Expanding Enzyme Function through Data-Guided Evolution

Thesis by

Ravi Goel Lal

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2026

(Defended July 18, 2025)

Ravi Goel Lal

ORCID: 0000-0001-6943-4147

ACKNOWLEDGEMENTS

I am deeply grateful to have reached this milestone in my academic journey. Completing this Ph.D. thesis has been both a challenging and profoundly rewarding experience. In all honesty, I could not have completed this work without through the support, generosity, and encouragement of many individuals. I owe sincere thanks to all those who contributed their time, insight, and friendship along the way; this work stands not only as a reflection of my own efforts, but as a testament to the community that carried me through it.

Firstly, I would like to convey my deepest gratitude to my advisor, Prof. Frances H. Arnold, for all of the insight and guidance she has given me throughout my time at Caltech. No single person has been as influential to my thinking as a scientist as Frances has. Every meeting with Frances is a privilege; you may not always learn something new, but you will certainly think a little differently afterwards. I have greatly benefited from the environment that Frances fosters in which we are taught to continuously ask, “Why are we doing this?” Frances, thank you for always letting me pursue my ideas while gently nudging me in the right direction. I hope to never forget what you have taught me: that our job as scientists is to help the Earth and the people that live on it.

Next, thank you to the members of my thesis committee: Prof. Mikhail G. Shapiro, Prof. Stephen L. Mayo, Prof. Joe Parker, and Prof. Brian M. Stoltz. I appreciate the precious time and counsel you have given me over the last six years. I recall my candidacy exam during the early days of the pandemic, when Zoom was still slightly confusing.

I want to express my deepest appreciation for two members of the Chemical Engineering department: Prof. Mike Vicic and Allison Kinard. Together the both of you have made the department an incredibly welcoming environment. Mike, I have thoroughly enjoyed your mentorship and friendship during my time at Caltech. I always look forward to running into you around campus so that we can talk about life, goings-on in science, and gossip about the department. Allison, I want to thank you for making me always feel like there is someone looking out for me in the department. I always feel safe in knowing that I can come to you with any question I may have with regards to completing my PhD. Whenever I get to speak with either of you I feel a sense of ease. While the sweets and treats at Sweets and Treats are very nice, I come for the chance to chat with both of you.

There is one person who has influenced my growth not just as a scientist, but also as an adult, as much as Frances: Dr. Sabine Brinkmann-Chen. Sabine, I am so, so grateful to have had you in my life throughout my PhD. I know that when I walk into our office every day that I will be met with a friend and a mentor who I can come to with any problem, and I know that you will always give me your honest opinion. I can say without a doubt that I would not have made it through graduate school without your continuous guidance. Also, thank you for reading every part of this thesis and making it a stronger, more rigorous work. I wouldn’t have given up my seat in our office for anything.

Often, when I attend socials at Caltech, I wonder why I don't know anyone there. I believe that this is because the people in the Arnold lab have consumed my attention. I have had the privilege to work with some of the coolest, smartest, nerdiest, fittest people I have ever met in the Arnold lab. Firstly, I would like to thank Dr. David C. Miller, Dr. Nick J. Porter, and Dr. Nat W. Goldberg. The teachings and insights you gave me in my early years at Caltech were operative to the way I think about protein engineering and biocatalysis now. Dave, thank you for teaching me how to think about and perform organic chemistry when I was basically a blank slate, and for taking me under your wing when I started in the Arnold lab. Nick and Nat thank you for continuously being willing to answer my most naïve questions about proteins and biocatalysis and making me want to understand more about the proteins we engineer.

Nearly every member of the Arnold lab has touched my life in one way or another. I want to extend a special thank you to Jae, Edwin, and Katie. You three are my family. The adventures that we have gone on in the past three years are some of my most cherished memories from graduate school. The Arnold lab has truly become a home to me over the last six years, and I ascribe this primarily to you.

Next, thank you to all of the friends who have supported me throughout my graduate studies. From the Tustin squad to all of the members of the group chat 'small group' I want to convey my gratitude to you in helping to maintain my mental health. In particular I want to thank all of the people I have gotten to live with these past six years. Alec and Kayla, you were so important to making Pasadena a home during the pandemic. Chirag, Vinay, and Anjan thank you for providing a roof when I was working in San Francisco. Ishaan and Mark, thank you both for founding the Lyndon Lounge with me.

Finally, I want to thank my family. To my mom, who is the strongest person I know, thank you for supporting me in basically everything I have ever done. To my dad, thank you for making me want to be a scientist since I was a child and providing a guiding hand through my PhD. To my brother, Rohit, thank you for always being a phone call away whenever I feel lonely.

ABSTRACT

Through the process of evolution, Nature has optimized enzymes for the chemical transformations which drive biology. The use of enzymes for chemical synthesis is enticing as these privileged catalysts can facilitate reactions in a highly sustainable and selective manner. Directed evolution (DE) is a strategy for engineering proteins to improve a desired function. This approach has been demonstrated to be of great effect for the development of enzymatic activities which have never been naturally observed. These 'new-to-nature' chemistries not only showcase the power of DE to unlock novel functions, but also the potential of biocatalysis to deliver high-value chemical compounds. This thesis details the exploration of new-to-nature biocatalytic reactions of hemoproteins using traditional DE and novel machine learning-assisted directed evolution (MLDE) methods. Chapter I provides an overview of how DE has enabled new-to-nature-biocatalysis, and the challenges associated with developing this novel biocatalytic reactions. In Chapter II, a cytochrome P411 is found to catalyze a carbene-mediated [1,2]-Stevens rearrangement to furnish azetidines via ring expansion, and is engineered to deliver these products with high yield and enantioselectivity. The evolution of this activity is a demonstration of how established DE techniques can be utilized to arrive at enzymes capable of unprecedented chemistry. Chapter III describes efforts towards advancing this ring-expansion chemistry for the synthesis of proline analogs. Though this activity could not be found with existing sequence diversity, several engineering insights were made about protoglobins (small, thermostable hemoproteins). Chapters II and III both highlight challenges associated with DE. In Chapter IV, active-learning assisted directed evolution (ALDE) is introduced as a workflow which leverages MLDE in an iterative fashion to greatly accelerate DE efforts. Alongside simulations on combinatorially complete datasets, ALDE was validated in the wet lab by simultaneously evolving a protoglobin-based cyclopropanation for improved yield and stereoselectivity. Finally, Chapter V describes the use of ALDE to engineer protoglobins with active sites which have been optimized to catalyze a broad scope of nitrene and carbene transfer reactions. This demonstration of enzyme ensemble engineering displays the power of diversity-oriented evolution to provide broad solutions in biocatalysis.

PUBLISHED CONTENT AND CONTRIBUTIONS

† denotes equal contribution

* denotes corresponding author

1. Miller, D.C.; **Lal, R.G.**; Marchetti, L.A.; Arnold, F.H.* Biocatalytic One-Carbon Ring Expansion of Aziridines to Azetidines via a Highly Enantioselective [1,2]-Stevens Rearrangement. *J. Am. Chem. Soc.* **2022**, *144*(11), 4739-4745.

   D.C.M. conceived the project. D.C.M. and L.A.M. discovered initial activity. D.C.M and R.G.L. conducted directed evolution and substrate scope studies. D.C.M. wrote initial draft. D.C.M., R.G.L., and F.H.A. edited and revised manuscript.

2. Yang, J.†; **Lal, R.G.**†; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.*; Arnold, F.H.* Active learning-assisted directed evolution. *Nat. Commun.* **2025**, *16*, 714.

   R.G.L. and J.Y. conceived the project. R.G.L. participated in all wet lab experimentation. J.Y., J.C.B., and R.A. performed software development and computational experimentation. M.A.H. participated in early wet lab investigations. S.K. and M.H. provided DNA synthesis resources. R.G.L. and J.Y. wrote the initial draft. R.G.L., J.Y., J.C.B., R.A., M.A.H, Y.Y., and F.H.A. edited and revised manuscript.

3. **Lal, R.G.**; Yang, J.; Zhang, Z.; Arnold, F.H.* Machine Learning-Assisted Evolution of Broadly Functional Enzyme Libraries. *ACS Central Science*. In Review.

   R.G.L. conceived the project. R.G.L. participated in all wet lab experimentation apart from chemical synthesis of authentic standards. J.Y. performed machine learning computaitons. R.G.L., J.Y., and F.L. performed data processing and analysis. Z.Z. synthesized authentic standards. R.G.L. wrote the initial draft. R.G.L., J.Y., F.L., Z.Z., and F.H.A. edited and revised the manuscript.

PUBLISHED CONTENT NOT INCLUDED IN THESIS

4. Das, A.; Gao, S.; **Lal, R.G**.; Hicks, M.H.; Oyala, P.H.; Arnold, F.H. Reaction Discovery Using Spectroscopic Insights from an Enzymatic C–H Amination Intermediate. *J. Am. Chem. Soc.* **2024**, *146*(30), 20556-20562.

   A.D. conceived the project. A.D., S.G., and R.G.L. performed biocatalytic reactions. A.D., M.H.H., and P.H.O. performed spectroscopic characterizations. A.D. wrote the initial draft. A.D., S.G., R.G.L., M.H.H., P.H.O., and F.H.A. edited and revised the manuscript.

TABLE OF CONTENTS

LIST OF FIGURES, TABLES, AND SCHEMES

Figure Number *Page*

Table Number *Page*

Scheme Number *Page*

ABBREVIATIONS

α-KG	α-ketoglutarate
ALA	5-aminolevulinic acid
ALDE	active learning-assisted directed evolution
Alloc	allyloxycarbonyl
*Ape*Pgb	*Aeropyrum pernix* protoglobin
ASR	ancestral sequence reconstruction
BO	Bayesian optimization
Boc	tert-butyloxycarbonyl
BSA	bovine serum albumin
Cbz	benzyloxycarbonyl
DE	directed evolution
DKL	deep kernel learning
DNN	deep neural network
E. coli	*Escherichia coli*
EDA	ethyl diazoacetate
epPCR	error-prone polymerase chain reaction
er	enantiomeric ratio
evSeq	every variant Sequencing
Fpro	4-fluoroproline
ftMLDE	focused training machine learning-assisted directed evolution
GB1	Immunoglobulin-binding domain B1 of streptococcal protein G

GC	gas chromatography
GC-FID	gas chromatography with flame ionization detection
GC-MS	liquid chromatography with mass spectrometric detection
GP	Gaussian process
HB	HyperBroth
HRMS	high-resolution mass spectrometry
IPTG	isopropyl-β-D-thiogalactoside
ITR	Initial Training Reaction
k_{cat}	rate constant
K_M	Michaelis constant
LB	Lysogeny Broth (Luria-Bertani medium)
LC	liquid chromatography
LC-MS	liquid chromatography with mass spectrometric detection
LevSeq	Long-read every variant sequencing
m/z	mass to charge ratio
M9-N	nitrogen-free M9 minimal media
MAE	mean absolute error
*Ma*Pgb	*Methanosarcina acetivorans* protoglobin
MISER	multiple injections in a single experimental run
ML	machine learning
MLDE	machine learning-assisted directed evolution
mr-ALDE	multi reaction active learning-assisted directed evolution

MSA	multiple sequence alignment
NGS	next-generation sequencing
NMR	nuclear magnetic resonance
OD600	optical density at 600 nm
P411	serine-ligated cytochrome P450 enzyme variant
P411-AzetS	P411 azetidine synthase
P450	cysteine-ligated cytochrome P450 monooxygenase
$P450_{BM3}$	*Bacillus megaterium* cytochrome P450
*Par*LQ	*Pyrobaculum arsenaticum* protoglobin W59L Y60Q
*Par*Pgb	*Pyrobaculum arsenaticum* protoglobin
PCR	polymerase chain reaction
PDB	Protein Data Bank
PLP	pyridoxal 5'-phosphate
PromPgb	Promiscuous Protoglobin
RmaNOD	*Rhodothermus marinus* nitric oxide dioxygenase
RNAP	RNA polymerase
RPM	revolutions per minute
RT	room temperature
σ	Spearman correlation
SiPro	silaproline
SSM	site-saturation mutagenesis
SUMS	Substrate Multiplexed Screening

T_{50}	incubation temperature at which a protein loses 50% of its activity
TB	Terrific Broth
TLC	thin-layer chromatography
TrpB	β-subunit of tryptophan synthase
TS	Thompson sampling
TTN	total turnover number
UCB	upper confidence bound

Chapter I

EXPLORING PROTEIN SEQUENCE SPACE FOR NEW-TO-NATURE BIOCATALYSIS

ABSTRACT

Directed evolution (DE) is a powerful strategy for engineering enzymes with improved or novel catalytic functions. DE has enabled the development of biocatalysts capable of mediating "new-to-nature" chemical transformations that have no counterparts in biological metabolism or synthesis. Despite these successes, engineering enzymes for non-native chemistry remains challenging. The discovery of starting points with even trace activity toward a desired reaction is difficult because protein function is hard to predict and promiscuous activities are rare. Furthermore, optimization of these activities is complicated by rugged protein fitness landscapes shaped by epistatic interactions between mutations, which makes exploration of combinatorial sequence space experimentally demanding. Machine learning-assisted directed evolution (MLDE) has emerged as a promising strategy to address these challenges by using sequence–function data to guide the selection of promising variants for experimental testing. This chapter reviews directed evolution in the context of new-to-nature biocatalysis, discusses the challenges associated with discovering and optimizing novel enzymatic functions, and introduces machine learning-guided strategies for navigating protein fitness landscapes to accelerate enzyme engineering.

1.1. Introduction

Enzymes are remarkable catalysts, capable of accelerating chemical transformations with extraordinary rates and selectivities under mild conditions.[1,2] While natural enzymes evolved to support the metabolic and survival needs of living organisms, decades of protein engineering have demonstrated that their catalytic capabilities can extend far beyond native biological functions.[3] In particular, directed evolution (DE), a protein engineering strategy, has enabled the development of enzymes which are capable of mediating "new-to-nature" biocatalytic transformations; chemical reactions for which no natural counterpart exists. These advances have established biocatalysis as a powerful strategy for accessing complex molecular architectures with high chemo-, regio-, and stereoselectivity.[4]

Directed evolution operates through iterative cycles of protein mutation and screening to identify variants with improved activity. Beneficial mutations are accumulated in this process until a functional goal has been achieved (**Figure 1-1**). There are several challenges associated with the development of new-to-nature enzymatic functions through DE. Engineering a novel biocatalytic activity generally requires an initial protein sequence that displays measurable, even if weak, activity toward the desired transformation, enabling subsequent rounds of mutagenesis and screening to identify improved catalysts. However, identifying such starting points typically requires screening enzyme libraries assembled based on biochemical intuition and prior engineering experience, without assurance that the desired activity will be discovered.[5] Furthermore, optimization of a single reaction does not necessarily confer broader substrate scope or mechanistic flexibility to a final enzyme variant, both of which are often desirable traits.

Protein fitness optimization can be viewed as traversing a fitness landscape, a mapping between amino acid sequences and their associated fitness values, to identify variants with improved performance. The challenges associated with the DE of new-to-nature functions stem, in part, from the immense size of protein sequence space (a protein of length N can take on 20^N unique sequences) and the fact that functional proteins in this space are vanishingly rare.[6] While protein space is vast, John Maynard Smith argued that, for natural evolution to proceed, a functional protein in this landscape must be neighbored by at least one functional sequence.[7] Thus, DE relies on the premise that protein fitness landscapes

can be locally traversed around functional sequences through one or a small number of mutations at a time. However, in practice these local landscapes are rarely smooth, as a result of epistasis, the phenomenon whereby the effects of mutations introduced in conjunction may be nonadditive relative to their individual contributions.[8] As a result, the engineering of enzymes for new-to-nature chemistry often becomes a complex search problem in a vast and sparsely functional sequence space.

This thesis examines strategies for navigating two of the primary challenges in the development of new-to-nature functions in enzymes: 1) the identification of activities for which there is no evolutionary precedent often requires screening of large libraries of enzymes and 2) after a starting sequence has been uncovered, optimization through DE can be laborious, as real protein fitness landscapes are often rugged and challenging to traverse. By combining experimental interrogation of sequence–function landscapes with machine learning-guided variant selection, we seek to not only improve the outcomes of engineering enzymes toward a single biocatalytic function, but also the breadth of chemical transformations accessible to a protein scaffold.

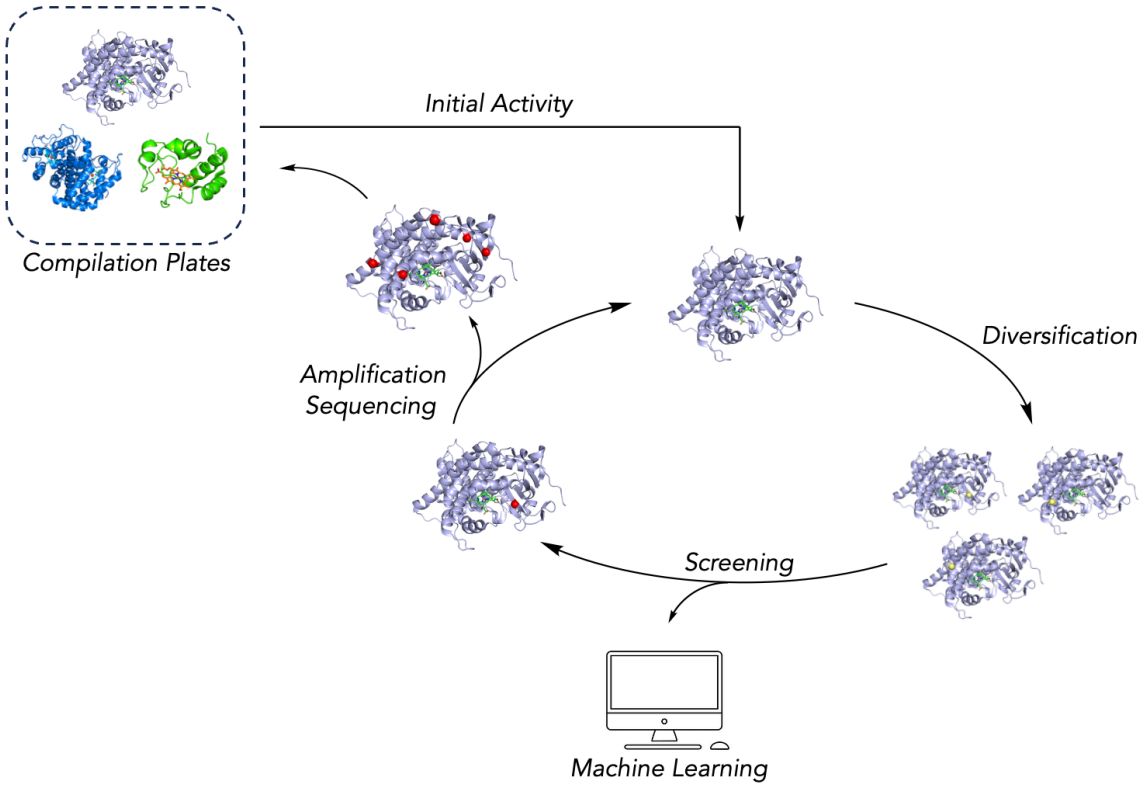


Figure 1-1. The cycle of direct evolution. After a desired function has been found within existing sequence space, a parent enzyme is diversified and assessed for improved activity. Mutations are accumulated through repeated cycles until a desired functional outcome is achieved. Recently, data collected in the process of screening variant libraries have been used to train machine learning models to aid in variant selection for subsequent rounds of engineering.

1.2. New-to-Nature Biocatalysis

Among the most striking demonstrations of the power of directed evolution has been the development of enzymes capable of catalyzing chemical transformations that have no known counterparts in biological metabolism. These "new-to-nature" reactions expand the scope of biocatalysis beyond the repertoire of natural enzymatic chemistry, enabling enzymes to access reaction mechanisms and bond constructions traditionally associated with synthetic organo-, photo-, or organometallic catalysis.[9-12] Thanks to the substrate-binding and transition-state stabilizing capabilities of enzyme active sites, protein engineers and chemists have achieved levels of stereo- and regioselectivity in biocatalytic systems that far exceed those of traditional small-molecule catalysts. Moreover, in recent years there have been several exciting demonstrations of enzymes which can access

reaction mechanisms and reaction modes which have not been reported in the synthetic organic literature.[13,14]

A prominent class of such new-to-nature enzymatic transformations involves the transfer of highly reactive carbene and nitrene intermediates to organic substrates. Enzymatic nitrene transfer was first reported by Breslow and coworkers, who found that microsomal cytochrome P450 isolated from rabbit liver could activate tosylimides for intramolecular amination via a putative nitrene intermediate, albeit with only 2 turnovers.[15] Inspired by this result, and emboldened by subsequent studies that abiological iron porphyrin complexes could catalyze carbene and nitrene transfer reactions,[16,17] researchers in the Arnold lab discovered that a cytochrome P450 from *Bacillus megaterium*[18] could catalyze carbene and nitrene transfer in the presence of the appropriate precursors.[19,20] An early breakthrough in engineering these cytochromes P450 was the mutation of the axial heme-coordinating cysteine to serine, which substantially improved carbene and nitrene transfer activity and enabled further DE toward high yields and stereoselectivity.[21] In the subsequent decade, these cytochromes P411, so named for their blue-shifted Soret peak at 411 nm, were rapidly diversified to enable a host of nitrene- and carbene-mediated transformations, primarily focused on X–H bond functionalizations and additions to π-systems (**Figure 1-2A**).[22-25] Recent advances in this field have focused on developing carbene and nitrene transferases that are better suited for industrial applications. These efforts include engineering more thermo- and chemostable catalysts, such as those based on protoglobin scaffolds,[26] as well as identifying carbene and nitrene precursors that are safer and more economical for use at scale (**Figure 1-2B**).[27-30]

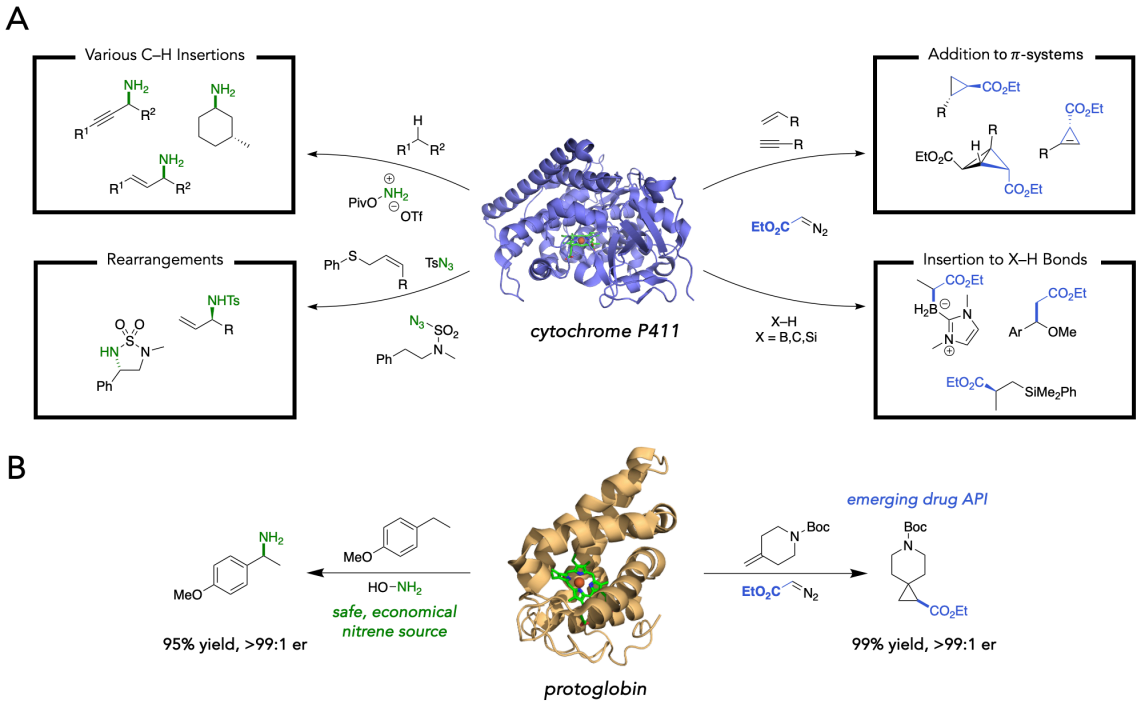


Figure 1-2. (A) Cytochromes P411 proved to be an excellent starting point for exploring the scope of carbene and nitrene activities available to biocatalysis. **(B)** By moving to the smaller, more thermostable protoglobin scaffold, more synthetically useful activities could be accessed.

From a protein engineering perspective, the laboratory evolution of new-to-nature biocatalytic functions presents several challenges. Firstly, as there is inherently no information about fitness towards new-to-nature chemistries encoded in the diversity of natural protein sequences, bioinformatic methods for inferring protein function using multiple sequence alignment-based (MSA) methods tend to fail for identifying starting points for DE campaigns and for identifying potentially beneficial mutations.[31] Additionally, the products of new-to-nature reactions typically require chromatographic analysis for quantification, which limits screening throughput relative to assays that can be coupled to optical readouts. As a result, not only is it challenging to scale the size of screened mutant libraries, but it is also difficult to quantify the effects of mutations on catalytic parameters such as K_M and k_{cat}. Improvements in apparent activity may arise from changes in multiple underlying factors such as substrate binding, turnover rate, or protein expression and stability (**Figure 1-3**). Without this information, it can be difficult to determine an effective protein engineering strategy. Finally, there are very few new-to-nature function data which are well-annotated and publicly available.[32] This is due to the

low-throughput nature of these engineering campaigns, the tendency for most campaigns to focus on activity for a single substrate, and the lack of standardized formats for reporting such data within the protein engineering literature. Given these challenges, DE campaigns aimed at new-to-nature enzymatic activities are often slow and resource intensive, leaving substantial opportunity for methodological improvement.

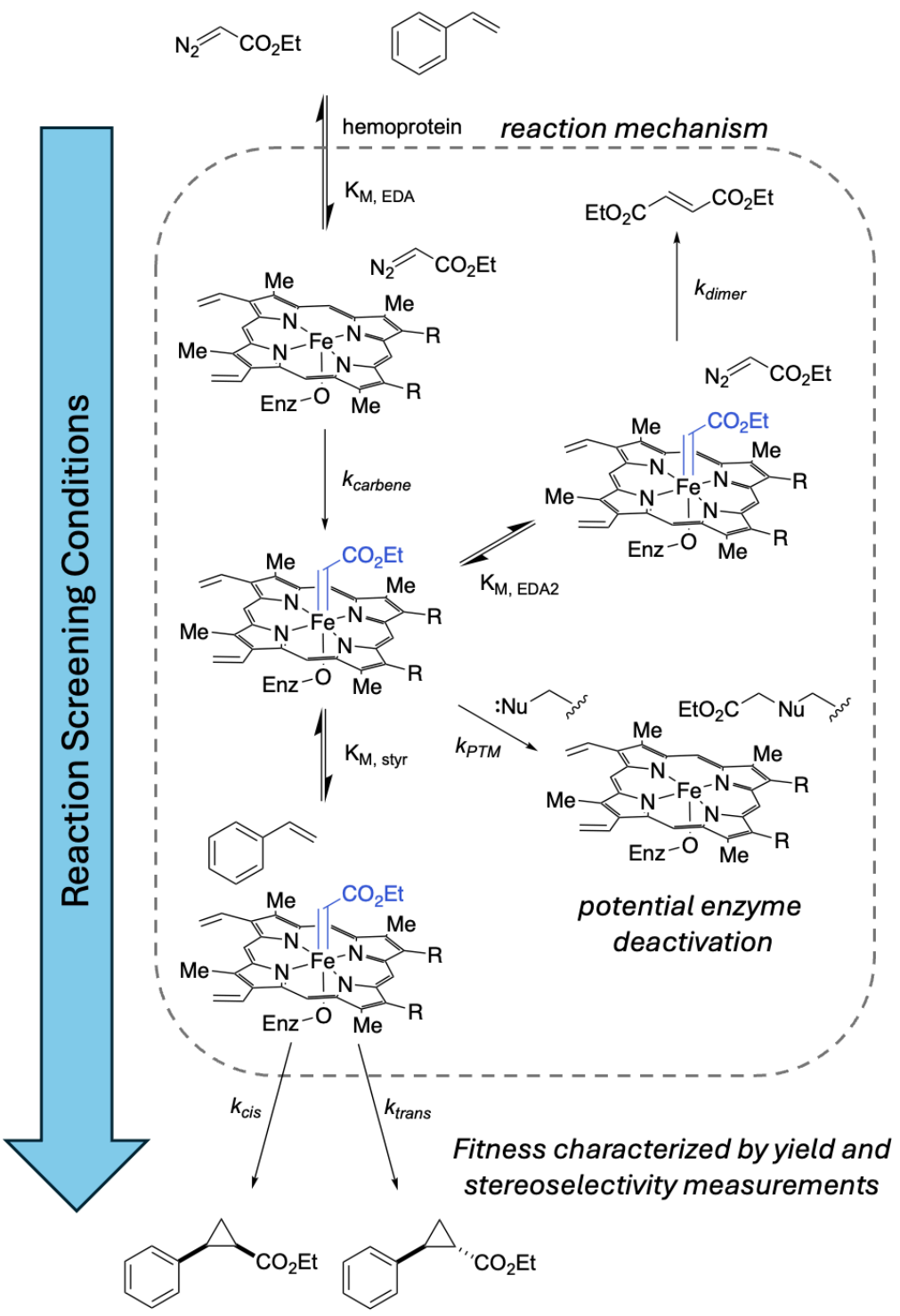


Figure 1-3. In directed evolution, screening for variants which have an improved, defined fitness value allows for a mechanism-ignorant improvement of a desired function. Shown here, a plausible mechanistic pathway for enzymatic cyclopropanation has eight kinetic parameters. Presumably some set of these values is optimized through DE.

1.3. Enzyme Promiscuity

New-to-nature activities are discovered by exposing existing protein sequences–selected based on chemical intuition and prior biochemical characterizations–to abiological substrates and screening for desired reaction products. This is possible because enzymes are promiscuous; they can serendipitously catalyze reactions for which they have not previously evolved.[33] Advanced by Jensen in 1976, promiscuity is a key mechanism by which modern enzymes were specialized from primordial proteins which were capable of enacting several suboptimal functions.[34] **Figure 1-4** presents a graphical view of how enzyme promiscuity can be exploited for the discovery of novel functions.

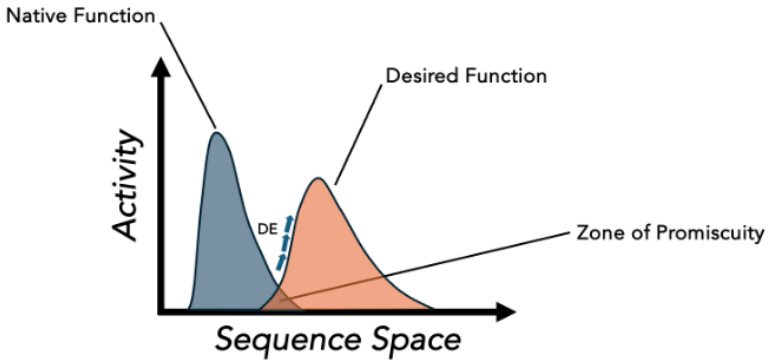


Figure 1-4. Novel enzymatic function is discovered by leveraging promiscuous activity.

Importantly, a new-to-nature DE campaign cannot begin without the determination of an enzyme sequence that exhibits at least trace levels of the desired activity. The discovery of such starting points remains a central bottleneck in the development of novel enzymatic functions as it is challenging to predict the promiscuous functions of proteins from their sequence alone.[35,36] Given the dearth of public non-natural protein function data which could be used to infer promiscuous functions, researchers must often compile and screen diverse enzyme libraries based on biochemical intuition and prior protein engineering experience.[5] When screening these enzyme libraries protein engineers have to acquire the DNA sequences encoding member proteins and contend with the expression criteria associated with each protein, all with no guarantee that the desired activity will be uncovered upon screening.

Owing to the challenges associated with the discovery of novel enzymatic functions, there have been efforts to design enzyme libraries which may be well suited for initial reaction screening efforts, both through computational and wet-lab methods. Ancestral sequence reconstruction (ASR), a bioinformatic method which infers and "resurrects" ancient protein sequences through phylogenetic analysis, is believed to generate proteins which are more promiscuous because they likely served as generalist catalysts before specialization through evolution to their modern descendants.[37,38] ASR, however, would not be expected to generate enzymes which are competent towards new-to-nature mechanisms, as there is presumably no functional signal for these transformations in evolutionary history. One other promising computational method for generating enzyme libraries with heightened promiscuity is FuncLib, a software tool developed by Kheronsky *et al.*[39] FuncLib operates by diversifying the active site of an enzyme scaffold at user-defined residues and selecting combinations of mutations which 1) are not believed to destabilize the protein based on physics-based calculations and 2) are believed to be probable based on MSA assessment. Application of FuncLib to different enzymatic systems has shown promising results for generating promiscuous enzyme libraries;[40,41] however, it remains unclear whether this approach will be effective for improving promiscuity toward new-to-nature biocatalytic reactions, as the method relies on information derived from multiple sequence alignments.

In addition to computational strategies, wet-lab approaches have been developed to generate enzyme libraries with enhanced catalytic and substrate promiscuity. For example, Buller and coworkers introduced Substrate Multiplexed Screening (SUMS), in which enzyme variants are simultaneously subjected to several substrate analogs and subsequent reaction extracts are analyzed for multiple reaction products.[42] This protocol allows for the direct engineering of promiscuity, as mutations which allow for activity upon multiple substrates can be carried forward. To date, SUMS has only been demonstrated to work with PLP-dependent enzymes;[43] nevertheless, it remains a promising tool for developing promiscuous catalysts.

Another recently developed workflow for predicting promiscuous functions of enzymes is CATNIP, a machine learning (ML) model for predicting the substrate compatibility of α-ketoglutarate (α-KG)/Fe(II)-dependent enzymes.[44] The CATNIP model was trained using

a large dataset of function data from (α-KG)/Fe(II)-dependent enzyme and substrate pairs. While this method is inherently resource intensive, the authors demonstrate that CATNIP is capable of facilitating the discovery of enzymes which are functional towards a target substrate. It is possible that such data generation and dissemination will be necessary for members of the protein engineering community to directly predict which existing enzymes will be able to facilitate new-to-nature functions.

1.4. Epistasis

After a desired function has been found in a starting "parent" enzyme, directed evolution can commence through traversal of the sequence-function landscape via mutagenesis and screening. In an ideal world this landscape would bear a smooth, monotonic shape, bearing a global fitness maximum (**Figure 1-5A**). In this case, the optimal variant can theoretically be accessed by a number of search trajectories.[45] However, in reality, sequence-function landscapes are generally rugged, demonstrating local maxima and minima in fitness (**Figure 1-5B**).[46,47] Local optima act as traps from where improved function can only be achieved after an initial reduction in fitness.[48] This trait of real protein fitness landscapes makes starting point selection and overall engineering strategy crucial, as different paths can lead to vastly different engineering outcomes.

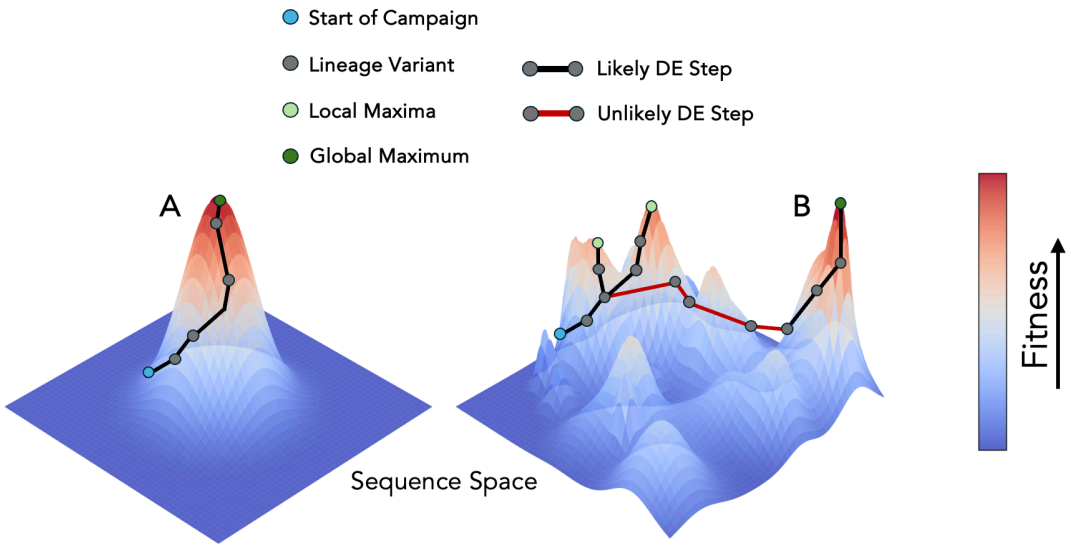


Figure 1-5. (A) When traversing an ideal, smooth landscape nearly any DE trajectory will arrive at the global maximum. **(B)** Searching a real, rugged fitness landscape is much more complex. Given a starting point, it is possible to arrive at different local maxima based on evolution strategy, and arrival at the global maximum may require steps which incur a loss in fitness.

Protein fitness landscape ruggedness comes from epistasis, the phenomenon whereby the effects of a mutation are dependent on the context in which it is introduced.[49,50] Epistasis presents a challenge in directed evolution because it means that many of the physical phenomena underlying an enzyme's function are the result of the identities of several interacting residues. Thus, decisions made in earlier rounds of DE can heavily influence the outcomes of later stages. One might envision that screening libraries in which multiple targeted sites are simultaneously mutated could more effectively search epistatic fitness landscapes. However, such multi-site saturation libraries rapidly run into a combinatorial design problem (**Figure 1-6**). A bacterial protein of typical length (270 amino acids) has 5,132 single mutants.[51] For the same protein, there are approximately 13 million possible double mutants. Thus, screening the double-mutant space of an enzyme would require nearly 1,000 times the screening resources needed to survey the local single-mutant space. Furthermore, because functional protein sequence space is sparse, the likelihood of discovering improved variants decreases rapidly as additional mutations move a sequence further from a known functional starting point.[6] These limitations highlight the need for

approaches that can more efficiently navigate epistatic fitness landscapes and prioritize promising multi-mutation variants without exhaustive experimental screening.

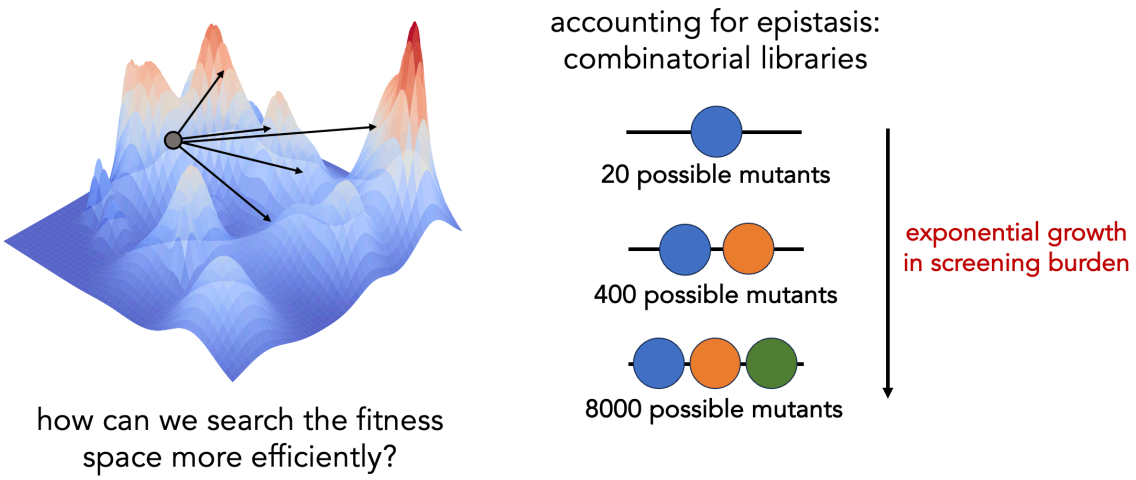


Figure 1-6. Screening combinatorial libraries would allow for further jumps in the sequence-function landscape; however, the exponential growth in screening burden makes randomly searching these design spaces inaccessible through traditional methods.

