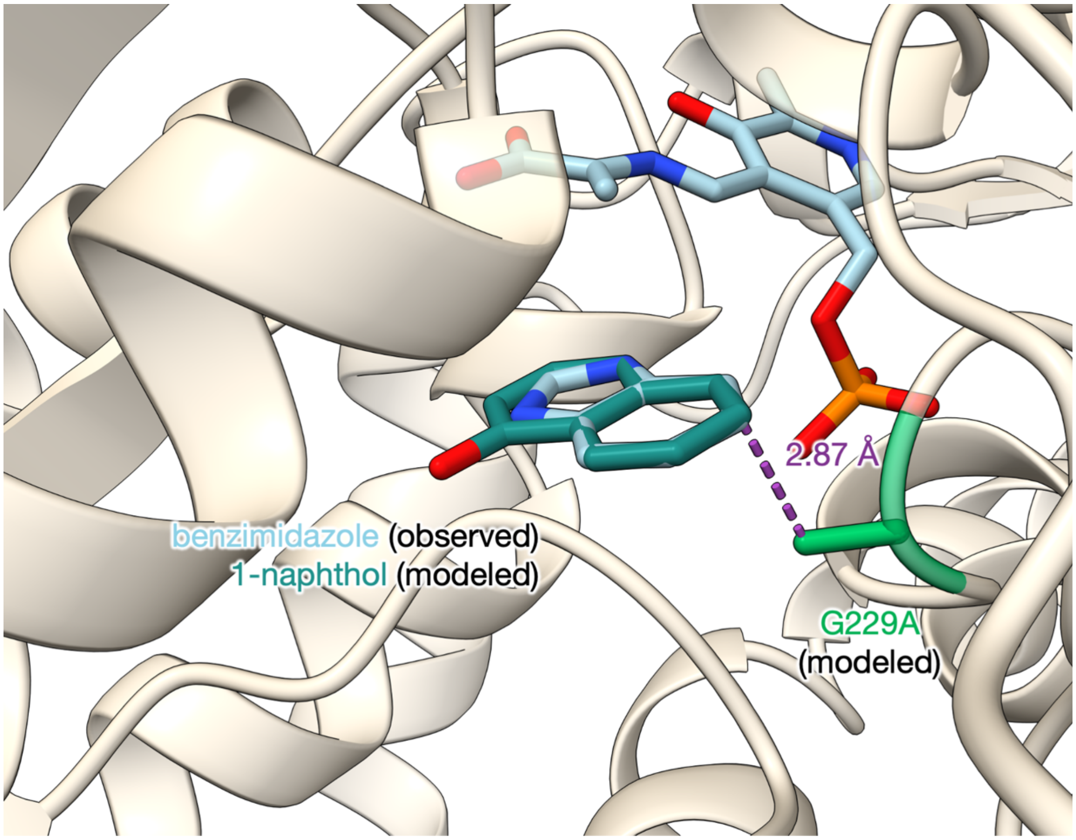


Figure E-8. Modeling the G229A mutation. Using the same model as in **Figure E-3**, glycine 229 was mutated to alanine (green), which would add a methyl group in close proximity to the benzene ring of indole and 1-naphthol when putatively bound for C–C bond formation.

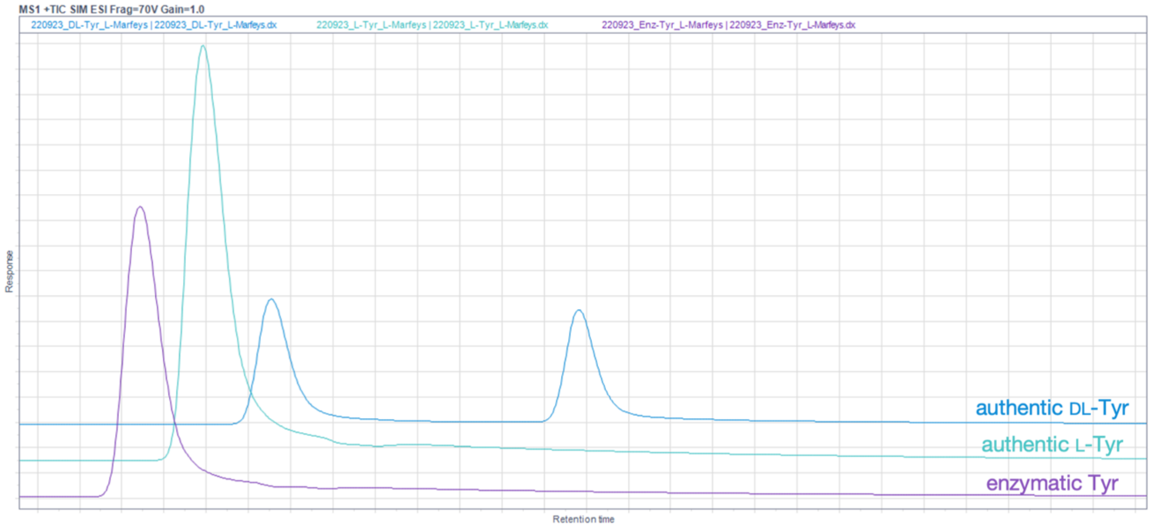


Figure E-9. Enantiopurity of enzymatic Tyr product. Amino acids are derivatized with Nα-(5-fluoro-2,4-dinitrophenyl)-L-alaninamide (Marfey's reagent) which affords chiral resolution of enantiomers. Authentic DL-tyrosine is shown in the upper (blue) trace and shows two peaks, one for each enantiomer. Authentic L-tyrosine is shown in the middle (teal) trace, with only a single peak. Enzymatically prepared Tyr is shown in the bottom (purple) trace, with only a single detectable peak consistent with that of L-tyrosine. Traces are offset slightly in both axes for clarity.

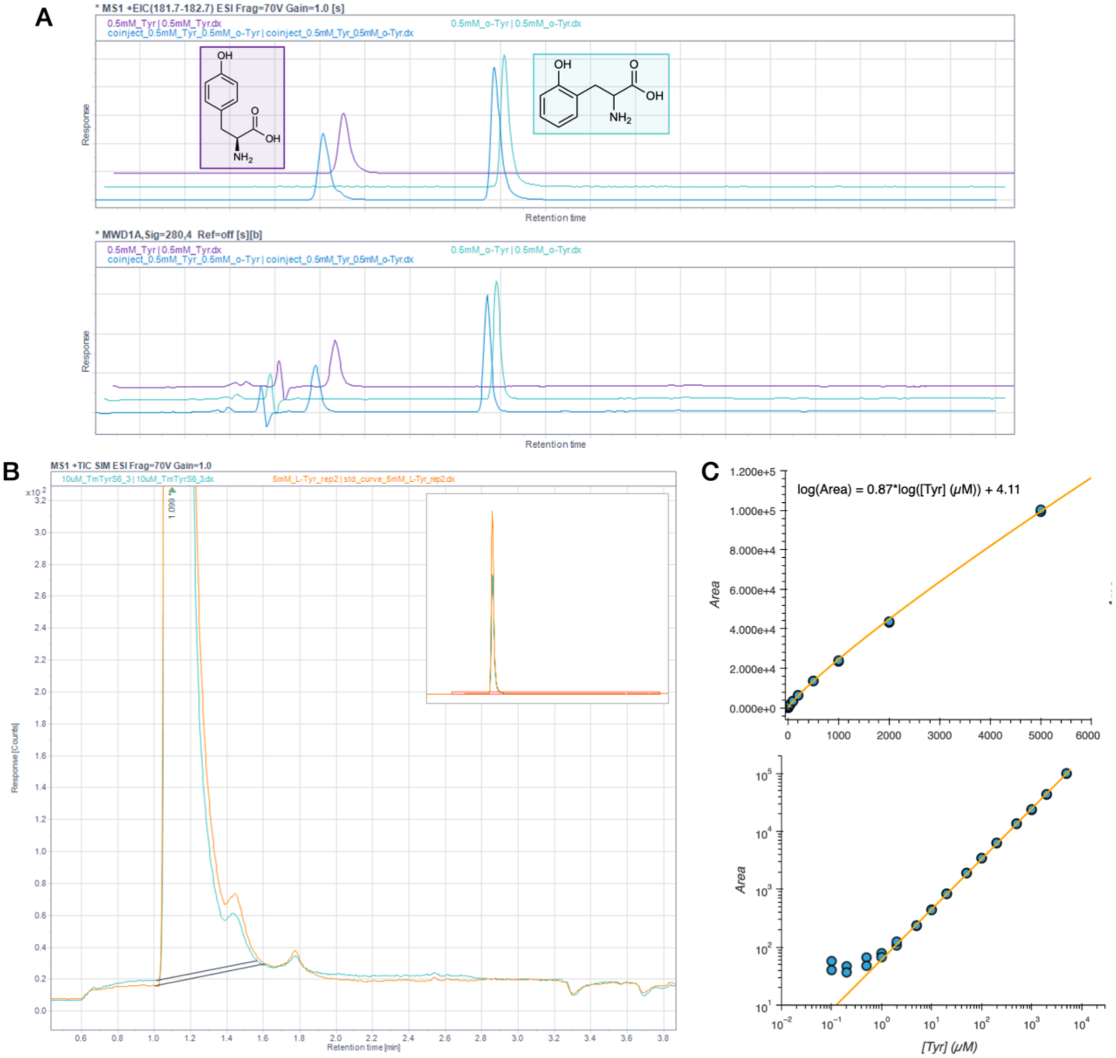


Figure E-10. LCMS characterization of authentic and enzymatic tyrosine. a. Comparison of authentic standards of Tyr (in purple), *ortho*-DL-tyrosine (*ortho*-Tyr, in teal), and a coinjection of both (in blue) on the same LCMS method used throughout this study. *Upper*: Extracted ion counts for 182 m/z, the molecular ion. *Lower*: Absorbance at 280 nm. Despite the small constitutional difference between these isomers, their retention times are separated by ~0.7 min (1.1 min vs. 1.8 min) on a standard reversed-phase column and four-minute method. Traces are offset slightly in both axes for clarity. **b.** Subset of a single-ion mode (SIM) trace of enzymatic Tyr (produced by 10 µM TmTyrS6 over 24 hours, in teal) and a 5-mM solution of authentic Tyr (yellow), showing nearly identical traces. *Inset*: the full trace. **c.** Calibration curve and analytical detection limit of Tyr. Solutions of Tyr at known concentration were analyzed in duplicate and the MS peaks were integrated. *Upper:* The concentrations can be related to the peak area by a logarithmic relationship. *Lower*: The peak areas deviate from the logarithmic relationship below 1 µM Tyr, at which point instrument noise and residual Tyr from non-exhaustive column washing makes quantification impossible and presents a lower limit of detection.