1.5. Machine Learning as a Navigation Tool for Directed Evolution

An attractive approach to accelerating directed evolution efforts is to incorporate machine learning (ML) methods into existing workflows. Machine learning-assisted directed evolution (MLDE) has the potential to ease the experimental screening burden of an evolutionary campaign by evaluating sequences *in silico* based on ML models that have been trained on existing data. MLDE refers to a specific implementation of supervised ML[52] for the fitness prediction of protein variants containing several mutations, aiming to capture details of epistasis (**Figure 1-7**). In early explorations of MLDE, supervised models were used to learn relationships between protein sequences and their associated function to approximate the fitness landscape.[53-55] Building on this body of work, Wittmann and coworkers noted that combinatorial libraries contain a high density of zero- or low-activity variants and incorporated zero-shot predictors into their model to predict high- activity mutants based on data collected from a library designed to be informative for model training.[56] In simulations on the GB1 landscape (a combinatorially complete empirical functional dataset for quadruple mutants of protein G domain B1),[57] this "focused training

MLDE" (ftMLDE) framework was able to arrive at the global optimum 99.70% of the time, in comparison to traditional MLDE which achieved this optimum 8.85% of the time. The rapidly evolving field of MLDE and other applications of machine learning to protein engineering efforts is well reviewed by Yang and Li.[58]

Crucially, machine learning has been shown to integrate into existing protein engineering workflows effectively.[59-61] MLDE was first experimentally validated on the carbon-silicon bond-forming landscape of nitric oxide dioxygenase from *Rhodothermus marinus* (RmaNOD), enabling access to stereodivergent variants with minimal screening.[62] A unifying theme of these experimental validations is that ML methods are most effective when trained on large, rich sequence-function datasets. Given that machine learning has the capability to extract value from enzyme variants which do not display improved activity, it is essential that the sequences of all tested variants are collected. Furthermore, if more sequences can be interrogated in training then ML models can more effectively learn epistasis in the fitness landscape.

Since the 1990s, advances in DNA technologies—driven in part by large-scale sequencing efforts such as the Human Genome Project—have dramatically expanded the capabilities of protein engineering.[63] These advances have also made MLDE a viable tool in the modern biotechnological age. Next-generation sequencing (NGS) methods now allow the function data of variants in directed evolution libraries to be labeled with their sequences, for example through barcoding strategies that enable sequencing of spatially segregated mutants.[64-66] These approaches generate large sequence–function datasets that are valuable not only for biochemical analysis of engineered enzymes but also for training machine learning models.[67,68] In parallel, advances in DNA synthesis have substantially reduced the cost and increased the scale of producing designed DNA sequences. Innovations such as Twist Bioscience's silicon-based phosphoramidite synthesis and enzymatic oligonucleotide synthesis methods based on terminal deoxynucleotidyl transferase have enabled efficient and scalable DNA production.[69-74] These developments allow researchers to readily synthesize model-predicted variants and assemble large gene libraries using oligo pool–based methods,[75,76] thereby making computationally guided directed evolution approaches such as MLDE experimentally accessible. Together, these advances in

sequencing and DNA synthesis have created the experimental infrastructure necessary to generate, analyze, and test the large variant libraries required for machine learning–guided protein engineering. These developments create an opportunity to apply machine learning not only to accelerate directed evolution, but also to design enzyme libraries capable of accessing diverse new-to-nature chemistries.

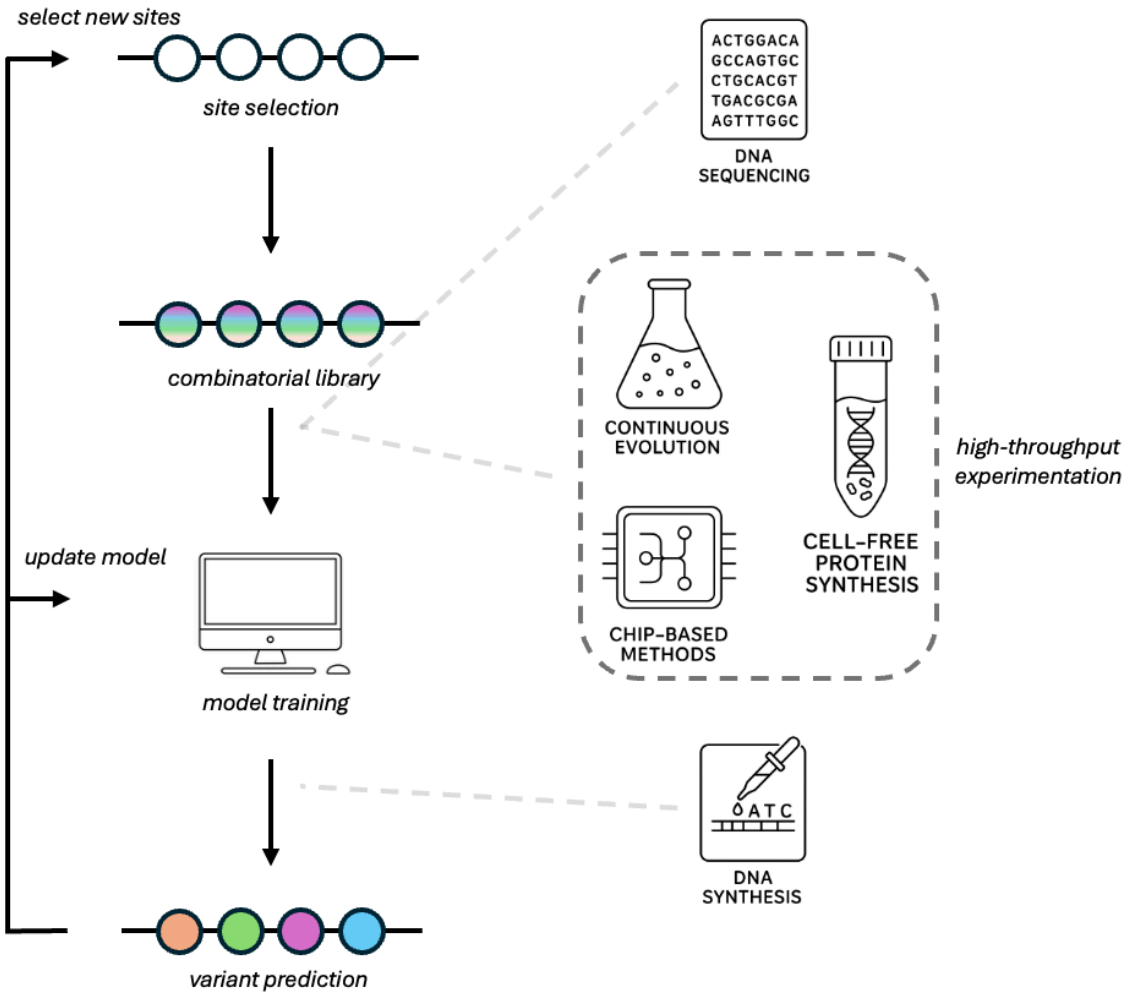


Figure 1-7. In a general MLDE workflow, first sites relevant to activity are selected. Combinatorial libraries of mutants at these positions are assessed for model training to predict variants with improved function. Modern biotechnologies present an opportunity to accelerate MLDE efforts.

1.6. Toward Broadly Functional Enzyme Libraries

This thesis investigates strategies for navigating protein sequence–function landscapes using machine learning methods to enable the discovery and engineering of enzymes capable of catalyzing diverse non-native reactions. In **Chapter 2**, directed evolution is applied to develop cytochrome P411 variants capable of mediating aziridine ring expansion to form azetidines, demonstrating the potential of engineered hemoproteins to access valuable synthetic transformations. However, this work also highlights limitations of conventional directed evolution, including slow optimization and restricted substrate scope. **Chapter 3** explores the challenges associated with discovering new enzymatic activities, revealing that functional variants for certain transformations may be rare or difficult to identify even when screening diverse protein libraries.

To address these challenges, **Chapter 4** introduces active learning-assisted directed evolution (ALDE), a machine learning framework that iteratively integrates experimental sequence–function data with predictive modeling to guide exploration of protein sequence space. This approach enables efficient identification of beneficial multi-mutation variants while accounting for epistatic interactions between residues. Building on this strategy, **Chapter 5** presents multi-reaction active learning-assisted directed evolution (mr-ALDE), in which machine learning models trained on activity data from multiple reactions are used to design enzyme libraries optimized for broad catalytic functionality. Variants predicted using this approach exhibit improved activity across diverse carbene and nitrene transfer reactions and enable access to transformations not catalyzed by the parent enzyme.

Together, these studies demonstrate how integrating experimental data with machine learning can accelerate enzyme engineering and facilitate the development of broadly functional biocatalysts for new-to-nature chemistry.

Chapter I Bibliography

1. Sheldon, R.A. E factors, green chemistry and catalysis: an odyssey. *Chem. Commun.* **2008**, *29*, 3352-3365.
2. Sheldon, R.A.; Woodley, J.M. Role of Biocatalysis in Sustainable Chemistry. *Chem. Rev.* **2018**, *118*, 801-838.
3. Miller, D.C.; Athavale, S.V.; Arnold, F.H. Combining chemistry and protein engineering for new-to-nature biocatalysis. *Nature Synth.* **2022**, *1*, 18-23.
4. Winkler, C.K.; Schrittwieser, J.H.; Kroutil, W. Power of Biocatalysis for Organic Synthesis. *ACS Cent. Sci.* **2021**, *7*, 55-71.
5. Arnold, F.H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed.* **2017**, *57*, 4143-4148.
6. Arnold, F.H. The Library of Maynard-Smith: My Search for Meaning in the Protein Universe. *Microbe Mag.* **2011**, *6*, 316-318.
7. Smith, J.M. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225*, 563-564.
8. Miton, C.M.; Buda, K.; Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 160-168.
9. Emmanuel, M.A.; Bender, S.G.; Bilideau, C.; Carceller, J.M.; DeHovitz, J.S.; Fu, H.; Liu, Y.; Nicholls, B.T.; Ouyang, Y.; Page, C.G.; Qiao, T.; Raps, F.C.; Sorigué, D.R.; Sun, S.-Z.; Turek-Herman, J.; Ye, Y.; Rivas-Souchet, A.; Cao, J.; Hyster, T.K. Photobiocatalytic Strategies for Organic Synthesis. *Chem. Rev.* **2023**, *9*, 5459-5520.
10. Zetzsche, L.E.; Yazarians, J.A.; Chakrabarty, S.; Hinze, M.E.; Murray, L.A.M.; Lukowski, A.L.; Joyce, L.A.; Narayan, A.R.H. Biocatalytic oxidative cross-coupling reactions for biaryl bond formation. *Nature* **2022**, *603*, 79-85.
11. Zhao, Q.; Chen, Z.; Soler, J.; Chen, X.; Rui, J.; Ji, N.T.; Yu, Q.E.; Yang, Y.; Garcia-Borràs, M.; Huang, X. Engineering non-haem iron enzymes for enantioselective C(sp^3)–F bond formation via radical fluorine transfer. *Nat. Synth.* **2024**, *3*, 958-966.
12. Zhou, Q.; Chin, M.; Fu, Y.; Liu, P.; Yang, Y. Stereodivergent atom-transfer radical cyclization by engineered cytochromes P450. *Science* **2021**, *374*, 1612-1616.
13. Raps, F.C.; Rivas-Souchet, A.; Jones, C.M.; Hyster, T.K. Emergence of a distinct mechanism of C–N bond formation in photoenzymes. *Nature* **2025**, *637*, 362-368.
14. Raps, F.C.; Hyster, T.K. Emergent Mechanisms in Biocatalysis. *ACS Cent. Sci.* **2025**, *11*, 1029-1040.
15. Svastits, E.W.; Dawson, J.H.; Breslow, R.; Gellman, S.H. Functionalized Nitrogen Atom Transfer Catalyzed by Cytochrome P-450. *J. Am. Chem. Soc.* **1985**, *107*, 6427-6428.
16. Guo, M.; Corona, T.; Ray, K.; Nam, W. Heme and Nonheme High-Valent Iron and Manganese Oxo Cores in Biological and Abiological Oxidation Reactions. *ACS Cent. Sci.* **2019**, *5*, 13-28.
17. Wolf, J.R.; Hamaker, C.G.; Djukic, J.-P.; Kodadek, T.; Woo, L.K. Shape and stereoselective cyclopropanation of alkenes catalyzed by iron porphyrins. *J. Am. Chem. Soc.* **1995**, *117*, 9194-9199.
18. Whitehouse, C.J.C.; Bell, S.G.; Wong, L.-L. $P450_{BM3}$(CYP102A1): connecting the dots. *Chem. Soc. Rev.* **2012**, *41*, 1218-1260.

19. Coelho, P.S.; Brustad, E.M.; Kannan, A.; Arnold, F.H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* **2013**, *339*, 307-310.
20. McIntosh, J.A.; Coelho, P.S.; Farwell, C.C.; Wang, Z.J.; Lewis, J.C. Brown, T.R.; Arnold, F.H. Enantioselective Intramolecular C-H Amination Catalyzed by Engineered Cytochrome P450 Enzymes in vitro and in vivo. *Angew. Chem. Int. Ed.* **2013**, *52*, 9309-9312.
21. Coelho, P.S.; Wang, Z.J.; Ener, M.E.; Baril, S.A.; Kannan, A.; Arnold, F.H.; Brustad, E.M. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **2013**, *9*, 485-487.
22. Prier, C.K.; Zhang, R.K.; Buller, A.R.; Brinkmann-Chen, S.; Arnold, F.H. Enantioselective, intermolecular benzylic C–H amination catalysed by an engineered iron-haem enzyme. *Nat. Chem.* **2017**, *9* 629-634.
23. Chen, K.; Huang, X.; Kan, S.B.J.; Zhang, R.K.; Arnold, F.H. Enzymatic construction of highly strained carbocycles. *Science* **2018**, *360*, 71-75.
24. Huang, X.; Garcia-Borràs, M.; Miao, K.; Kan, S.B.J.; Zutshi, A.; Houk, K.N.; Arnold, F.H. A Biocatalytic Platform for Synthesis of Chiral α-Trifluoromethylated Organoborons. *ACS Cent. Sci.* **2019**, *5* 270-276.
25. Zhang, R.K.; Chen, K.; Huang, X.; Wohlschlager, L.; Renata, H.; Arnold, F.H. Enzymatic assembly of carbon–carbon bonds via iron-catalysed sp^3 C–H functionalization. *Nature* **2019**, *565*, 67-72.
26. Pesce, A.; Bolognesi, M.; Nardini, M. Protoglobin: structure and ligand-binding properties. *Microbial Globins – Status and Opportunities*; Poole, R.K., Ed.; Elsevier, 2013; pp 79-96. DOI: 10.1016/B978-0-12-407693-8.00003-0
27. Knight, A.M.; Kan, S.B.J.; Lewis, R.D.; Brandenberg, O.F.; Chen, K.; Arnold, F.H. Diverse engineered heme proteins enable stereodivergent cyclopropanation of unactivated alkenes. *ACS Cent. Sci.* **2018**, *4*, 372-377.
28. Porter, N.J.; Danelius, E.; Gonen, T.; Arnold, F.H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **2022**, *144*, 8892-8896.
29. Gao, S.; Das, A.; Alfonzo, E.; Sicinski, K.M.; Rieger, D.; Arnold, F.H. Enzymatic Nitrogen Incorporation Using Hydroxylamine. *J. Am. Chem. Soc.* **2023**, *145*, 20196-20201.
30. Kennemur, J.L.; Long, Y.; Ko, C.J.; Das, A.; Arnold, F.H. Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]alkanes. *J. Am. Chem. Soc.* **2025**, *147*, 27165-27171.
31. Yang, J.; Lal, R.G.; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.; Arnold, F.H. Active learning-assisted directed evolution. *Nat. Commun.* **2025**, *16*, 714.
32. Long, Y.; Abbasinejad, F.; Li, F.-Z.; Reinprecht, P.; Wittmann, B.; Kennemur, J.L.; Carder, H.; Yang, J.; Lambert, T.; O'Meara, R.; Radtke, L.; Qin, Z.; Brinkmann-Chen, S.; Arnold, F.,; Mora, A. Enzyme Engineering Database (EnzEngDB): a platform for sharing and interpreting sequence–function relationships across protein engineering campaigns. *Nucleic Acids Research* **2026**, *D1*, D564-D571.
33. Kheronsky, O.; Roodveldt, C.; Tawfik, D.S. Enzyme promiscuity: evolution and mechanistic aspects. *Curr Op. Chem. Biol.* **2006**, *10*, 498.

34. Jensen, R.A. ENZYME RECRUITMENT IN EVOLUTION OF NEW FUNCTION. *Annu. Rev. Microbiol.* **1976**, *30*, 409.
35. Jeffery, C.J. Current successes and remaining challenges in protein function prediction. *Front. Bioinform.* **2023**, *3*, 1222182.
36. Marshall, J.R.; Mangas-Sanchez, J.; Turner, N.J. Expanding the synthetic scope of biocatalysis by enzyme discovery and protein engineering. *Tetrahedron* **2021**, *82*, 131926.
37. Prakinee, K.; Phaisan, S.; Kongjaroon, S.; Chaiyen, P. Ancestral Sequence Reconstruction for Designing Biocatalysts and Investigating their Functional Mechanisms. *JACS Au* **2024**, *4*, 4571-4591.
38. Chiang, C.-H.; Wang, Y.; Hussain, A.; Brooks III, C.L.; Narayan, A.R.H. Ancestral Sequence Reconstruction to Enable Biocatalytic Synthesis of Azaphilones. *J. Am. Chem. Soc.* **2024**, *146*, 30194-30203.
39. Kheronsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D.L.; Prilusky, J.; Amengual-Rigo, P.; Guallar, V.; Tawfik, D.S.; Fleishman, S.J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Molecular Cell* **2018**, *72*, 178-186.
40. Gomez de Santos, P.; Mateljak, I.; Hoang, M.D.; Fleishman, S.J.; Hollmann, F.; Alcalde, M. Repertoire of Computationally Designed Peroxygenases for Enantiodivergent C–H Oxyfunctionalization Reactions. *J. Am. Chem. Soc.* **2023**, *145*, 3443-3453.
41. Listov, D.; Vos, E.; Hoffka, G.; Hoch, S.Y.; Berg, A.; Hamer-Rogotner, S.; Dym, O.; Kamerlin, S.C.L.; Fleishman, S.J. Complete computational design of high-efficiency Kemp elimination enzymes. *Nature* **2025**, *643*, 1421-1427.
42. McDonald, A.D.; Higgins, P.M.; Buller, A.R. Substrate multiplexed protein engineering facilitates promiscuous biocatalytic synthesis. *Nat. Commun.* **2022**, *13*, 5242.
43. Campbell, M.E.; Ohler, A.R.; McGill, M.J.; Buller, A.R. Promiscuity Guided Evolution of Decarboxylative Aldolases for Synthesis of Tertiary γ-Hydroxy Amino Acids. *Angew. Chem. Int. Ed.* **2025**, *64*, e202422109.
44. Paton, A.E.; Boiko, D.A.; Perkins, J.C.; Cemalovic, N.I.; Reschützegger, T.; Gomes, G.; Narayan, A.R.H. Connecting chemical and protein sequence space to predict biocatalytic reactions. *Nature* **2025**, *646*, 108-116.
45. Lobkovsky, A.E.; Wolf, Y.I.; Koonin, E.V. Quantifying the similarity of monotonic trajectories in rough and smooth fitness landscapes. *Mol. BioSyst.* **2013**, *9***,** 1627.
46. Macken, C.A.; Perelson, A.S. Protein evolution on rugged landscapes. *Proc. Nat. Acad. Sci.* **1989**, *86*, 6191.
47. Sandhu, M.; Chen, J.Z.; Matthews, D.S.; Spence, M.A.; Pulsford, S.B.; Gall, B.; Kaczmarski, J.A.; Nichols, J.; Tokuriki, N.; Jackson, C.J. Computational and Experimental Exploration of Protein Fitness Landscapes: Navigating Smooth and Rugged Terrains. *Biochemistry* **2025**, *64*, 1673.
48. Starr, T.N.; Thornton, J.W. Epistasis in protein evolution. *Protein Sci.* **2016**, *25*, 1204.
49. Yang, G.; Anderson, D.W.; Baier, F.; Dohmen, E.; Hong, N.; Carr, P.D.; Caroline, S.; Kamerlin, L.; Jackson, C.J.; Bornberg-Bauer, E.; Tokuriki, N. Higher-order

epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat. Chem. Biol.* **2019**, *15*, 1120.
50. Otwinowski, J.; McCandlish, D.M.; Plotkin, J.B. Inferring the shape of global epistasis. *Proc. Nat. Acad. Sci.* **2018**, *115*, E7550.
51. Nevers, Y.; Glover, N.M.; Dessimoz, C.; Lecompte, O. Protein length distribution is remarkably uniform across the tree of life. *Genome Biol.* **2023**, *24*, 135.
52. Jiang, T.; Gradus, J.L.; Rosellini, A.J. Supervised Machine Learning: A Brief Primer. *Behav. Ther.* **2020**, *51*, 675.
53. Romero, P.A.; Krause, A.; Arnold, F.H. Navigating the Protein Fitness Landscape with Gaussian Processes. *Proc. Natl. Acad. Sci.* **2012**, *110*, E193.
54. Bryant, D.H.; Bashir, A.; Sinai, S.; Jain, N.K.; Ogden, P.J.; Riley, P.F.; Church, G.M.; Colwell, L.J.; Kelsic, E.D. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **2021**, *39*, 691.
55. Saito, Y.; Oikawa, M.; Nakzawa, H.; Niide, T.; Kameda, T.; Tsuda, K.; Umetsu, M. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth. Biol.* **2018**, *7*, 2014.
56. Wittmann, B.J.; Yue, Y.; Arnold, F.H. Informed training set design enables efficient machine-learning assisted directed protein evolution. *Cell Systems* **2021**, *12*, 1026.
57. 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* **2016**, *5*, e16965.
58. Yang, J.; Li, F.-Z.; Arnold, F.H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **2024**, *10*, 226.
59. Bedbrook, C.N.; Yang, K.K.; Rice, A.J.; Gradinaru, V.; Arnold, F.H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* **2017**, *13*, e1005786.
60. Thomas, N.; Belanger, D.; Xu, C.; Lee, H.; Hirano, K.; Iwai, K.; Polic, V.; Nyberg, K.D.; Hoff, K.G.; Frenz, L.; Emrich, C.A.; Kim, J.W.; Chavarha, M.; Ramanan, A.; Agresti, J.J.; Colwell, L.J. Engineering highly active nuclease enzymes with machine learning and high-throughput screening. Cell *Systems* **2025**, *16*, 101236.
61. Xie, W.J.; Liu, D.; Wang, X.; Zhang, A.; Wei, Q.; Nandi, A.; Dong, S.; Warshel, A. *Proc. Natl. Acad. Sci.* **2023**, *120*, e2312848120.
62. 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.* **2019**, *116*, 8852.
63. Gibbs, R.A. The Human Genome Project changed everything. *Nat. Rev. Genet.* **2020**, *21*, 575.
64. 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. *Molec. Ecol. Res.* **2015**, *15*, 855.
65. 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.* **2022**, *11*, 1313.
66. Long, Y.; Mora, A.; Li, F.-Z.; Gürsoy, E.; Johnston, K.E.; Arnold, F.H. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *ACS Synth. Biol.* **2025**, *14*, 230.

67. Alfonzo, E.; Hanley, D.; Li, Z.-Q.; Sicinski, K.M.; Gao, S.; Arnold, F.H. Biocatalytic Synthesis of α-Amino Esters via Nitrene C–H Insertion. *J. Am. Chem. Soc.* **2024**, *146*, 27267.
68. D'Costa, S.; Hinds, E.C.; Freschlin, C.R.; Song, H.; Romero, P.A. Inferring protein fitness landscapes from laboratory evolution experiments. *PLoS Comput. Biol.* **2023**, *19*, e1010956.
69. Arnaud, C.H. Twist Bioscience. *C&EN* **2015**, *93*.
70. Roy, S.; Caruthers, M. Synthesis of DNA/RNA and Their Analogs via Phosphoramidite and H-Phosphonate Chemistries. *Molecules* **2013**, *18*, 14268.
71. Sarac, I.; Hollenstein, M. Terminal Deoxynucleotidyl Transferase in the Synthesis and Modification of Nucleic Acids. *ChemBioChem* **2018**, *20*, 860.
72. Lu, X.; Li, J.; Li, C.; Lou, Q.; Peng, K.; Cai, B.; Liu, Y.; Yao, Y.; Lu, L.; Tian, Z.; Ma, H.; Wang, W.; Cheng, J.; Guo, X.; Jiang, H.; Ma, Y. Enzymatic DNA Synthesis by Engineering Terminal Deoxynucleotidyl Transferase. *ACS Catal.* **2022**, *12*, 2988.
73. Barthel, S.; Palluk, S.; Hillson, N.J.; Keasling, J.D.; Arlow, D.H. Enhancing Terminal Deoxynucleotidyl Transferase Activity on Substrates with 3' Terminal Structures for Enzymatic De Novo DNA Synthesis. *Genes* **2020**, *11*, 102.
74. Bomgardner, M.M. Ansa raises funds to make DNA with enzymes. *C&EN* **2020**, *98*.
75. Freschlin, C.R.; Yang, K.K.; Romero, P.A. Scalable and cost-efficient custom gene library assembly from oligopools. *bioRxiv* **2025**, DOI: 10.1101/2025.03.22.644747
76. Robinson, N.E; Zhang, W.; Ghosh, R.; Gerber, B.; Zhang, H.; Sanfiorenzo, C.; Wang, S.; Di Carlo, D.; Wang, K. Construction of complex and diverse DNA sequences using DNA three-way junctions. *Nature* **2026**, DOI: 10.1038/s41586-025-10006-0

Chapter II

BIOCATALYTIC ONE-CARBON RING EXPANSION OF AZIRIDINES TO AZETIDINES VIA A HIGHLY ENANTIOSELECTIVE [1,2]-STEVENS REARRANGEMENT

Material from this chapter appears in: "Miller, D.C.; **Lal, R.G.**; Marchetti, L.A.; Arnold, F. H. * Biocatalyic One-Carbon Ring Expansion of Aziridines to Azetidines via a Highly Enantioselective [1,2]-Stevens Rearrangement. *J. Am. Chem. Soc.* **2022**, *144*(11), 4739–4745."

R.G.L participated in the execution of the research, including enzymatic reactions, substrate scope studies, and synthetic applications. R.G.L participated in writing and reviewing the manuscript.

ABSTRACT

We report enantioselective one-carbon ring expansion of aziridines to make azetidines as a new-to-nature activity of engineered “carbene transferase” enzymes. A laboratory-evolved variant of cytochrome $P450_{BM3}$, P411-AzetS, not only exerts unparalleled stereocontrol (99:1 er) over a [1,2]-Stevens rearrangement but also overrides the inherent reactivity of aziridinium ylides, cheletropic extrusion of olefins, to perform a [1,2]-Stevens rearrangement. By controlling the fate of the highly reactive aziridinium ylide intermediates, these evolvable biocatalysts promote a transformation which cannot currently be performed using other catalyst classes.

2.1. Introduction

Ring-size manipulation has emerged as a powerful strategy to convert readily available cyclic structures into ring-expanded or ring-contracted compounds that are more difficult to synthesize using conventional means.[1] In particular, "cut and sew" strategies relying on transition-metal catalyzed oxidative addition across C–C bonds are useful approaches for insertion of carbon monoxide or two-carbon fragments such as olefins and alkynes to effect one- or two-carbon ring expansions, respectively.[2] For nitrogen-containing heterocycles, one possible strategy for ring expansion is to induce a [1,2]-Stevens rearrangement by formation of an ammonium ylide, resulting in one-carbon ring expansion.[3] Pioneering works by Hata, West, and Couty demonstrated this approach for 4- to 5-membered ring expansions, wherein treatment of an azetidine with a diazo compound in the presence of a copper catalyst provided facile access to the corresponding pyrrolidine.[4] Conceptually, carbene transfer followed by an intramolecular [1,2]-Stevens rearrangement complements "cut and sew" reactions for non-carbonylative, one-carbon homologation of nitrogen-containing compounds. Given the prevalence of nitrogen heterocycles across numerous sectors of the chemical industry, especially pharmaceuticals,[5] extending these methodologies to other saturated *N*-heterocycles would represent a new approach for the synthesis of important chiral amine building blocks.

Despite their promising properties,[6] azetidines are underrepresented relative to closely related nitrogen-containing heterocycles: this is due to a lack of robust synthetic methods to access these species,[7-8] especially using asymmetric catalysis.[9-10] Application of a ring-expansion strategy for the asymmetric, one-carbon homologation of readily prepared aziridines via carbene insertion would be an attractive new entry towards the enantioselective synthesis of azetidines (**Figure 2-1**). However, this approach comes with two major selectivity challenges. The first is the innate reactivity of the intermediate aziridinium ylides, which undergo highly favorable cheletropic extrusion of olefins in many contexts.[11]

Schomaker and others have demonstrated that these reactive intermediates can be harnessed in [2,3]-Stevens rearrangements and other ring-opening reactions.[12] However, we are unaware of any examples of a one-carbon ring expansion of aziridines through a

[1,2]-Stevens rearrangement strategy. Secondly, the diradical mechanism of the [1,2]-Stevens rearrangement[13] has made it a challenging reaction class for asymmetric catalysis: few asymmetric variations have been reported.[14] Enantiopure quaternary ammonium salts can undergo [1,2]-Stevens rearrangements with *N*-to-*C* chirality transfer;[15] however, escape of the radical pair from the solvent cage is often competitive with radical recombination,[16] and erosion of enantiopurity is often observed. General strategies for stereocontrol over these rearrangements are an unmet challenge facing the field of asymmetric catalysis.

The joint selectivity challenges presented by the asymmetric one-carbon ring expansion of aziridines into azetidines requires a potential catalyst not only to select for the [1,2]-Stevens rearrangement in preference to cheletropic extrusion of olefins, but also to exert enantiocontrol over potential radical intermediates. Nature utilizes ring-size manipulation in the biosynthesis of natural products, with common strategies for biocatalytic one-carbon ring expansion including oxidative ring expansions[17] and carbocation rearrangements.[18] Furthermore, enzymes derived from cytochrome $P450_{BM3}$, such as cytochromes P411, and other hemoproteins have emerged as powerful catalysts for carbene transfer reactions,[19] and formation of strained rings such as cyclopropanes and cyclopropenes with excellent stereoselectivities has been reported.[20] The most common reactions of enzymatic iron-carbenoid intermediates are additions across π-systems[19-20] or X–H bond insertions:[21-22] biocatalytic C–N bond insertion through Stevens rearrangements of any kind have yet to be reported. We envisioned that a carbene transfer enzyme could potentially achieve the requisite chemo- and stereoselection necessary to perform this challenging reaction (**Figure 2-1**).

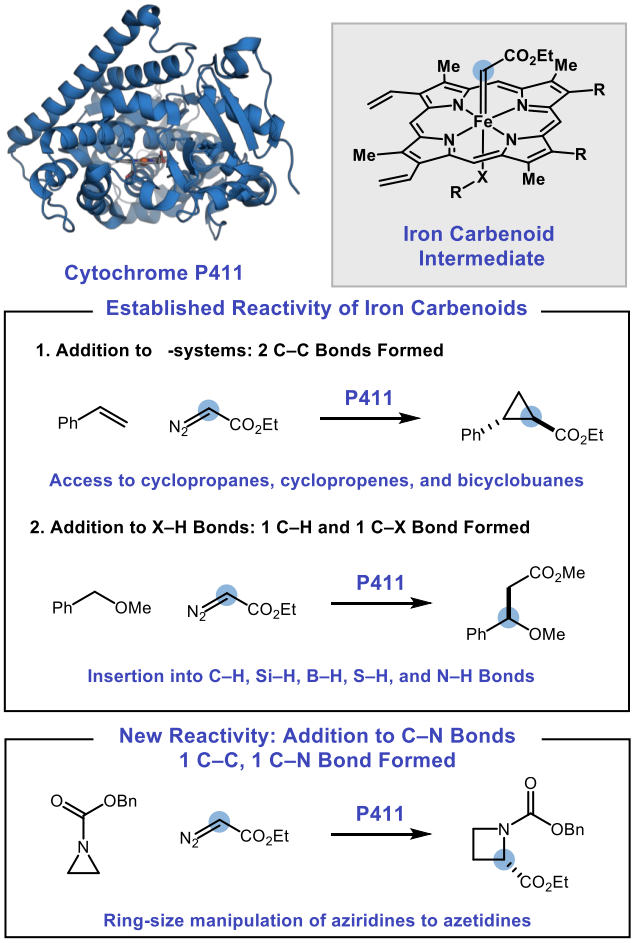


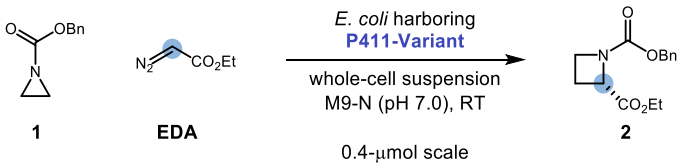
Figure 2-1. Classification of enzyme-mediated carbene transfer reactions for various bond disconnections.

2.2 Results and Discussion

We initiated our studies by screening a panel of hemoproteins for the model reaction of benzyl aziridine-1-carboxylate **1** with ethyl diazoacetate (EDA) as a carbene precursor to provide enantioenriched azetidine **2** (**Table 2-1**) in suspensions of *Escherichia coli* (*E. coli*) whole cells. We were delighted to find that a variant of P411$_{BM3}$-CIS[23] with the additional mutations P248T, I263G, and L437F ("Parent F2"), provided the product with 3.6% yield, 73 total turnover numbers (TTNs), and 90:10 er favoring the (*S*)-enantiomer (Entry 1). Parent F2 is derived from hemoproteins originally engineered for the cyclopropanation of heteroatom-substituted olefins[24] and is 17 mutations away from its wild-type progenitor, cytochrome P450$_{BM3}$ from *Bacillus megaterium*, which natively catalyzes the oxidation of

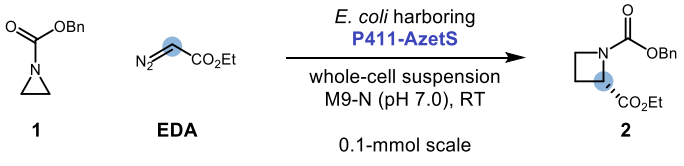
long-chain fatty acids.[25] Control experiments revealed that hemin is unable to catalyze this reaction (see **Appendix A** for details). Further control reactions indicated that the observed formation of the ring-opened hydrolysis product of **1** is not an enzyme-dependent process. No other aziridine-derived byproducts (e.g., cheletropic extrusion products[11], carbene insertion into the benzylic C–H[21c], or α-N–H bonds of the substrate[21e]) were identified, including a second ring expansion to form the corresponding pyrrolidine.[4] Further experiments demonstrated that neither **2** nor the unsubstituted benzyl azetidine-1-carboxylate underwent ring expansion under the disclosed conditions. Chemoselectivity for aziridine ring expansion over azetidine ring expansion in this system can be attributed to the increased pyramidalization at nitrogen observed for acylaziridines and related compounds, which increases their *N*-nucleophilicity relative to less strained amides.[26]

Table 2-1. Lineage and Reaction Optimization[a]



1 **EDA** **2**

Entry	Variant	Mutations Relative to Prior Generation	TTN	Yield (%)	e.r.
1	Parent F2	None	73	3.6	90:10
2	F2.1	G263Y	70	3.5	75:25
3	F2.2	T327V	126	6.3	56:44
4	F2.3	A330T	193	9.6	59:41
5	F2.4	H266P	394	19.7	62:38
6	F2.5	M177Q	699	34.9	94:6
7	F2.6	T436G	945	47.3	93:7
8	F2.7	L233F	997	49.8	94:6
9	F2.8	T149M	1040	52.0	99:1
10	F2.9	R47Q	1190	59.7	99:1
11	P411-AzetS	M118K	1200	59.9	99:1



1 **EDA** **2**

Entry	Change from Conditions Above	TTN	Yield (%)	e.r.
12	None	1580	79.1	99:1
13	20 mM [**1**]; 30 mM [**EDA**]	2200	55.0	99:1
14	Lysate	1090	54.4	99:1
15	Lysate; 20 mM [**1**]; 30 mM [**EDA**]	1570	39.3	99:1
16	4 °C	1610	80.2	99:1
17	Lysate; 4 °C	1380	68.7	99:1

[a]Reactions were performed on the designated scale and run for 16 h with 10 mM of **1**, 15 mM of EDA, and 5 μM of protein. TTN and yields were determined via GC analysis of crude reaction mixtures relative to an internal standard and represent the average of three experiments. The enantiomeric ratio (er) of the product was determined by chiral GC.

Encouraged by this promising initial activity and high enantioselectivity, we chose Parent F2 as a starting point for directed evolution to improve enzyme performance using iterative site-saturation mutagenesis (SSM) of residues located in the heme domain (Entries 2–11), screening for improved azetidine yield by gas chromatography. Sites were selected for mutagenesis based on success in previous directed evolution campaigns of $P450_{BM3}$ as well

as prior knowledge of residues responsible for substrate binding and catalysis in the heme domain of this protein scaffold.[17a] Ten beneficial mutations were identified during this campaign, resulting in a more efficient 'azetidine synthase' (P411-AzetS) with a net improvement of 16-fold in TTN and improved enantioselection (99:1 er). With P411-AzetS in hand, we next examined the impact of varying the reaction conditions on the product yield (Entries 12–17). Notably, increasing the scale from 4 μmol to 100 μmol resulted in an increase in the reaction yield. When the concentrations of **1** and EDA were doubled to 20 mM and 30 mM, respectively, a decrease in reaction yield was observed (although TTN increased). The ring expansion reaction also proceeded in clarified cell lysate, albeit with decreased yields when compared to analogous reactions performed with whole-cell suspensions. Lastly, decreasing the reaction temperature from 22 to 4 °C did not have a meaningful impact on the reaction yields when run in whole-cell suspensions.

Next, we sought to examine the substrate scope of this reaction and whether or not the new selectivities we observed could be extended to other substrates (**Scheme 2-1**). When this reaction was run at 0.5-mmol scale, azetidine **2** could be formed in 75% yield, 1490 TTN, 67% isolated yield, and 99:1 er. Other aromatic groups could be used in lieu of a phenyl group with uniformly high enantioselection observed in all cases. Notably, a thiophene-bearing aziridine could undergo chemoselective ring expansion to azetidine **3** with no observed cyclopropanation byproducts. This selectivity is notable not only because thiophenes are known to react with EDA-derived metal carbenoids under mild conditions,[27] but also because Parent F2 was originally engineered to perform cyclopropanation of heteroatom-substituted olefins.[24] Fluorine substituents were also tolerated on the arene ring at the *para*, *meta*, and *ortho* positions to furnish fluorinated products **4**–**6**. In addition to EDA, other diazoacetate compounds could participate in one-carbon ring expansion with at least 99:1 er (**7**–**8**). When methyl diazoacetate was used as the carbene precursor to yield **9**, a notable decrease in er (81:19) was observed. One hypothesis for this decrease in enantiopurity is that the smaller aliphatic chain allows for greater conformational freedom of the iron porphyrin carbene intermediate or the putative diradical intermediate. This explanation is consistent with prior work on enzyme-mediated carbene transfer reactions using perfluoroalkyl-stabilized diazo compounds as carbene precursors, where the substrate chain length has a profound influence on the absolute stereochemical

configuration of the reaction product.[21e] The reaction could also be scaled up from 0.5-mmol scale to 10-mmol scale to furnish **2** in 1220 TTN, 61% yield, and 99:1 er with an isolated yield of 1.44 g (55% isolated yield), demonstrating that gram-scale production of enantioenriched azetidines is viable using this platform and that extension of this activity could be a powerful tool for the asymmetric synthesis of chiral heterocycles.