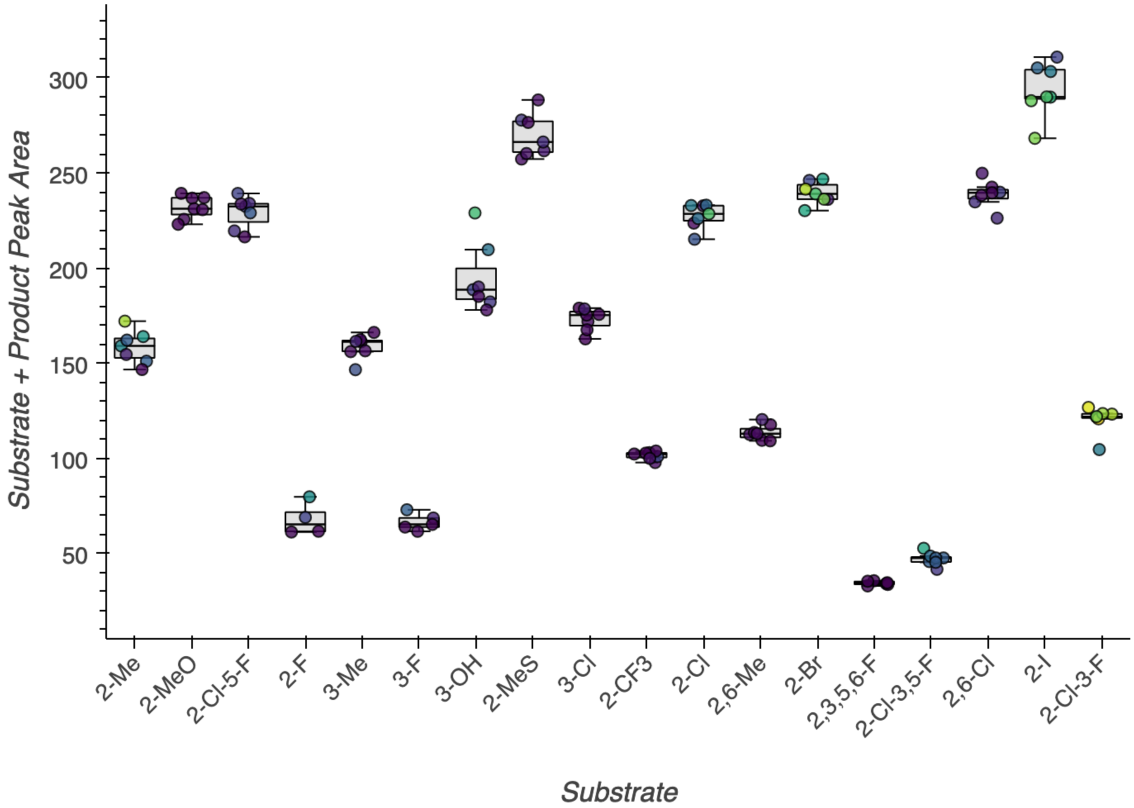


Figure E-11. Comparison of total HPLC areas from reactions in Figure 6-3a. See **Section E.2.3**, *Estimating product formation by HPLC peak area percentages* for details. Generally, the change in absorbance as the substrate is alkylated to product is minimal. The substrate and product peak areas are summed for each reaction and plotted along the y-axis, and then colored by product peak area percent. The total absorbance in the system changes minimally with product formation even when the reactions are high yielding, demonstrating that the substrate and product species absorb similarly and therefore their HPLC peak areas are proportional to their real concentrations, allowing estimation of product formation. Possible exceptions are 3-OH (3-hydroxyphenol) and 2-I (2-iodophenol), which result in an increase and decrease in total area with higher product formation, respectively. Shapes correspond to replicates (circles = replicate 1, squares = replicate 2).

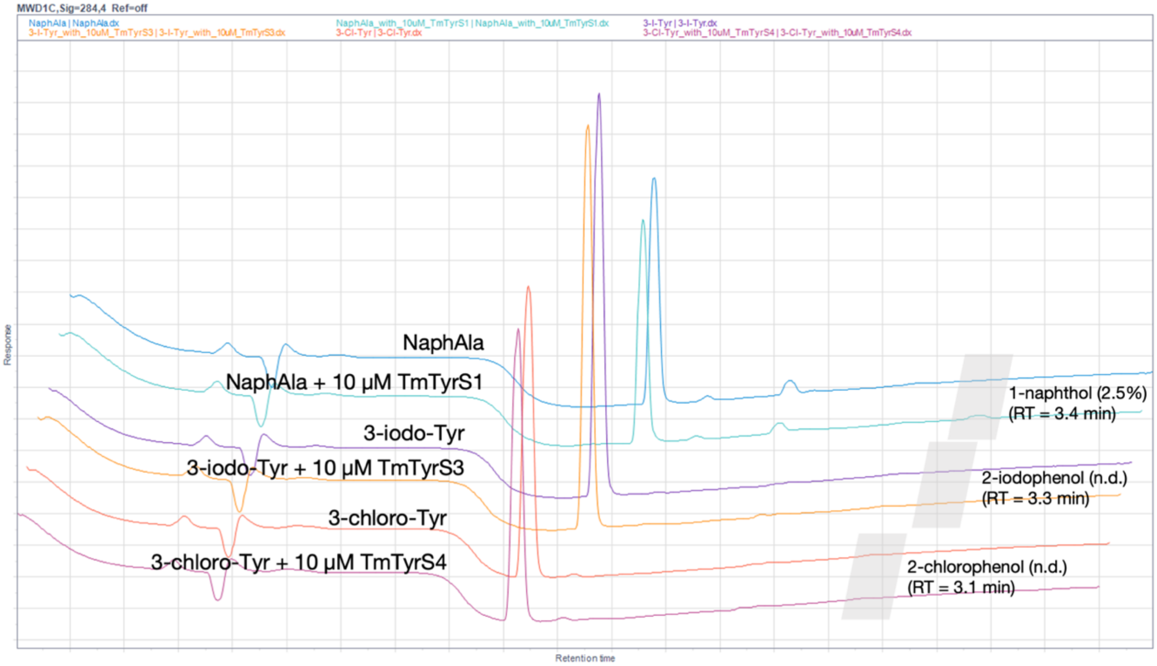


Figure E-12. TyrS enzymes are effectively irreversible. Saturated solutions of NaphAla, 3-iodo-Tyr, and 3-chloro-Tyr were incubated with 10 µM of the enzyme that has the highest activity toward each respective phenol analog. Under thermodynamic control, as with TPL, this would also be the enzyme that best degrades the product. Approximate areas in which the phenol analogs elute are designated in gray boxes. Only NaphAla shows very minor degradation at a rate far less than its rate of NaphAla synthesis, indicating that these reactions remain under kinetic control and are effectively irreversible.