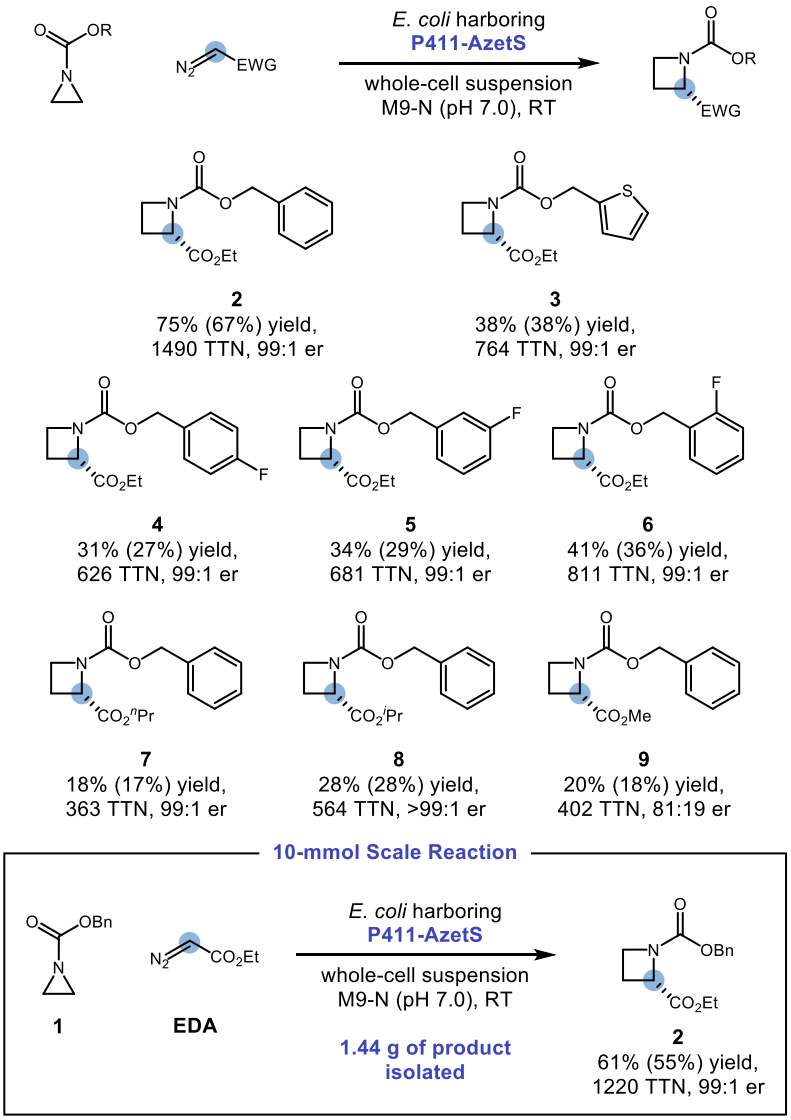
The current P411-AzetS lineage performs poorly with other substrate classes. Aziridine substrates with substituents on the carbon backbone of the ring were unable to undergo ring expansion due to their pronounced capacity for ring opening by hydrolysis relative to unsubstituted aziridine rings. This limitation also prevented *N*-alkyl or *N*-aryl aziridines from serving as viable substrates. Other classes of nitrogen protecting groups (e.g., amides and sulfonamides) demonstrated poor activity; one explanation is that the decreased *N*-nucleophilicity of these species hinders their ability to form aziridinium ylides. Finally, other carbamate-protecting groups (e.g., -Boc, -Alloc, and -CO_2Me) did not form the desired products, suggesting that the arene may be necessary for proper substrate binding with this lineage of enzymes. With respect to the diazo coupling partner, diazoacetates were uniquely effective: when other diazo coupling partners were subjected to the reaction conditions, only unreacted diazo starting materials or dimerization products were recovered.

While the limited substrate scope of P411-AzetS hampers its synthetic utility, this result highlights the power of directed evolution to deliver an incredibly selective catalyst. Over the course of this campaign, the cytochrome $P450_{BM3}$ scaffold was trained to recognize a non-natural substrate with atom-level selectivity. P411-AzetS demonstrates no activity with N-hydrocinnamoyl-aziridine (**10**)**,** which only differs from **1** by a single heavy atom. To assess if this selectivity was developed throughout evolution, several aziridines which demonstrated no activity with P411-AzetS were assayed against variants F2.5 to F2.9 (**Table 2-2**). We were excited to find that earlier variants were capable of driving azetidine formation from aziridines **10** and **11**, while aziridines **12** and **13** showed no ring expansion activity. This finding underlines an important phenomenon which occurs during directed evolution campaigns: while optimization on a model system can lead to a highly selective enzyme, the sequences generated through the course of evolution will often be proficient

at related tasks. Efforts to expand the observed, unprecedented reactivity and selectivity to the synthesis of other classes of azetidines are ongoing.

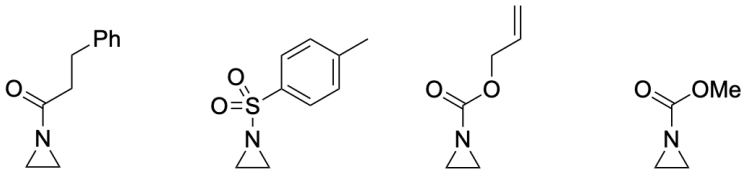
We note that in the course of evolving the P411-AzetS lineage for this novel activity, late-stage variants exhibited reduced thermostability. For example, the variant F2.9 has a T_{50} of approximately 40 °C (**Figure A-2** of **Appendix A**). Although the enzymes are capable of facilitating unprecedented chemistry, it is also important to consider their suitability for downstream applications. Firstly, as most mutations to a protein scaffold are destabilizing, the reduced stability of these enzymes makes them less favorable candidates for directed evolution campaigns (such as to improve their substrate promiscuity).[28] Furthermore, compromised thermostability would likely limit the utility of P411-AzetS in industrial biocatalytic processes.[29] Indeed, purified P411-AzetS was found to precipitate following overnight incubation under impelled dialysis conditions at 4°C. These findings underscore the importance of evolving enzymes not only for improved catalytic function, but also properties that support their practical deployment.

Scheme 2-1. Substrate Scope[a]



[a]Reactions were performed on 0.5-mmol scale unless otherwise specified. Analytical yields and TTN were determined by GC-FID. Yields for isolated and purified material are designated in parentheses. The er was determined by Chiral GC. For 0.5-mmol scale reactions, all numbers reported represent the average of two trials. For the 10-mmol scale reaction, the reported numbers represent one run.

Table 2-2. Screening of substrates which demonstrated no activity with P411-AzetS[a]



Variant	10	11	12	13
F2.5	<5%	trace	n.d.	n.d.
F2.6	<5%	trace	n.d.	n.d.
F2.7	<5%	trace	n.d.	n.d.
F2.8	<5%	n.d.	n.d.	n.d.
F2.9	n.d.	n.d.	n.d.	n.d.
P411-AzetS	n.d.	n.d.	n.d.	n.d.

Yield

[a]Reactions were performed on 4 μmol scale and run for 16 h with 10 mM of **1**, 15 mM of EDA, and 5 μM of protein. TTN and yields were determined via GC analysis of crude reaction mixtures relative to an internal standard and represent the average of three experiments. The enantiomeric ratio (er) of the product was determined by chiral GC.

A hypothetical mechanism for the one-carbon ring expansion of aziridines is shown in **Figure 2-2**. The reaction of a hemoprotein with a suitable carbene precursor forms an electrophilic iron-carbenoid intermediate, which could be trapped by a sufficiently nucleophilic aziridine. Ammonium ylides are commonly proposed as intermediates in hemoprotein-catalyzed N–H insertion reactions,[22] and Schomaker has reported numerous examples where carbamate-protected aziridines react with metal-carbenoid electrophiles to form aziridinium ylides.[11,12c-f] At the present time, it is not clear whether this intermediate would exist as a "free" or metal-bound ylide, although computational analysis of enzymatic N–H insertion reactions suggests that ammonium ylide intermediates react after dissociation from the iron center.[22f] Finally, the aziridinium ylide could undergo the desired [1,2]-Stevens rearrangement preferentially over cheletropic extrusion of ethylene, liberating the desired product and regenerating the hemoprotein. We envisioned that the

active site of an enzyme could mimic solvent caging effects, which are known to exert selectivity over radical recombination in [1,2]-Stevens rearrangements, to achieve asymmetric induction during ring expansion.[15-16] Such effects may also explain why free hemin is unable to catalyze the reaction in the absence of the specific confinement provided by the active sites of this enzyme lineage. Additionally, hemoproteins demonstrate high stereoselectivity in radical reactions, both in their native reactivity[30] as well as in new-to-nature activity cultivated through protein engineering,[31] lending further support to this hypothesis.

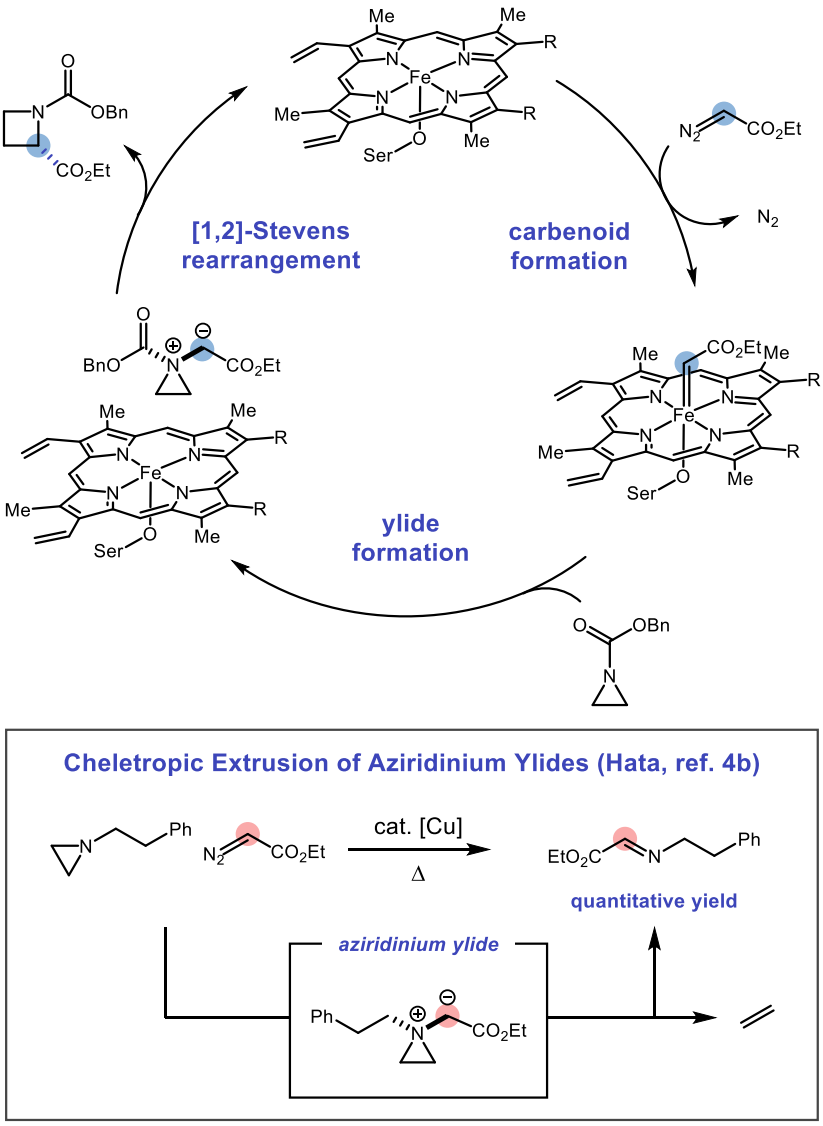


Figure 2-2. Possible catalytic cycle for one-carbon ring expansion of aziridines to furnish chiral azetidines, with cheletropic extrusion of ethylene as a possible side reaction. The ammonium ylide may also remain iron-bound.

In summary, we have demonstrated unprecedented hemoprotein-catalyzed [1,2]-Stevens rearrangement in the context of a one-carbon ring expansion of aziridines to azetidines. This system not only represents a rare example of a highly enantioselective [1,2]-Stevens rearrangement of ammonium ylides, but also selectivity for the [1,2]-Stevens rearrangement of aziridinium ylides over cheletropic extrusion of ethylene. This work highlights a key strength of biocatalysis: enzymatic stabilization of reactive intermediates enables not just enhanced stereoselectivity, but also unprecedented regio- and

chemoselectivity. We are optimistic that observed selectivities can be extended to other types of [1,2]-Stevens rearrangements, providing the grounds for future work in this area toward the synthesis of enantioenriched heterocycles and other chiral amines.

Chapter II Bibliography

1. a) Zhao, K.; Yamashita, K.; Carpenter, J.E.; Sherwood, T.C.; Ewing, W.R.; Cheng, P.T.W.; Knowles, R.R. Catalytic Ring Expansions of Cyclic Alcohols Enabled by Proton-Coupled Electron Transfer. *J. Am. Chem. Soc*. **2019**, *141*, 8752. b) Dherange, B.D.; Kelly, P.Q.; Liles, J.P.; Sigman, M.S.; Levin, M.D. Carbon Atom Insertion into Pyrroles and Indoles Promoted by Chlorodiazirines. *J. Am. Chem. Soc.* **2021**. *143*, 11337. c) Kennedy, S.H.; Dherange, B.D.; Berger, K.J.; Levin, M.D. Skeletal editing through direct nitrogen deletion of secondary amines. *Nature* **2021**, *593*, 223. d) Donald, J.R.; Unsworth, W.P. Ring-Expansion Reactions in the Synthesis of Macrocycles and Medium-Sized Rings. *Chem. Eur. J.* **2017**, *23*, 8780. e) Dowd, P.; Zhang, W. Free radical-mediated ring expansion and related annulations. *Chem. Rev.* **1993**, *93*, 2091.
2. For reviews, see: a) Chen, P.-H.; Billett, B.A.; Tsukamoto, T.; Dong, G. "Cut and Sew" Transformations via Transition-Metal-Catalyzed Carbon–Carbon Bond Activation. *ACS Catal.* **2017**, *7*, 1340. b) Xu, T.; Dermenci, A.; Dong, G. Transition metal-catalyzed C-C bond activation of four-membered cyclic ketones. *Top. Curr. Chem.* **2014**, *346*, 233. c) Gao, Y.; Fu, X.-F.; Yu, Z.-X. Transition Metal-Catalyzed Cycloadditions of Cyclopropanes for the Synthesis of Carbocycles: C–C Activation in Cyclopropanes. *Top. Curr. Chem.* **2014**, *346*, 195. d) Xia, Y.; Dong, G. Temporary or removable directing groups enable activation of unstrained C–C bonds. *Nat. Rev. Chem.* **2020**, *4*, 600. e) Jun, C.-H. Transition metal-catalyzed carbon–carbon bond activation. *Chem. Soc. Rev.* **2004**, *33*, 610. f) Chen, F.; Wang, T.; Jiao, N. Recent Advances in Transition-Metal-Catalyzed Functionalization of Unstrained Carbon–Carbon Bonds. *Chem. Rev.* **2014**, *114*, 8613. g) Souillart, L.; Cramer, N. Catalytic C–C Bond Activations via Oxidative Addition to Transition Metals. *Chem. Rev.* **2015**, *115*, 9410.
3. For representative examples, see: a) Tayama, E. Ring-Substitution, Enlargement, and Contraction by Base-Induced Rearrangements of N-Heterocyclic Ammonium Salts. *Heterocycles* **2016**, *92*, 793. b) Wittig, G.; Tenhaeff, H.; Schoch, W.; Koenig, G. Einige Synthesen über Ylide. *Liebigs Ann. Chem.* **1951**, *572*, 1. c) Chicharro, R.; de Castro, S.; Reino, J.; Arán, V.J. Synthesis of Tri- and Tetracyclic Condensed Quinoxalin-2-ones Fused Across the C-3−N-4 Bond. *Eur. J. Org. Chem.* **2003**, *2003*, 2314. d) Pedrosa, R.; Andrés, C.; Delgado, M. Stereocontrolled Ring Enlargement by Diastereoselective Stevens Rearrangement in Chiral 1,3-Oxazolidinium Salts. A Novel Entry to Enantiopure Morpholines. *Synlett* **2000**, *6*, 893.e) Harthong, S.; Bach, R.; Besnard, C.; Guénée, L.; Lacour, J. Ring-Expansion Reactions of Binaphthyl Azepines and Ferrocenophanes through Metal-Catalyzed [1,2]-Stevens Rearrangements. *Synthesis* **2013**, *45*, 2070. f) Vanecko, J.A.; West, F.G. A Novel, Stereoselective Silyl-Directed Stevens [1,2]-Shift of Ammonium Ylides. *Org. Lett.* **2002**, *4*, 2813. g) Hanessian, S.; Mauduit, M. Highly Diastereoselective Intramolecular [1,2]-Stevens Rearrangements—Asymmetric Syntheses of Functionalized Isopavines as Morphinomimetics. *Angew. Chem. Int. Ed.* **2001**, *40*, 3810. h) Liou, J.-P.; Cheng, C.-Y. Total synthesis of (±)-desoxycodeine-D: a novel route to the morphine skeleton. *Tetrahedron Letters* **2000**, *41*, 915. i) Sharma, A.; Besnard, C.; Guénée, L.; Lacour, J. Asymmetric

synthesis of ethano-Tröger bases using CuTC-catalyzed diazo decomposition reactions. *Org. Biomol. Chem.* **2012**, *10*, 966. j) Vanecko, J.A.; Wan, H.; West, F.G. Recent advances in the Stevens rearrangement of ammonium ylides. Application to the synthesis of alkaloid natural products. *Tetrahedron* **2006**, *62*, 1043. k) Kowalkowska, A.; Jończyk, A. [1,2] Stevens sigmatropic rearrangement of pyrrolidinium ylides—simple synthesis of 3-aryl-2-cyano-1-methylpiperidines. *Tetrahedron* **2015**, *71*, 9630. l) Lahm, G.; Pacheco, J.C.O.; Opatz, T. Rearrangements of Nitrile-Stabilized Ammonium Ylides. *Synthesis* **2014**, *46*, 2413. m) Empel, C.; Jana, S.; Koenigs, R.M. Advances in [1,2]-Sigmatropic Rearrangements of Onium Ylides via Carbene Transfer Reactions. *Synthesis* **2021**, *53*, 4567.

4. a) Hata, Y.; Watanable, M. Fragmentation reaction of aziridinium ylids. *Tetrahedron Letters* **1972**, *13*, 3827. b) Hata, Y.; Watanabe, M. Fragmentation reaction of aziridinium ylids. II. *Tetrahedron Letters* **1972**, *13*, 4659. c) Bott, T.M.; Vanecko, J.A.; West, F.G. One-Carbon Ring Expansion of Azetidines via Ammonium Ylide [1,2]-Shifts: A Simple Route to Substituted Pyrrolidines. *J. Org. Chem.* **2009**, *74*, 2832. d) Drouillat, B.; d'Aboville, E.; Bourdreux, F.; Couty, F. Synthesis of 2-Phenyl- and 2,2-Diarylpyrrolidines through Stevens Rearrangement Performed on Azetidinium Ions. *Eur. J. Org. Chem.* **2014**, *2014*, 1103. e) Couty, F.; Durrat, F.; Evano, G.; Prim, D. Synthesis and reactivity of enantiomerically pure N-alkyl-2-alkenyl azetidinium salts. *Tetrahedron Letters* **2004**, *45*, 7525.
5. Approximately 59% of all small-molecule drugs contain at least one nitrogen-containing heterocycle: Vitaku, E.; Smith, D.T.; Njardarson, J.T. Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J. Med. Chem.* **2014**, *57*, 10257.
6. a) St. Jean, D. J.; Fotsch, C. Mitigating Heterocycle Metabolism in Drug Discovery. *J. Med. Chem.* **2012**, *55*, 6002. b) Wang, D.X.; Booth, H.; Lerner-Marmarosh, N.; Osdene, T.S.; Abood, L.G. Structure–Activity Relationships for Nicotine Analogs Comparing Competition for [^{3}H]Nicotine Binding and Psychotropic Potency. *Drug Dev. Res.* **1998**, *45*, 10.
7. a) Brandi, A.; Cicchi, S.; Cordero, F.M. Novel Syntheses of Azetidines and Azetidinones. *Chem. Rev.* **2008**, *108*, 3988. (b) Mehra, V.; Lumb, I.; Anand, A.; Kumar, V. Recent advances in synthetic facets of immensely reactive azetidines. *RSC Adv.* **2017**, *7*, 45763.
8. For recent photochemical [2+2] strategies to access azetidines: a) Becker, M.R.; Richardson, A.D.; Schindler, C.S. Functionalized azetidines via visible light-enabled aza Paternò-Büchi reactions. *Nat. Commun.* **2019**, *10*, 5095. b) Becker, M.R.; Wearing, E.R.; Schindler, C.S. Synthesis of azetidines via visible-light-mediated intermolecular [2+2] photocycloadditions. *Nat. Chem.* **2020**, *12*, 898. c) Richardson, A.D.; Becker, M.R.; Schindler, C.S. Synthesis of azetidines by aza Paternò–Büchi reactions. *Chem. Sci.* **2020**, *11*, 7553. d) Sakamoto, R.; Inada, T.; Sakura, S.; Maruoka, K. [2 + 2] Photocycloadditions between the Carbon–Nitrogen Double Bonds of Imines and Carbon–Carbon Double Bonds. *Org. Lett.* **2016**, *18*, 6252. e) Flores, D.; Neville, M.; Schmidt, V. Intermolecular 2+2 Imine-Olefin Photocycloadditions Enabled by Cu(I)-Alkene MLCT. *Nat. Commun.* **2022**, *13*, 2764.

9. For selected examples, see: a) Malik, S.; Nadir, U.K. A Facile Synthesis of 1-Arenesulfonylazetidines through Reaction of 1-Arenesulfonylaziridines with Dimethylsulfoxonium Methylide Generated under Microwave Irradiation. *Synlett* **2008**, *1*, 108. b) Han, J.-Q.; Zhang. H.-H.; Xu, P.-F.; Luo, Y.-C. Lewis Acid and (Hypo)iodite Relay Catalysis Allows a Strategy for the Synthesis of Polysubstituted Azetidines and Tetrahydroquinolines. *Org. Lett.* **2016**, *18*, 5212. c) He, G.; Zhao, Y.; Zhang, S.; Lu, C.; Chen, G. Highly Efficient Syntheses of Azetidines, Pyrrolidines, and Indolines via Palladium Catalyzed Intramolecular Amination of C(sp^3)–H and C(sp^2)–H Bonds at γ and δ Positions. *J. Am. Chem. Soc.* **2012**, *134*, 3. d) Zhang. H-H.; Luo, Y.-C.; Wang, H.-P.; Chen. W.; Xu, P.-F. $TiCl_4$ Promoted Formal [3 + 3] Cycloaddition of Cyclopropane 1,1-Diesters with Azides: Synthesis of Highly Functionalized Triazinines and Azetidines. *Org. Lett.* **2016**, *16*, 4896. e) Lowe, J.T.; Lee, M.D.; Akella, L.B.; Davoine, E.; Donckele, E.J.; Durak, L.; Duvall, J.R.; Gerard, B.; Holson, E.B.; Joliton, A.; Kesavan, S.; Lemercier, B.C.; Liu, H.; Marié, J.-C.; Mulrooney, C.A.; Muncipinto, G.; Welzel-O'Shea, M.; Panko, L.M.; Rowley, A.; Suh, B.-C.; Thomas, M.; Wanger, F.F.; Wei, J.; Foley, M.A.; Marcaurelle, L.A. Synthesis and Profiling of a Diverse Collection of Azetidine-Based Scaffolds for the Development of CNS-Focused Lead-like Libraries. *J. Org. Chem.* **2012**, *77*, 7187.
10. For synthetic sequences leading to enantioenriched aziridines, see: a) Kapoor, R.; Chawla, R.; Singh, S.; Yadav, L.D.S. Organocatalytic Asymmetric Synthesis of 1,2,4-Trisubstituted Azetidines by Reductive Cyclization of Aza-Michael Adducts of Enones. *Synlett* **2012**, *23*, 1321. b) Hanessian, S.; Bernstein, N.; Yang, R.Y.; Maguire, R. Asymmetric synthesis of L-azetidine-2-carboxylic acid and 3-substituted congeners--conformationally constrained analogs of phenylalanine, naphthylalanine, and leucine. *Bioorg. Med. Chem. Lett.* **1999**, *17*, 1437. c) Marichev, K.O.; Wang, K.; Dong, K.; Greco, N.; Massey, L.A.; Deng, Y.; Arman, H.; Doyle, M.P. Synthesis of Chiral Tetrasubstituted Azetidines from Donor–Acceptor Azetines via Asymmetric Copper(I)-Catalyzed Imido-Ylide [3+1]-Cycloaddition with Metallo-Enolcarbenes. *Angew. Chem. Int. Ed.* **2019**, *58*, 16188. d) Singh, G.S. Advances in synthesis and chemistry of azetidines. In *Advances in Heterocyclic Chemistry.* Academic Press, 2001; pp 1-74. e) Ma, X.; Zhao, H.; Binayeva, M.; Ralph, G.; Diane, M.; Zhao, S.; Wang, C.-Y.; Biscoe, M.R. A General Approach to Stereospecific Cross-Coupling Reactions of Nitrogen-Containing Stereocenters. *Chem* **2020**, *6*, 781.
11. Dequina, H.J.; Schomaker, J.M. Aziridinium Ylides: Underused Intermediates for Complex Amine Synthesis. *Trends in Chemistry* **2020**, *2*, 874.
12. a) Clark, J.S.; Hodgson, P.B.; Goldsmith, M.D.; Blake, A.J.; Cooke, P.A.; Street, L.J. Rearrangement of ammonium ylides produced by intramolecular reaction of catalytically generated metal carbenoids. Part 2. Stereoselective synthesis of bicyclic amines. *J. Chem. Soc. Perkin Trans. 1* **2001**, 3325. b) Rowlands, G.J.; Barnes, W.K. Studies on the [2,3]-Stevens rearrangement of aziridinium ions. *Tetrahedron Letters* **2004**, *45*, 5347. c) Nicastri, K.A.; Zappia, S.A.; Pratt, J.C.; Duncan, J.M.; Guzei, I.A.; Fernández, I.; Schomaker, J.M. Tunable Aziridinium Ylide Reactivity: Noncovalent Interactions Enable Divergent Product Outcomes. *ACS Catal.* **2022**, *12*, 1572. d) Schmid, S.C.; Guzei, I.A.; Fernández, I.; Schomaker,

J.M. Ring Expansion of Bicyclic Methyleneaziridines via Concerted, Near-Barrierless [2,3]-Stevens Rearrangements of Aziridinium Ylides. *ACS Catal.* **2018**, *8*, 7907. e) Schmid, S.C.; Guzei, I.A.; Schomaker, J.M. A Stereoselective [3+1] Ring Expansion for the Synthesis of Highly Substituted Methylene Azetidines. *Angew. Chem. Int. Ed.* **2017**, *56*, 12229. f) Eshon, J.; Nicastri, K.A.; Schmid, S.C.; Raskopf, W.T.; Guzei, I.A.; Fernández, I.; Schomaker, J.M. Intermolecular [3+3] ring expansion of aziridines to dehydropiperidines through the intermediacy of aziridinium ylides. *Nat. Commun.* **2020**, *11*, 1273.

13. a) Bach, R.; Harthong, S.; Lacour, J. Nitrogen- and Sulfur-Based Stevens and Related Rearrangements. *Comprehensive Organic Synthesis II*, **2014**, *3*, 992. b) Lepley, A.R.; Becker, R.H.; Giumanini, A.G. Benzyne addition to *N,N*-dimethylbenzylamine. *J. Org. Chem.* **1971**, *36*, 1222.
14. a) Qu, J.-P.; Xu, Z.-H.; Zhou, J.; Cao, C.-L.; Sun, X.-L.; Dai, L.-X.; Tang, Y. Ligand-Accelerated Asymmetric [1,2]-Stevens Rearrangment of Sulfur Ylides *via* Decomposition of Diazomalonates Catalyzed by Chiral Bisoxazoline/Copper Complex. *Adv. Synth. Catal.* **2009**, *351*, 308. b) Tomooka, K.; Sakamaki, J.; Harada, M.; Wada, R. Enantioselective [1,2]-Stevens Rearrangement Using Sugar-Derived Alkoxides as Chiral Promoters. *Synlett* **2008**, 5, 683. c) Hong, F.-L.; Shi, C.-Y.; Hong, P.; Zhai, T.-Y. Zhu, X.-Q. Lu, X. Ye, L.-W. Copper-Catalyzed Asymmetric Diyne Cyclization via [1,2]-Stevens-Type Rearrangement for the Synthesis of Chiral Chromeno[3,4-c]pyrroles. *Angew. Chem. Int. Ed.* **2022**, *61*, e202115554.
15. a) Tayama, E.; Nanbara, S.; Nakai, T. Asymmetric [1,2] Stevens Rearrangement of (*S*)-*N*-Benzylic Proline-derived Ammonium Salts under Biphasic Conditions. *Chem. Lett.* **2006**, *35*, 478. b) Gonçalves-Farbos, M-H.; Vial, L.; Lacour, J. Enantioselective [1,2]-Stevens rearrangement of quaternary ammonium salts. A mechanistic evaluation. *Chem. Commun.* **2008**, 829. c) Palombi, L. The first electro-induced asymmetric Stevens rearrangement of (*S*)- and (*R*)-N-benzyl proline-derived ammonium salts. *Catalysis Communications* **2011**, *12*, 485. d) Glaeske, K.W.; West, F.G. Chirality Transfer from Carbon to Nitrogen to Carbon via Cyclic Ammonium Ylides. *Org. Lett.* **1999**, *1*, 31. e) Vial, L.; Gonçalves, M.-H.; Morgantini, P.-Y.; Weber, J.; Bernardinelli, G.; Lacour, J. Unusual Regio- and Enantioselective [1,2]-Stevens Rearrangement of a Spirobi[dibenzazepinium] Cation. *Synlett* **2004**, *9*, 1565
16. a) Woodward, J.R. Radical Pairs in Solution. *Prog. React. Kinet. Mec.* **2002**, *27*, 165. b) Franck, J.; Rainbowitsch, E. Some remarks about free radicals and the photochemistry of solutions. *Trans. Faraday Soc.* **1934**, *30*, 120. c) Braden, D.A.; Parrack, E.E.; Tyler, D.R. Solvent cage effects. I. Effect of radical mass and size on radical cage pair recombination efficiency. II. Is geminate recombination of polar radicals sensitive to solvent polarity? *Coord. Chem. Rev.* **2001**, *211*, 279.
17. For reviews and representative examples, see: a) Whitehouse, C.J.C.; Bell, S.G.; Wong, L.-L. $P450_{BM3}$ (CYP102A1): connecting the dots. *Chem. Soc. Rev.* **2011**, *41*, 1218. b) Thiel, D.; Doknić, D.; Deska, J. Enzymatic aerobic ring rearrangement of optically active furylcarbinols. *Nat. Commun.* **2014**, *5*, 5278. c) Tang, M.-C.; Zou, Y.; Watanabe, K.; Walsh, C.T.; Tang, Y. Oxidative Cyclization in Natural Product Biosynthesis. *Chem. Rev.* **2017**, *117*, 5226. d) Bat-Erdene, U.; Kanayama, D.; Tan, D.; Turner, W.C.; Houk, K.N.; Ohashi, M.; Tang, Y. Iterative Catalysis in the

Biosynthesis of Mitochondrial Complex II Inhibitors Harzianopyridone and Atpenin B. *J. Am. Chem. Soc.* **2000**, *142*, 8550. e) Fürst, M.J.L.; Gran-Scheuch, A.; Aalbers, F.S.; Fraaije, M.W. Baeyer–Villiger Monooxygenases: Tunable Oxidative Biocatalysts. *ACS Catal.* **2019**, *9*, 11207. f) Leisch, H.; Morley, K.; Lau, P.C.K. Baeyer−Villiger Monooxygenases: More Than Just Green Chemistry. *Chem. Rev.* **2011**, *111*, 4165. g) Deska, J.; Thiel, D.; Gianolio, E. The Achmatowicz Rearrangement – Oxidative Ring Expansion of Furfuryl Alcohols. *Synthesis* **2015**, *47*, 3435.

18. For reviews and representative examples, see: a) Christianson, D.W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **2017**, *117*, 11570. b) Hoshino, T.; Kouda, M.; Abe, T.; Ohashi, S. New Cyclization Mechanism for Squalene: a Ring-expansion Step for the Five-membered C-ring Intermediate in Hopene Biosynthesis. *Biosci. Biotechnol. Biochem.* **1999**, *63*, 2038. c) Xu, M.; Jia, M.; Hong, Y.J.; Yin. X.; Tantillo, D.J.; Proteau, P.J.; Peters, R.J. Premutilin Synthase: Ring Rearrangement by a Class II Diterpene Cyclase. *Org. Lett.* **2018**, *20*, 1200. d) Quan, Z.; Dickschat, J.S. Biosynthetic Gene Cluster for Asperterpenols A and B and the Cyclization Mechanism of Asperterpenol A Synthase. *Org. Lett.* **2020**, *22*, 7552. e) Xu, R.; Fazio, G.C.; Matsuda, S.P.T. On the origins of triterpenoid skeletal diversity. *Phytochemistry* **2004**, *65*, 261. f) Rudolf, J.D.; Chang, C.-Y. Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. *Nat. Prod. Rep.* **2020,** *37*, 425. g) Dickschat, J.S. Bacterial Diterpene Biosynthesis. *Angew. Chem. Int. Ed.* **2019**, *58*, 15964.
19. a) Brandenberg, O.F.; Fasan, R.; Arnold, F.H. Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **2017**, *47*, 102. b) Yang, Y.; Arnold, F.H. Navigating the Unnatural Reaction Space: Directed Evolution of Heme Proteins for Selective Carbene and Nitrene Transfer. *Acc. Chem. Res.* **2021**, *54*, 1209. c) Liu, Z.; Arnold, F.H. New-to-nature chemistry from old protein machinery: carbene and nitrene transferases. *Curr. Opin. Biotechnol.* **2021**, *69*, 43. d) Dunham, N.P.; Arnold, F.H. Nature's Machinery, Repurposed: Expanding the Repertoire of Iron-Dependent Oxygenases. *ACS Catal.* **2020**, *10*, 12239.
20. a) Chen, K.C.; Arnold, F.H. Engineering Cytochrome P450s for Enantioselective Cyclopropenation of Internal Alkynes. *J. Am. Chem. Soc.* **2020**, *142*, 6891. b) Chen, K.C.; Huang, X.; Kan, S.B.J.; Zhang, R.K.; Arnold, F.H. Enzymatic construction of highly strained carbocycles. *Science* **2018**, *360*, 71.
21. a) Kan, S.B.J.; Lewis, R.D.; Chen, K.; Arnold, F.H. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science* **2016**, *354*, 1048. b) Kan, S.B.J.; Huang, X.; Gumulya, Y.; Chen, K.; Arnold, F.H. Genetically programmed chiral organoborane synthesis. *Nature* **2017**, *552*, 132. c) Zhang, R.K.; Chen. K.; Huang, X.; Wohlschlager, L.; Renata, H.; Arnold, F.H. Enzymatic assembly of carbon–carbon bonds via iron-catalysed sp^3 C–H functionalization. *Nature* **2019**, *565*, 67. d) Chen, K.; Zhang, S.-Q.; Brandenberg, O.F.; Hong, X.; Arnold, F.H. Alternate Heme Ligation Steers Activity and Selectivity in Engineered Cytochrome P450-Catalyzed Carbene-Transfer Reactions. *J. Am. Chem. Soc.* **2018**, *140*, 16402. e) Zhang, J.; Huang, X.; Zhang,

R.K.; Arnold, F.H. Enantiodivergent α-Amino C–H Fluoroalkylation Catalyzed by Engineered Cytochrome P450s. *J. Am. Chem. Soc.* **2019**, *141*, 9798.

22. a) Wang, Z.J.; Peck. N.E.; Renata, H.; Arnold, F.H. Cytochrome P450-catalyzed insertion of carbenoids into N–H bonds. *Chem. Sci.* **2014**, *5*, 598. b) Steck, V.; Carminati, D.M.; Johnson, N.R.; Fasan, R. Enantioselective Synthesis of Chiral Amines via Biocatalytic Carbene N–H Insertion. *ACS Catal.* **2020**, *10*, 10967. c) Sreenilayam, G.; Fasan, R. Myoglobin-catalyzed intermolecular carbene N–H insertion with arylamine substrates. *Chem. Commun.* **2015**, *15*, 1532. d) Sreenilayam, G.; Moore, E.J. Steck, V.; Fasan, R. Metal Substitution Modulates the Reactivity and Extends the Reaction Scope of Myoglobin Carbene Transfer Catalysts. *Adv. Synth. Catal.* **2017**, *359*, 2076. e) Steck, V.; Sreenilayam, G.; Fasan, R. Selective Functionalization of Aliphatic Amines via Myoglobin-Catalyzed Carbene N–H Insertion. *Synlett* **2020**, *31*, 224. f) Liu, Z.; Calvó-Tusell, C.; Zhou, A.Z.; Chen, K.; Garcia-Borràs, M.; Arnold, F.H. Dual-Function Enzyme Catalysis for Enantioselective Carbon–Nitrogen Bond Formation. *Nature Chemistry* **2021**, *13*, 1166.
23. Coelho, P.S.; Wang, Z.J.; Ener, M.E.; Baril, S.A.; Kannan, A.; Arnold, F.H.; Brustad, E.M. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **2013**, *9*, 485.
24. Brandenberg, O.F.; Prier, C.K.; Chen, K.; Knight, A.M.; Wu, Z.; Arnold, F.H. Stereoselective Enzymatic Synthesis of Heteroatom-Substituted Cyclopropanes. *ACS Catal.* **2018**, *8*, 2629.
25. Narhi, L.O.; Fulco, A.J. Characterization of a catalytically self-sufficient 119,000-dalton cytochrome P-450 monooxygenase induced by barbiturates in *Bacillus megaterium*. *J. Biol. Chem.* **1986**, *261*, 7160.
26. a) Ohwada, T.; Okamoto, I.; Shudo, K.; Yamaguchi, K. Intrinsic pyramidal nitrogen of *N*-sulfonamides. *Tetrahedron Letters*, **1998**, *39*, 7877. b) Ferraris, D.; Drury III, W.J.; Cox, C.; Lectka, T. "Orthogonal" Lewis Acids: Catalyzed Ring Opening and Rearrangement of Acylaziridines. *J. Org. Chem.* **1998**, *63*, 4568. c) Cho, S.J.; Cui, C.; Lee, J.Y.; Park, J.K.; Suh, S.B.; Park, J.; Kim, B.H.; Kim, K.S. *N*-Protonation vs *O*-Protonation in Strained Amides: *Ab Initio* Study. J. *Org. Chem.* **1997**, *62*, 4068.
27. Waser, M.; Moher, E.D.; Borders, S.S.K.; Hansen, M.M.; Hoard, D.W.; Laurila, M.E.; LeTourneau, M.E.; Miller, R.D.; Phillips, M.L.; Sullivan, K.A.; Ward, J.A.; Xie, C.; Bye, C.A.; Leitner, T.J.; Herzog-Krimbacher, B.; Kordian, M.; Müllner, M. Process Development for a Key Synthetic Intermediate of LY2140023, a Clinical Candidate for the Treatment of Schizophrenia. *Org. Process Res. Dev.* **2011**, *15*, 1266.
28. a) Bloom, J.D.; Labthavikul, S.T.; Oter, C.R.; Arnold, F.H. Protein stability promotes evolvability. *Proc. Nat. Acad. Sci*. **2006**, *103*, 5869. b) Tokuriki, N.; Tawfik, D.S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **2009**, *19*, 596.
29. Gomes, M.D.; Woodley, J.M. Considerations when Measuring Biocatalyst Performance. *Molecules* **2019**, *24*, 3573.
30. Ortiz de Montellano, P.R. Hydrocarbon Hydroxylation by Cytochrome P450 Enzymes. *Chem. Rev.* **2010**, 110, 932.

31. Yang, Y.; Cho, I.; Qi, X.; Liu, P.; Arnold, F.H. An enzymatic platform for the asymmetric amination of primary, secondary and tertiary C(sp^3)–H bonds. *Nat. Chem.* **2019**, *11*, 987.