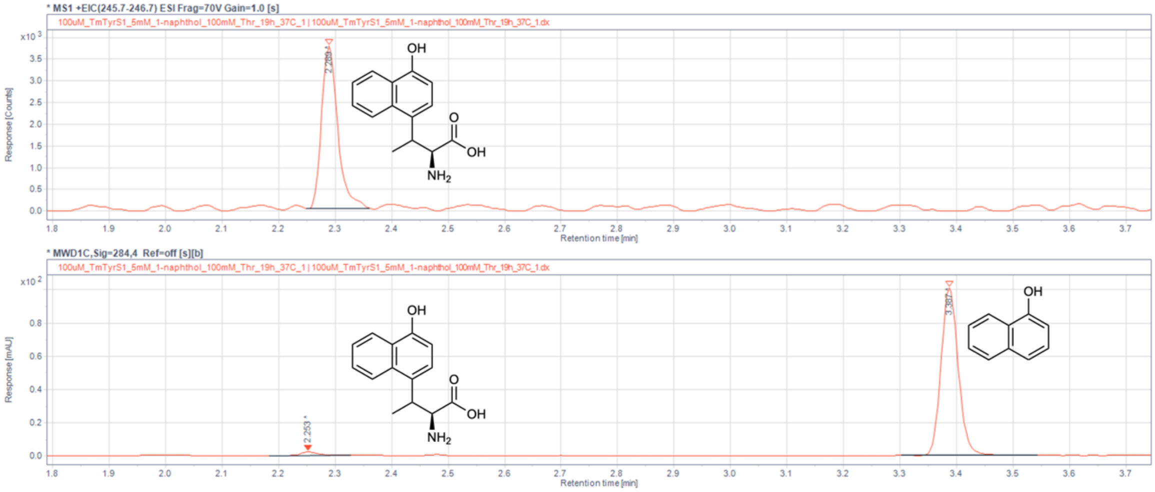
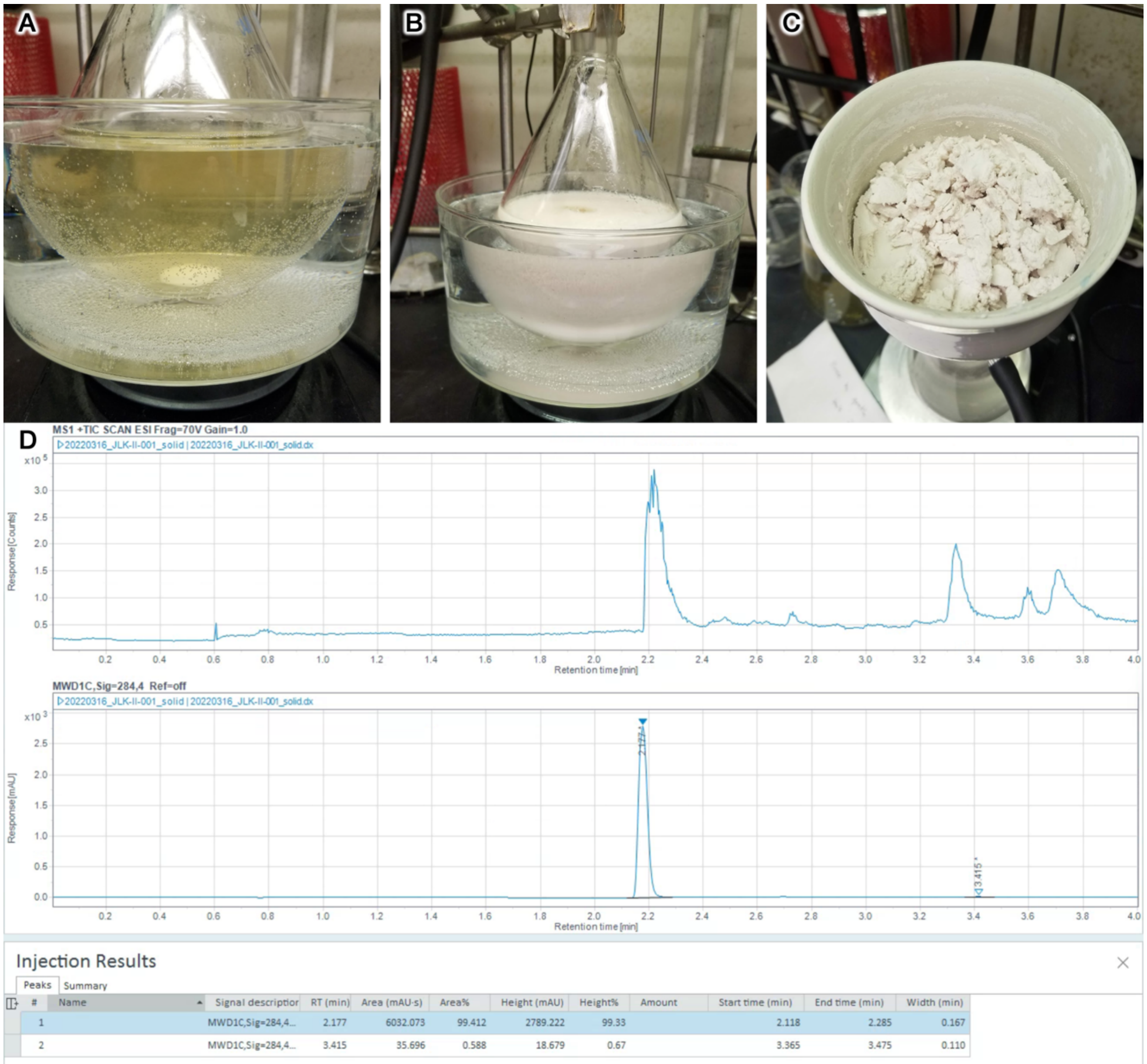
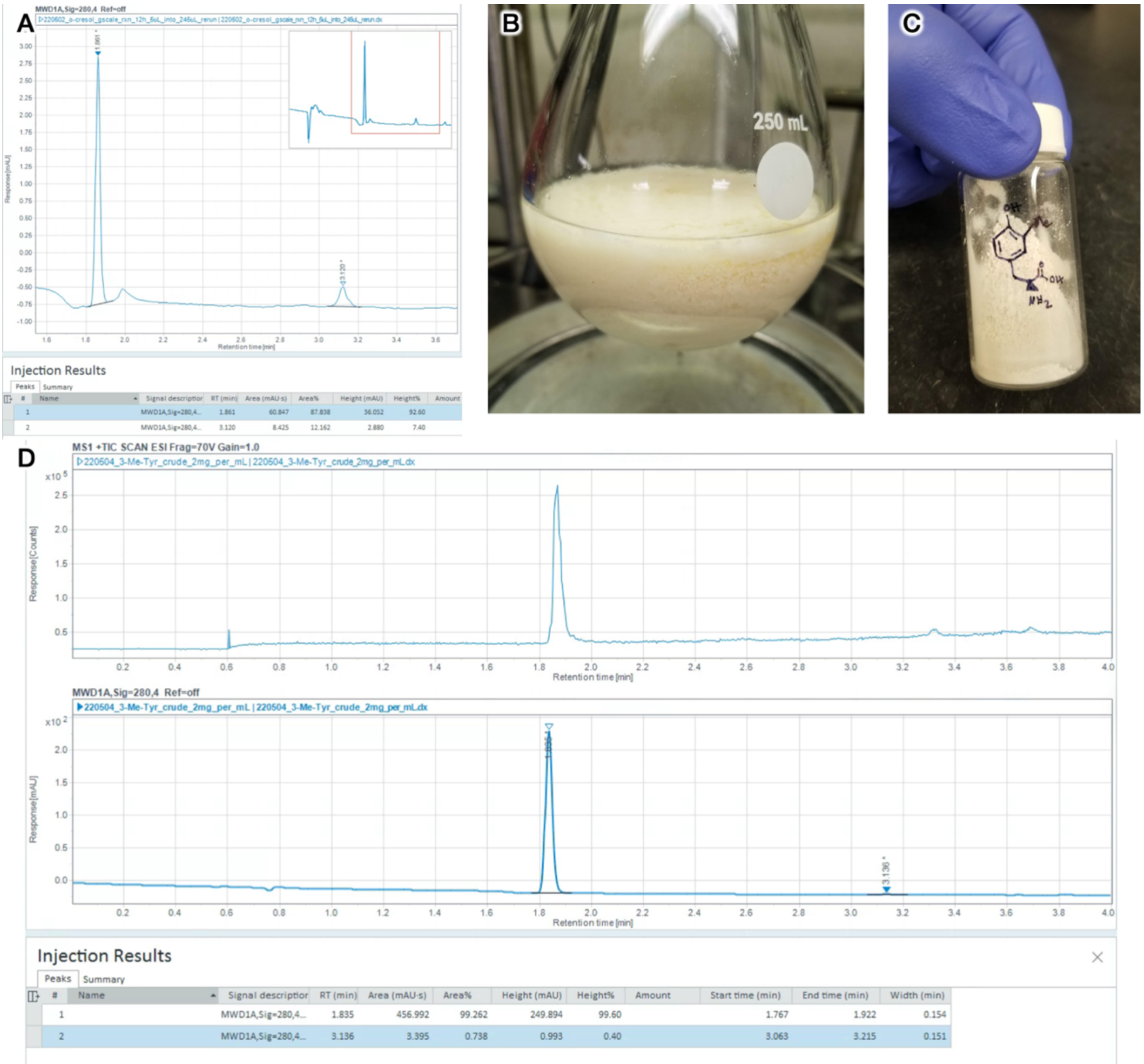


Figure E-13. Biocatalytic production of β-methyl-NaphAla in a single step from L-threonine (Thr) and NaphAla. Reactions run according to conditions given in Fig. 3B and analyzed by LCMS. *Upper*: Extracted ion counts for the condensation product of Thr and NaphAla (β-methyl-NaphAla, 246 m/z). *Lower*: Absorbance at 284 nm. The peak for β-methyl-NaphAla is integrated at 2.253 min (2.289 min on the MS, due to the 0.03 mL delay between UV detector and MS), 0.07 min later than NaphAla (retention time = 2.18 min). Integration at the isosbestic point between 1-naphthol and an amino acid product (284 nm) gives an average HPLC yield of 2% over three technical replicates.



#	Name	Signal description	RT (min)	Area (mAU·s)	Area%	Height (mAU)	Height%	Amount	Start time (min)	End time (min)	Width (min)
1		MWD1C,Sig=284,4...	2.177	6032.073	99.412	2789.222	99.33		2.118	2.285	0.167
2		MWD1C,Sig=284,4...	3.415	35.696	0.588	18.679	0.67		3.365	3.475	0.110

Figure E-14. Multi-gram-scale NaphAla preparation and characterization. a. Solution of TmTyrS1 and Ser in KPi, prior to the addition of 1-naphthol. **b.** The reaction after 1 day of continuous 1-naphthol addition, at which point the NaphAla product has crashed out of solution. **c.** The product can be isolated by simply collecting it in a filter, washing with ice-cold water and ethyl acetate, and then drying. **d.** Example of the purity of the isolate, which is >99% pure relative to 1-naphthol by HPLC.



Injection Results

Peaks Summary

#	Name	Signal descriptior	RT (min)	Area (mAU·s)	Area%	Height (mAU)	Height%	Amount
1		MWD1A,Sig=280,4...	1.861	60.847	87.838	36.052	92.60	
2		MWD1A,Sig=280,4...	3.120	8.425	12.162	2.880	7.40	

Injection Results

Peaks Summary

#	Name	Signal descriptior	RT (min)	Area (mAU·s)	Area%	Height (mAU)	Height%	Amount	Start time (min)	End time (min)	Width (min)
1		MWD1A,Sig=280,4...	1.835	456.992	99.262	249.894	99.60		1.767	1.922	0.154
2		MWD1A,Sig=280,4...	3.136	3.395	0.738	0.993	0.40		3.063	3.215	0.151

Figure E-15. Gram-scale 3-methyl-Tyr preparation and characterization. a. After 12 hours of reaction time at 50 mM 2-methylphenol, ~90% of the substrate had been converted to product. *Inset:* the entire HPLC trace. **b.** After a second addition of 50 mM 2-methylphenol and 10 hours, the product crashed out of solution. **c.** The product was isolated over a filter, washed, and dried to afford 1.13 g of pure 3-methyl-Tyr. **d.** HPLC purity of 3-methyl-Tyr to 2-methylphenol is >99%.

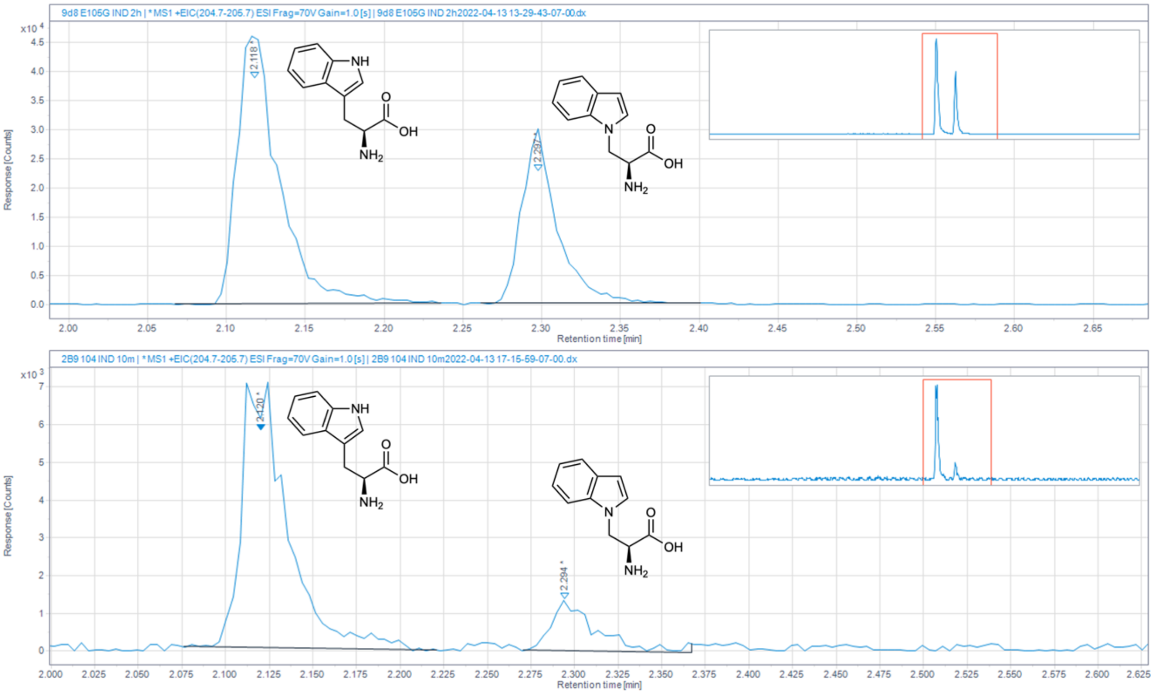


Figure E-16. Reactions of Tm9D8* E105G and *Pf*2B9 E104G with indole. Traces shown are extracted ion counts for 205 m/z (the molecular ion of Trp) which shows two peaks, one assigned to Trp (2.1 min) and one assigned to isoTrp (2.3 min), the N-alkylation product. *Upper*: 10 µM Tm9D8* E105G with 10 mM indole for 2 hours at 37 °C. *Lower*: 10 µM *Pf*2B9 E104G with 10 mM indole for 10 minutes. *Inset*: Full trace.

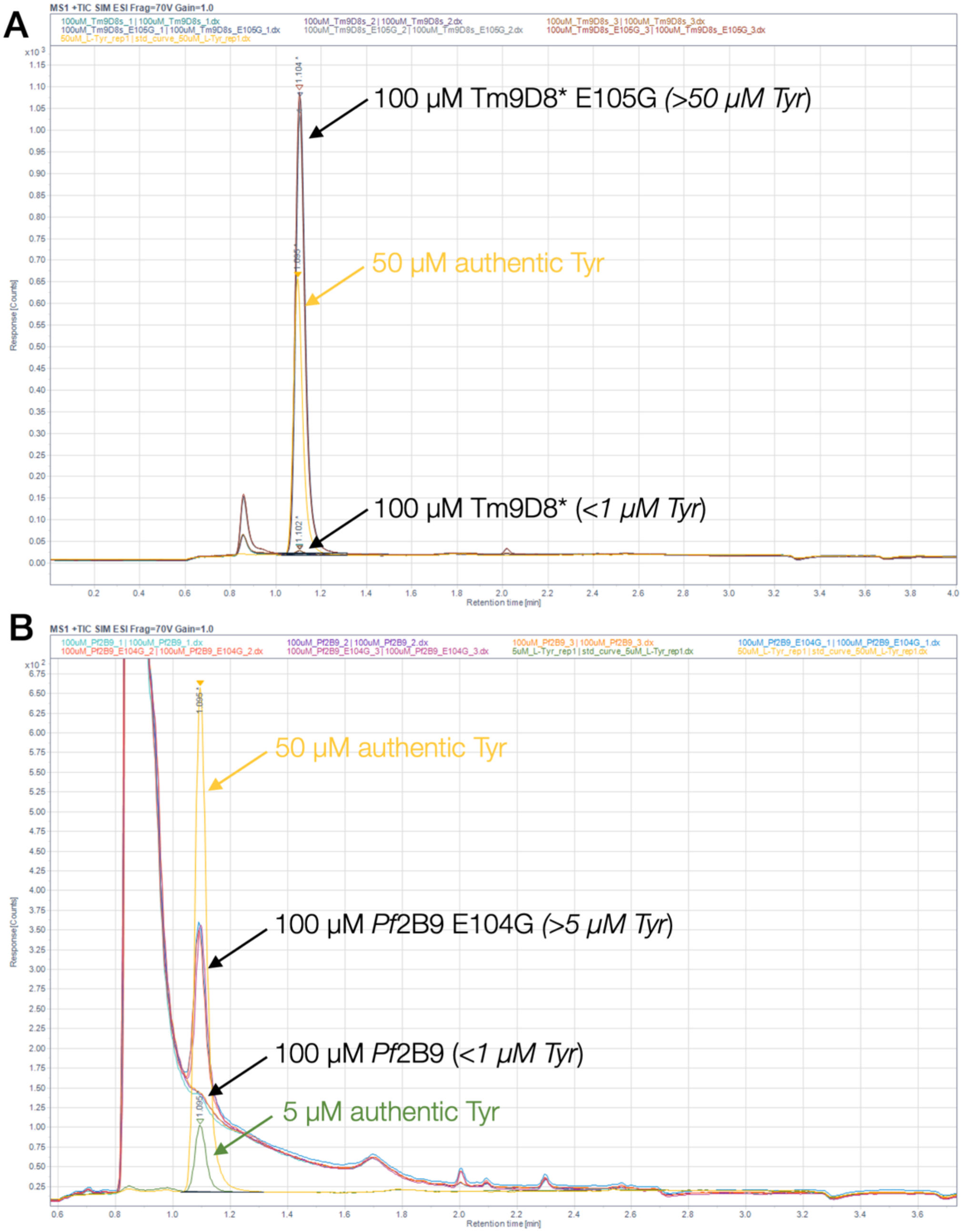


Figure E-17. Tyr formation in Tm9D8* E105G and *Pf*2B9 E104G. A reaction containing 100 µM enzyme with 50 mM phenol and 50 mM Ser was incubated at either **a.** 37 °C (Tm9D8* variants) or **b.** 75 °C (*Pf*2B9 variants) for 24 hours in technical triplicate, then analyzed by LCMS (single-ion mode for 182 m/z). A trace of 50 µM Tyr standard is shown for reference (yellow). Tm9D8* E105G exceeds 50 µM Tyr per 100 µM enzyme per 24 hours. A small peak is integrated for Tm9D8*, but this is below the threshold of reliable detection (see **Figure E-10c**; this is likely due to residual Tyr on the column). *Pf*2B9 E104G forms a small amount of Tyr. *Pf*2B9 shows only a small, unreliable peak (again, see **Figure E-10c**). A trace of 5 µM Tyr (~5x above the limit of detection) is shown for reference in green.

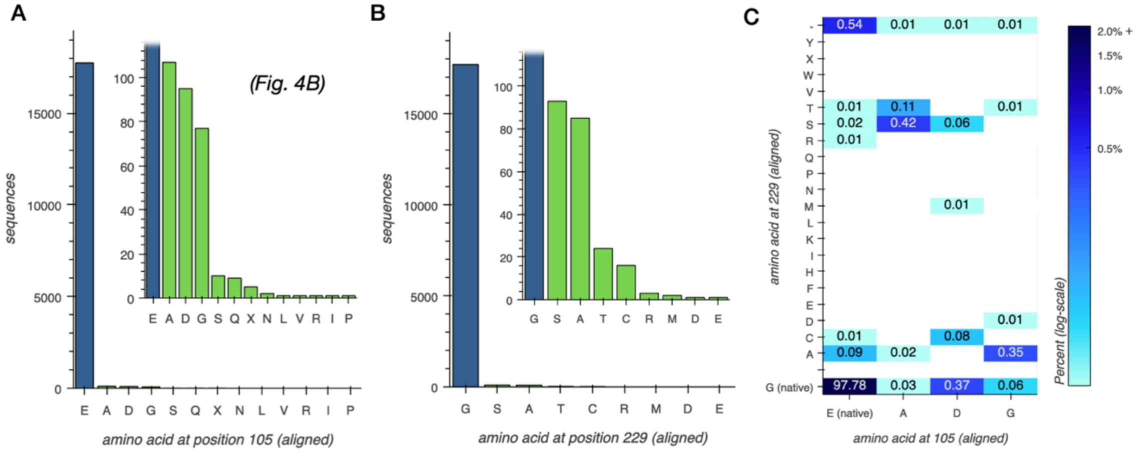


Figure E-18. Conservation of E105 and G229 in aligned TrpB-like sequences. a. Number of sequences with a given amino acid at position 105. *Inset*: Axis-adjust view. The is the same plot as **Figure 6-4b**. A, D, and G are the only other amino acids with significant frequencies. **b.** Number of sequences with a given amino acid at position 229. *Inset*: Axis-adjust view. S, A, and T are the only other amino acids with significant frequencies. **c.** Correlations of amino acids found at position 105 and 229. While the vast majority of sequences contain the residues native to *Tm*TrpB (E105, G229; 97.78%), those with non-carboxylate (e.g., not E or D) sidechains at position 105 are correlated with different amino acid identities at 229: E105A + G229S/T (0.53%); and E105G + G229A (0.35%). The latter are the same mutations identified in this study that proved most beneficial for activity with phenol.