Appendix A

SUPPLEMENTARY INFORMATION FOR CHAPTER II

A.1. General Information and Protocols

A.1.1. Safety Statement

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure. For protein concentration determination experiments, carbon monoxide (CO) is used. CO is flammable, highly toxic, and can be lethal at high doses and must be used in a fume hood equipped with a CO detector to avoid accidental exposure. Other than that, no unexpected or unusually high safety concerns were raised with these methods. Safety notes for individual synthetic procedures will be documented alongside the procedure.

A.1.2. General Information

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure to chemicals. Reagents and solvents were obtained commercially (Sigma-Aldrich, Alfa Aesar, VWR, Fischer, Matrix Scientific, Oakwood Chemical, TCI America, and other suppliers) and used without prior purification unless otherwise stated. Organic solutions were concentrated under reduced pressure on an IKA RV 10 rotary evaporator. Thin-layer chromatography (TLC) was performed on commercial Millipore Silica Gel 60 plates containing the F254 fluorescent indicator. Visualization of the developed chromatographs was performed by irradiation with UV light, or treating with and appropriate TLC staining solution (e.g., Ceric Ammonium Molybdate, $KMnO_4$, or Bromocresol Green) followed by heating if necessary. Chromatographic purification was accomplished by flash chromatography on Silacycle F60 silica gel according to the method of Still[1] or using a Biotage Isolera One instrument.

A.1.3. Spectral Data

All NMR spectra were obtained at the Caltech Liquid NMR Facility. For azetidine products, ^{1}H and ^{13}C NMR were recorded on a Bruker Prodigy 400 MHz instrument (400 MHz and 101 MHz). ^{19}F NMR spectra were recorded on a Varian 300 MHz spectrometer (282 MHz). For intermediates, ^{1}H spectra were also recorded using a Varian 300 MHz spectrometer (300 MHz), a Varian 500 MHz spectrometer (500 MHz), and a Varian 600 MHz spectrometer (600 MHz). ^{1}H and ^{13}C spectra are referred to residual $CDCl_3$ solvent signals referenced at δ 7.26 and 77.0 ppm, respectively. For spectra taken in DMSO, the residual ^{1}H and ^{13}C solvent signals are referenced at δ 2.50 and 39.51 ppm, respectively. ^{19}F spectra are referenced by addition of the appropriate internal reference standard, using either fluorobenzene (referenced at δ –113.15 ppm) or hexafluorobenzene (referenced at δ –161.90) and are clearly labeled when shown. Data for ^{1}H NMR are reported as follows: chemical shift (δ ppm), integration, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentad, sext = sextet, hept = heptet, m = multiplet, br s = broad singlet), and coupling constant (Hz). Data for ^{13}C NMR are reported in terms of chemical shift, multiplicity, and coupling constant: no special nomenclature is used for equivalent carbons. Data for ^{19}F NMR are reported in terms of chemical shift, multiplicity, and coupling constant. High-resolution mass spectra (HRMS) were obtained at Caltech Mass Spectrometry Facility.

A.1.4. Gas Chromatography Data

Gas chromatography (GC) was performed on an Agilent Technologies 7892A GC system equipped with a split-mode capillary injection system and flame-ionization detectors. For achiral analyses, an Agilent J&W HP-5 Column was used as the stationary phase. For chiral analyses, the specific stationary phase is provided along with the chiral traces.

A.1.5. Cloning, Site-Saturation Mutagenesis (SSM), and Plasmid Isolation

Electrocompetent *Escherichia coli* (*E. coli*) cells were prepared following the protocol of Sambrook and Russell.[2] Phusion polymerase and *Dpn*I were purchased from New England

Biolabs (NEB, Ipswich, MA). SSM experiments were performed using primers bearing degenerate codons (NDT, VHG, TGG) as per the "22 codon trick" using a modified QuikChange™ protocol.[3] The PCR conditions were as follows (final concentrations): Phusion HF Buffer 1x, 0.2 mM dNTPs each, 0.5 μM of forward primers, 0.5 μM reverse primer, and 0.02 U/μL of Phusion polymerase. Upon completion of PCRs, the remaining template was digested with *Dpn*I. Gel purification was performed with a Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA). The purified PCR product was then assembled using the Gibson assembly protocol.[4]

The assembly products obtained were used to transform electrocompetent *E. cloni*® EXPRESS BL21(DE3) cells (Lucigen, Middleton, WI) with a MicroPulser Electroporator (Bio-Rad, Hercules, CA). Luria-Bertani medium (LB; 0.6 mL) was added to electroporated cells and they were incubated at 37 ºC with shaking at 220 rpm for 45 minutes before being plated on LB agar plates supplemented with 100 μg/mL ampicillin (LB-amp agar plates). Plates were incubated at 37 ºC overnight. Single colonies from these plates were used to inoculate flask cultures, prepare glycerol stocks, and isolate plasmids for sequencing. Plasmids were isolated using a QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany), and the genes were sequence-verified (Laragen, Inc.).

A.1.6. 96-Well Plate Library Expression

96-well deep-well plates are shaken in an INFORS HT Multitron Shaker in all instances. Single colonies from LB-agar plates were picked using sterilized toothpicks, which were used to inoculate 400 μL of LB containing 100 μg/mL of ampicillin (LB-amp) in 2 mL 96-well deep-well plates. The plates are incubated at 37 ºC and 220 rpm overnight. For expression cultures, 50 μL of these precultures are used to inoculate 900 μL of Hyperbroth (AthenaES) with 100 μg/mL of ampicillin (HB-amp) per well in 96-well deep-well plates. The remaining overnight culture plates are sealed are stored in a 4 ºC refrigerator until needed again. The expression cultures are initially incubated at 37 ºC and 225 rpm for 2.5 hours, at which point they are allowed to cool on ice for 45 minutes. Expression of proteins was induced with isopropyl-β-D-thiogalactoside (IPTG) and cellular heme production was increased with 5-aminolevulinic acid (ALA). An induction mixture containing IPTG and

ALA in HB-amp (50 μL) was added to each well such that the final concentrations of IPTG and ALA were 0.5 mM and 1.0 mM respectively. The total culture volumes were 1 mL. The plates were then incubated at 22 ºC and 225 rpm overnight.

A.1.7. 96-Well Plate Library Reactions and Screening

Expression cultures containing *E. coli* expressing hemoproteins of interest were centrifuged at 5000 × g for 5 minutes at 4 ºC. The supernatant was discarded, and nitrogen-free M9 minimal media (M9-N, 380 μL) was added to each well. The plates were then put into a vinyl Coy anaerobic chamber (0 – 30 ppm O_2) and the pellets are resuspended. To each well were added 20 μL of a MeCN solution with 200 mM of the desired aziridine substrate and 300 mM of ethyl diazoacetate (EDA). The final reaction volume was 400 μL, and the final concentrations of the desired aziridine and EDA were 10 mM and 15 mM, respectively. The plates were then sealed carefully with a foil cover, removed from the Coy chamber, and shaken at room temperature for 16 hours. Once complete, plates were worked up for analysis by adding 600 μL of a 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (10 mM concentration). A silicone sealing mat (AWSM1003S, ArcticWhite) was used to cover the plate, and the two layers were thoroughly mixed. The plate was then centrifuged (5000 × g for 5 minutes at room temperature) to separate the phases. Afterwards, an aliquot of the organic layer was transferred to a GC vial insert in a GC vial and the samples were assayed by GC.

If wells were identified which showed improved activity over control wells containing the parent variant, these were subjected to sequence identification and validation of activity. The corresponding wells, as well as a parent control well, in the overnight culture plate were streaked out on an LB-Amp plate. Single colonies from these plates were then subjected to the small-scale protein expression conditions below (**A.1.8**), followed by the small-scale biocatalytic reaction protocol described below (**A.1.9**).

A.1.8. Small-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 6 mL of LB-amp in a 15 mL plastic culture tube. Cultures are incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 incubator. Two mL of these overnight cultures were used to inoculate 100 mL of HB-amp (1% v/v starter culture in expression culture) in 250-mL Erlenmeyer flasks. The remainder of the overnight culture was subjected to sequence identification (for new variants) and verification (for parent control wells). The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaker, at which point they were held on ice for 45 minutes. Protein expression was then induced by direct addition of 100 µL of stock solutions containing 500 mM IPTG and 1.0 mM ALA such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16 hours in an Innova 42 shaker.

A.1.9. Small-Scale Biocatalytic Reactions for Lineage Validation

The corresponding 100 mL expression cultures were pelleted (5000 × g for 5 minutes at 4 ºC) and resuspended in 6 mL of M9-N buffer. The protein concentration of this sample was determined by CO binding (*vide infra*), and 380 µL portions of whole cell suspension (WCS) were prepared in GC vials such that the protein concentration is 5.25 µM. The whole cell suspensions were put into a vinyl Coy anaerobic chamber, at which point 10 µL of a 400 mM solution of aziridine in MeCN followed by 10 µL of a 600 mM solution of EDA in MeCN were added such that the final reaction concentrations were 5.0 µM of the protein variant, 10 mM of the desired aziridine, and 15 mM of EDA. The GC vials were tightly capped with screwcaps with a septum, were brought out of the Coy chamber, and were allowed to shake at RT for 16 hours. Once complete, the reactions were transferred to a 1.7 mL Eppendorf tube and 600 µL of a 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (10 mM concentration) were added. The layers were thoroughly mixed, and the sample was centrifuged (14000 × g for 10 minutes at RT) to separate the phases. Afterwards, an aliquot of the organic layer was subjected to GC analysis and the samples are assayed by GC.

A.1.10. Large-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 25 mL of LB-amp in a 125-mL unbaffled Erlenmeyer flask. Cultures were incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 incubator. These overnight cultures (20 mL) were used to inoculate 1000 mL of HB-amp (1% v/v starter culture in expression culture) in a 2.8 L- Erlenmeyer flask. The remainder of the overnight culture was subjected to sequence identification. The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaker, at which point they were held on ice for 45 minutes. Protein expression was then induced by direct addition of 1.0 mL of stock solutions containing 500 mM IPTG and 1.0 mM ALA such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16 hours in an Innova 42 shaker.

A.1.11. Processing of Large Scale Expression Cultures for Preparative Biocatalytic Reactions

The corresponding 1L cultures were pelleted (5000 × g for 5 minutes at 4 ºC) and resuspended in 40 mL of M9-N buffer. The whole cell suspensions were held on ice until the protein concentration could be determined (*vide infra*) and reactions could be run.

A.1.12. Lysis of Whole Cell Suspensions

Whole cell suspensions from either small scale protein expression or large scale protein expression are lysed as follows. An aliquot of the whole cell suspension (3 mL) was diluted in 3 mL of M9-N buffer, and the cells were lysed by sonication on ice for 2 minutes at 25% amplitude (1 second on, 2 second off) using a QSonica Q500 Sonicator and a ½-inch tip. The sonicated cell mixture was clarified by centrifugation (14000 × g for 10 minutes at 4 ºC), siphoning off the supernatant from the cellular debris into a fresh container.

A.1.13. Protein Concentration Determination via CO-Binding Assay

The CO-binding assay was performed with clarified lysate as described above using a modified literature procedure.[5] The extinction coefficient for $\varepsilon_{410\text{-}490}$ of 0.103 M^{-1} cm^{-1} as measured for $P411_{BM3}$-CIS is used to estimate the concentration of the P411 enzymes disclosed in this work. UV-Vis spectra are taken using a Tecan SPARK instrument using untreated, 96-well flat-bottom polystyrene microplates (Evergreen Scientific, 290-8115-01F).

Clarified lyase (90 μL) with the hemoprotein of interest was diluted with 90 μL of M9-N buffer prior to the addition of 20 μL of a 300 mM solution of sodium dithionite in M9-N. The mixture was thoroughly mixed, and a UV-Vis absorbance measurement was taken at the peak maximum at 410 nm and at 490 nm as a baseline measurement. The plate was then transferred to a vacuum chamber, which was evacuated and backfilled with an atmosphere of CO. The plates were allowed to incubate under an atmosphere of CO for 30 minutes. Once incubation was complete, a second UV-Vis absorbance measurement was taken at 410 nm and at 490 nm. Beer's law was used to determine the hemoprotein concentration of the solution using the $\Delta A_{411\text{-}490}$ between the CO-bound and reduced samples, the $\varepsilon_{411\text{-}490}$ value above, and the pathlength.

A.2. Directed Evolution for Aziridine Ring Expansion

Table A-1. Sequence differences between relevant P450/P411 variants and wild-type $P450_{BM3}$.

Protein Variant	**Mutations Relative to Wild-type $P450_{BM3}$**
$P450_{BM3}$	None
$P450_{BM3}$-CIS	V78A F87V P142S T175I A184V S226R H236Q E252G T268A A290V L353V I366V E442K
$P411_{BM3}$-CIS	V78A F87V P142S T175I A184V S226R H236Q E252G T268A A290V L353V I366V C400S E442K
$P411_{BM3}$-CIS P248T I263G L437F (Parent F.2)	V78A F87V P142S T175I A184V S226R H236Q P248T E252G I263G T268A A290V L353V I366V C400S L437F E442K

Table A-2. Evolutionary trajectory of P411 variants involved in this study.

Protein Variant	**Mutations Relative to Wild-type P450$_{BM3}$** (New Mutations in Bold)
Parent F2	None
Parent F2.1	**G263Y**
Parent F2.2	G263Y **T327V**
Parent F2.3	G263Y T327V **A330T**
Parent F2.4	G263Y **H266P** T327V A330T
Parent F2.5	**M177Q** G263Y H266P T327V A330T
Parent F2.6	M177Q G263Y H266P T327V A330T **T436G**
Parent F2.7	M177Q **L233F** G263Y H266P T327V A330T T436G
Parent F2.8	**T149M** M177Q L233F G263Y H266P T327V A330T T436G
Parent F2.9	**R47Q** T149M M177Q L233F G263Y H266P T327V A330T T436G
P411-AzetS	R47Q **M118K** T149M M177Q L233F G263Y H266P T327V A330T T436G

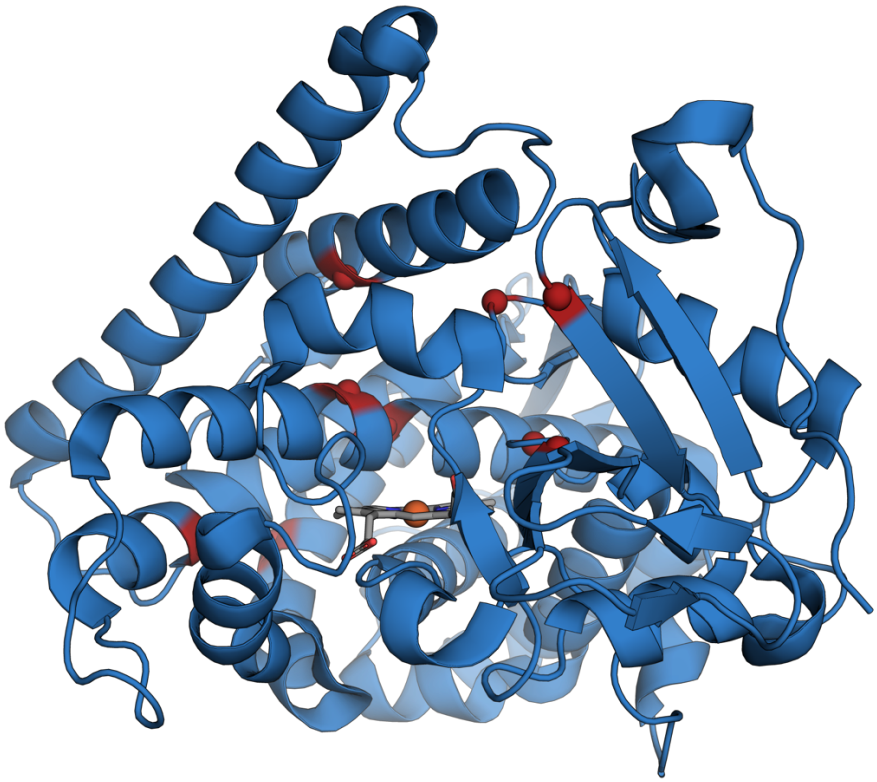


Figure A-1. Homology model of Parent F2 with mutated sites shown in red.

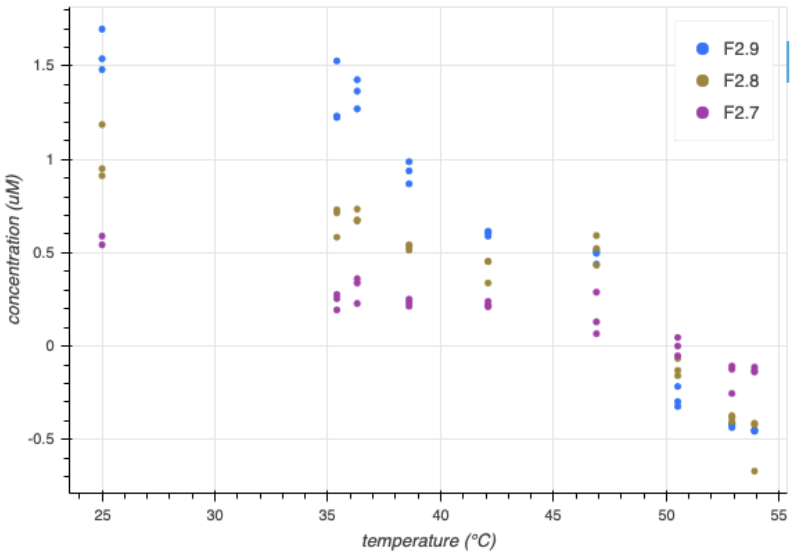


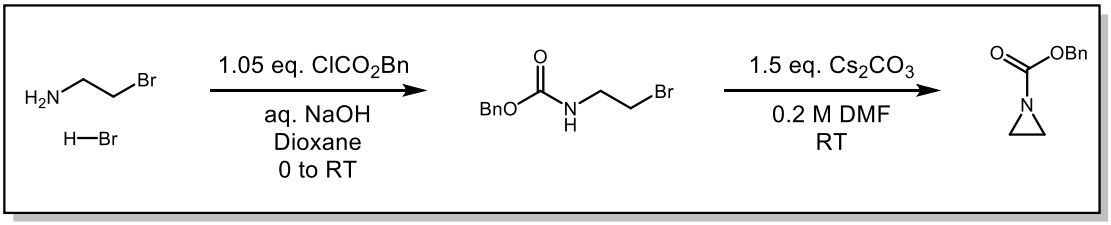
Figure A-2. Thermostability measurements of the late-stage variants en route to P411-AzetS. Stability was assessed through use of a CO-binding assay to assess the difference in concentration of heme-loaded protein (as a proxy for folded protein) in samples of purified protein incubated at varied temperatures for 10 minutes.

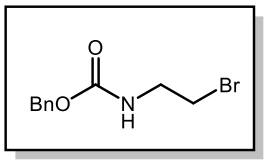
A.3. Control Experiments

Table A-3. Control experiments performed to assess potential background reactivity. Reactions were performed on 4-μmol scale as outlined in the main body of the manuscript.

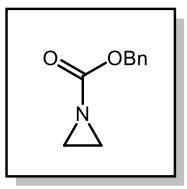
Entry	Changes to Conditions Above	Product Observed?
1	None	Yes
2	No Enzyme, 10 μM Hemin	No
3	No Enzyme, 10 μM Hemin, 300 mM sodium dithionite	No
4	No Enzyme, 1 mg/mL BSA	No
5	No Enzyme, 1 mg/mL BSA 300 mM sodium dithionite	No
6	No Enzyme, 10 μM Hemin, 1 mg/mL BSA	No
7	No Enzyme, 10 μM Hemin, 1 mg/mL BSA, 300 mM sodium dithionite	No
8	Buffer Only	No
9	Buffer with 300 mM sodium dithionite	No

A.4. Preparation of Substrates



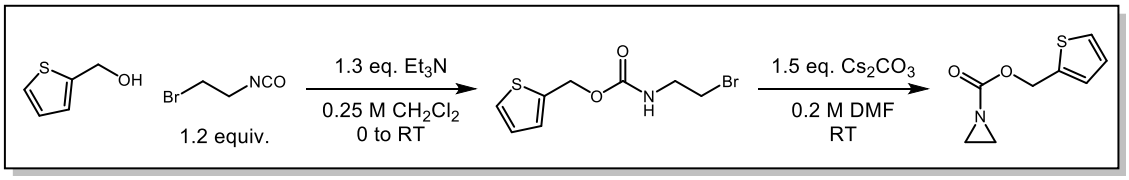


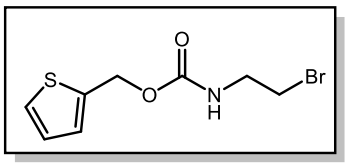
Benzyl (2-bromoethyl)carbamate. To a round-bottom flask equipped with a stir bar: 2-bromoethylamine hydrobromide (5.00 g, 24.4 mmol, 1.0 equiv.) was suspended in dioxane (24 mL). The suspension was chilled in an ice bath and 1.0 M aq. NaOH solution (30 mL) was added slowly. Once addition was complete, benzyl chloroformate (4.37 g, 3.66 mL, 25.6 mmol, 1.05 equiv.) was added dropwise. The solution was allowed to ambiently warm to room temperature with stirring overnight. Once complete, the reaction was partitioned into 150 mL of diethyl ether. The aqueous layer was drained, and the organics were washed with an additional portion of water (30 mL) and brine (30 mL) before drying over sodium sulfate. Once dry, the organics were decanted from the drying agent and the volatiles were stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 15% EtOAc in hexanes) to afford 4.65 g (74% yield) of the titled compound as a colorless oil that gradually solidifies upon standing. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.44 – 7.29 (m, 5H), 5.17 (br s, 1H), 5.12 (s, 2H), 3.62 (q, J = 5.9 Hz, 2H), 3.48 (t, J = 5.8 Hz, 2H). The spectrum obtained is in accord with prior literature reports.[6]



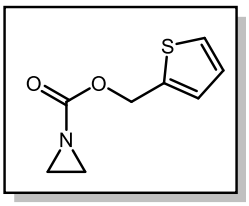
Benzyl (2-bromoethyl)carbamate. To a round-bottom flask equipped with a stir bar: benzyl (2-bromoethyl)carbamate (4.65 g, 18 mmol, 1 equiv.) was dissolved in DMF (90 mL). Cesium carbonate (8.80 g, 27 mmol, 1.5 equiv.) was added in a single portion. The reaction was allowed to stir vigorously at room temperature until judged complete by TLC analysis. Once

complete, the reaction was partitioned into 180 mL of ethyl acetate and 90 mL of water. The phases were thoroughly mixed, and the aqueous layer was drained. The organics were washed with 5% aq. LiCl (w/w) solution (3x 30 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted from the drying agent and the volatiles were stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 2.44 g (77% yield) of the titled compound as a colorless oil. ^{1}H NMR (600 MHz, $CDCl_3$) δ 7.39 – 7.31 (m, 5H), 5.14 (s, 2H), 2.23 (s, 4H). The spectrum obtained is in accord with prior literature reports.[7]

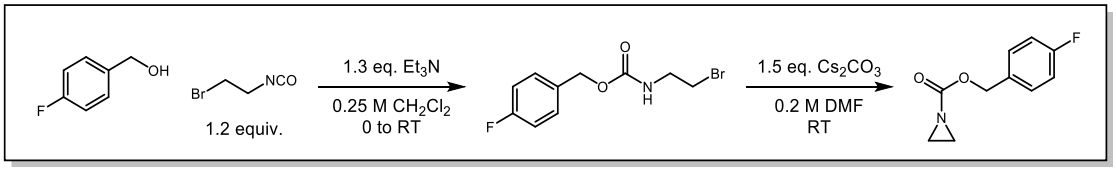


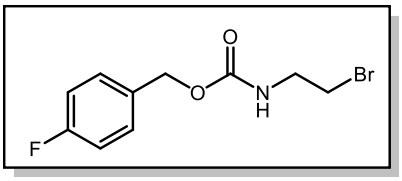


Thiophen-2-ylmethyl (2-bromoethyl)carbamate. Synthesized using the procedure of Marlin.8 Thiophen-2-ylmethanol (457 mg, 0.38 mL, 4 mmol, 1.0 equiv.) and triethylamine (526 mg, 0.73 mL, 5.2 mmol, 1.3 equiv.) were dissolved in methylene chloride (16 mL). The flask was chilled in an ice bath and 2-bromoethyl isocyanate (720 mg, 0.43 mL, 4.8 mmol, 1.2 equiv.) were added dropwise. The reaction was allowed to ambiently warm to RT with stirring and is allowed to stir until judged complete by TLC. Once complete, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from 100% hexanes to 25% EtOAc in hexanes) to afford 1.02 g (97% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (dd, J = 5.1, 1.2 Hz, 1H), 7.12 – 7.06 (m, 1H), 6.99 (dd, J = 5.1, 3.5 Hz, 1H), 5.26 (s, 2H), 5.19 (s, 1H), 3.60 (q, J = 5.9 Hz, 2H), 3.46 (t, J = 5.9 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 155.82, 138.21, 128.05, 126.83, 126.82, 61.15, 42.72, 32.35.



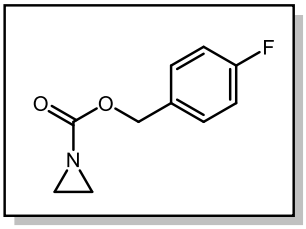
Thiophen-2-ylmethyl aziridine-1-carboxylate. Thiophen-2-ylmethyl (2-bromoethyl)carbamate (1.02 g, 3.86 mmol, 1 equiv.) was dissolved in DMF (19 mL). Cesium carbonate (1.89 g, 5.79 mmol, 1.5 equiv.) was added in a single portion and the reaction was vigorously stirred at RT until judged complete by TLC. Once finished, the reaction was partitioned into 40 mL of EtOAc and 20 mL of water. The layers were thoroughly mixed and the aqueous layer was drained. The organics were further washed with 5% aq. LiCl solution (w/w) (3x 10 mL) prior to drying the organics over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 707 mg (63% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (dd, J = 5.1, 1.2 Hz, 1H), 7.11 (ddt, J = 3.5, 1.3, 0.7 Hz, 1H), 6.98 (dd, J = 5.1, 3.5 Hz, 1H), 5.28 (d, J = 0.7 Hz, 2H), 2.22 (s, 4H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.35, 137.60, 128.33, 126.97, 126.81, 62.41, 25.88. HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_8H_{10}NO_2S$) requires *m/z* 183.0354, found *m/z* 183.0334.



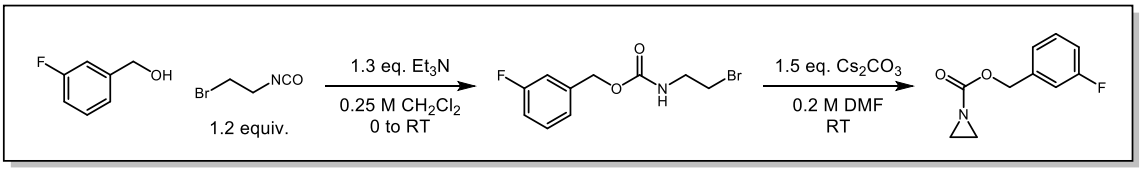


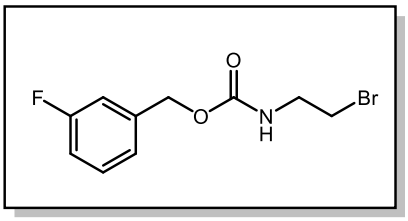
4-Fluorobenzyl (2-bromoethyl)carbamate. Synthesized using the procedure of Marlin.[8] 4-fluorobenzyl alcohol (505 mg, 0.43 mL, 4 mmol, 1.0 equiv.) and triethylamine (526 mg, 0.73 mL, 5.2 mmol, 1.3 equiv.) were dissolved in methylene chloride (16 mL). The flask was chilled in an ice bath and 2-bromoethyl isocyanate (720 mg, 0.43 mL, 4.8 mmol, 1.2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring and was allowed to stir until judged complete by TLC. Once complete, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from

100% hexanes to 25% EtOAc in hexanes) to afford 830 mg (75% yield) of the titled compound as a white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.38 – 7.30 (m, 2H), 7.09 – 6.99 (m, 2H), 5.18 (br s, 1H), 5.07 (s, 2H), 3.61 (q, *J* = 5.9 Hz, 2H), 3.47 (t, *J* = 5.8 Hz, 2H).^{13}C NMR (101 MHz, $CDCl_3$) δ 162.62 (d, *J* = 246.8 Hz), 156.00, 132.06 (d, *J* = 3.3 Hz), 130.13 (d, *J* = 8.2 Hz), 115.46 (d, *J* = 21.5 Hz), 66.24, 42.72, 32.45. ^{19}F NMR (282 MHz, $CDCl_3$) δ -113.76 (m). HRMS (ES+) exact mass calculated for $[M+H+CH_3CN]^+$ ($C_{12}H_{15}BrFN_2O_2$) requires *m/z* 317.0301, found *m/z* 317.0328.

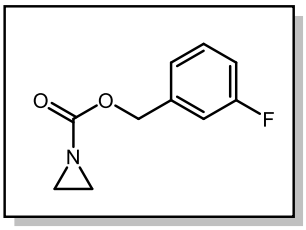


4-Fluorobenzyl aziridine-1-carboxylate. 4-fluorobenzyl (2-bromoethyl)carbamate (830 mg, 3.00 mmol, 1 equiv.) was dissolved in DMF (15 mL). Cesium carbonate (1.47 g, 4.50 mmol, 1.5 equiv.) was added in a single portion and the reaction was vigorously stirred at RT until judged complete by TLC. Once finished, the reaction was partitioned into 30 mL of EtOAc and 15 mL of water. The layers were thoroughly mixed, and the aqueous layer was drained. The organics were further washed with 5% aq. LiCl solution (w/w) (3x 10 mL) prior to drying the organics over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 414 mg (71% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 – 7.32 (m, 2H), 7.13 – 6.94 (m, 2H), 5.09 (s, 2H), 2.22 (s, 4H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.50, 162.71 (d, *J* = 247.1 Hz), 131.54 (d, *J* = 3.3 Hz), 130.25 (d, *J* = 8.3 Hz), 115.49 (d, *J* = 21.6 Hz), 67.50, 25.86. ^{19}F NMR (282 MHz, $CDCl_3$) δ -112.79 (tt, *J* = 9.0, 5.7 Hz). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{10}H_{11}FNO_2$) requires *m/z* 196.0774, found *m/z* 196.0793.



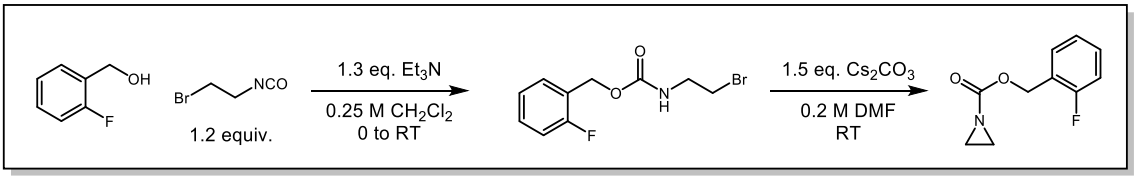


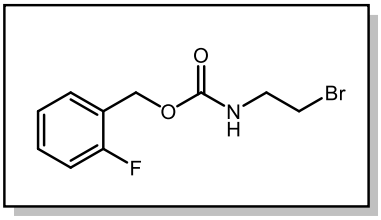
3-Fluorobenzyl (2-bromoethyl)carbamate. Synthesized using the procedure of Marlin.[8] 3-Fluorobenzyl alcohol (505 mg, 0.43 mL, 4 mmol, 1.0 equiv.) and triethylamine (526 mg, 0.73 mL, 5.2 mmol, 1.3 equiv.) were dissolved in methylene chloride (16 mL). The flask was chilled in an ice bath and 2-bromoethyl isocyanate (720 mg, 0.43 mL, 4.8 mmol, 1.2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring and was allowed to stir until judged complete by TLC. Once complete, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from 100% hexanes to 25% EtOAc in hexanes) to afford 980 mg (89% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (td, *J* = 8.0, 5.8 Hz, 1H), 7.15 – 7.08 (m, 2H), 7.06 (dt, *J* = 9.5, 2.1 Hz, 1H), 7.04 – 6.97 (m, 1H), 5.24 (s, 1H), 5.10 (s, 2H), 3.61 (q, *J* = 5.9 Hz, 2H), 3.47 (t, *J* = 5.8 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 162.81 (d, *J* = 246.3 Hz), 155.89, 138.76 (d, *J* = 7.3 Hz), 130.09 (d, *J* = 8.1 Hz), 123.34 (d, *J* = 3.0 Hz), 115.04 (d, *J* = 21.1 Hz), 114.74 (d, *J* = 21.9 Hz), 66.03, 42.74, 32.38. ^{19}F NMR (282 MHz, $CDCl_3$) δ -113.02 (td, *J* = 9.0, 5.7 Hz). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{10}H_{12}BrFNO_2$) requires *m/z* 276.0035, found *m/z* 276.0047.



3-Fluorobenzyl aziridine-1-carboxylate. 3-Fluorobenzyl (2-bromoethyl)carbamate (980 mg, 3.55 mmol, 1 equiv.) was dissolved in DMF (18 mL). Cesium carbonate (1.73 g, 5.33 mmol, 1.5 equiv.) was added in a single portion and the reaction was vigorously stirred at RT until judged complete by TLC. Once finished, the reaction was partitioned into 30 mL of EtOAc and 15 mL of water. The

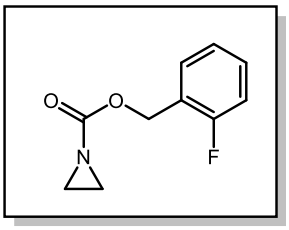
layers were thoroughly mixed, and the aqueous layer was drained. The organics were further washed with 5% aq. LiCl solution (w/w) (3x 10 mL) prior to drying the organics over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 525 mg (76% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (td, *J* = 7.9, 5.8 Hz, 1H), 7.13 (ddt, *J* = 7.6, 1.6, 0.8 Hz, 1H), 7.08 (dt, *J* = 9.6, 2.1 Hz, 1H), 7.05 – 6.98 (m, 1H), 5.12 (s, 2H), 2.24 (s, 4H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.38, 162.80 (d, *J* = 246.5 Hz), 138.17 (d, *J* = 7.4 Hz), 130.13 (d, *J* = 8.1 Hz), 123.46 (d, *J* = 3.0 Hz), 115.20 (d, *J* = 21.0 Hz), 114.85 (d, *J* = 22.0 Hz), 67.26 (d, *J* = 1.9 Hz), 25.88. ^{19}F NMR (282 MHz, $CDCl_3$) δ -112.94 (td, *J* = 9.0, 5.8 Hz). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{10}H_{11}FNO_2$) requires *m/z* 196.0774, found *m/z* 196.0748.



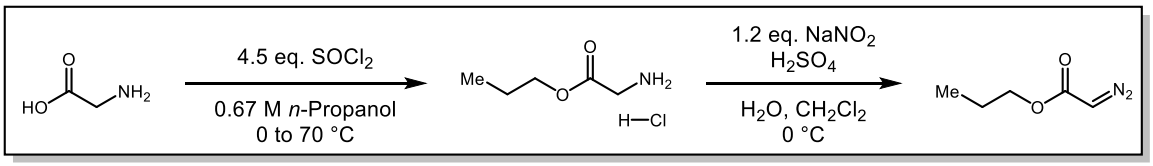


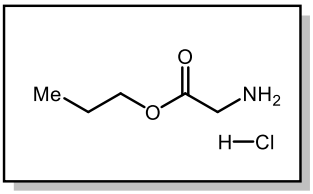
2-Fluorobenzyl (2-bromoethyl)carbamate. Synthesized using the procedure of Marlin.[8] 2-Fluorobenzyl alcohol (505 mg, 0.43 mL, 4 mmol, 1.0 equiv.) and triethylamine (526 mg, 0.73 mL, 5.2 mmol, 1.3 equiv.) was dissolved in methylene chloride (16 mL). The flask was chilled in an ice bath and 2-bromoethyl isocyanate (720 mg, 0.43 mL, 4.8 mmol, 1.2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring and was allowed to stir until judged complete by TLC. Once complete, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from 100% hexanes to 25% EtOAc in hexanes) to afford 970 mg (88% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 (td, *J* = 7.5, 1.8 Hz, 1H), 7.32 (tdd, *J* = 7.4, 5.3, 1.8 Hz, 1H), 7.14 (td, *J* = 7.5, 1.2 Hz, 1H), 7.07 (ddd, *J* = 9.7, 8.2,

1.2 Hz, 1H), 5.19 (s, 2H), 5.19 (br s, 1H -- overlaps with benzylic protons), 3.61 (q, J = 5.9 Hz, 2H), 3.47 (t, J = 5.9 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 160.97 (d, J = 248.5 Hz), 155.95, 130.59 (d, J = 3.9 Hz), 130.21 (d, J = 8.2 Hz), 124.14 (d, J = 3.7 Hz), 123.38 (d, J = 14.4 Hz), 115.46 (d, J = 21.2 Hz), 60.86 (d, J = 4.3 Hz), 42.75, 32.42. ^{19}F NMR (282 MHz, $CDCl_3$) δ -118.37 (m). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{10}H_{12}BrFNO_2$) requires *m/z* 276.0035, found *m/z* 276.0042.

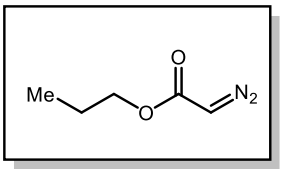


2-Fluorobenzyl aziridine-1-carboxylate. 2-Fluorobenzyl (2-bromoethyl)carbamate (970 mg, 3.51 mmol, 1 equiv.) was dissolved in DMF (18 mL). Cesium carbonate (1.72 g, 5.27 mmol, 1.5 equiv.) was added in a single portion and the reaction was vigorously stirred at RT until judged complete by TLC. Once finished, the reaction was partitioned into 30 mL of EtOAc and 15 mL of water. The layers were thoroughly mixed, and the aqueous layer was drained. The organics were further washed with 5% aq. LiCl solution (w/w) (3x 10 mL) prior to drying the organics over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 495 mg (72% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 (td, J = 7.5, 1.9 Hz, 1H), 7.33 (tdd, J = 7.5, 5.3, 1.8 Hz, 1H), 7.14 (td, J = 7.5, 1.2 Hz, 1H), 7.07 (ddd, J = 9.6, 8.2, 1.1 Hz, 1H), 5.20 (s, 2H), 2.24 (s, 4H).^{13}C NMR (101 MHz, $CDCl_3$) δ 163.47, 160.98 (d, J = 248.5 Hz), 130.57 (d, J = 3.7 Hz), 130.34 (d, J = 8.2 Hz), 124.15 (d, J = 3.7 Hz), 122.87 (d, J = 14.6 Hz), 115.48 (d, J = 21.2 Hz), 62.21 (d, J = 4.2 Hz), 25.90. ^{19}F NMR (282 MHz, $CDCl_3$) δ -118.17 (m). HRMS (ES+)



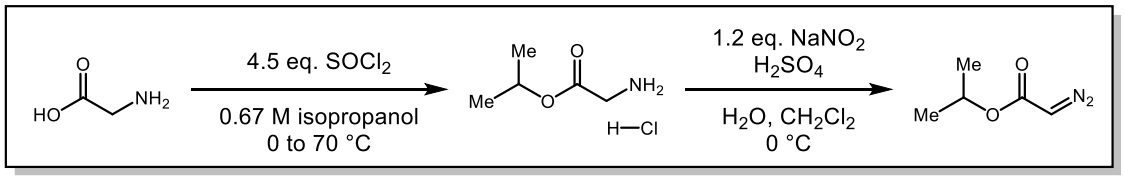


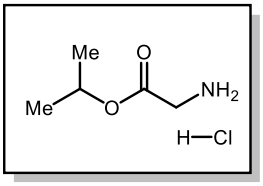
Propyl glycinate hydrochloride. To a 100-mL round-bottom flask equipped with a stir bar, septum, and drying tube: glycine (1.50 g, 20 mmol, 1.0 equiv.) was suspended in *n*-propanol (30 mL). The suspension was chilled in an ice bath with stirring prior to the dropwise addition of thionyl chloride (10.7 g, 6.56 mL, 90 mmol, 4.5 equiv.). Once addition was complete and the initial exotherm had subsided, the septum with drying tube was replaced with a reflux condenser with a drying tube and the solution was heated to 70 °C with stirring for 20 hours. When done, the reaction was allowed to cool to RT and the volatiles were stripped under vacuum. The product was precipitated by blanketing the crude oil with about 10 mL of diethyl ether and gently scratching to induce precipitation. The resultant solid was isolated by vacuum filtration, washing liberally with additional diethyl ether, and was thoroughly dried under vacuum to afford 2.42 g (79% isolated yield) of the titled compound as a white solid. ^{1}H NMR (400 MHz, DMSO) δ 8.75 – 8.38 (br s, 3H), 4.10 (t, J = 6.6 Hz, 2H), 3.77 (s, 2H), 1.62 (h, J = 7.4 Hz, 2H), 0.90 (t, J = 7.4 Hz, 3H). ^{13}C NMR (101 MHz, DMSO-*d*6) δ 167.65, 66.82, 39.10 (obscured by DMSO-*d*6; assigned via HSQC), 21.42, 10.17. HRMS (ES+) exact mass calculated for $[M+H+CH_3CN]^+$ ($C_{12}H_{15}BrFN_2O_2$) requires *m/z* 317.0301, found *m/z* 317.0328. HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_5H_{12}NO_2$) requires *m/z* 118.0868, found *m/z* 118.0898.