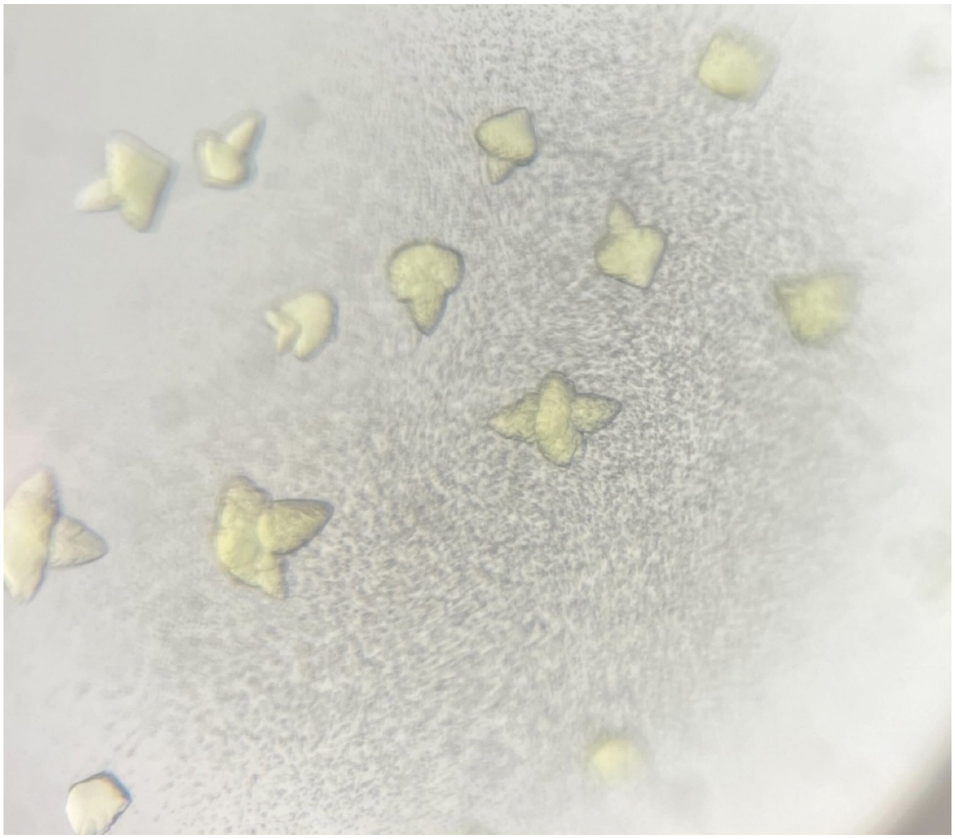


Figure E-19. Crystals of TmTyrS1. The morphology is fairly amorphous but yielded excellent diffraction and were amenable to small-molecule soaking experiments.

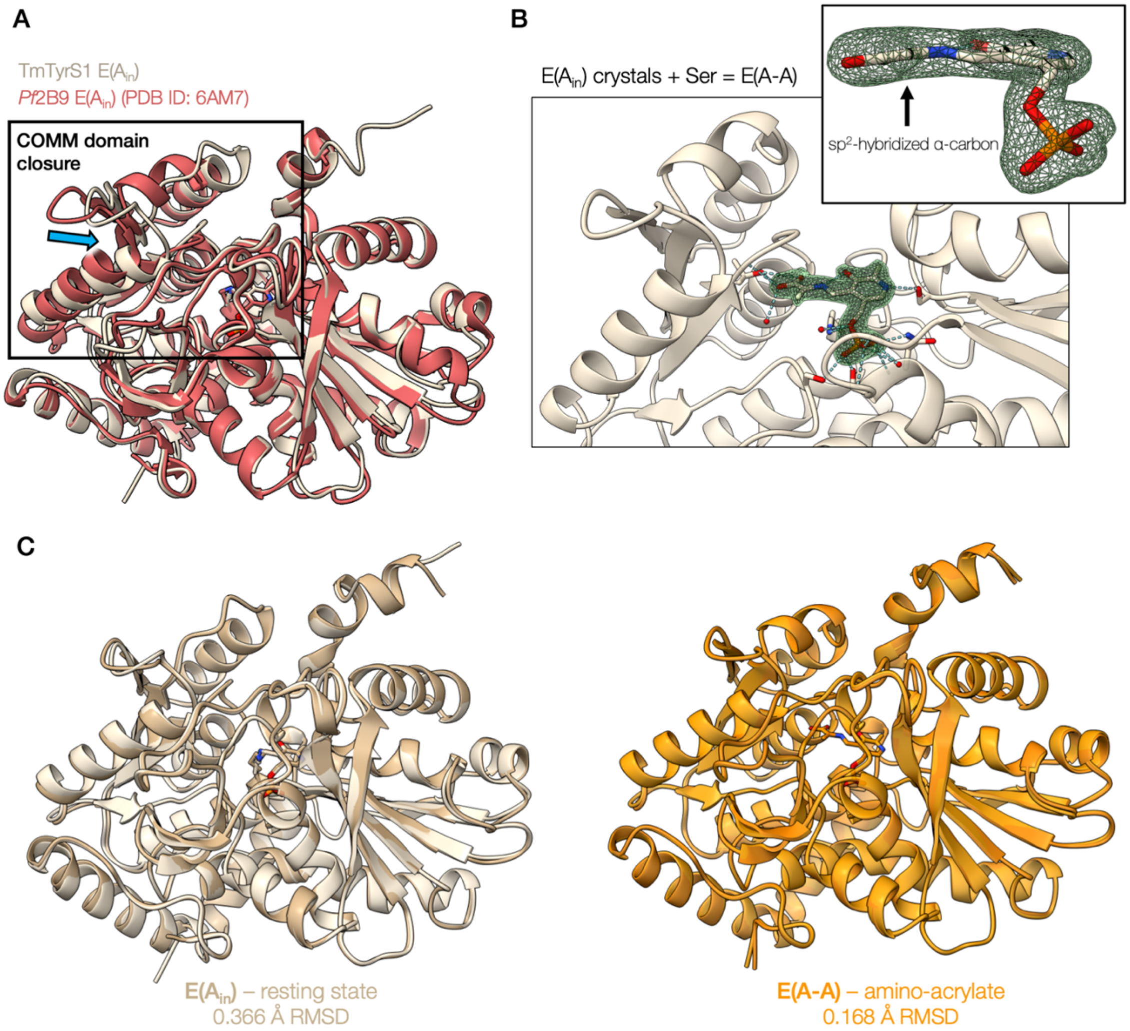


Figure E-20. Crystal structures of TmTyrS1. a. The internal aldimine E(A_{in}) resting state (beige) compared to the E(A_{in}) state of a different stand-alone TrpB variant, *Pf*2B9 (red, PDB: 6AM7). **b.** The active amino-acrylate-bound E(A-A) state can be obtained by simple soaking of E(A_{in}) crystal with Ser. *Inset*: Head-on view of the amino-acrylate, with the polder omit map contoured at 5σ, demonstrating the planarity of the sp^2-hybridized Cα. **c.** Structural changes resulting from the addition of Ser to E(A_{in}) crystals, reducing the conformational heterogeneity between the two TmTyrS1 subunits in the asymmetric unit (each subunit aligned and colored lighter or darker).

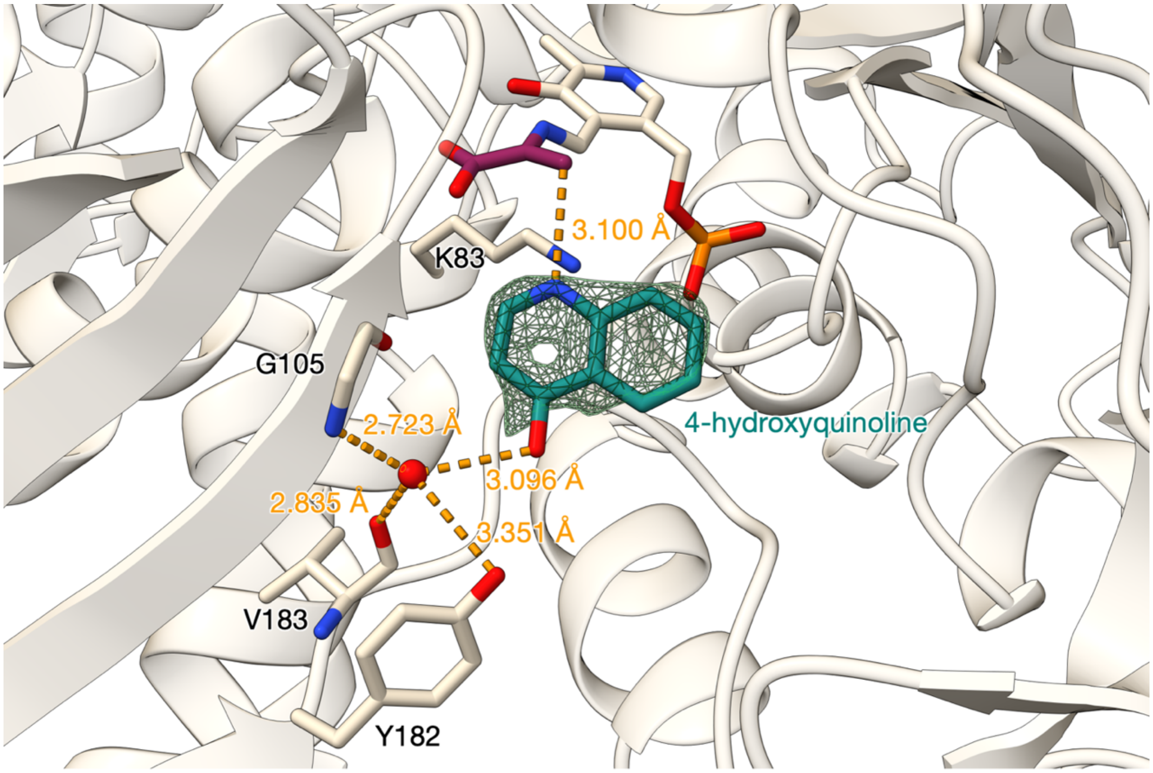


Figure E-21. TmTyrS1 in the E(A-A) state bound with 4-hydroxylquinoline. Distances for hydrogen bonding and other electrostatic interactions are shown. Polder omit map contoured at 7σ.

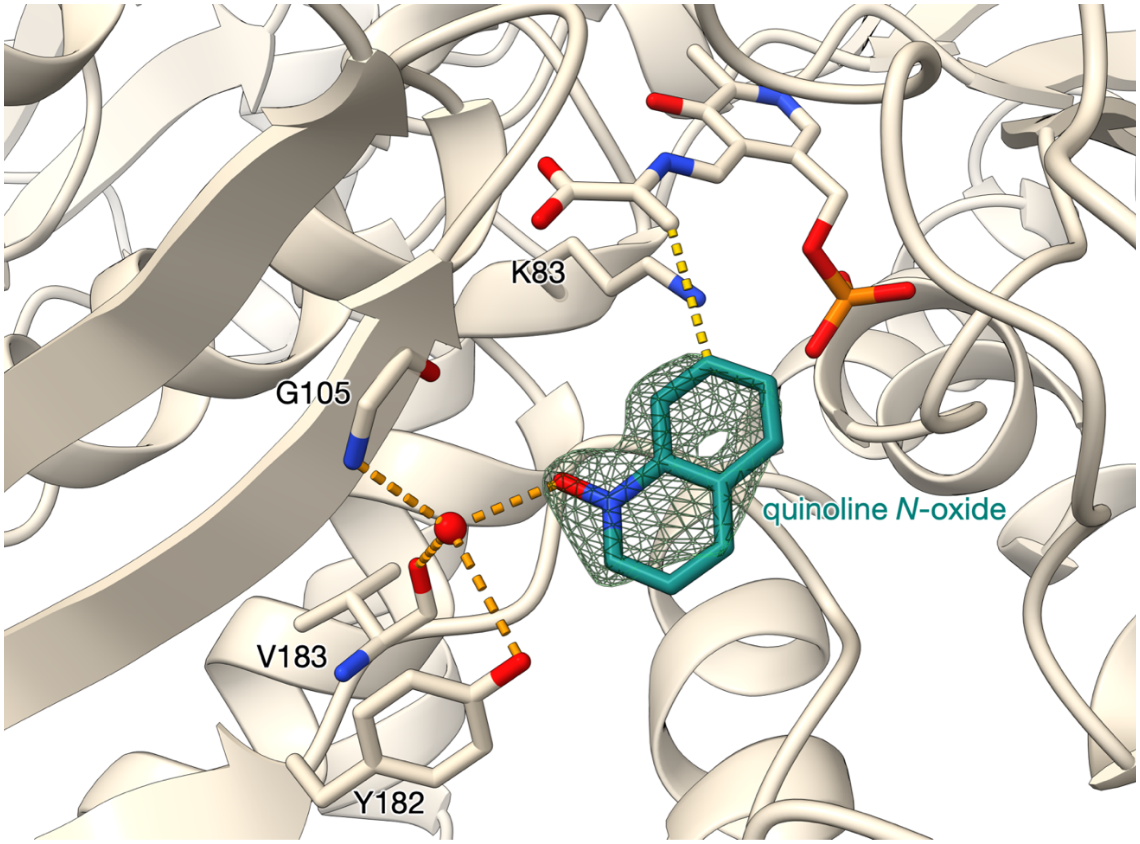


Figure E-22. TmTyrS1 in the E(A-A) state bound with quinoline *N*-oxide. Polder omit map contoured at 7.6σ.

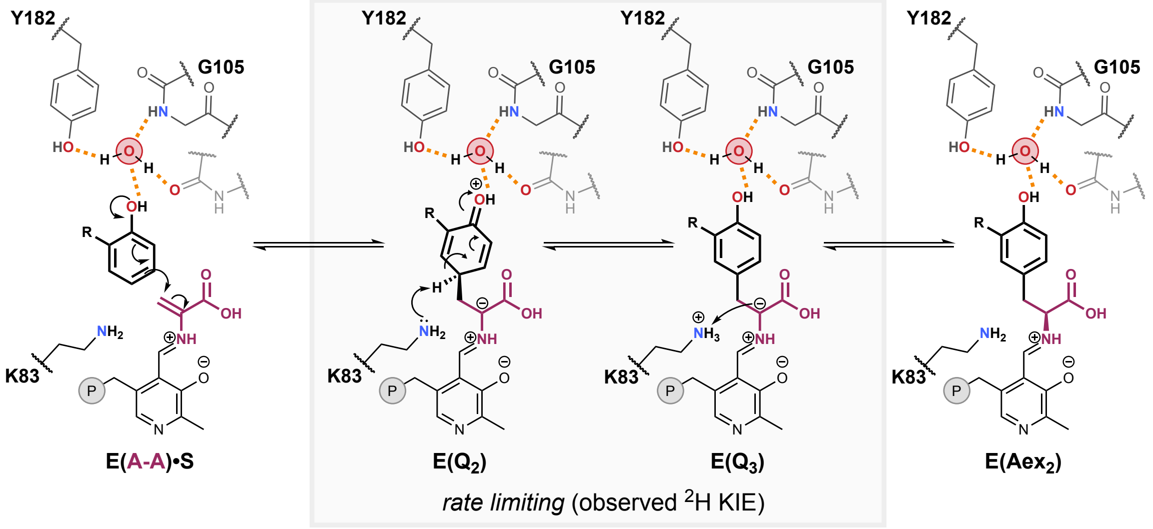


Figure E-23. Proposed mechanism of step-wise alkylation of phenolic substrates by TyrS variants. Based on observed hydrogen bonding interactions and kinetic isotope effects (KIEs), a catalytic water coordinates phenol after the amino-acrylate E(A-A) state is reached to form a Michaelis complex (E(A-A)•S) that facilitates regioselective alkylation. Deprotonation after C–C bond formation (converting from quinonoid intermediate $E(Q_2)$ to $E(Q_3)$) is rate limiting based on observed primary KIEs. A final protonation event produces the covalently bound L-Tyr analog as an external aldimine, $E(Aex_2)$. The hydrogen bonding arrangement to Y182 is inferred based on the binding of 4-hydroxyquinoline over its tautomer 4-quinolone, but this lacks more rigorous evidence.

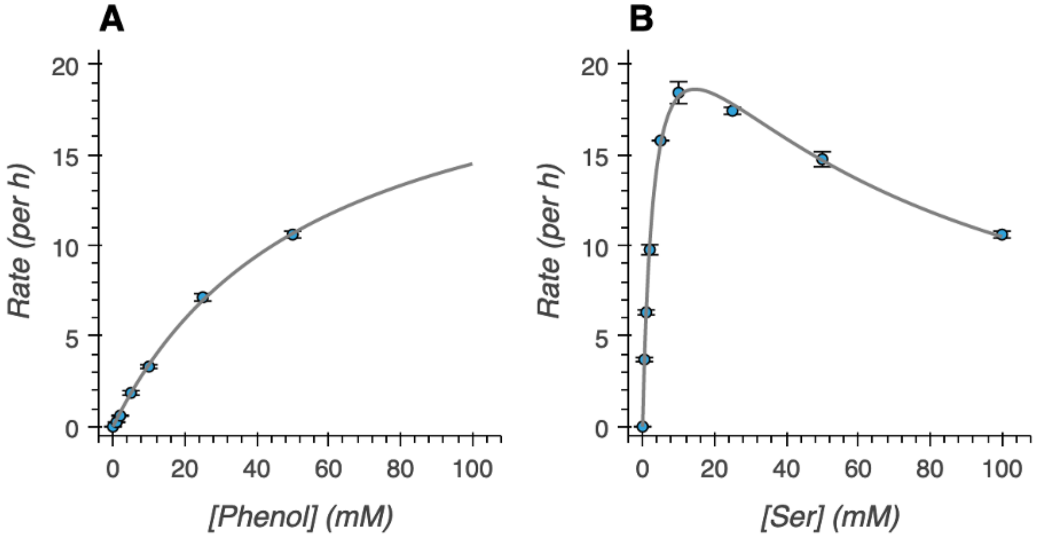


Figure E-24. Kinetics of TmTyrS6. a. Initial rates at varying phenol concentration, holding Ser at 100 mM. The apparent K_M is 56 mM, and k_{cat} is 23 per hour. Saturation behavior would be observed at a concentration that far exceeds the solubility of phenol, likely due to poor binding interactions with the small, symmetric substrate. **b.** Initial rates at varying Ser concentration, holding phenol at 50 mM. A moderate inhibitory effect is observed, and the data is modeled according to the substrate-inhibition model. The apparent K_M is 3.3 mM, k_{cat} is 27 per hour, and K_i is 65 mM. Error bars represent the standard error of the mean for the initial rate estimations. These rates were determined from time courses that were observed to be under steady-state conditions for n=3 time points (0, 4, and 8 hours, using an implicit zero time point) for all samples except 0.5, 1, and 2 mM Ser which used n=4 time points (0, 0.5, 1, and 2 hours, using an implicit zero). See **Section E.1.11**, *Michaelis-Menten kinetics* for more details.