Propyl 2-diazoacetae (propyl diazoacetate). This procedure was adapted from that of Searle for the synthesis of ethyl diazoacetate.[9] <u>*Safety note:*</u> the procedure reported by Searle was performed on a 1-mole scale: our procedure is a hundred-fold reduction in scale as a matter of safety. Work should be performed in a well-ventilated fume hood.

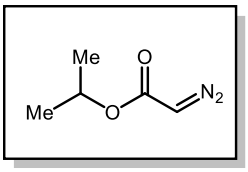
To a 40-mL vial: propyl glycinate hydrochloride (1.54 g, 10 mmol, 1 equiv.) was added to a suspension of methylene chloride (2.5 mL) in water (6 mL). The solution was chilled in an ice bath prior to the addition of sodium nitrite (828 mg, 12 mmol, 1.2 equiv.) in a single portion. Once dissolved, a single drop of concentrated sulfuric acid was added, resulting in bubbling of the solution and the development of a yellow color. The solution was allowed to stir for 30 minutes on ice. Once complete, the reaction was transferred to a separatory funnel and diluted with 20 mL of methylene chloride. The aqueous layer was separated, and the organics were washed with saturated sodium bicarbonate (2x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles were stripped under vacuum, going no lower than 130 torr and spinning the flask in an ice-water bath. The resulting residue was purified by silica gel column chromatography (gradient from pentanes to 10% diethyl ether in pentanes, concentrating fractions as before) to afford 1.27 g (99% yield) of the titled compound as a yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 4.73 (s, 1H), 4.12 (t, J = 6.7 Hz, 2H), 1.67 (h, J = 7.4 Hz, 1H), 0.94 (t, J = 7.4 Hz, 3H). The chemical shifts are in accord with prior literature reports.[10]





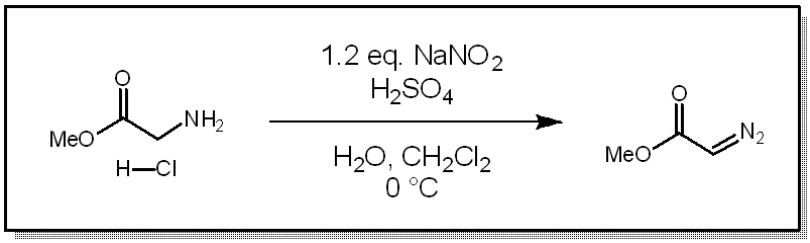
Isopropyl glycinate hydrochloride. To a 100-mL round-bottom flask equipped with a stir bar, septum, and drying tube: glycine (1.50 g, 20 mmol, 1.0 equiv.) was suspended in isopropanol (30 mL). The suspension was chilled in an ice bath with stirring prior to the dropwise addition of thionyl chloride (10.7 g, 6.56 mL, 90 mmol, 4.5 equiv.). Once addition was complete and the initial exotherm had subsided, the septum with drying tube was replaced with a reflux condenser with a drying tube and the solution was heated to 70 °C with stirring for 20 hours. When done, the reaction was allowed to cool to RT and the volatiles were stripped under vacuum. The product was precipitated by blanketing the

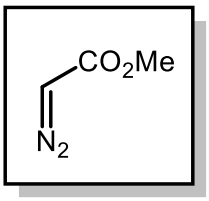
crude oil with about 10 mL of diethyl ether and gently scratching to induce precipitation. The resultant solid was isolated by vacuum filtration, washing liberally with additional diethyl ether, and was thoroughly dried under vacuum to afford 2.33 g (76% isolated yield) of the titled compound as a white solid. ^{1}H NMR (400 MHz, DMSO) δ 8.53 (br s, 3H), 4.99 (hept, J = 6.3 Hz, 1H), 3.71 (q, J = 6.2 Hz, 2H), 1.23 (d, J = 6.3 Hz, 6H). The chemical shifts are in accord with prior literature reports.[11]



Isopropyl 2-diazoacetae (isopropyl diazoacetate). This procedure was adapted from that of Searle for the synthesis of ethyl diazoacetate.[9] *Safety note:* the procedure reported by Searle was performed on a 1-mole scale: our procedure is a hundred-fold reduction in scale as a matter of safety. Work should be performed in a well-ventilated fume hood.

To a 40-mL vial: isopropyl glycinate hydrochloride (1.54 g, 10 mmol, 1 equiv.) was added to a suspension of methylene chloride (2.5 mL) in water (6 mL). The solution was chilled in an ice bath prior to the addition of sodium nitrite (828 mg, 12 mmol, 1.2 equiv.) in a single portion. Once dissolved, a single drop of concentrated sulfuric acid was added, resulting in bubbling of the solution and the development of a yellow color. The solution was allowed to stir for 30 minutes on ice. Once complete, the reaction was transferred to a separatory funnel and diluted with 20 mL of methylene chloride. The aqueous layer was separated, and the organics were washed with saturated sodium bicarbonate (2x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles were stripped under vacuum, going no lower than 130 torr and spinning the flask in an ice-water bath. The resulting residue was purified by silica gel column chromatography (gradient from pentanes to 10% diethyl ether in pentanes) to afford 738 mg (58% yield) of the titled compound as a yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.09 (hept, J = 6.3 Hz, 1H), 4.69 (s, 1H), 1.26 (d, J = 6.3 Hz, 6H). The chemical shifts are in accord with prior literature reports.[10]



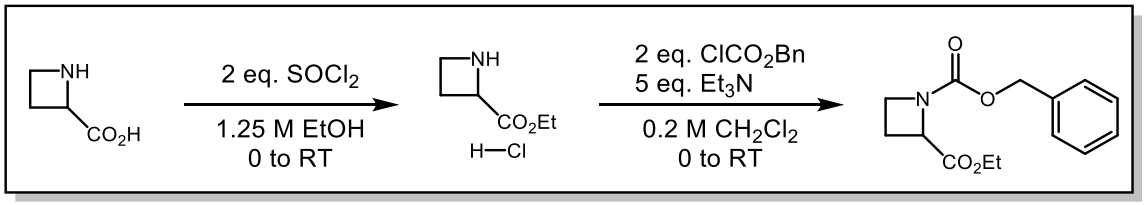


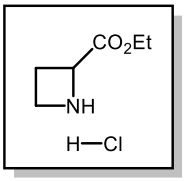
Methyl 2-diazoacetae (Methyl diazoacetate).

This procedure was adapted from that of Searle for the synthesis of ethyl diazoacetate.[9] *Safety note:* the procedure reported by Searle was performed on a 1-mole scale: our procedure is a hundred-fold reduction in scale as a matter of safety. Work should be performed in a well-ventilated fumehood.

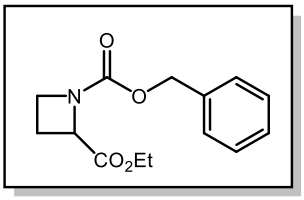
To a 40-mL vial: glycine methyl ester hydrochloride (1.26 g, 10 mmol, 1 equiv.) was added to a suspension of methylene chloride (2.5 mL) in water (6 mL). The solution was chilled on an ice bath prior to the addition of sodium nitrite (828 mg, 12 mmol, 1.2 equiv.) in a single portion. Once dissolved, a single drop of concentrated sulfuric acid was added, resulting in bubbling of the solution and the development of a yellow color. The solution was allowed to stir for 30 minutes on ice. Once complete, the reaction was transferred to a separatory funnel and diluted with 20 mL of methylene chloride. The aqueous layer was separated, and the organics were washed with saturated sodium bicarbonate (2x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles were stripped under vacuum, going no lower than 200 torr and spinning the flask in an ice-water bath. The resulting residue was purified by silica gel column chromatography (gradient from pentanes to 10% diethyl ether in pentanes) to afford 584 mg (58% yield) of the titled compound as a yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 4.75 (s, 1H), 3.76 (s, 3). The chemical shifts are in accord with prior literature reports.[11]

A.5. Preparation of Authentic Standards



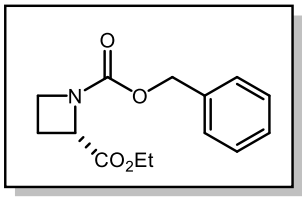


(±)-Ethyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in ethanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.

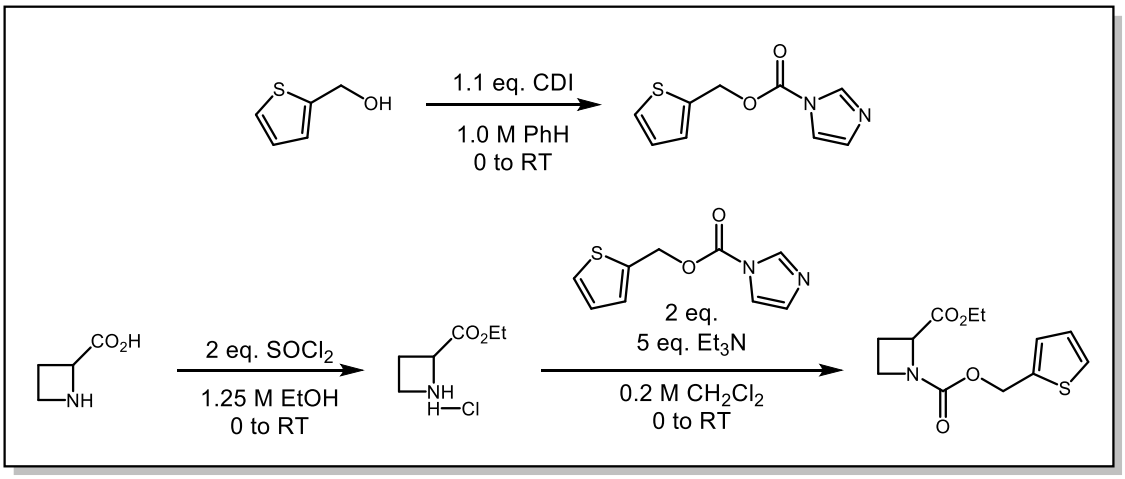


(±)-1-Benzyl 2-ethyl azetidine-1,2-dicarboxylate. The crude (±)-ethyl azetidine-2-carboxylate hydrochloride (165.62 mg, 1 mmol, 1 equiv.) synthesized above is suspended in methylene chloride (5 mL) and the dram vial is equipped with a stir bar, a septum, and a drying tube. The vial is chilled in an ice bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition is complete, benzyl chloroformate (0.29 mL, 2 mmol, 2 equiv.) is added dropwise. The reaction is allowed to warm to RT and stirring is continued until the reaction is judged complete by TLC analysis. Once complete, the reaction is transferred to a separatory funnel and quenched with 5 mL of saturated sodium bicarbonate solution. The layers are thoroughly mixed, and the aqueous layer is removed. The organic layer is washed with water (1x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics are decanted and the

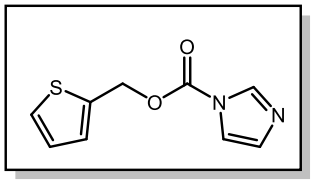
volatiles are removed under vacuum. The crude product is purified by silica gel column chromatography (gradient from hexanes to 33% EtOAc in hexanes) to afford 165 mg (63% yield) of the titled compound a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.39 – 7.27 (m, 5H), 5.10 (m, 2H), 4.68 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.19 (br s, 2H), 4.12 (m, 1H), 3.97 (m, 1H), 2.57 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.22 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.30 – 1.17 (br m, 3H).



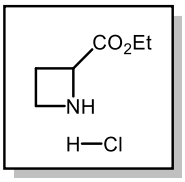
1-Benzyl 2-ethyl (*S*)-azetidine-1,2-dicarboxylate. This compound was synthesized using the analogous procedure to the racemate, except *(S)*-azetidine-2-carboxylic acid was used instead of racemic azetidine-2-carboxylic acid. This procedure yielded 82 mg (32% isolated yield) of the titled compound as a colorless oil over two steps. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.39 – 7.27 (m, 5H), 5.10 (m, 2H), 4.68 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.19 (br s, 2H), 4.12 (m, 1H), 3.97 (m, 1H), 2.57 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.22 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.30 – 1.17 (br m, 3H).



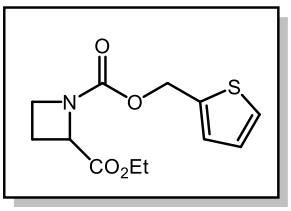
Thiophen-2-ylmethyl 1H-imidazole-1-carboxylate. Synthesized using the procedure reported by Snaddon.[12] To an oven-dried 8-mL dram vial equipped with a stir bar, septum,



and drying tube: 2-thiophenemethanol (343 mg, 0.28 mL, 3 mmol, 1 equiv.) was dissolved in benzene (3 mL) and the solution was chilled in an ice bath. Carbonyldiimidazole (CDI) (503 mg, 3.3 mmol, 1.1 equiv.) was added by quickly removing the septum and pouring in the solid reagent in a single portion. Once the initial exotherm subsided, the reaction was allowed to stir at RT until judged complete by TLC. Once finished, a small amount of methylene chloride was added to fully dissolve all solids. The reaction mixture was washed with water (2x 1.5 mL) and brine (1.5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted, and the volatiles were removed under vacuum to afford 500 mg (80% yield) of the titled compound as a colorless oil. The purity of the crude was sufficient to carry forward without further purification. ^{1}H NMR (600 MHz, $CDCl_3$) δ 8.15 (2, 1H), 7.43 (t, J = 1.4 Hz, 1H), 7.41 (dd, J = 5.1, 1.3 Hz, 1H), 7.23 (dd, J = 3.3, 1.3 Hz, 1H), 7.06 (m, 1H), 7.04 (m, 1H), 5.58 (s, 2H). The spectral data is consistent with prior literature reports.[13]

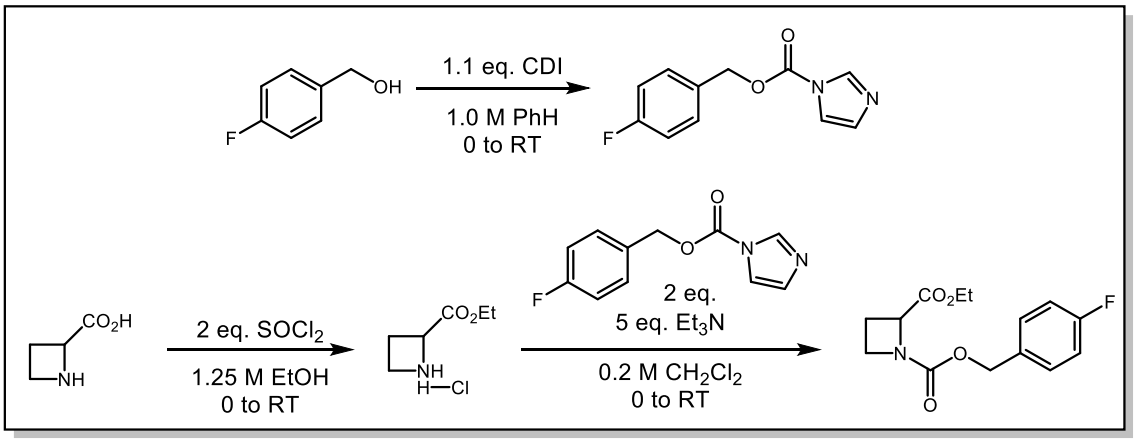


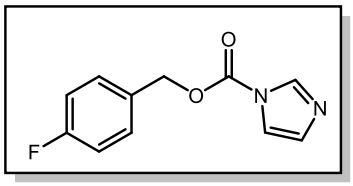
(±)-Ethyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in ethanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



(±)-2-Ethyl-1-(thiophen-2-ylmethyl) azetidine-1,2-dicarboxylate. The crude (±)-ethyl azetidine-2-carboxylate hydrochloride (165.62 mg, 1 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (3 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice

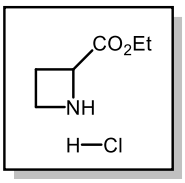
bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition was complete, thiophen-2-ylmethyl 1H-imidazole-1-carboxylate (416 mg, 2 mmol, 2 equiv.) was added to the mixture in 2 mL of methylene chloride. The reaction was allowed to warm to RT and stirring was continued until the reaction is judged complete by TLC analysis. When done, the reaction was transferred to a separatory funnel and diluted with 5 mL of methylene chloride. The reaction is washed with 1.0 M HCl (5 mL), sat. aq. $NaHCO_3$ (5 mL), and brine (5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude residue was purified by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes) to afford 93 mg (35% yield) of the titled compound as a colorless oil. ^{1}H NMR (600 MHz, $CDCl_3$) δ 7.29 (d, J = 5.1 Hz, 1H), 7.09 – 7.04 (m, 1H), 6.96 (t, J = 4.3 Hz, 1H), 5.21 (m, 2H), 4.66 (dd, J = 9.4, 5.3 Hz, 1H), 4.20 (br s, 2H), 4.16 – 4.06 (m, 1H), 3.96 (m, 1H), 2.63 – 2.48 (m, 1H), 2.20 (ddt, J = 10.8, 8.4, 5.5 Hz, 1H), 1.32 – 1.12 (m, 4H).



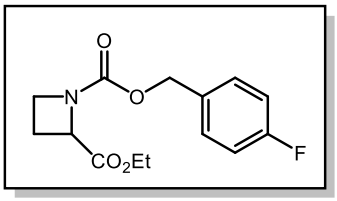


4-Fluorobenzyl 1H-imidazole-1-carboxylate. Synthesized using the procedure reported by Snaddon.[12] To an oven-dried 8 mL dram vial equipped with a stir bar, septum, and drying tube: 4-fluorobenzyl alcohol (378 mg, 0.33 mL, 3 mmol, 1 equiv.) is dissolved in benzene (3 mL) and the solution is chilled in an ice bath. Carbonyldiimidazole (CDI) (503 mg, 3.3 mmol, 1.1 equiv.) is added by quickly removing

the septum and pouring in the solid reagent in a single portion. Once the initial exotherm subsided, the reaction is allowed to stir at RT until judged complete by TLC. Once finished, a small amount of methylene chloride is added to fully dissolve all solids. The reaction mixture is washed with water (2x 1.5 mL) and brine (1.5 mL) prior to drying over sodium sulfate. Once dry, the organics are decanted and the volatiles are removed under vacuum to afford 560 mg (85% yield) of the titled compound as a colorless oil. The purity of the crude is sufficient to carry forward without further purification. ^{1}H NMR (600 MHz, $CDCl_3$) δ 8.13 (s, 1H), 7.47 – 7.41 (m, 3H), 7.14 – 7.01 (m, 3H), 5.38 (s, 2H). The spectral data is consistent with prior literature reports.[14]

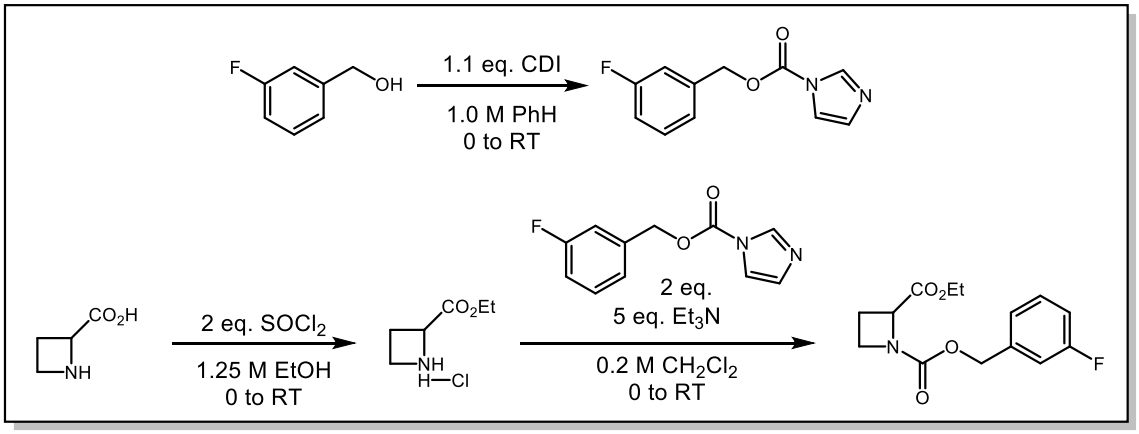


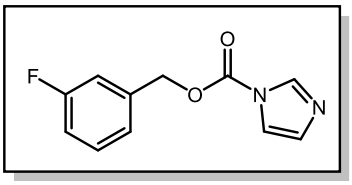
(±)-Ethyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in ethanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



(±)-2-ethyl-1-(4-fluorobenzyl) azetidine-1,2-dicarboxylate. The crude (±)-ethyl azetidine-2-carboxylate hydrochloride (165.62 mg, 1 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (3 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition was complete, 4-fluorobenzyl 1H-imidazole-1-carboxylate (440 mg, 2 mmol, 2 equiv.) was added to the mixture in 2 mL of methylene chloride. The reaction was allowed to warm to RT and stirring is continued until the reaction was judged complete by TLC analysis. When done, the reaction was transferred to a separatory funnel and diluted with 5 mL of

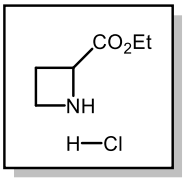
methylene chloride. The reaction was washed with 1.0 M HCl (5 mL), sat. aq. $NaHCO_3$ (5 mL), and brine (5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude residue was purified by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes) to afford 71 mg (25% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 (m, 2H), 7.08 – 6.97 (m, 2H), 5.14 – 4.96 (m, 2H), 4.74 – 4.61 (m, 1H), 4.18 (br s, 2H), 4.10 (m, 1H), 4.02 – 3.86 (m, 1H), 2.56 (dddd, *J* = 11.5, 10.5, 8.5, 6.3 Hz, 1H), 2.29 – 2.13 (m, 1H), 1.33 – 1.13 (br s, 3H).



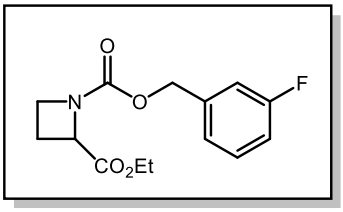


3-Fluorobenzyl 1H-imidazole-1-carboxylate. Synthesized using the procedure reported by Snaddon.[12] To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: 3-fluorobenzyl alcohol (378 mg, 0.33 mL, 3 mmol, 1 equiv.) was dissolved in benzene (3 mL) and the solution was chilled in an ice bath. Carbonyldiimidazole (CDI) (503 mg, 3.3 mmol, 1.1 equiv.) was added by quickly removing the septum and pouring in the solid reagent in a single portion. Once the initial exotherm subsided, the reaction was allowed to stir at RT until judged complete by TLC. Once finished, a small amount of methylene chloride was added to fully dissolve all solids and the mixture was diluted in 5 mL of EtOAc. The reaction mixture was washed with water (2x 1.5 mL) and brine (1.5 mL) prior to drying over sodium sulfate. Once dry, the

organics were decanted, and the volatiles removed under vacuum to afford 660 mg (quantiative yield) of the titled compound as a colorless oil. The purity of the crude was sufficient to carry forward without further purification. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.15 (t, *J* = 1.1 Hz, 1H), 7.44 (t, *J* = 1.5 Hz, 1H), 7.39 (td, *J* = 8.0, 5.8 Hz, 1H), 7.32 (m, 1H), 7.22 (m, 1H), 7.18 – 7.05 (m, 3H), 5.40 (s, 2H). The spectral data is consistent with prior literature reports.[15]

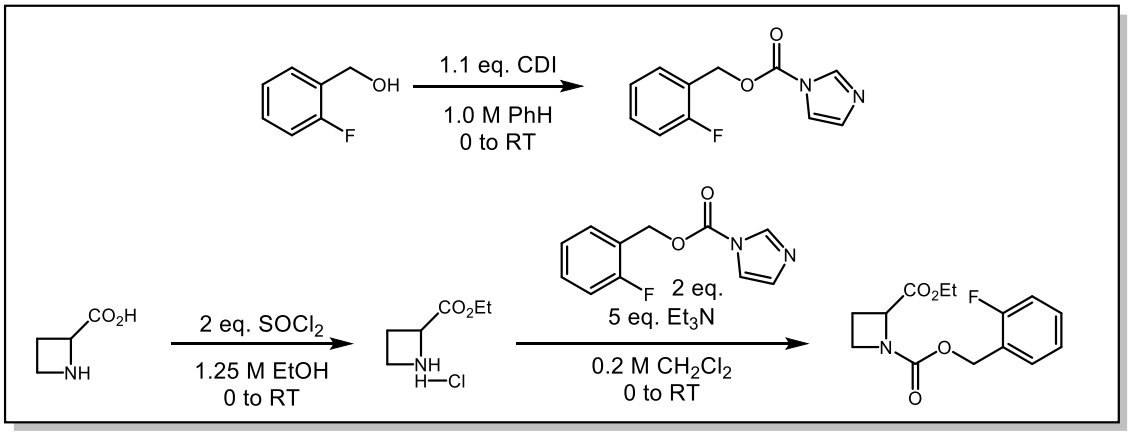


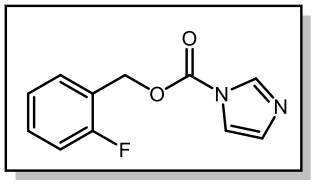
(±)-Ethyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in ethanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



(±)-2-Ethyl-1-(3-fluorobenzyl) azetidine-1,2-dicarboxylate. The crude (±)-ethyl azetidine-2-carboxylate hydrochloride (165.62 mg, 1 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (3 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition was complete, 3-fluorobenzyl 1H-imidazole-1-carboxylate (440 mg, 2 mmol, 2 equiv.) was added to the mixture in 2 mL of methylene chloride. The reaction was allowed to warm to RT and stirring was continued until the reaction was judged complete by TLC analysis. When done, the reaction was transferred to a separatory funnel and diluted with 5 mL of methylene chloride. The reaction was washed with 1.0 M HCl (5 mL), sat. aq. $NaHCO_3$ (5 mL), and brine (5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude residue was

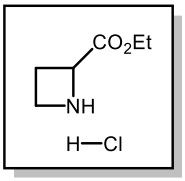
purified by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes) to afford 104 mg (37% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.30 (td, *J* = 7.9, 5.8 Hz, 1H), 7.13 – 6.94 (m, 3H), 5.22 – 4.98 (m, 2H), 4.70 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.27 – 4.17 (m, 2H), 4.16 – 4.08 (m, 1H), 3.99 (ddd, *J* = 9.1, 8.1, 5.6 Hz, 1H), 2.59 (dtd, *J* = 11.6, 9.2, 6.3 Hz, 1H), 2.23 (ddt, *J* = 11.4, 9.1, 5.5 Hz, 1H), 1.30 – 1.18 (m, 3H).



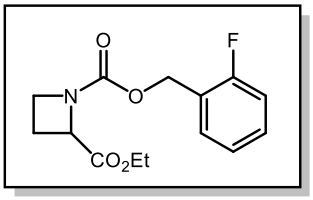


2-Fluorobenzyl 1H-imidazole-1-carboxylate. Synthesized using the procedure reported by Snaddon.[12] To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: 2-fluorobenzyl alcohol (378 mg, 0.32 mL, 3 mmol, 1 equiv.) was dissolved in benzene (3 mL) and the solution was chilled in an ice bath. Carbonyldiimidazole (CDI) (503 mg, 3.3 mmol, 1.1 equiv.) was added by quickly removing the septum and pouring in the solid reagent in a single portion. Once the initial exotherm subsided, the reaction was allowed to stir at RT until judged complete by TLC. Once finished, a small amount of methylene chloride was added to fully dissolve all solids and the mixture is diluted in 5 mL of EtOAc. The reaction mixture was washed with water (2x 1.5 mL) and brine (1.5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and the volatiles were removed under vacuum to afford 580 mg (88% yield) of the titled compound as a colorless oil. The purity of the crude was sufficient to carry

forward without further purification. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.15 (t, *J* = 1.1 Hz, 1H), 7.50 – 7.37 (m, 3H), 7.23 – 7.09 (m, 2H), 7.06 (dd, *J* = 1.7, 0.8 Hz, 1H), 5.49 (s, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 161.25 (d, *J* = 249.9 Hz), 148.49, 137.13, 131.39 (d, *J* = 8.3 Hz), 131.15 (d, *J* = 3.3 Hz), 130.63, 124.42 (d, *J* = 3.7 Hz), 121.22 (d, *J* = 14.5 Hz), 117.15, 115.83 (d, *J* = 20.9 Hz), 63.80 (d, *J* = 4.2 Hz). ^{19}F NMR (282 MHz, $CDCl_3$) δ -117.51 (m). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{11}H_{10}FN_2O_2$) requires *m/z* 211.0726, found *m/z* 211.0726.

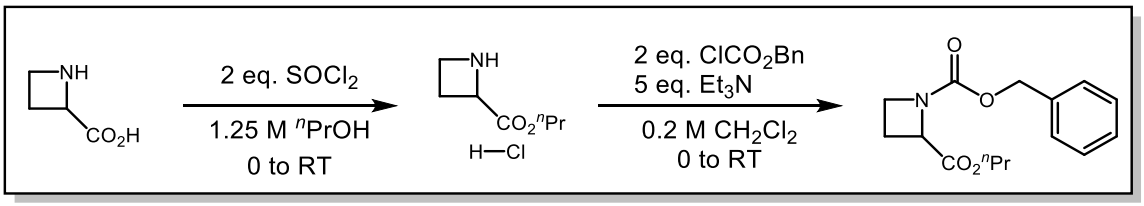


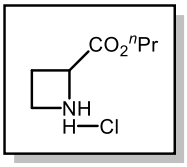
(±)-Ethyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in ethanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



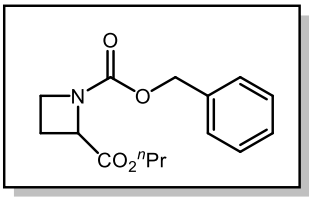
(±)-2-Ethyl-1-(2-fluorobenzyl) azetidine-1,2-dicarboxylate. The crude (±)-ethyl azetidine-2-carboxylate hydrochloride (165.62 mg, 1 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (3 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition was complete, 2-fluorobenzyl 1H-imidazole-1-carboxylate (440 mg, 2 mmol, 2 equiv.) was added to the mixture in 2 mL of methylene chloride. The reaction was allowed to warm to RT and stirring was continued until the reaction was judged complete by TLC analysis. When done, the reaction was transferred to a separatory funnel and diluted with 5 mL of methylene chloride. The reaction was washed with 1.0 M HCl (5 mL), sat. aq. $NaHCO_3$ (5 mL), and brine (5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted and

the volatiles stripped under vacuum. The crude residue was purified by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes) to afford 102 mg (36% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.38 (m, 1H), 7.29 (m, 1H), 7.12 (m, 1H), 7.04 (m, 1H), 5.17 (q, J = 12.7 Hz, 2H), 4.68 (dd, J = 9.3, 5.3 Hz, 1H), 4.19 (m, 2H), 4.11 (td, J = 8.6, 6.3 Hz, 1H), 3.97 (td, J = 8.6, 8.2, 5.7 Hz, 1H), 2.57 (dtd, J = 11.6, 9.2, 6.3 Hz, 1H), 2.21 (ddt, J = 11.3, 9.0, 5.5 Hz, 1H), 1.24 (m, 3H).



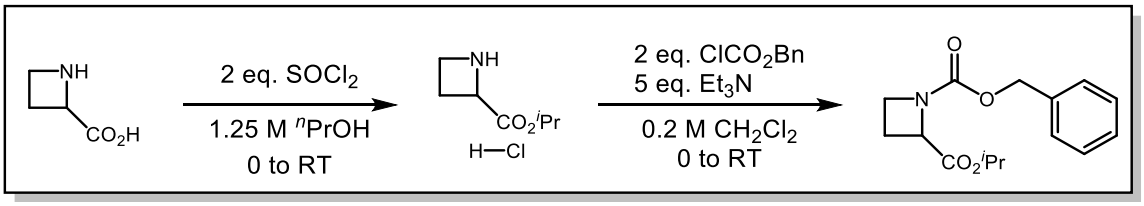


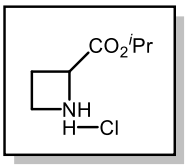
(±)-Propyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (50 mg, 1 mmol, 1 equiv.) was suspended in *n*-propanol (0.6 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.07 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the brown solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



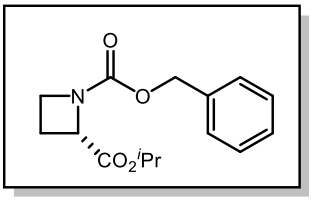
(±)-1-Benzyl 2-propyl azetidine-1,2-dicarboxylate. The crude (±)-propyl azetidine-2-carboxylate hydrochloride (89.9 mg, 0.5 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (2.5 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.35 mL, 2.5 mmol, 5 equiv.). Once addition was complete, benzyl

chloroformate (0.14 mL, 1 mmol, 2 equiv.) was added dropwise. The reaction was allowed to warm to RT and stirring was continued until the reaction is judged complete by TLC analysis. Once complete, the reaction was transferred to a separatory funnel and quenched with 5 mL of saturated sodium bicarbonate solution. The layers were thoroughly mixed, and the aqueous layer was removed. The organic layer was washed with water (1x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted, and the volatiles were removed under vacuum. The crude product was purified by silica gel column chromatography (gradient from hexanes to 33% EtOAc in hexanes) to afford 25 mg (18% yield) of the titled compound a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 – 7.27 (m, 5H), 5.17 – 5.04 (m, 2H), 4.69 (dd, J = 9.3, 5.2 Hz, 1H), 4.20 – 4.06 (m, 3H), 3.97 (ddd, J = 9.1, 8.1, 5.6 Hz, 1H), 2.57 (dtd, J = 11.5, 9.2, 6.3 Hz, 1H), 2.21 (ddt, J = 11.3, 8.9, 5.5 Hz, 1H), 1.64 (m, 2H – overlaps with HOD peak), 0.91 (m, 3H).

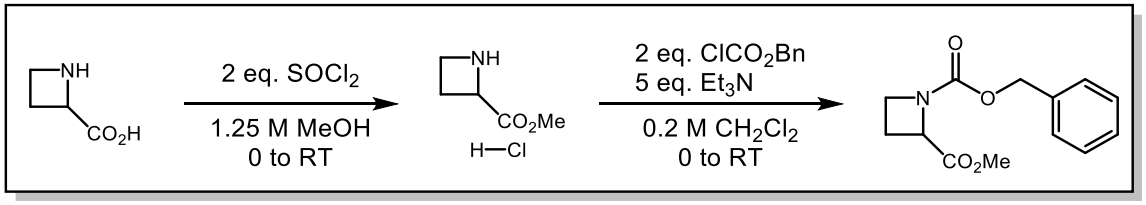


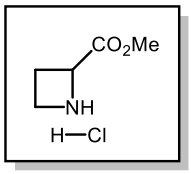


(±)-Isopropyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (50 mg, 1 mmol, 1 equiv.) was suspended in isopropanol (0.6 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.07 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the brown solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



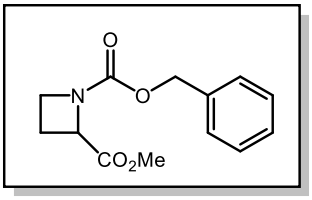
(±)-1-Benzyl 2-isopropyl azetidine-1,2-dicarboxylate. The crude (±)-isopropyl azetidine-2-carboxylate hydrochloride (89.9 mg, 0.5 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (2.5 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.35 mL, 2.5 mmol, 5 equiv.). Once addition is complete, benzyl chloroformate (0.14 mL, 1 mmol, 2 equiv.) was added dropwise. The reaction was allowed to warm to RT and stirring was continued until the reaction was judged complete by TLC analysis. Once complete, the reaction was transferred to a separatory funnel and quenched with 5 mL of saturated sodium bicarbonate solution. The layers were thoroughly mixed, and the aqueous layer was removed. The organic layer was washed with water (1x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted, and the volatiles removed under vacuum. The crude product was purified by silica gel column chromatography (gradient from hexanes to 33% EtOAc in hexanes) to afford 65 mg (47% yield) of the titled compound a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.39 – 7.27 (m, 5H), 5.18 – 4.98 (m, 3H), 4.64 (dd, J = 9.3, 5.2 Hz, 1H), 4.11 (m, 1H), 3.97 (m, 1H), 2.56 (dtd, J = 11.5, 9.2, 6.3 Hz, 1H), 2.18 (ddt, J = 11.2, 9.0, 5.5 Hz, 1H), 1.22 (m, 6H).





(±)-Methyl azetidine-2-carboxylate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: (±)-azetidine-2-carboxylic acid (100 mg, 1 mmol, 1 equiv.) was suspended in methanol (1.25 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.15 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution

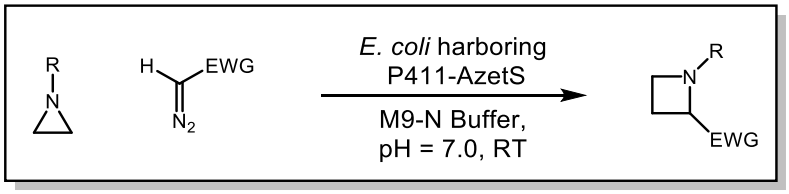
was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



(±)-1-Benzyl 2-methyl azetidine-1,2-dicarboxylate. The crude methyl azetidine-2-carboxylate hydrochloride (151.59 mg, 1 mmol, 1 equiv.) synthesized above was suspended in methylene chloride (5 mL) and the dram vial was equipped with a stir bar, a septum, and a drying tube. The vial was chilled in an ice bath prior to the addition of triethylamine (0.69 mL, 5 mmol, 5 equiv.). Once addition was complete, benzyl chloroformate (0.29 mL, 2 mmol, 2 equiv.) was added dropwise. The reaction was allowed to warm to RT and stirring was continued until the reaction was judged complete by TLC analysis. Once complete, the reaction was transferred to a separatory funnel and quenched with 5 mL of saturated sodium bicarbonate solution. The layers were thoroughly mixed, and the aqueous layer was removed. The organic layer was washed with water (1x 5 mL) and brine (1x 5 mL) prior to drying over sodium sulfate. Once dry, the organics were decanted, and the volatiles removed under vacuum. The crude product was purified by silica gel column chromatography (gradient from hexanes to 40% EtOAc in hexanes) to afford 123 mg (49% yield) of the titled compound a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 – 7.28 (m, 5H), 5.25 – 5.02 (m, 2H), 4.71 (dd, J = 9.3, 5.3 Hz, 1H), 4.11 m, 1H), 3.98 (m, 1H), 3.76 (br s, 3H), 2.57 (dtd, J = 11.6, 9.2, 6.3 Hz, 1H), 2.23 (ddt, J = 11.4, 9.0, 5.5 Hz, 1H).

A.6. Enzymatic Reactions & Product Characterization

A.6.1. General Procedure for Preparative Scale Aziridine Ring Expansion



To a 250-mL screw-cap Erlenmeyer flask: 47.5 mL of *E. coli* whole cell suspension harboring P411-AzetS ([P411-AzetS] = 5.25 μM, final reaction concentration 5.0 μM, 0.25 μmol, $5 \cdot 10^{-4}$ equiv.) was added. The whole cell suspension ws degassed with nitrogen and put into an anerobic Coy chamber. Under inert atmosphere, 1.25 mL of an acetonitrile solution of the corresponding aziridine ([Aziridine] = 400 mM, final reaction concentration 10 mM, 0.5 mmol, 1 equiv.) and 1.25 mL of an acetonitrile solution of the corresponding diazo compound ([Diazo] = 600 mM, final reaction concentration 15 mM, 0.75 mmol, 1 equiv.) were added in sequence. The vial was securely capped to exclude oxygen, the flask was removed from the anaerobic chamber, and the reaction was allowed to shake at RT until judged complete (ca. 4–16 hours).

TTN and Analytical Yield Determination

A 400 μL aliquot of the above reaction solution was mixed with 600 μL of a 1:1 solution of ethyl acetate:cyclohexane with 10 mM of the appropriate internal standard in a 1.7-mL Eppendorf tube. The layers were thoroughly mixed, and the phases were separated by centrifugation (14000 g, 10 minutes, RT). Once separated, 100 μL of the organic layer was siphoned off and diluted in 900 μL of ethyl acetate. This solution was subjected to GC-FID analysis to determine the analytical yield and TTN of the reaction.

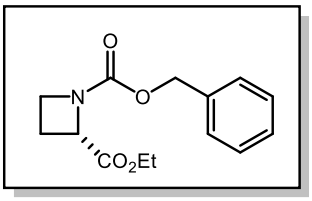
Isolated Yield:

Once complete, the reaction was split into two 25 mL aliquots across two 50 mL falcon tubes. The suspensions are extracted three times with ethyl acetate as follows: 25 mL of ethyl acetate were added to each tube and the phases were mixed by hand-shaking the tubes. Upon thorough mixing, the phases were separated by centrifugation (5000 g, 5 minutes, RT) and the organic phase was siphoned off. The combined organics were dried over sodium sulfate; once dry, the drying agent was decanted, and the volatiles removed under vacuum. The crude residue was purified by silica gel column chromatography to yield the titled compounds.

Note About Stereochemical Assignments:

The absolute stereochemistry for the model substrate **2** was determined by synthesis of the *S*-enantiomer, which corresponds to the major product produced in the reaction. The stereochemistry of products **3-9** are assigned by analogy.

A.6.2. Isolated and Analytical Yields of Enzymatic Ring Expansions

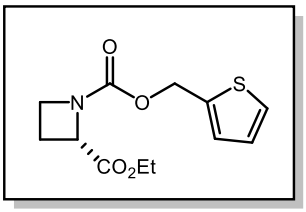


1-Benzyl 2-ethyl (*S*)-azetidine-1,2-dicarboxylate (2). Synthesized using the general procedure for preparative scale reactions starting with benzyl aziridine-1-carboxylate and ethyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 40% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	1560	1430	1490
%Yield (GC)	78%	72%	75%
Yield (Isolated)	91 mg (69%)	86 mg (65%)	67%

er (GC)	99:1	99:1	--/--

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.39 – 7.27 (m, 5H), 5.10 (m, 2H), 4.68 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.19 (br s, 2H), 4.12 (m, 1H), 3.97 (m, 1H), 2.57 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.22 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.30 – 1.17 (br m, 3H).^{13}C NMR (101 MHz, $CDCl_3$) δ 171.12, 155.71, 136.41, 128.41, 128.00, 127.88, 66.78, 61.28, 60.57, 47.68, 20.70, 14.08. HRMS (FAB+) exact mass calculated for $[M+H]^+$ ($C_{14}H_{18}O_4N$) requires *m/z* 264.1236, found *m/z* 264.1249.

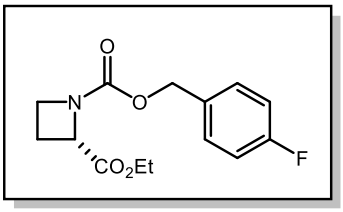


2-Ethyl 1-(thiophen-2-ylmethyl) (*S*)-azetidine-1,2-dicarboxylate (3). Synthesized using the general procedure for preparative scale reactions starting with thiophen-2-ylmethyl aziridine-1-carboxylate and ethyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	760	767	764
%Yield (GC)	38.0%	38.3%	38%
Yield (Isolated)	49.4 mg (37% yield)	51.4 mg (38% yield)	38%
er (GC)	99:1	99:1	--/--

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.29 (dd, *J* = 5.0, 1.3 Hz, 1H), 7.07 (d, *J* = 3.5 Hz, 1H), 6.96 (dd, *J* = 5.1, 3.5 Hz, 1H), 5.24 (m, 2H), 4.66 (dd, *J* = 9.4, 5.3 Hz, 1H), 4.20 (br s, 2H), 4.10 (m, 1H), 3.95 (m, 1H), 2.56 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.20 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.24 (br s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.03, 155.41, 138.48, 127.82,

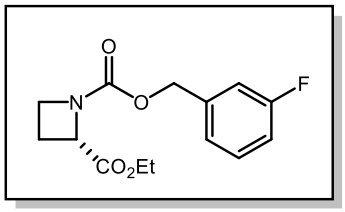
126.68, 126.61, 61.29, 61.19, 60.57, 47.71, 20.70, 14.07. HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{12}H_{16}NO_4S$) requires *m/z* 270.0800, found *m/z* 270.0792.



2-Ethyl 1-(4-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (4). Synthesized using the general procedure for preparative scale reactions starting with 4-fluorobenzyl aziridine-1-carboxylate and ethyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 25% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	613	640	626
%Yield (GC)	30.6%	32.0%	31%
Yield (Isolated)	41 mg (29% yield)	41 mg (29% yield)	29%
er (GC)	99:1	99:1	--/--

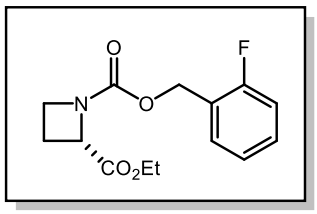
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 (dd, *J* = 8.4, 5.4 Hz, 2H), 7.02 (m, 2H), 5.06 (m, 2H), 4.67 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.19 (br s, 2H), 4.10 (m, 1H), 3.96 (m, 1H), 2.57 (dtd, *J* = 11.5, 9.2, 6.4 Hz, 1H), 2.21 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.30 – 1.18 (br s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.05, 162.52 (d, *J* = 246.4 Hz), 155.58, 132.26 (d, *J* = 3.3 Hz), 129.90 (d, *J* = 8.4 Hz), 115.31 (d, *J* = 21.4 Hz), 66.07, 61.28, 60.37, 47.70, 20.69, 14.07. ^{19}F NMR (282 MHz, $CDCl_3$) δ -114.28 (br s). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{14}H_{17}FNO_4$) requires *m/z* 282.1142, found *m/z* 282.1148.



2-Ethyl 1-(3-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (5). Synthesized using the general procedure for preparative scale reactions using 3-fluorobenzyl aziridine-1-carboxylate and ethyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	650	712	681
%Yield (GC)	33%	36%	34%
Yield (Isolated)	41.7 mg (30% yield)	39.7 mg (29% yield)	29%
er (GC)	99:1	99:1	--/--

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.30 (td, *J* = 7.9, 5.8 Hz, 1H), 7.13 – 6.94 (m, 3H), 5.22 – 4.98 (m, 2H), 4.70 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.27 – 4.17 (m, 2H), 4.16 – 4.08 (m, 1H), 3.99 (ddd, *J* = 9.1, 8.1, 5.6 Hz, 1H), 2.59 (dtd, *J* = 11.6, 9.2, 6.3 Hz, 1H), 2.23 (ddt, *J* = 11.4, 9.1, 5.5 Hz, 1H), 1.30 – 1.18 (m, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.03, 162.79 (d, *J* = 246.1 Hz), 155.39, 138.96 (d, *J* = 7.4 Hz), 129.95 (d, *J* = 8.2 Hz), 123.09, 114.82 (d, *J* = 21.1 Hz), 114.48 (d, *J* = 22.0 Hz), 65.84, 61.35, 60.63, 47.59, 20.72, 14.06. ^{19}F NMR (282 MHz, $CDCl_3$) δ -113.22 (td, *J* = 10.1, 6.5 Hz). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{14}H_{17}FNO_4$) requires *m/z* 282.1142, found *m/z* 282.1154.

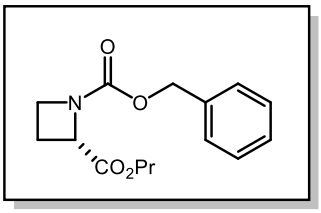


2-Ethyl 1-(2-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (6). Synthesized using the general procedure for preparative scale reactions using 2-fluorobenzyl aziridine-1-carboxylate and ethyl diazoacetate as substrates. The product was purified

by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	826	796	811
%Yield (GC)	41.3%	39.8%	41%
Yield (Isolated)	50.5 mg (36% yield)	50.2 mg (36% yield)	36%
er (GC)	99:1	99:1	--/--

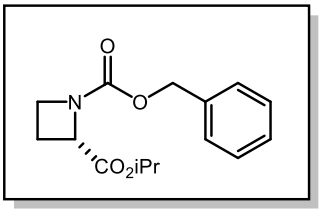
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.38 (m, 1H), 7.29 (m, 1H), 7.12 (m, 1H), 7.04 (m, 1H), 5.17 (q, *J* = 12.7 Hz, 2H), 4.68 (dd, *J* = 9.3, 5.3 Hz, 1H), 4.19 (m, 2H), 4.11 (td, *J* = 8.6, 6.3 Hz, 1H), 3.97 (td, *J* = 8.6, 8.2, 5.7 Hz, 1H), 2.57 (dtd, *J* = 11.6, 9.2, 6.3 Hz, 1H), 2.21 (ddt, *J* = 11.3, 9.0, 5.5 Hz, 1H), 1.24 (m, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.05, 160.81 (d, *J* = 248.3 Hz), 155.52, 130.29, 129.90 (d, *J* = 8.1 Hz), 124.02 (d, *J* = 3.7 Hz), 123.56 (d, *J* = 14.4 Hz), 115.28 (d, *J* = 21.2 Hz), 61.28, 60.76 (d, *J* = 4.4 Hz), 60.64 (br, obscured by doublet at 60.76), 47.70, 20.70, 14.05. ^{19}F NMR (282 MHz, $CDCl_3$) δ -118.40 (m). HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{14}H_{17}FNO_4$) requires *m/z* 282.1142, found *m/z* 282.1114.



1-Benzyl 2-propyl (*S*)-azetidine-1,2-dicarboxylate (7). Synthesized using the general procedure for preparative scale reactions using benzyl aziridine-1-carboxylate and propyl diazoacetate as substrates. The product is purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	314	411	363
%Yield (GC)	16%	21%	18%
Yield (Isolated)	19.6 mg (14% yield)	27.1 mg (20% yield)	17%
er (GC)	99:1	99:1	--/--

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.40 – 7.27 (m, 5H), 5.17 – 5.04 (m, 2H), 4.69 (dd, *J* = 9.3, 5.2 Hz, 1H), 4.20 – 4.06 (m, 3H), 3.97 (ddd, *J* = 9.1, 8.1, 5.6 Hz, 1H), 2.57 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.21 (ddt, *J* = 11.3, 8.9, 5.5 Hz, 1H), 1.64 (m, 2H – overlaps with HOD peak), 0.91 (m, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) 171.20, 155.70, 136.40, 128.40, 127.99, 127.89, 66.79, 66.78, 60.69, 47.62, 21.86, 20.74, 10.22. HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{15}H_{20}O_4N$) requires *m/z* 278.1392, found *m/z* 278.1400.

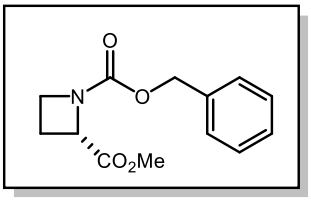


1-Benzyl 2-isopropyl (*S*)-azetidine-1,2-dicarboxylate (8).

Synthesized using the general procedure for preparative scale reactions using benzyl aziridine-1-carboxylate and propyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	511	617	564
%Yield (GC)	26%	31%	28%
Yield (Isolated)	34.8 mg (25% yield)	41.5 mg (30% yield)	28%
er (GC)	>99:1	>99:1	--/--

^{1}H NMR (400 MHz, $CDCl_3$) δ ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.39 – 7.27 (m, 5H), 5.18 – 4.98 (m, 3H), 4.64 (dd, *J* = 9.3, 5.2 Hz, 1H), 4.11 (m, 1H), 3.97 (m, 1H), 2.56 (dtd, *J* = 11.5, 9.2, 6.3 Hz, 1H), 2.18 (ddt, *J* = 11.2, 9.0, 5.5 Hz, 1H), 1.22 (m, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.63, 155.70, 136.41, 128.40, 127.97, 127.86, 68.84, 66.75, 60.85, 47.62, 21.65, 21.60, 20.68. HRMS (ES+) exact mass calculated for $[M+H]^+$ ($C_{15}H_{20}O_4N$) requires *m/z* 278.1392, found *m/z* 278.1367.



1-Benzyl 2-methyl (*S*)-azetidine-1,2-dicarboxylate (9). Synthesized using the general procedure for preparative scale reactions starting with benzyl aziridine-1-carboxylate and methyl diazoacetate as substrates. The product was purified by silica gel column chromatography (gradient from 100% hexanes to 40% EtOAc in hexanes) to afford the titled compound as a colorless oil.

Value	Run 1	Run 2	Average
TTN (GC)	405	399	402
%Yield (GC)	20.2	20.0	20
Yield (Isolated)	24 mg (19% yield)	21 mg (17% yield)	18%
er (GC)	81:19	81:19	

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.52 – 7.29 (m, 5H), 5.26 – 5.01 (m, 2H), 4.71 (dd, *J* = 9.1, 5.2 Hz, 1H), 4.11 (m, 1H), 3.98 (m, 1H), 3.73 (br s, 3H), 2.57 (dtd, *J* = 11.3, 9.2, 6.4 Hz, 1H), 2.23 (ddt, *J* = 11.4, 8.9, 5.4 Hz, 1H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.58, 155.73, 136.40, 128.43, 128.03, 127.91, 66.82, 60.51, 52.27, 47.75, 20.69. HRMS (FAB+) exact mass calculated for $[M+H]^+$ ($C_{13}H_{16}O_4N$) requires *m/z* 250.1079, found *m/z* 250.1067.

A.7. Procedure for 10-mmol Scale Reaction

Starter Culture & Expression Culture:

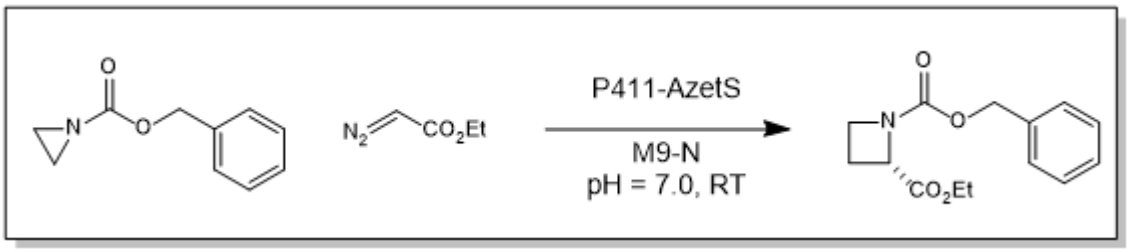
A single colony of *E. coli* harboring plasmids encoding P411-AzetS was used to inoculate 200 mL of LB-amp. The culture was allowed to incubate overnight at 37 °C, shaking at 240 RPM.

The following day: six expression cultures (1L) were inoculated with 20 mL of the above overnight culture. The expression cultures were incubated at 37 °C, shaking at 240 RPM, for 2.5 hours and the OD_{600} is measured to be 1.0. At this point, the six expression cultures were held on ice for 30 minutes wherein 1 mL of 1000x stocks of ALA and IPTG are added such that the final concentrations were 1.0 mM of ALA and 0.5 mM of IPTG, respectively. The cultures were allowed to incubate for 20 hours at 22 °C, shaking at 140 RPM.

Preparation of Clarified Lysate for Protein Concentration Determination:

The expression cultures were pelleted by centrifugation (5000 g, 5 minutes, 4 °C). The cell pellets were pumped into the anaerobic Coy chamber, wherein they were resuspended in 300 mL of rigorously degassed M9-N buffer adjusted to pH = 7.0 and combined in a screw-top Erlenmeyer. For protein concentration determination, 3 mL of the whole cell suspension were siphoned off, diluted in 3 mL of additional M9-N, and the protein concentration was measured as described in the “General Information and Procedures” section of the SI. The Erlenmeyer with the whole cell suspension was capped with an air-tight cap and was held on ice until the protein concentration can be accurately determined.

Reaction Setup:



Safety Note: 15 mmol of ethyl diazoacetate can be expected to liberate 15 mmol of N_2, or 336 mL of N_2. Use of a flask with excess headspace prevents overpressurization of the reaction vessel in this reaction.

Once the protein concentration was determined, the whole cell suspension was diluted in a 2.0 L screw-top Erlenmeyer flask in the anaerobic Coy chamber with rigorously degassed M9-N buffer adjusted to pH = 7.0 such that 950 mL of whole cell suspension has a protein concentration of 5.25 μM (final reaction concentration 5.0 μM, 5 μmol, $5 \cdot 10^{-4}$ equiv.). For this experiment, the OD_{600} was measured to be approximately 30. Once ready, 24 mL of a 417 mM solution of benzyl aziridine-1-carboxylate in MeCN (final reaction concentration, 10 mM, 10 mmol, 1.77 g, 1 equiv.) was added with agitation, followed by addition of 24 mL of a 625 mM solution of ethyl diazoacetate (final reaction concentration 15 mM, 1.71 g, 1.5 equiv.) with heavy agitation. The flask was tightly capped to exclude oxygen, pumped out of the anaerobic chamber, and was allowed to shake at 220 RPM until judged complete.

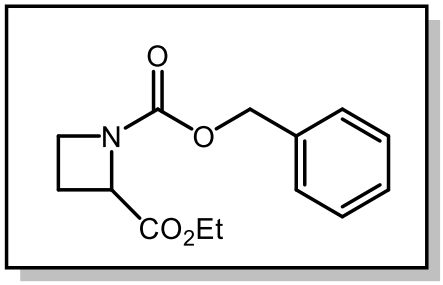
Workup and Isolation:

Once complete, three 400-μL aliquots were removed and used to determine the reaction TTN and yield by gas chromatography (1220 TTN, 61% yield). For preparative isolation, the reaction mixture was split into 2x 500 mL portions. Each portion was mixed with 500 mL of EtOAc; the layers were thoroughly mixed before the organic layer was siphoned off

(the layers may be separated by centrifugation, if necessary, at 5000 g for 5 minutes at RT). The combined organics were dried over sodium sulfate; once dried, the supernatant was decanted from the drying agent and the volatiles removed under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes) to afford 1.44 g (55% isolated yield) of the titled compound as a colorless oil with 99:1 er. Characterization data was completely identical to other samples of **2** synthesized in this manuscript.

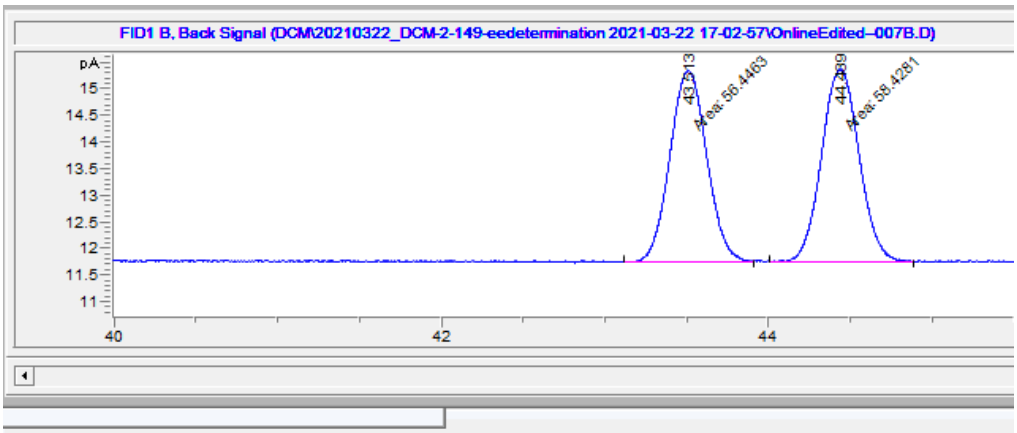
A.8. Chiral Traces for Determination of Enantiopurity: Preparative-Scale Reactions

The enantiopurity of all products was determined by gas chromatography using CP-Chirasil-Dex CB (Agilent) as a chiral stationary phase. All products could be separated using the same GC conditions, which are as follows: start at 50 °C, ramp at 10 °C/minute to 170 °C and hold for 38 minutes. Ramp at 30 °C/minute to 200 °C and hold for 9 minutes.



1-Benzyl 2-ethyl (*S*)-azetidine-1,2-dicarboxylate (2)

Racemic Standard:

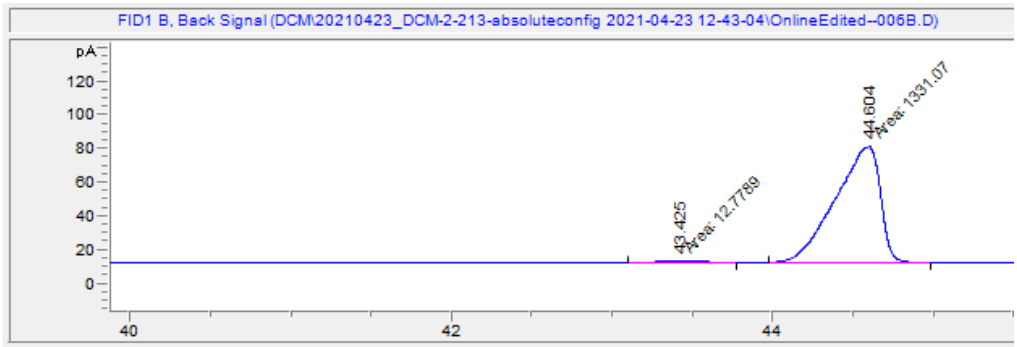


#	Time	Area	Height	Width	Area%	Symmetry
1	43.513	56.4	3.6	0.2618	49.137	0.954
2	44.439	58.4	3.6	0.2681	50.863	0.918

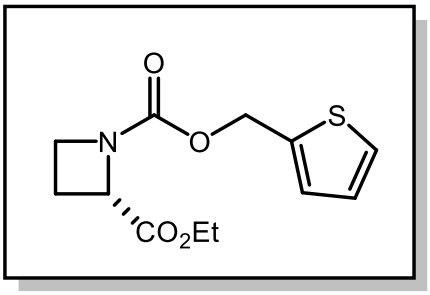
(*S*)-Enantiomer Standard:

#	Time	Area	Height	Width	Area%	Symmetry
1	44.721	1859.2	87.4	0.2682	100.000	4.229

Product from P411-AzetS:

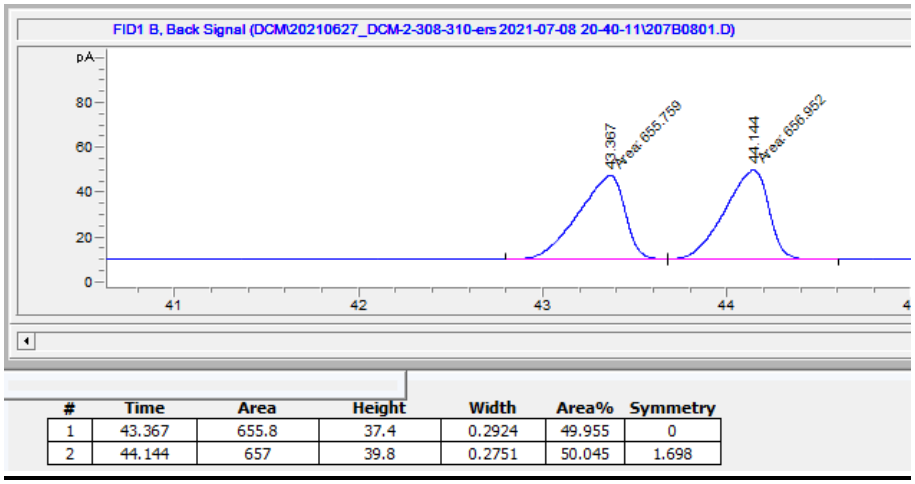


#	Time	Area	Height	Width	Area%	Symmetry
1	43.425	12.8	7.9E-1	0.2701	0.951	0.85
2	44.604	1331.1	68.6	0.3236	99.049	2.672



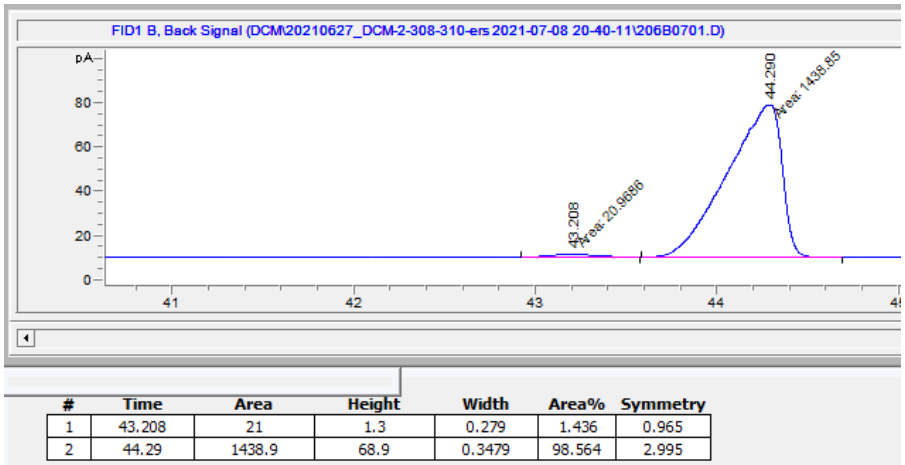
2-Ethyl 1-(thiophen-2-ylmethyl) (*S*)-azetidine-1,2-dicarboxylate (3)

Racemic Standard:

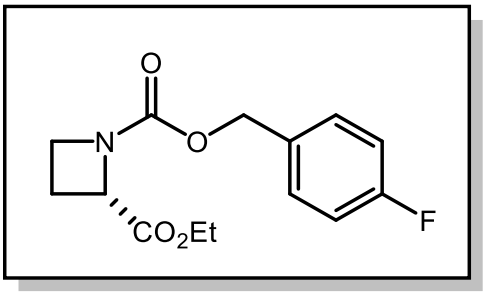


#	Time	Area	Height	Width	Area%	Symmetry
1	43.367	655.8	37.4	0.2924	49.955	0
2	44.144	657	39.8	0.2751	50.045	1.698

Product from P411-AzetS:

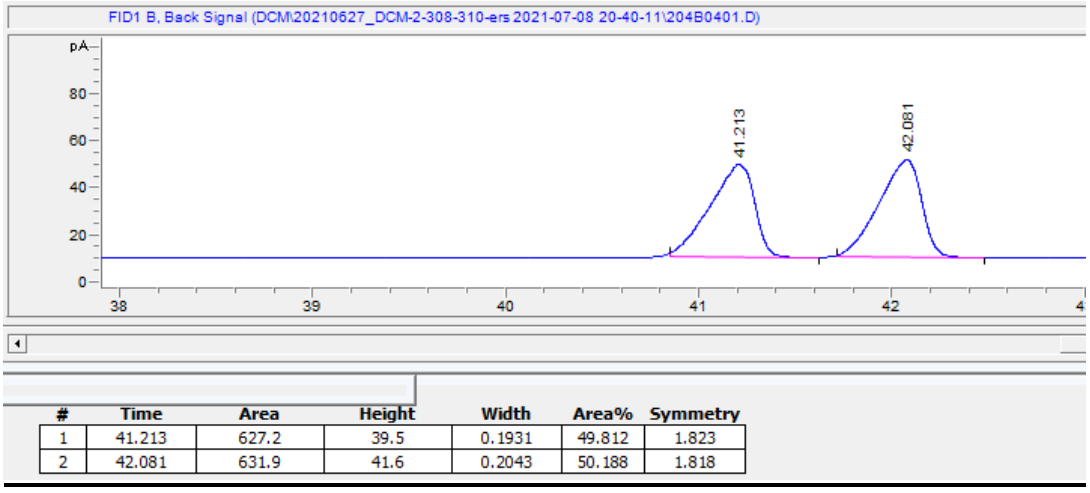


#	Time	Area	Height	Width	Area%	Symmetry
1	43.208	21	1.3	0.279	1.436	0.965
2	44.29	1438.9	68.9	0.3479	98.564	2.995



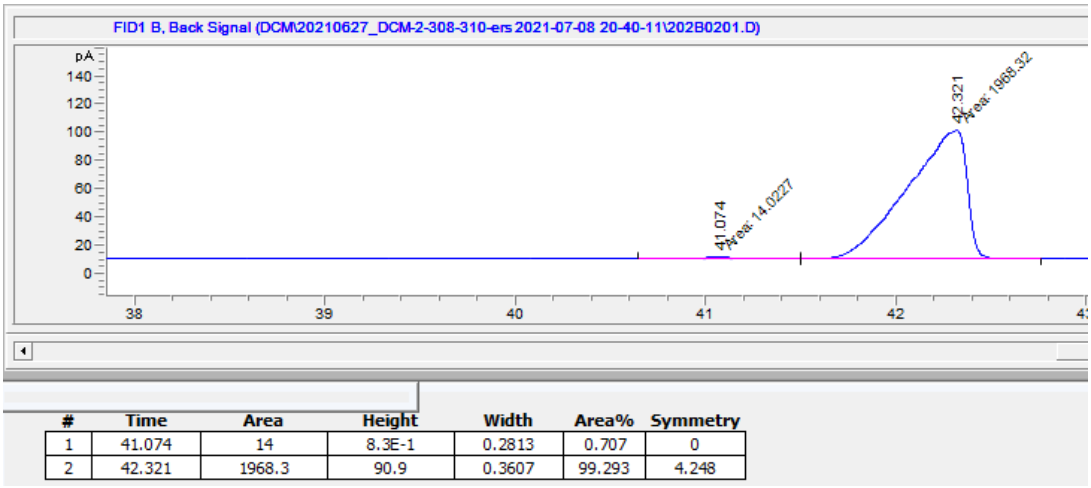
2-Ethyl 1-(4-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (4)

Racemic Standard:

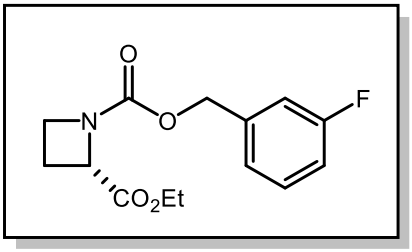


#	Time	Area	Height	Width	Area%	Symmetry
1	41.213	627.2	39.5	0.1931	49.812	1.823
2	42.081	631.9	41.6	0.2043	50.188	1.818

Product from P411-AzetS:

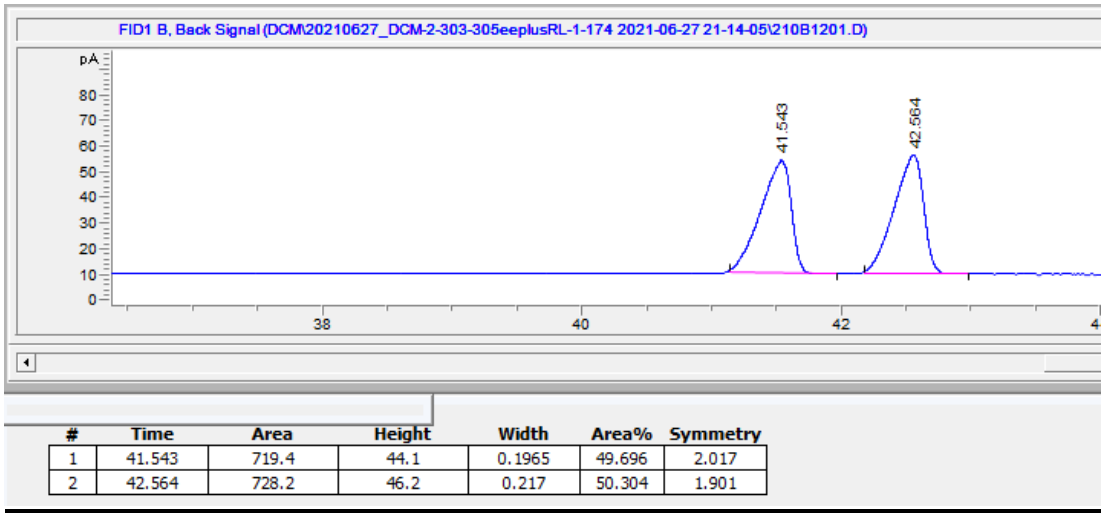


#	Time	Area	Height	Width	Area%	Symmetry
1	41.074	14	8.3E-1	0.2813	0.707	0
2	42.321	1968.3	90.9	0.3607	99.293	4.248



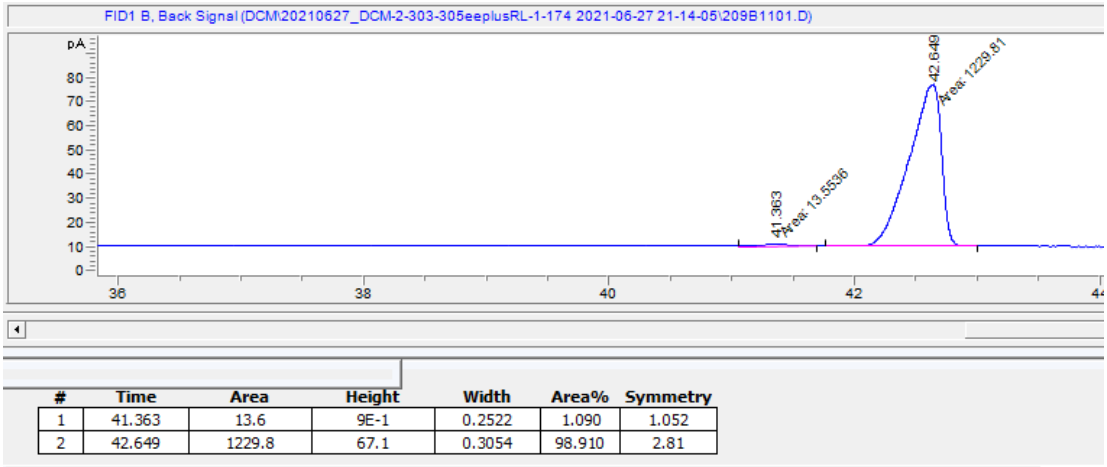
2-Ethyl 1-(3-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (5)

Racemic Standard:

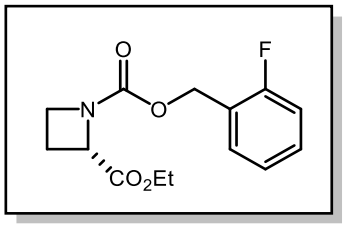


#	Time	Area	Height	Width	Area%	Symmetry
1	41.543	719.4	44.1	0.1965	49.696	2.017
2	42.564	728.2	46.2	0.217	50.304	1.901

Product from P411-AzetS:

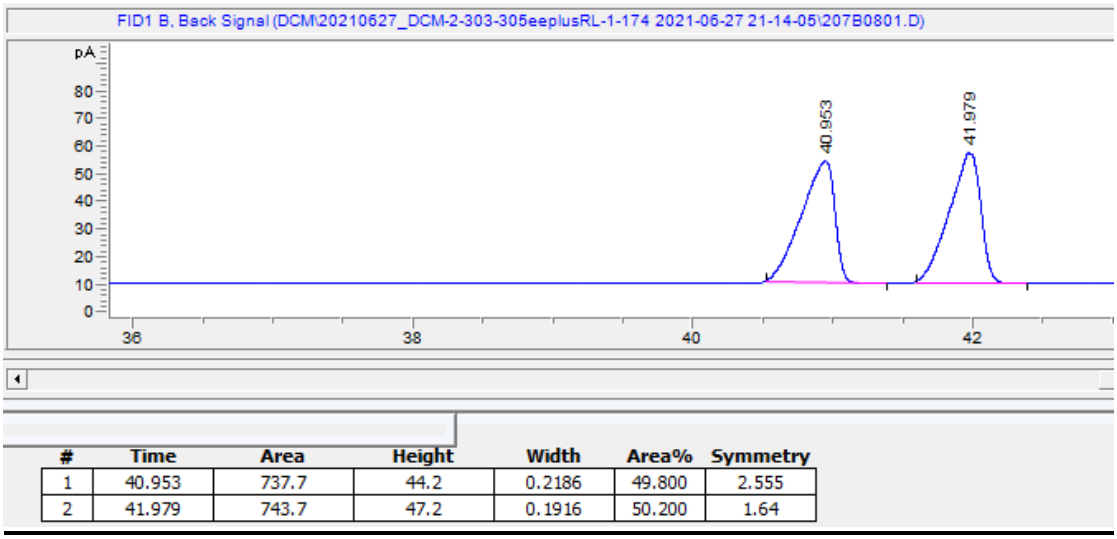


#	Time	Area	Height	Width	Area%	Symmetry
1	41.363	13.6	9E-1	0.2522	1.090	1.052
2	42.649	1229.8	67.1	0.3054	98.910	2.81



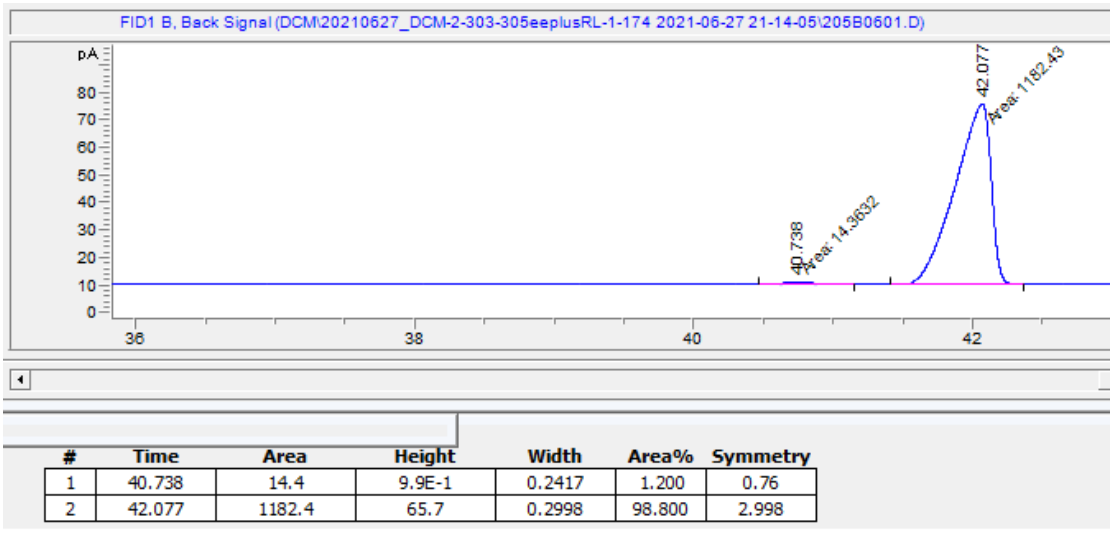
2-Ethyl 1-(2-fluorobenzyl) (*S*)-azetidine-1,2-dicarboxylate (6)

Racemic Standard:

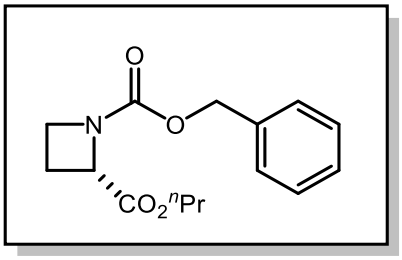


#	Time	Area	Height	Width	Area%	Symmetry
1	40.953	737.7	44.2	0.2186	49.800	2.555
2	41.979	743.7	47.2	0.1916	50.200	1.64

Product from P411-AzetS:

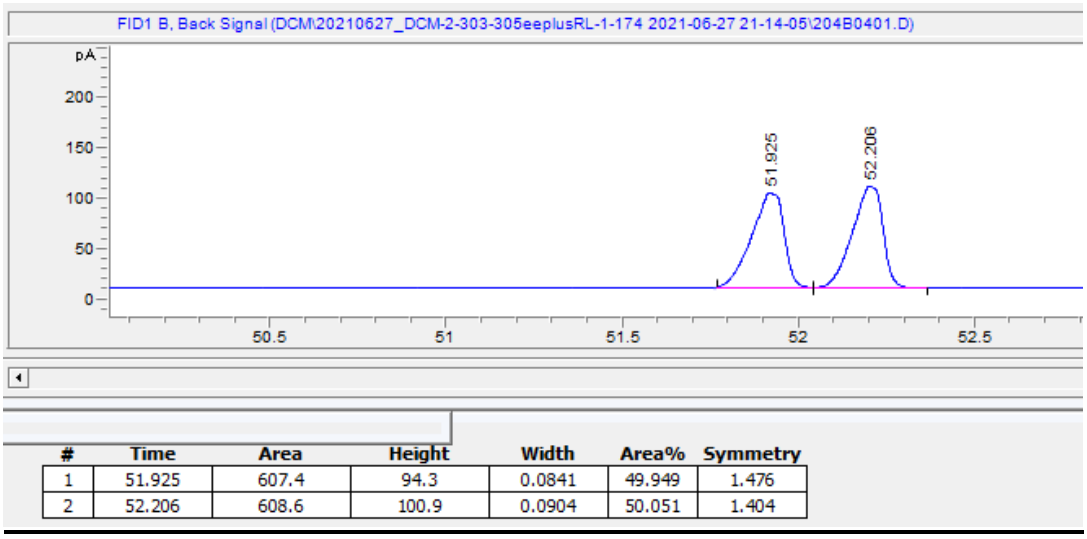


#	Time	Area	Height	Width	Area%	Symmetry
1	40.738	14.4	9.9E-1	0.2417	1.200	0.76
2	42.077	1182.4	65.7	0.2998	98.800	2.998



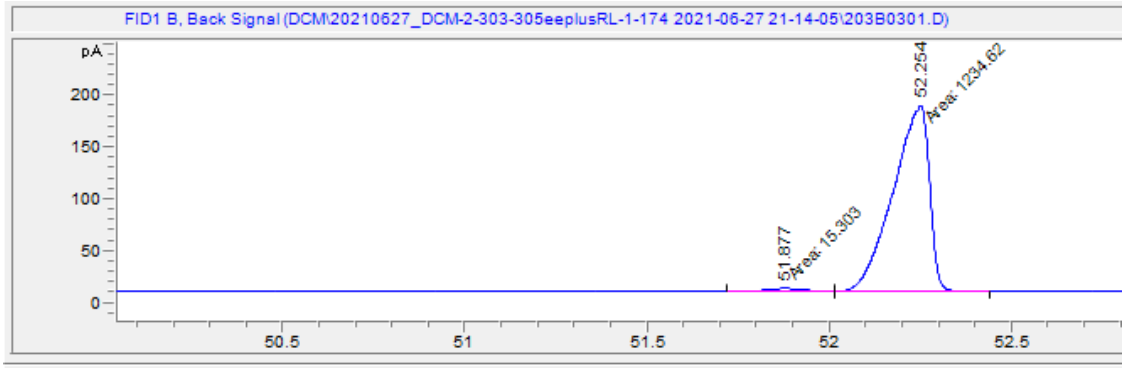
1-Benzyl 2-propyl azetidine-1,2-dicarboxylate (7)

Racemic Standard:

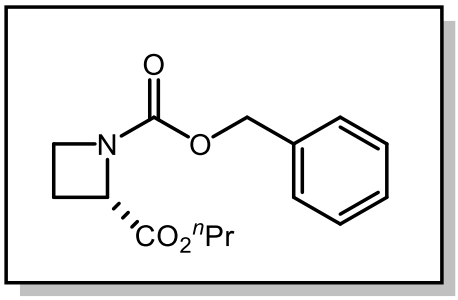


#	Time	Area	Height	Width	Area%	Symmetry
1	51.925	607.4	94.3	0.0841	49.949	1.476
2	52.206	608.6	100.9	0.0904	50.051	1.404

Product from P411-AzetS:

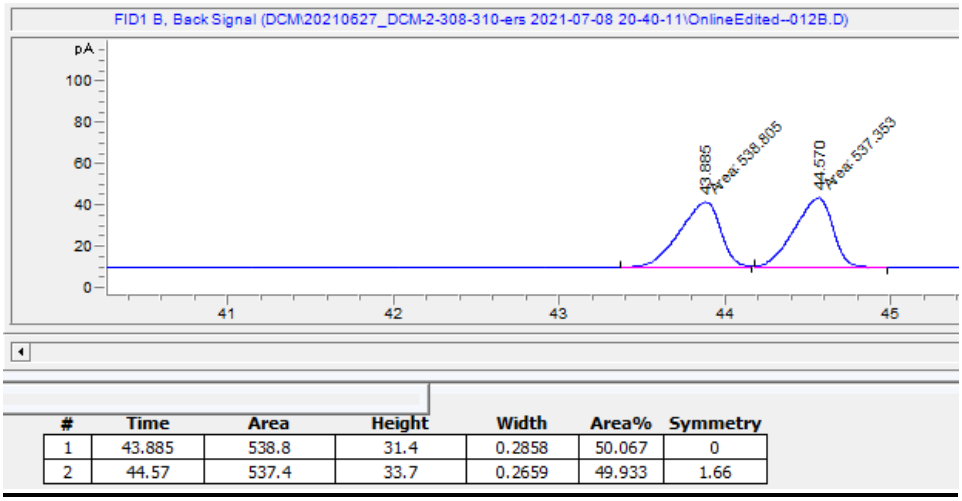


#	Time	Area	Height	Width	Area%	Symmetry
1	51.877	15.3	2.5	0.1033	1.224	0
2	52.254	1234.6	179.3	0.1148	98.776	2.966



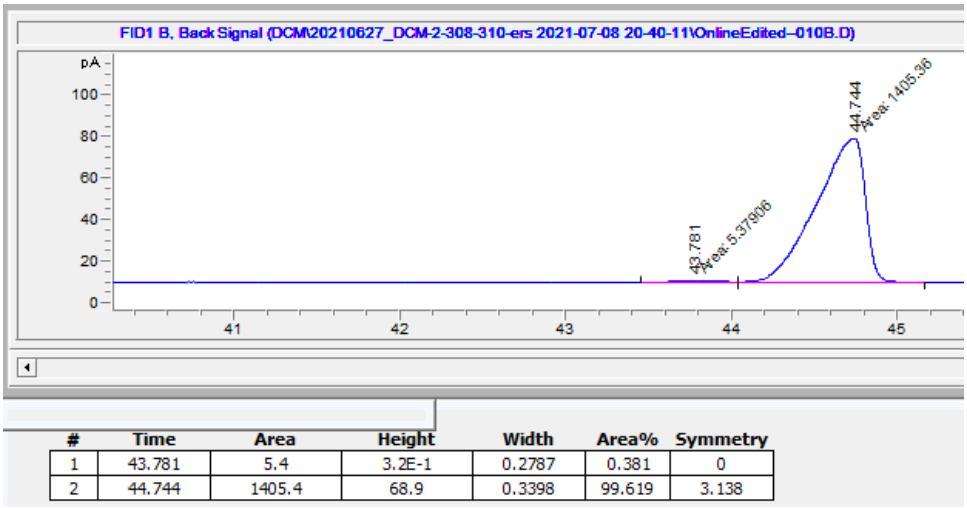
1-Benzyl 2-isopropyl azetidine-1,2-dicarboxylate (8)

Racemic Standard:

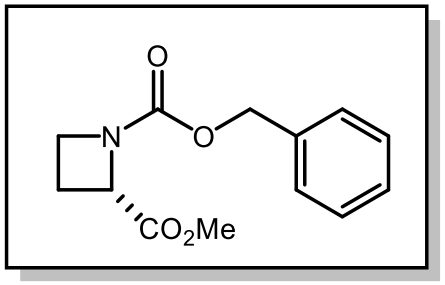


#	Time	Area	Height	Width	Area%	Symmetry
1	43.885	538.8	31.4	0.2858	50.067	0
2	44.57	537.4	33.7	0.2659	49.933	1.66

Product from P411-AzetS:

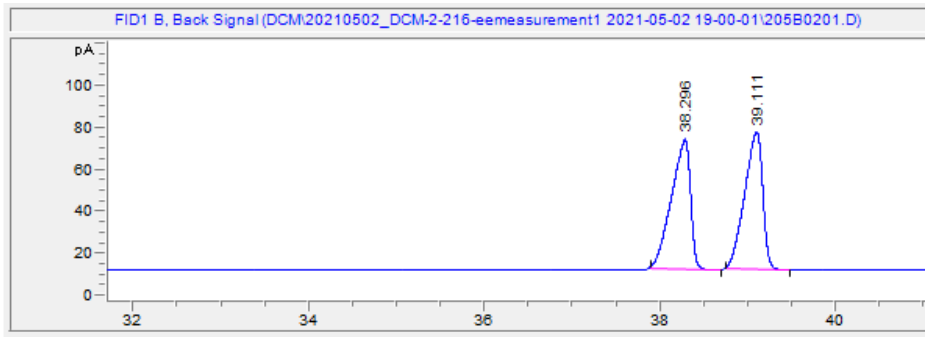


#	Time	Area	Height	Width	Area%	Symmetry
1	43.781	5.4	3.2E-1	0.2787	0.381	0
2	44.744	1405.4	68.9	0.3398	99.619	3.138



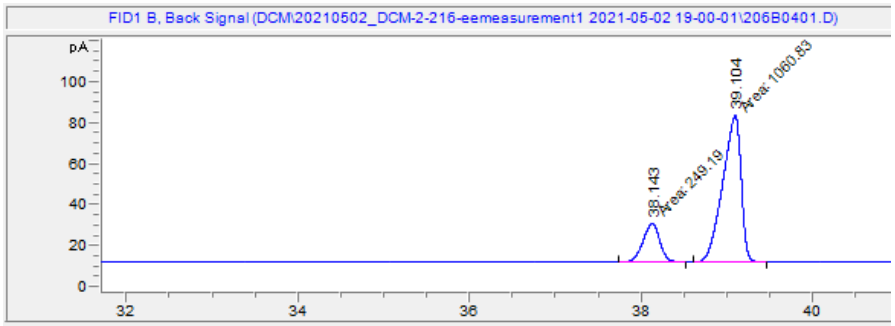
1-Benzyl 2-methyl azetidine-1,2-dicarboxylate (9)

Racemic Standard:



#	Time	Area	Height	Width	Area%	Symmetry
1	38.296	937.9	62	0.2005	50.010	2.45
2	39.111	937.6	65.5	0.177	49.990	1.84

Product from P411-AzetS:

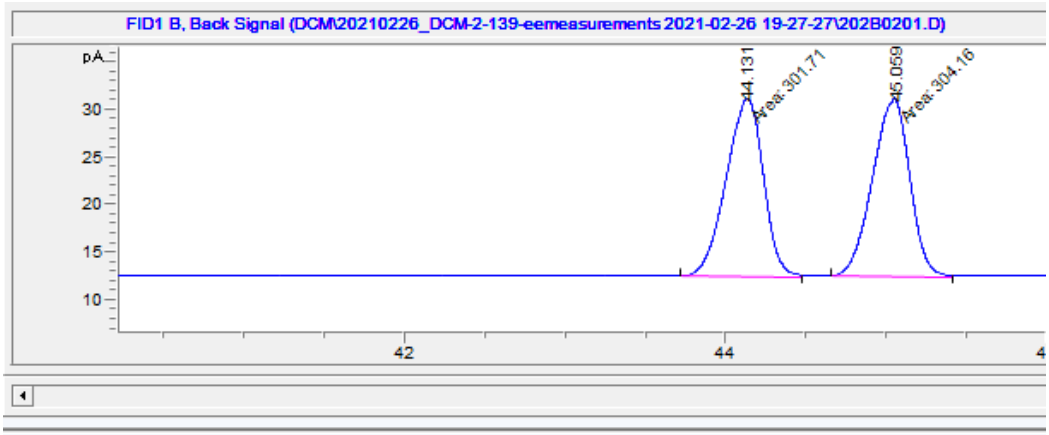


#	Time	Area	Height	Width	Area%	Symmetry
1	38.143	249.2	18.4	0.2263	19.022	1.262
2	39.104	1060.8	71.6	0.2469	80.978	1.946

A.9. Chiral Traces for Determination of Enantiopurity: Evolutionary Lineage

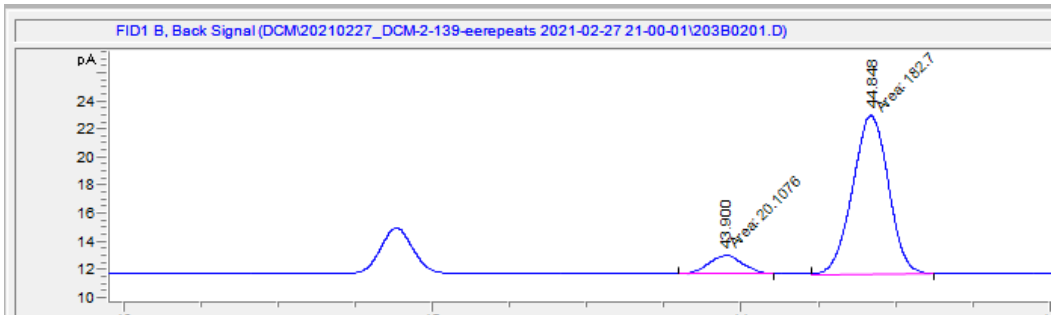
The enantiopurity was determined as outlined in the preceding section.

Racemic Standard:



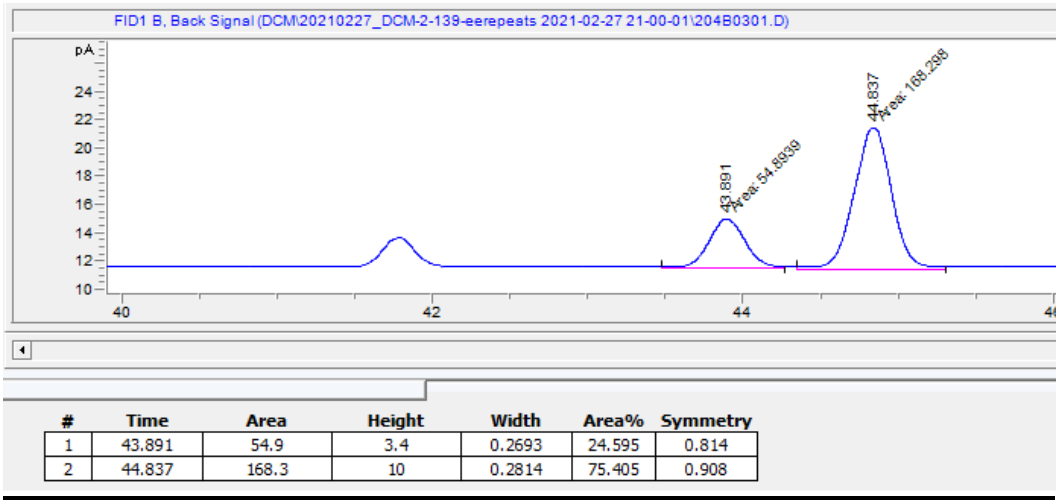
#	Time	Area	Height	Width	Area%	Symmetry
1	44.131	301.7	18.7	0.269	49.798	1.086
2	45.059	304.2	18.8	0.2695	50.202	1.399

Parent F2:



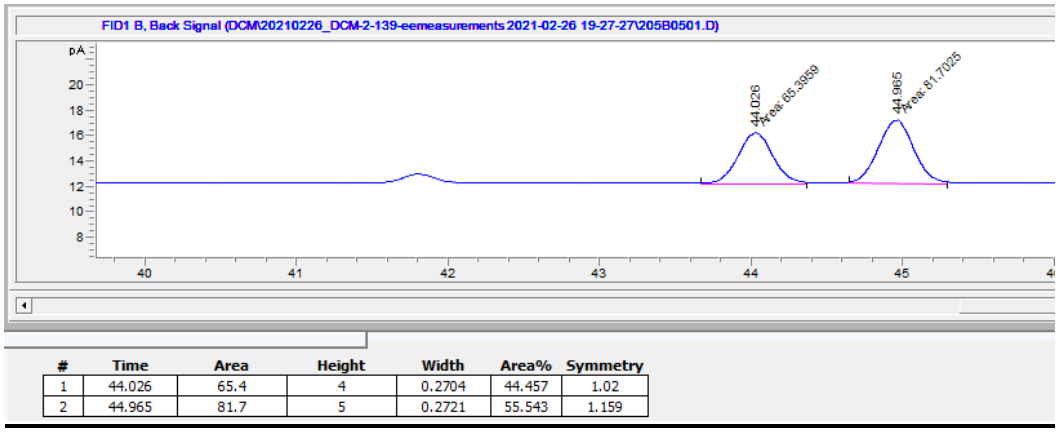
#	Time	Area	Height	Width	Area%	Symmetry
1	43.9	20.1	1.3	0.2576	9.915	0.932
2	44.848	182.7	11.3	0.2694	90.085	1.054

Parent F2.1:



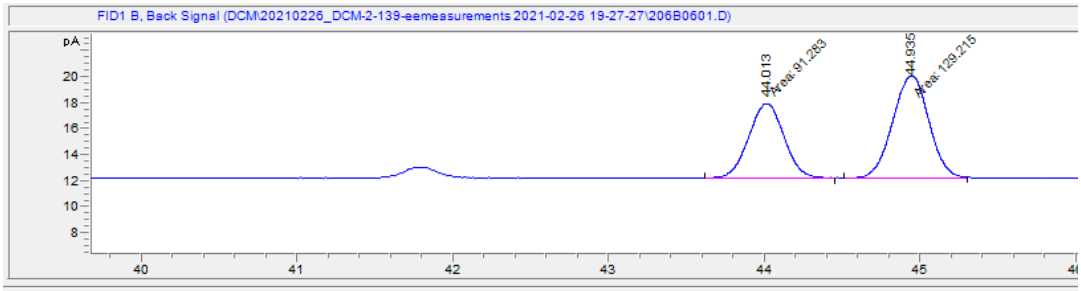
#	Time	Area	Height	Width	Area%	Symmetry
1	43.891	54.9	3.4	0.2693	24.595	0.814
2	44.837	168.3	10	0.2814	75.405	0.908

Parent F2.2:



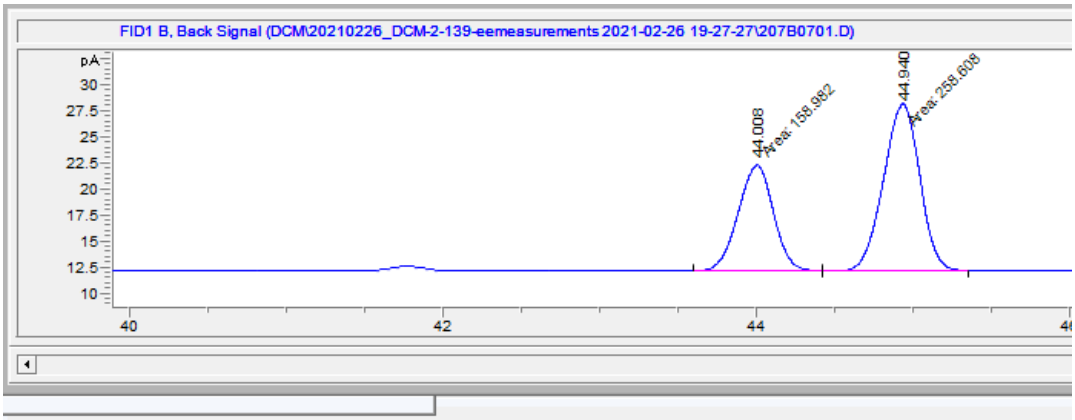
#	Time	Area	Height	Width	Area%	Symmetry
1	44.026	65.4	4	0.2704	44.457	1.02
2	44.965	81.7	5	0.2721	55.543	1.159

Parent F2.3:



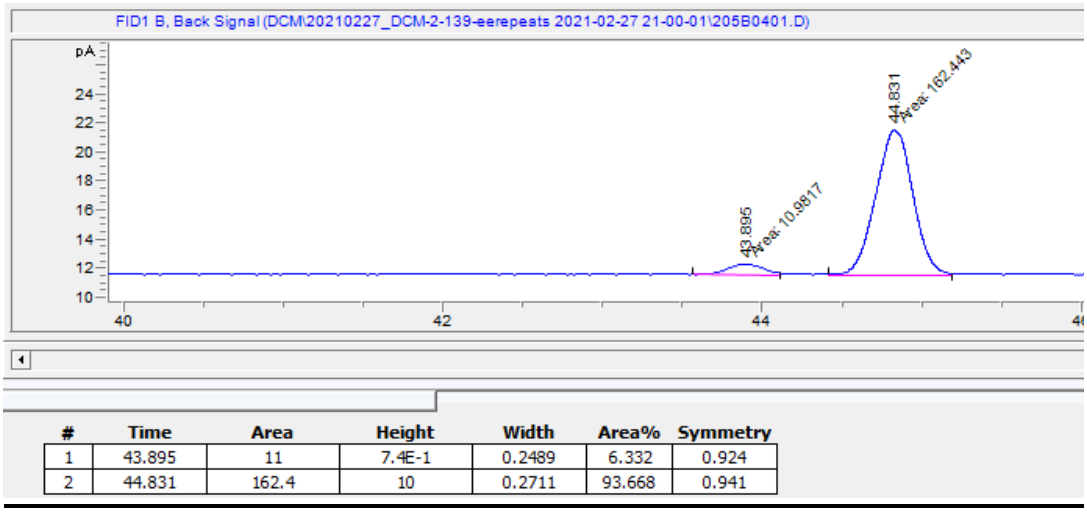
#	Time	Area	Height	Width	Area%	Symmetry
1	44.013	91.3	5.8	0.264	41.399	1.027
2	44.935	129.2	7.9	0.2724	58.601	0.983

Parent F2.4:



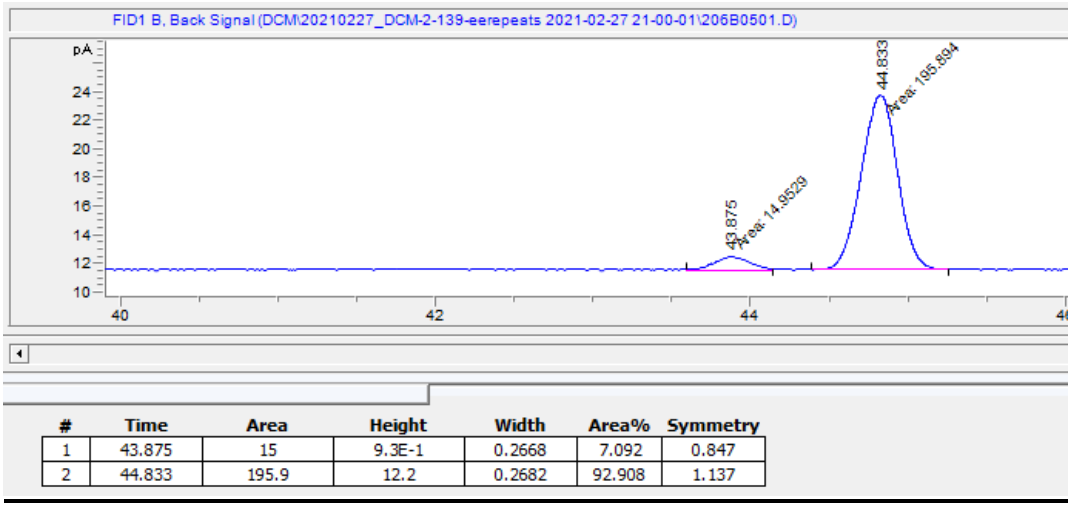
#	Time	Area	Height	Width	Area%	Symmetry
1	44.008	159	10.1	0.2617	38.071	1.079
2	44.94	258.6	16.1	0.2675	61.929	1.107

Parent F2.5:



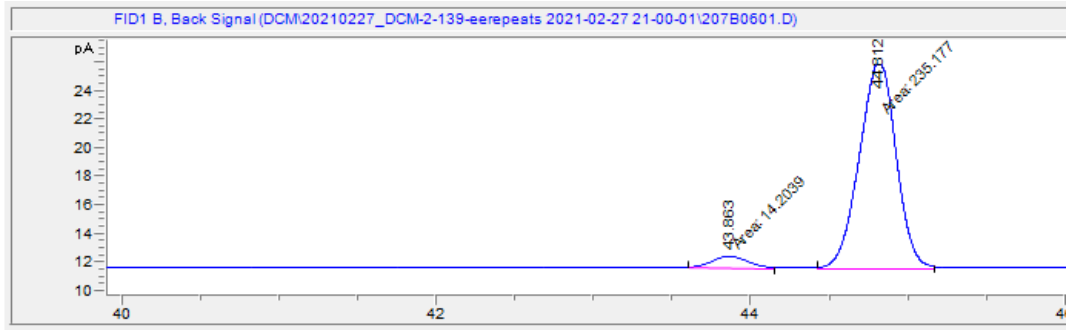
#	Time	Area	Height	Width	Area%	Symmetry
1	43.895	11	7.4E-1	0.2489	6.332	0.924
2	44.831	162.4	10	0.2711	93.668	0.941

Parent F2.6:



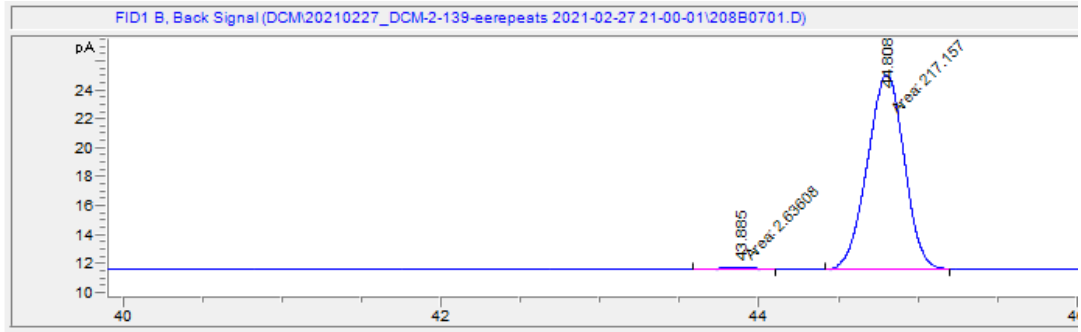
#	Time	Area	Height	Width	Area%	Symmetry
1	43.875	15	9.3E-1	0.2668	7.092	0.847
2	44.833	195.9	12.2	0.2682	92.908	1.137

Parent F2.7:



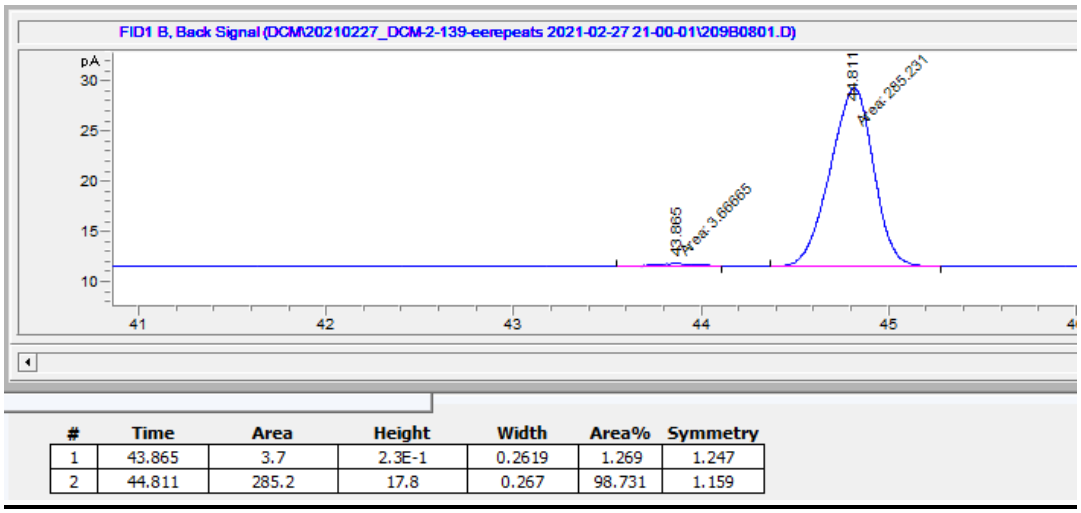
#	Time	Area	Height	Width	Area%	Symmetry
1	43.863	14.2	8.5E-1	0.2792	5.696	0.873
2	44.812	235.2	14.3	0.2739	94.304	0.956

Parent F2.8:



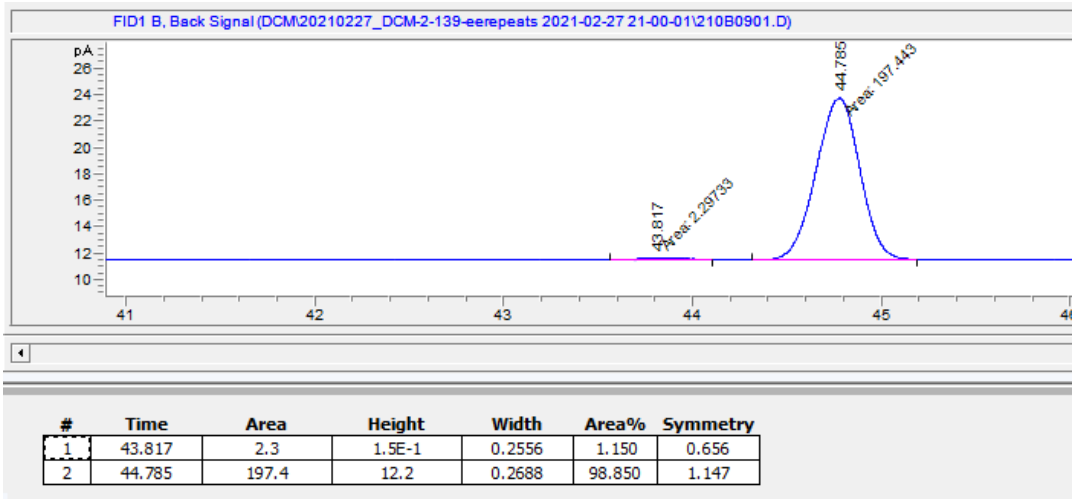
#	Time	Area	Height	Width	Area%	Symmetry
1	43.885	2.6	1.8E-1	0.2382	1.199	1.179
2	44.808	217.2	13.4	0.2702	98.801	1.068

Parent F2.9:



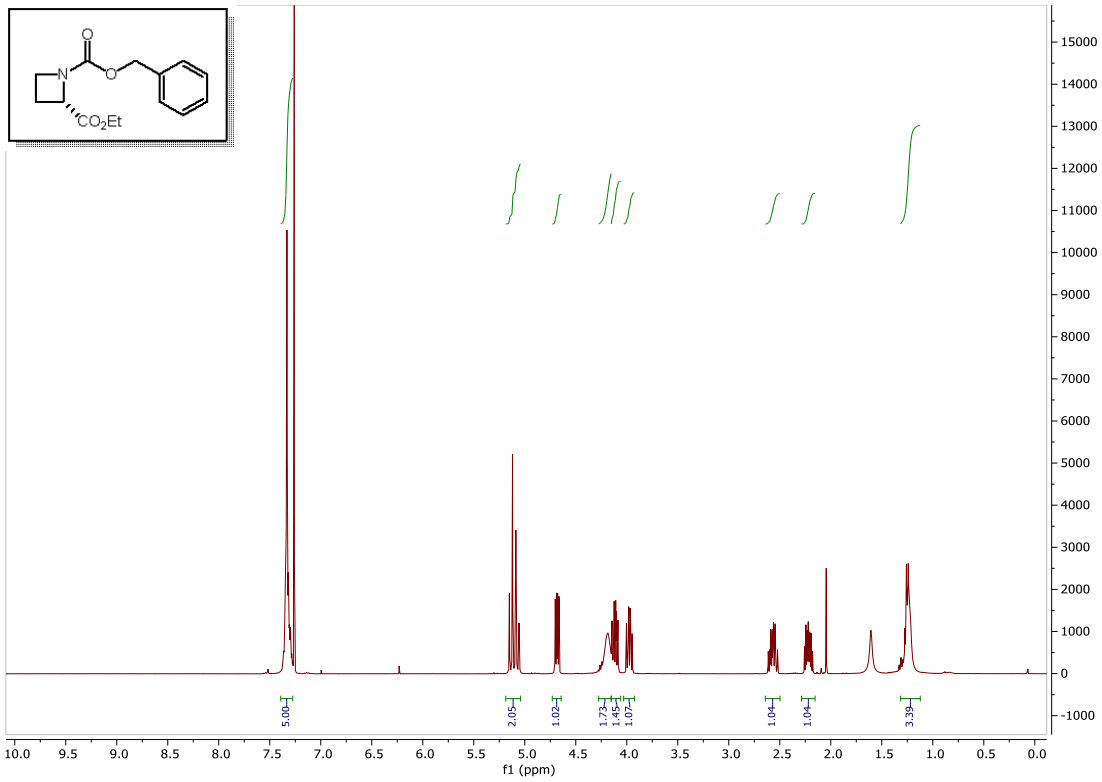
#	Time	Area	Height	Width	Area%	Symmetry
1	43.865	3.7	2.3E-1	0.2619	1.269	1.247
2	44.811	285.2	17.8	0.267	98.731	1.159

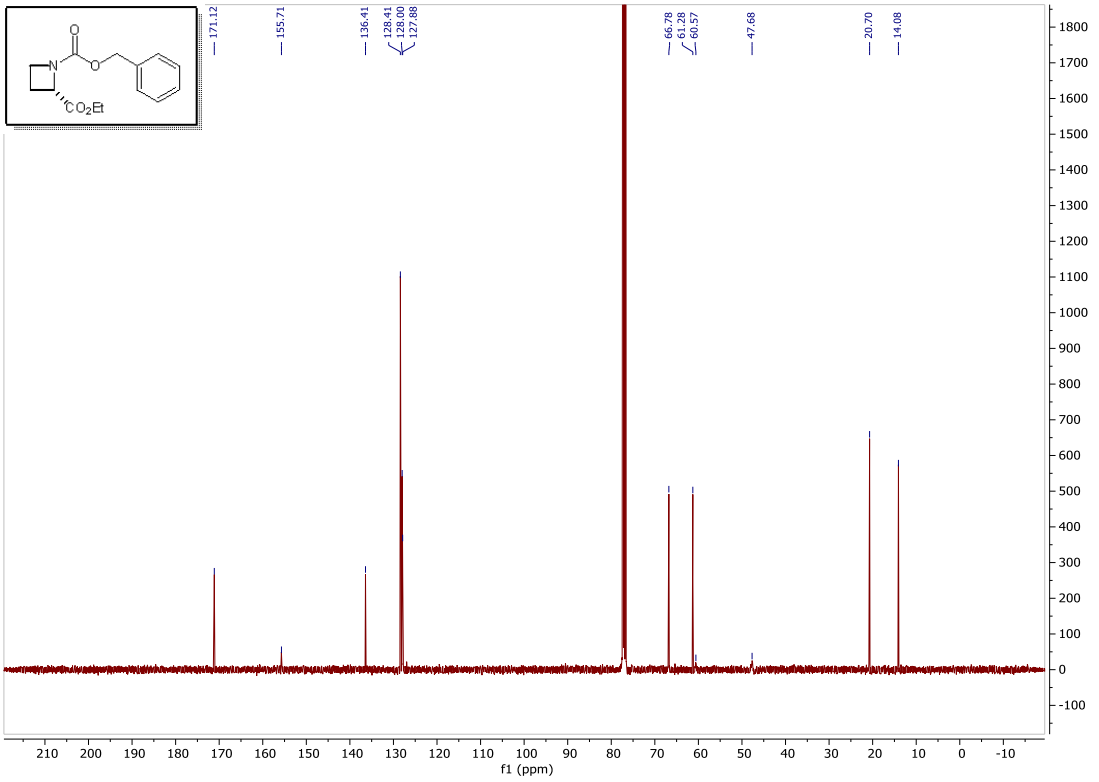
P411-AzetS:

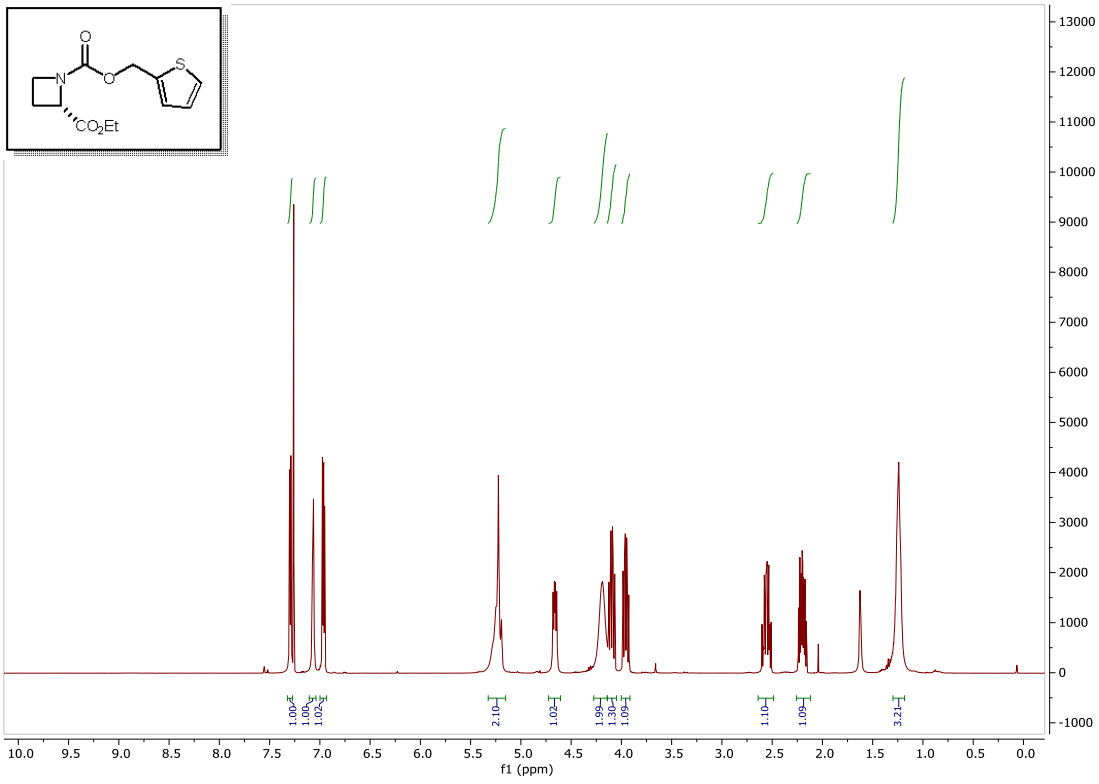


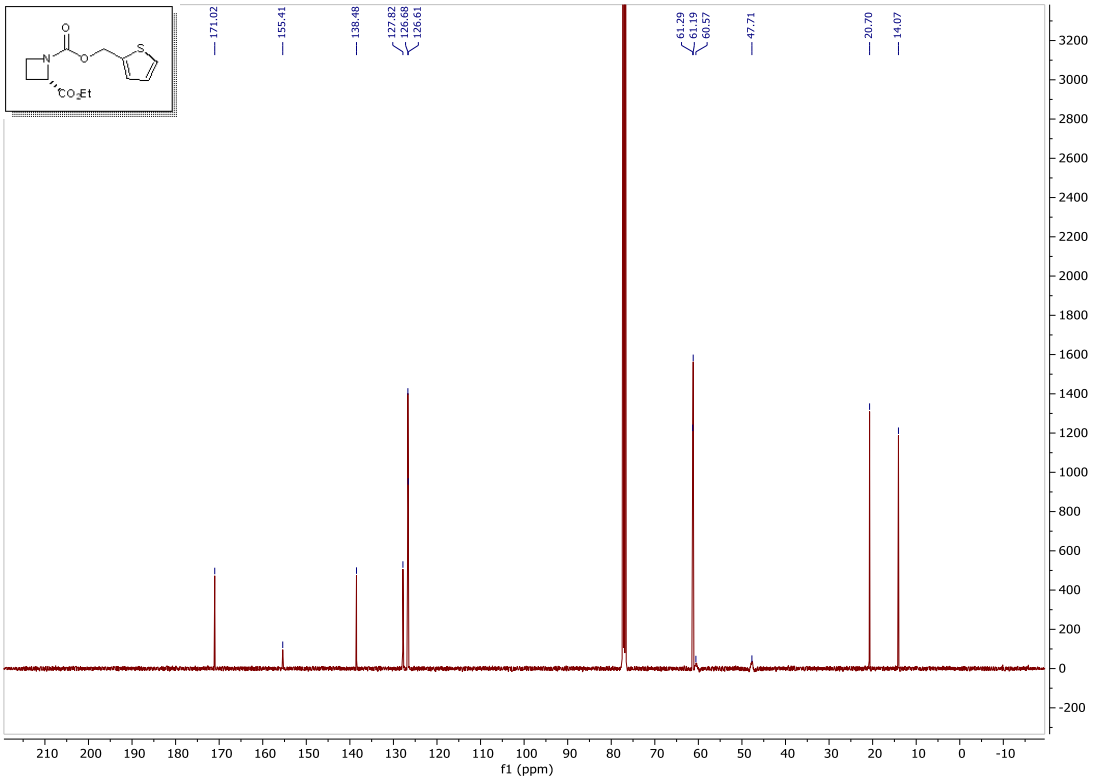
#	Time	Area	Height	Width	Area%	Symmetry
1	43.817	2.3	1.5E-1	0.2556	1.150	0.656
2	44.785	197.4	12.2	0.2688	98.850	1.147