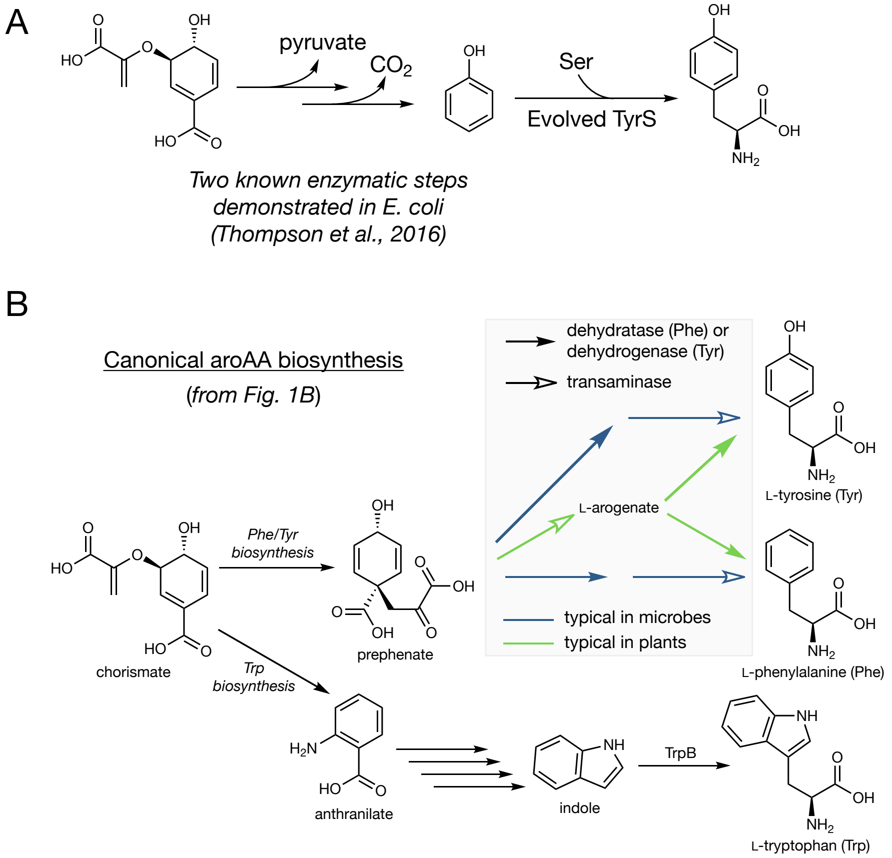


Figure E-25. A new logic for post-chorismate Tyr biosynthesis. a. Phenol is accessible from choristmate it two known enzymatic steps, and has been demonstrated in *E. coli*.[18] An evolved TyrS enzyme could complete this pathway to Tyr. **b.** This would expand the canonical pathways of aromatic amino acid (aroAA) biosynthesis, as life uses a universally conserved set of chemistries to produce aroAAs.

E.4 Supplementary Tables

Table E-1. Amino acid substitutions identified in this study.

Variant	Reference Variant	Amino acid substitutions
Tm9D8*	WT *Tm*TrpB	P19G, E30G, I69V, K96L, P140L, N167D, I184F, L213P, G228S, T292S
Tm9D8* E105G	Tm9D8*	E105G
TmTyrS1	Tm9D8* E105G	Y4N, Y12N, F41Y, I103T, F184P, V291A, S302P, R389H
TmTyrS2	TmTyrS1	A87T, I128V, H191Y
TmTyrS3	TmTyrS2	L147Q, V227M
TmTyrS4	TmTyrS3	I174T, P184A, S265P
TmTyrS5	TmTyrS4	A150V, Y181H, G229A, P265S[a]
TmTyrS6	TmTyrS5	K139R, L170F, W286G

[a] Reversion.

Table E-2. Turnover frequencies (TOFs) from Figure 6-2c and d.

Variant	[Enzyme] (μM)	[Phenol] (mM)	Time (h)	RT [min]	Area	[Tyr] (μM)	Turnovers	TOF (h^{-1})	Mean TOF (h^{-1})
Tm9D8*	100.0	50	24	1.102	28.7681	0.42	0.0	0.0	0.0
	100.0	50	24	1.105	26.2847	0.38	0.0	0.0	
	100.0	50	24	1.104	21.0719	0.29	0.0	0.0	
Tm9D8* E105G	100.0	50	24	1.107	3217.4171	96	0.96	0.04	0.04
	100.0	50	24	1.104	3172.4534	95	0.95	0.04	
	100.0	50	24	1.104	3301.2609	99	0.99	0.04	
TmTyrS1	100.0	50	24	1.101	2459.7951	70	0.70	0.03	0.03
	100.0	50	24	1.098	2575.7915	74	0.74	0.03	
	100.0	50	24	1.103	2556.0619	74	0.74	0.03	
TmTyrS2	50.0	50	24	1.095	3996.225	123	2.5	0.10	0.10
	50.0	50	24	1.097	4038.0278	125	2.5	0.10	
	50.0	50	24	1.101	4148.5208	129	2.6	0.11	
TmTyrS3	20.0	50	24	1.098	2952.4932	87	4.4	0.18	0.19
	20.0	50	24	1.102	3123.5107	93	4.6	0.19	
	20.0	50	24	1.098	3149.8636	94	4.7	0.2	
TmTyrS4	10.0	50	24	1.102	1869.0759	51	5.1	0.21	0.22
	10.0	50	24	1.095	1860.3236	51	5.1	0.21	
	10.0	50	24	1.098	1986.5306	55	5.5	0.23	
TmTyrS5	10.0	50	24	1.094	28267.4655	1170	117	4.9	5.0
	10.0	50	24	1.1	28779.4787	1200	120	5.0	
	10.0	50	24	1.098	29631.6258	1240	124	5.2	
TmTyrS6	10.0	50	24	1.096	72206.0343	3460	346	14	14
	10.0	50	24	1.099	71630.5227	3430	343	14	
	10.0	50	24	1.099	67582.34	3210	321	13	

Area corresponds to the integrated MS peak area of Tyr (RT = 1.1 min). These are converted to real concentrations (in μM) using the calibration curve from **Figure E-10c**, which can then be converted to enzymatic turnovers and turnover frequency (TOF) by normalizing to the enzyme loading and reaction time.

Table E-3. Data collection and refinement statistics for TmTyrS1 structures.

Structure	**TmTyrS1–E(A_{in})**	**TmTyrS1–E(A-A)**	**TmTyrS1–E(A-A) + quinoline *N*-oxide**	**TmTyrS1–E(A-A) + 4-hydroxylquinoline**
Unit cell				
Space group	*I*4	*I*4	*I*4	*I*4
a, b, c (Å)	164.6, 164.6, 83.1	164.6, 164.6, 84.3	164.4, 164.4, 84.3	164.9, 164.9, 84.06
α, β, ɣ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Data collection				
Wavelength (Å)	0.97946	0.97946	0.97946	0.97946
Resolution (Å)	39.15 – 2.05	45.97 – 1.90	38.74 – 1.70	38.86 – 2.00
Total/unique no. of reflections	949735/69702	1195353/88525	1659189/123166	1016930/75913
R_{merge}[a,b]	0.07 (2.09)	0.16 (2.43)	0.09 (1.99)	0.11 (1.73)
$R_{\text{p.i.m.}}$[a,c]	0.03 (0.58)	0.04 (0.67)	0.02 (0.56)	0.03 (0.48)
$CC_{1/2}$[a,d]	1.00 (0.62)	0.99 (0.58)	1.00 (0.59)	0.99 (0.69)
$I/\sigma(I)$[a]	21.0 (1.5)	12.0 (1.4)	17.6 (1.5)	15.9 (2.0)
Redundancy[a]	13.6 (14.0)	13.5 (14.0)	13.5 (13.6)	13.4 (13.8)
Completeness[a] (%)	100 (100)	100 (100)	100 (100)	99.8 (99.9)
Refinement				
No. of reflections used in refinement/test set	69678/3488	88500/4531	123151/12284	75899/7556
R_{work}[a,e]	0.216 (0.342)	0.170 (0.256)	0.167 (0.262)	0.187 (0.260)
R_{free}[a,e]	0.232 (0.363)	0.192 (0.294)	0.186 (0.300)	0.219 (0.318)
No. of nonhydrogen atoms				
protein	5956	5924	6016	5802
ligand	24	68	82	66
solvent	221	270	321	172
root-mean-square deviation from ideal geometry				
bonds (Å)	0.002	0.019	0.018	0.003
angles (°)	0.48	1.46	1.43	0.59
Ramachandran plot[f] (%)				
favored	97.09	98.17	97.78	97.90
allowed	2.91	1.83	1.96	1.97
disallowed	0.00	0.00	0.26	0.13
PDB accession code	8EGY	8EGZ	8EH0	8EH1

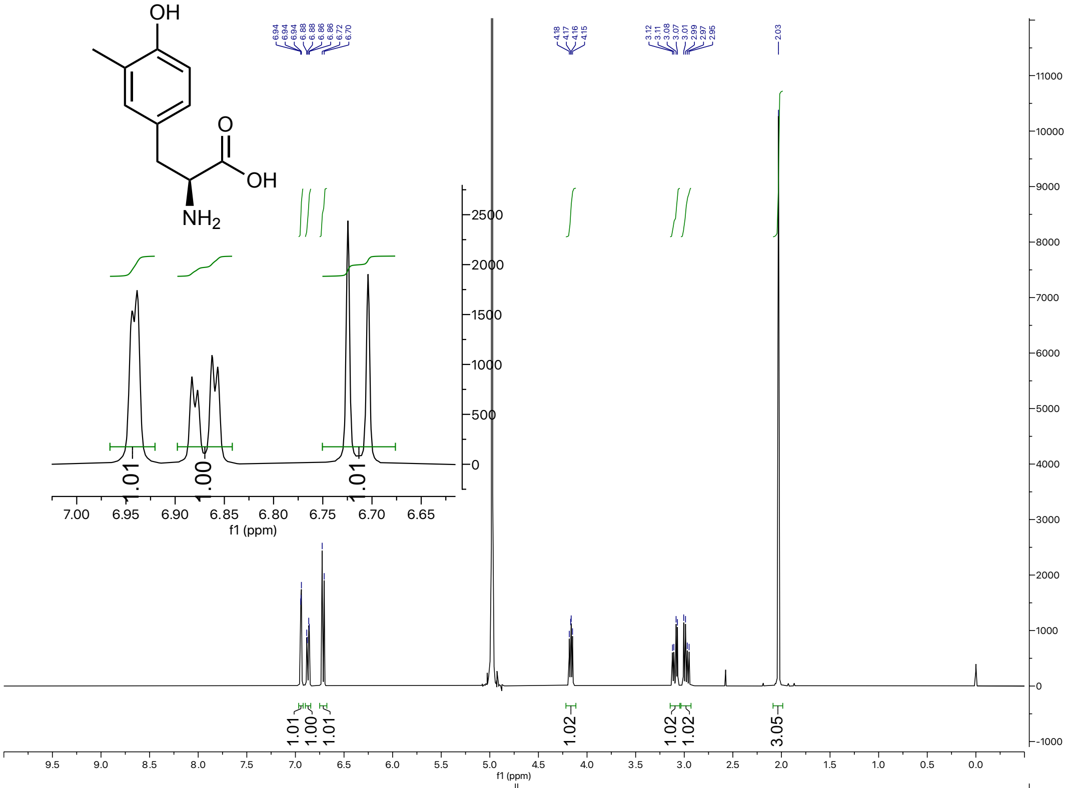
[a]Values in parentheses refer to data in the highest shell. [b]$R_{\text{merge}} = \sum_{hkl}\sum_{i}|I_{i,hkl} - \langle I\rangle_{hkl}|/\sum_{hkl}\sum_{i} I_{i,hkl}$, where $\langle I\rangle_{hkl}$ is the average intensity calculated for reflection *hkl* from replicate measurements. [c]$R_{\text{p.i.m.}} = (\sum_{hkl}(1/(N\text{-}1))^{1/2}\sum_{i}|I_{i,hkl} - \langle I\rangle_{hkl}|)/\sum_{hkl}\sum_{i} I_{i,hkl}$, where $\langle I\rangle_{hkl}$ is the average intensity calculated for reflection *hkl* from replicate measurements and N is the number of reflections. [d]Pearson correlation coefficient between random half-datasets. [e]$R_{\text{work}} = \sum||F_o| - |F_c||/\sum|F_o|$ for reflections contained in the working set. $|F_o|$ and $|F_c|$ are the observed and calculated structure factor amplitudes, respectively. R_{free} is calculated using the same expression for reflections contained in the test set held aside during refinement. [f]Calculated with PROCHECK.

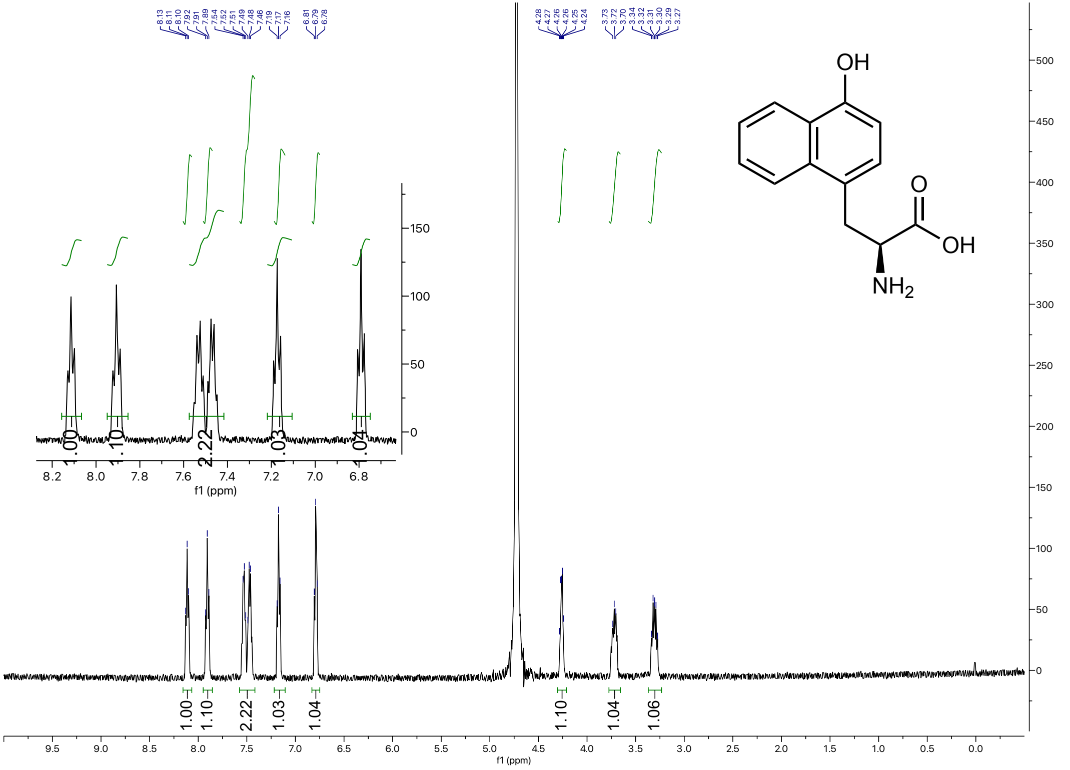
Table E-4. Competitive kinetic isotope effects (KIEs).

		Approximate KIE	
Variant	**Substrate**	**Average**	**Deviation**
Tm9D8* E105G	2-chlorophenol	1.620	0.088
	2-methylphenol	1.904	0.014
	Ser	0.930	0.028
TmTyrS1	2-chlorophenol	2.038	0.033
	2-methylphenol	2.787	0.020
	Ser	0.755	0.012
TmTyrS2	2-chlorophenol	2.371	0.157
	2-methylphenol	3.638	0.342
	Ser	0.882	0.002
TmTyrS3	2-chlorophenol	3.260	0.394
	2-methylphenol	4.005	0.491
	Ser	0.874	0.034
TmTyrS4	2-chlorophenol	2.810	0.022
	2-methylphenol	3.554	0.587
	Ser	0.893	0.034
TmTyrS5	2-chlorophenol	1.978	0.136
	2-methylphenol	2.405	0.043
	Ser	0.935	0.039
TmTyrS6	2-chlorophenol	2.101	0.363
	2-methylphenol	3.018	0.065
	Ser	0.882	0.010

Reactions performed in technical duplicate. KIEs measured in competition between the standard and deuterated substrate under as short of reaction times as possible to achieve 1–10% yield with minimal C–H proton exchange in the solvent. Ser reactions performed using 2-chlorophenol as the phenolic substrate; see **Section E.1.13**, *Measuring kinetic isotope effects* for more details. While these results unambiguously identify a primary KIE for deprotonation of C4 of phenolic substrates and no primary KIE for deprotonation of Cα of Ser, they should not be used for inferring trends occurring from evolution until more rigorous methodology (e.g., not in competition and/or a more sensitive method of quantification like single-ion mode) is used.

E.5 NMR Spectra





Appendix E Bibliography

1. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
2. Kille, S. *et al.* Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2**, 83–92 (2013).
3. Xia, Y., Chu, W., Qi, Q. & Xun, L. New insights into the QuikChange^{TM} process guide the use of Phusion DNA polymerase for site-directed mutagenesis. *Nucleic Acids Res.* **43**, e12 (2015).
4. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. & Arnold, F. H. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16**, 258–261 (1998).
5. Bhushan, R. & Brückner, H. Use of Marfey's reagent and analogs for chiral amino acid analysis: Assessment and applications to natural products and biological systems. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **879**, 3148–3161 (2011).
6. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
7. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
8. Buller, A. R. *et al.* Directed evolution of the tryptophan synthase β-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14599–14604 (2015).
9. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
10. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
11. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
12. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular

crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).

13. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).

14. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

15. Wittmann, B. J., Johnston, K. E., Almhjell, P. J. & Arnold, F. H. evSeq: Cost-effective amplicon sequencing of every variant in a protein library. *ACS Synth. Biol.* **11**, 1313–1324 (2022).

16. Milić, D. *et al.* Structures of apo- and holo-tyrosine phenol-lyase reveal a catalytically critical closed conformation and suggest a mechanism for activation by K^+ ions. *Biochemistry* **45**, 7544–7552 (2006).

17. Niks, D. *et al.* Allostery and substrate channeling in the tryptophan synthase bienzyme complex: Evidence for two subunit conformations and four quaternary states. *Biochemistry* **52**, 6396–6411 (2013).

18. Thompson, B., Machas, M. & Nielsen, D. R. Engineering and comparison of non-natural pathways for microbial phenol production. *Biotechnol. Bioeng.* **113**, 1745–1754 (2016).