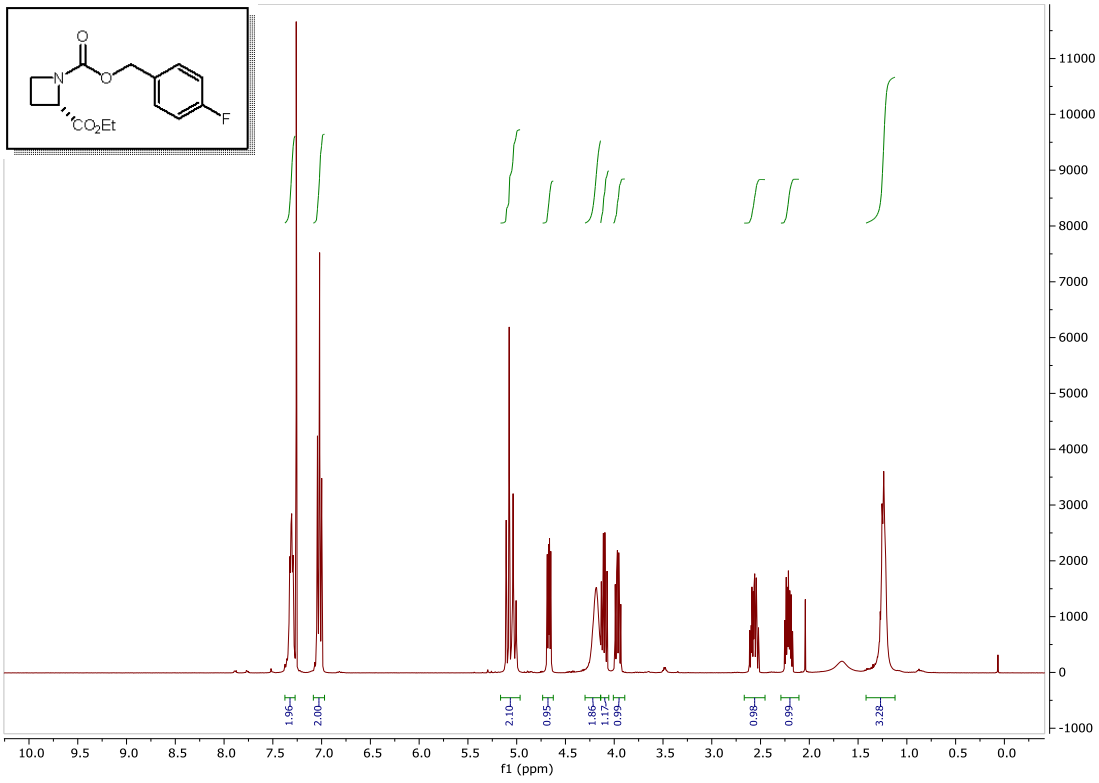
A.10. ^{1}H, ^{13}C, and ^{19}F NMR Spectra of Enzymatic Reaction Products

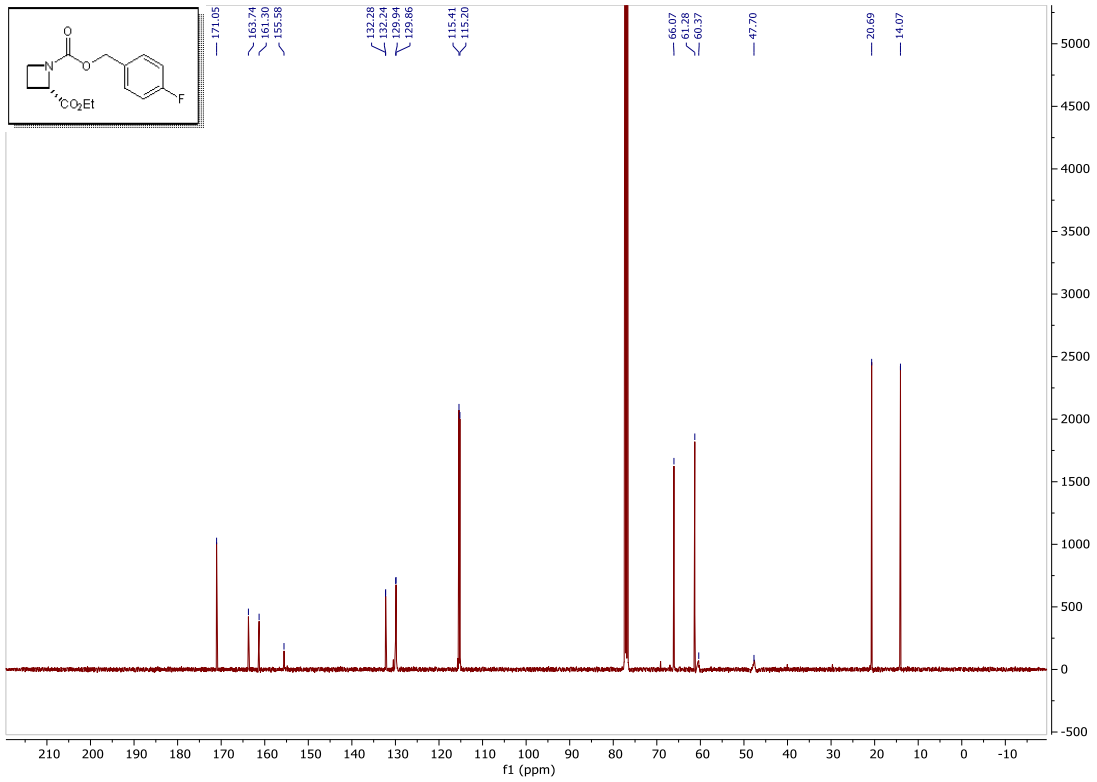


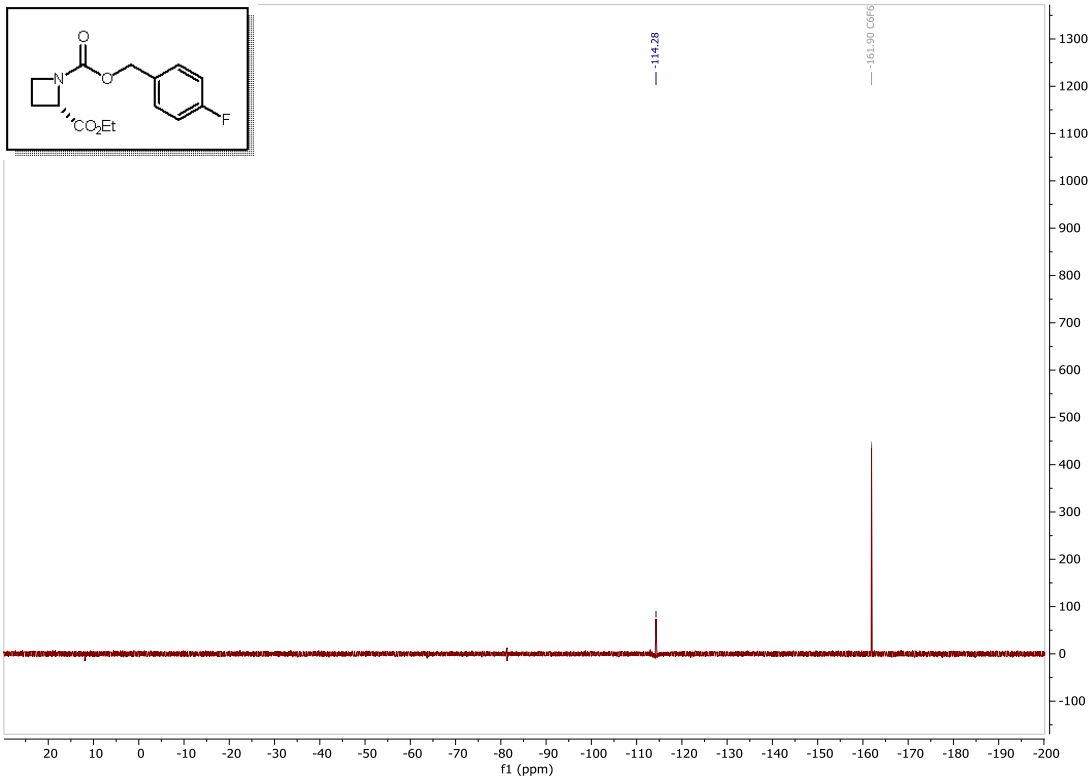

171.12
155.71
136.41
128.41
128.00
127.88
66.78
61.28
60.57
47.68
20.70
14.08
f1 (ppm)

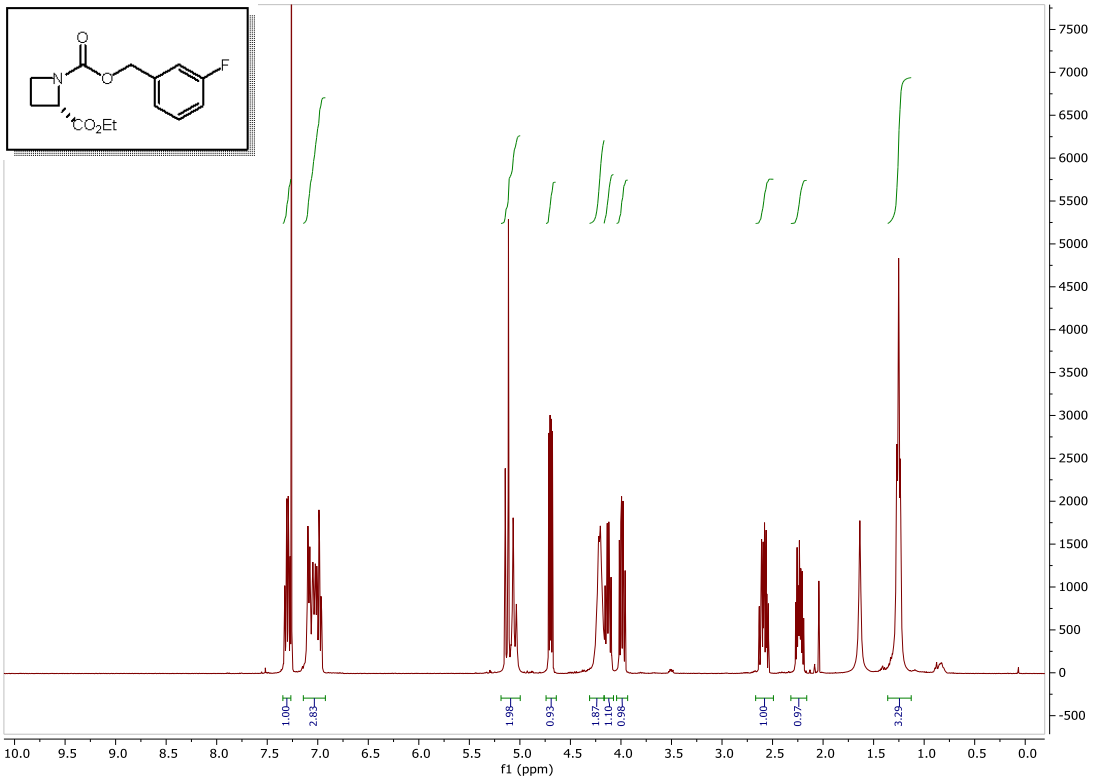

1.00
1.00
1.02
2.10
1.02
1.99
1.30
1.09
1.10
1.09
3.21
f1 (ppm)

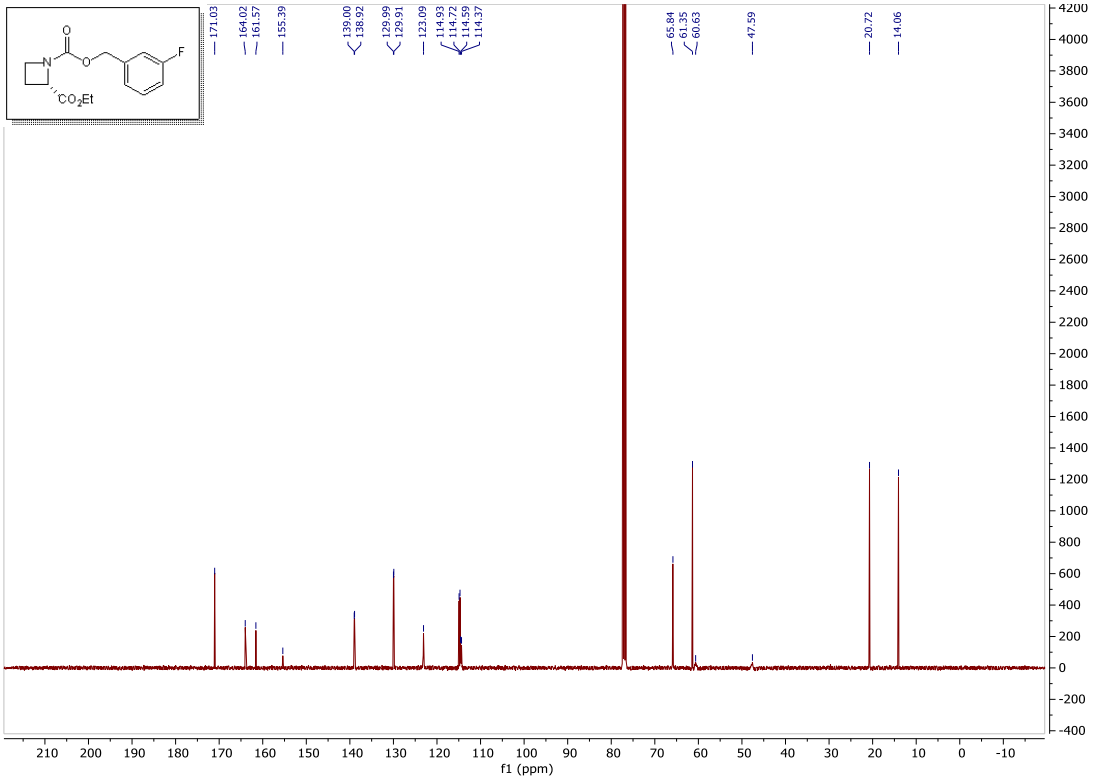

171.02
155.41
138.48
127.82
126.68
126.61
61.29
61.19
60.57
47.71
20.70
14.07
CO2Et
f1 (ppm)

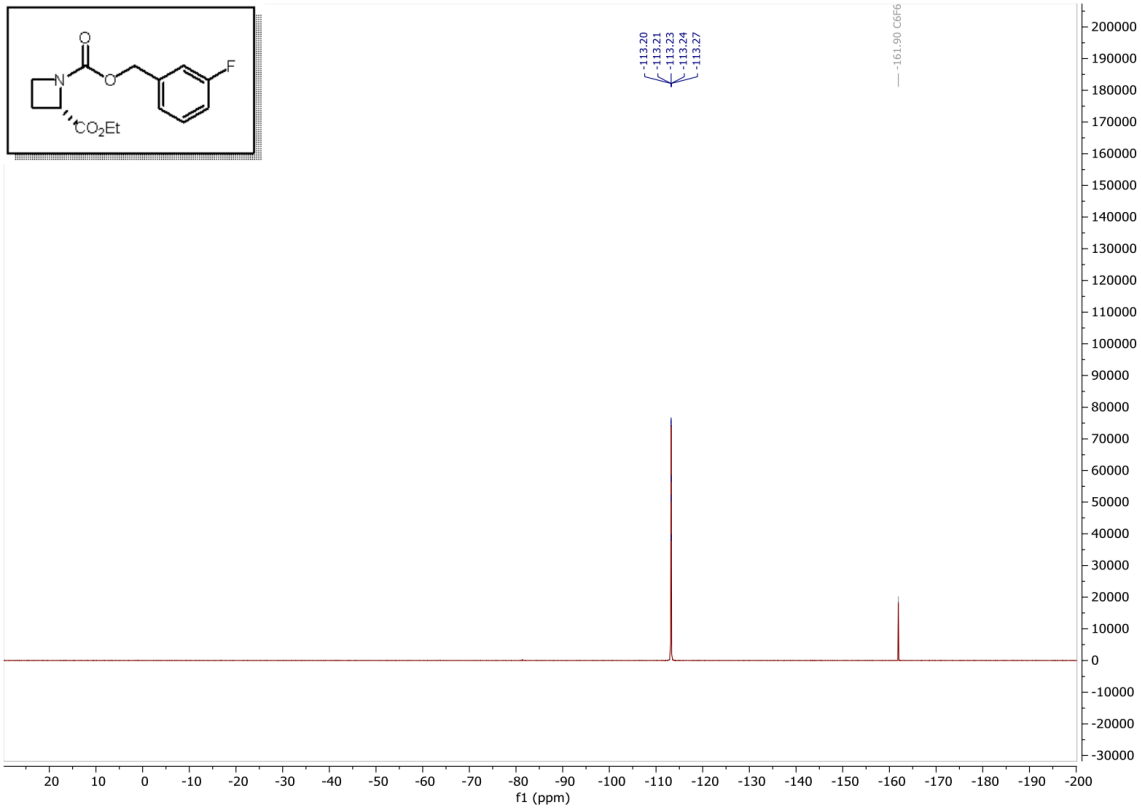

CO2Et
F
1.96
2.00
2.10
0.95
1.86
1.12
0.99
0.98
0.99
3.28
f1 (ppm)

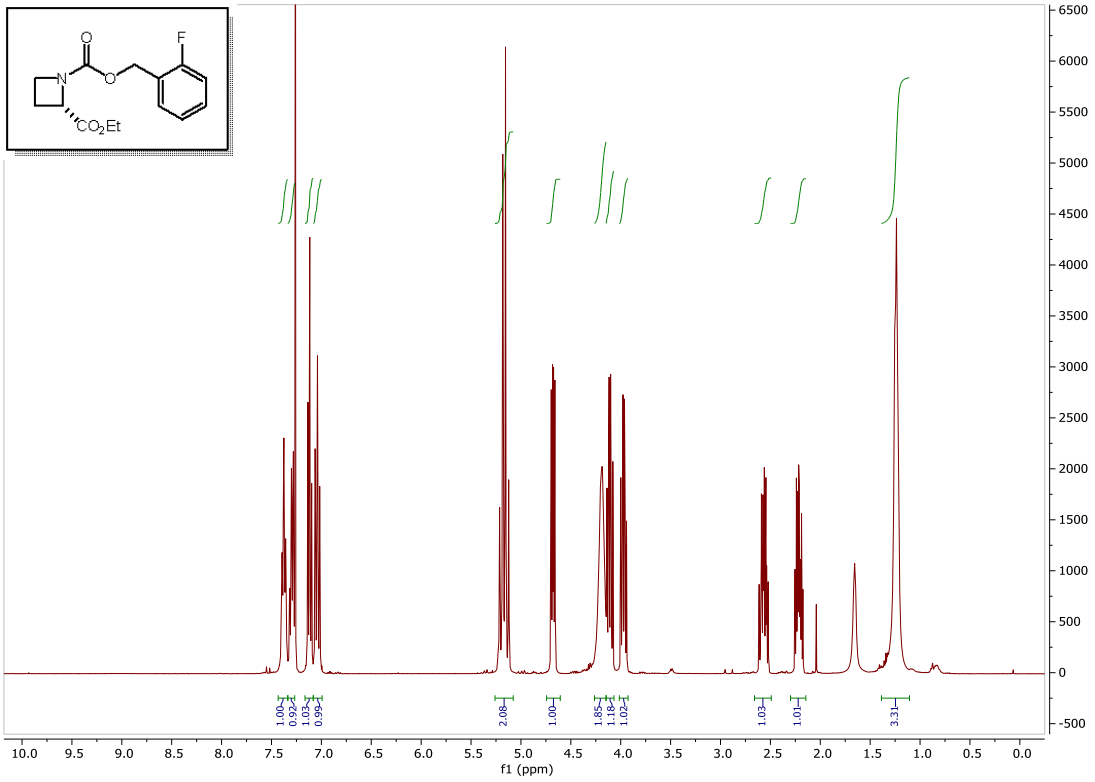

171.05
163.74
161.30
155.58
132.28
132.24
129.94
129.86
115.41
115.20
66.07
61.28
60.37
47.70
20.69
14.07
CO2Et
F
f1 (ppm)

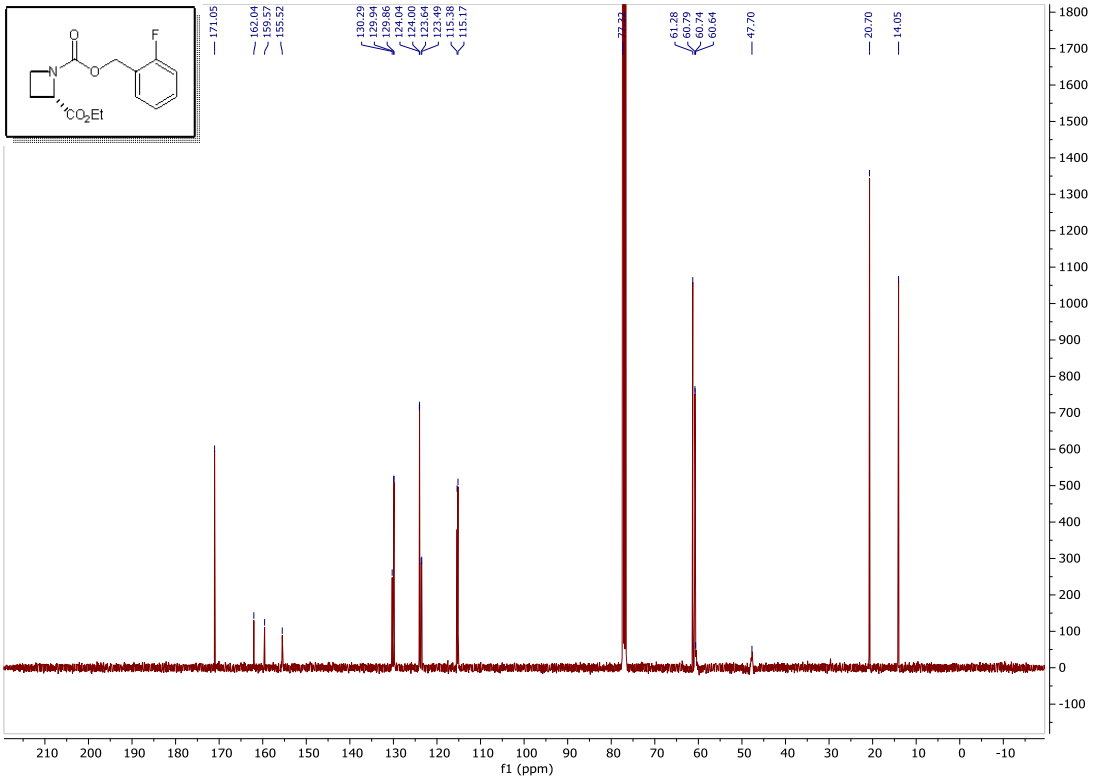

-114.28
-161.90 C6F6
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
-100
20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190 -200
f1 (ppm)

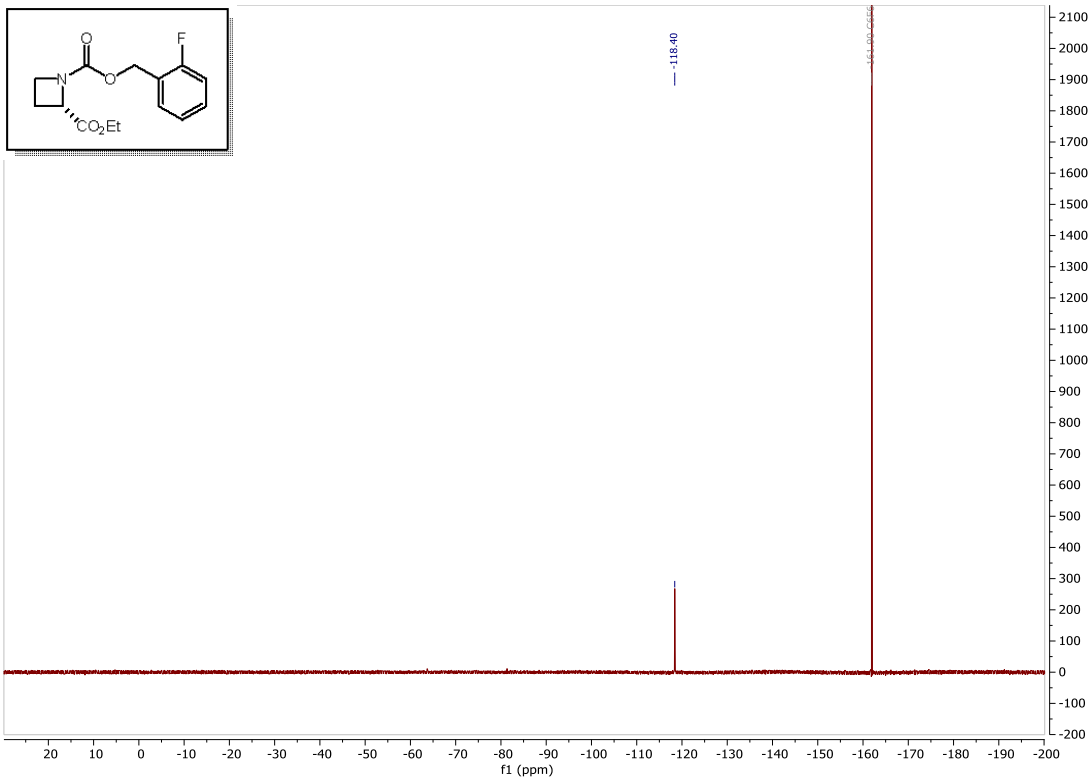

O
N
O
F
CO2Et
1.00
2.83
1.98
0.93
1.87
1.10
0.98
1.00
0.97
3.29
f1 (ppm)

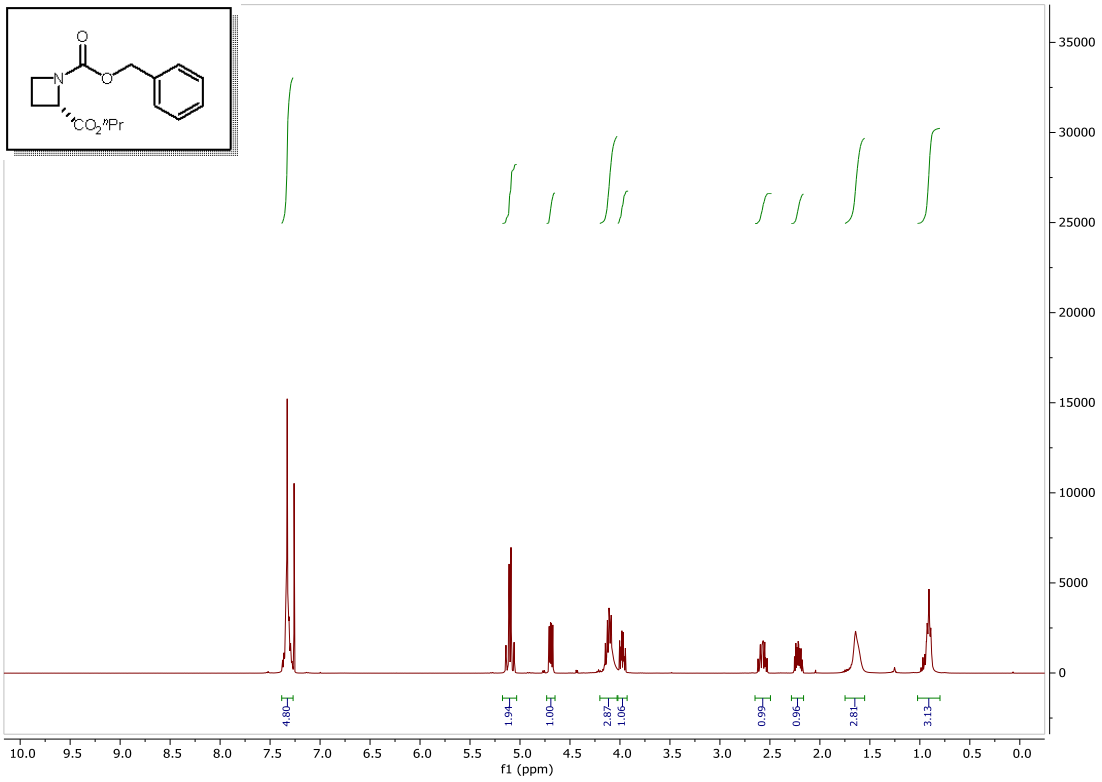

O
N
O
F
CO2Et
171.03
164.02
161.57
155.39
139.00
138.92
129.99
129.91
123.09
114.93
114.72
114.59
114.37
65.84
61.35
60.63
47.59
20.72
14.06
f1 (ppm)


-113.20
-113.21
-113.23
-113.24
-113.27
-161.90 C6F6
200000
190000
180000
170000
160000
150000
140000
130000
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
-10000
-20000
-30000
20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190 -200
f1 (ppm)

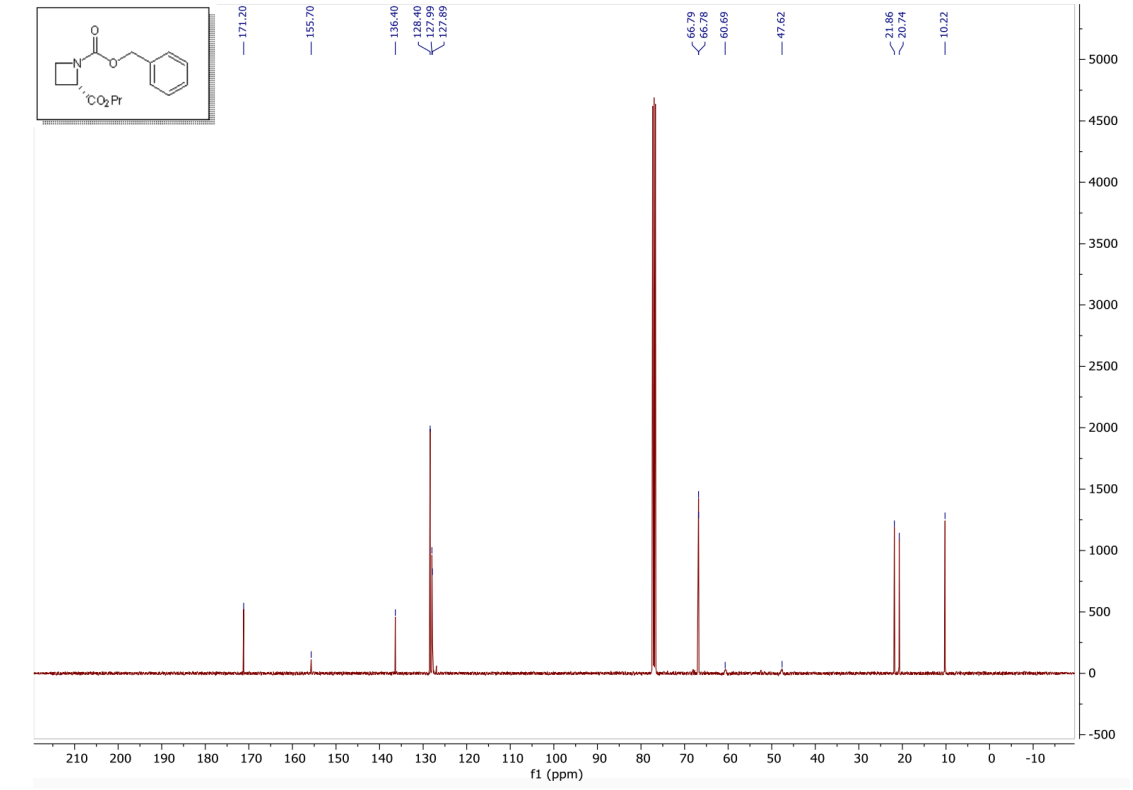

O
F
N
O
CO2Et
1.00
0.92
1.03
0.99
2.08
1.00
1.85
1.18
1.02
1.03
1.01
3.31
10.0 9.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0
f1 (ppm)
6500 6000 5500 5000 4500 4000 3500 3000 2500 2000 1500 1000 500 0 -500

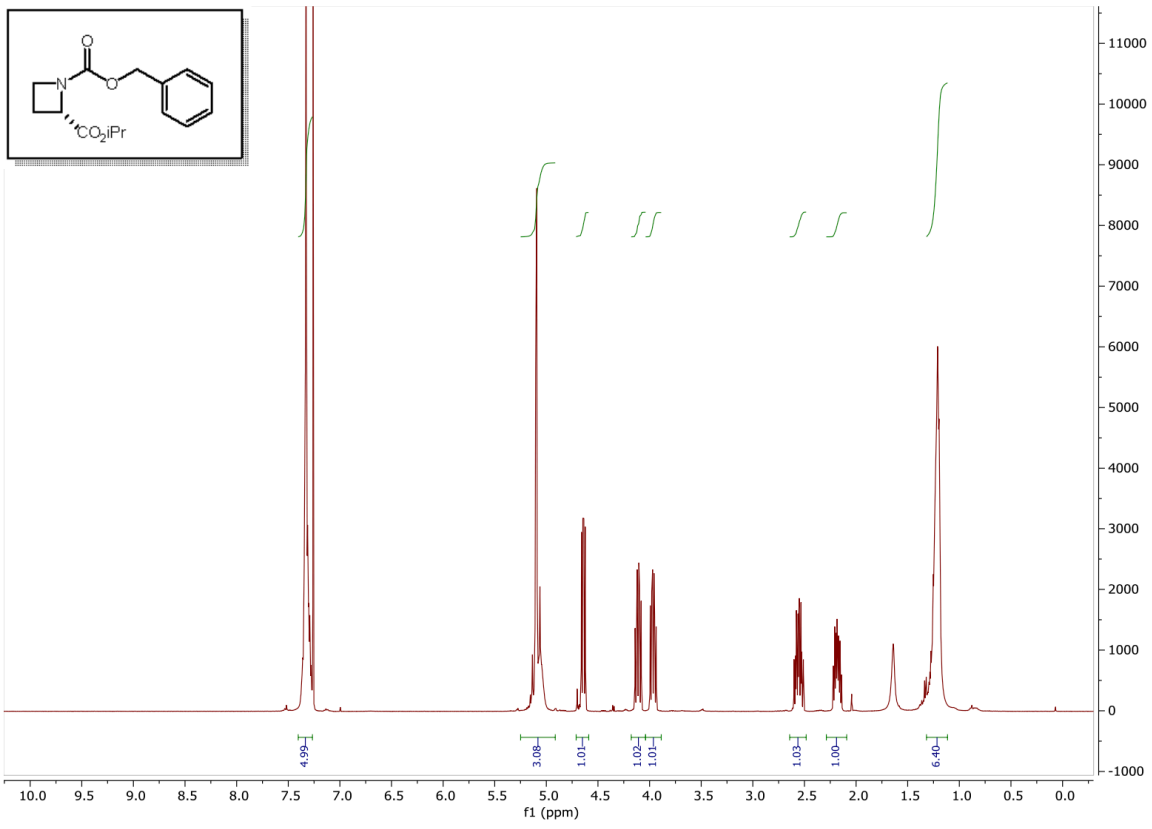

O
F
N
O
CO2Et
171.05
162.04
159.57
155.52
130.29
129.94
129.86
124.04
124.00
123.64
123.49
115.38
115.17
77.22
61.28
60.79
60.74
60.64
47.70
20.70
14.05
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10
f1 (ppm)
1800 1700 1600 1500 1400 1300 1200 1100 1000 900 800 700 600 500 400 300 200 100 0 -100

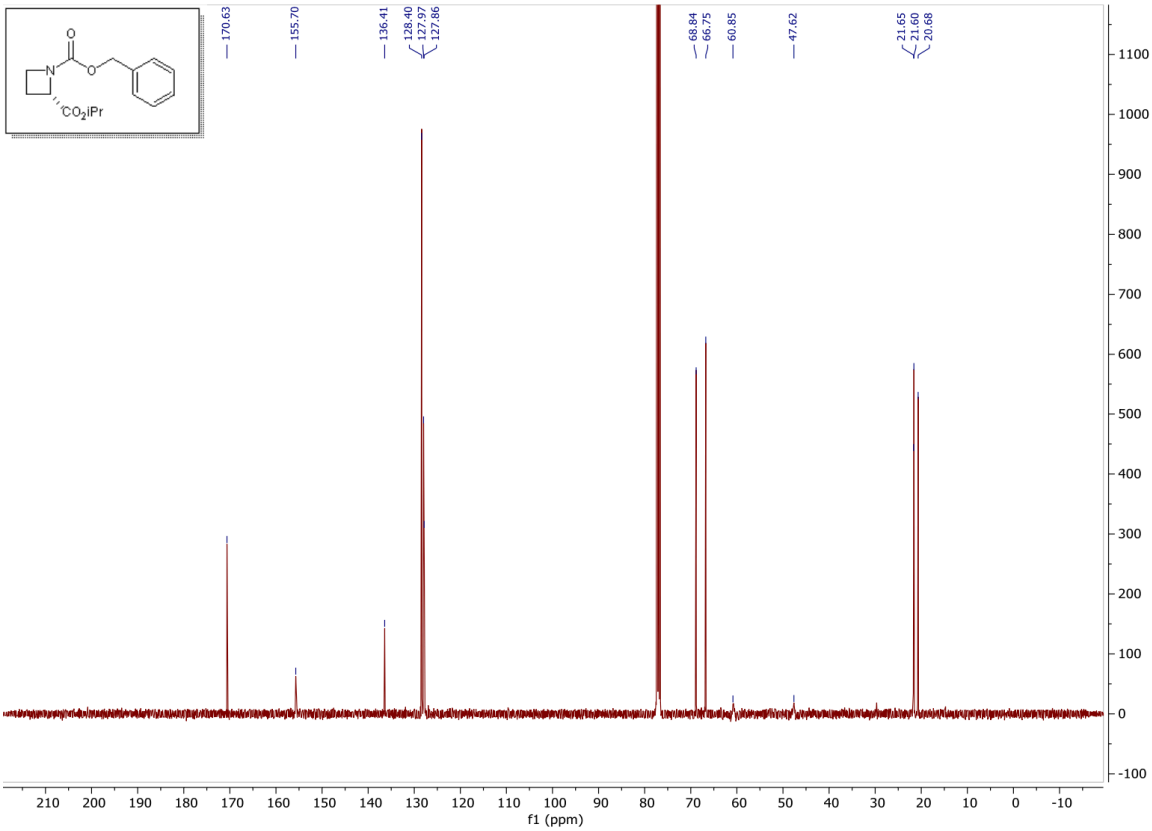

O
F
N
O
CO2Et
-118.40
-161.80
2100
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
-100
-200
20 10 0 -10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190 -200
f1 (ppm)

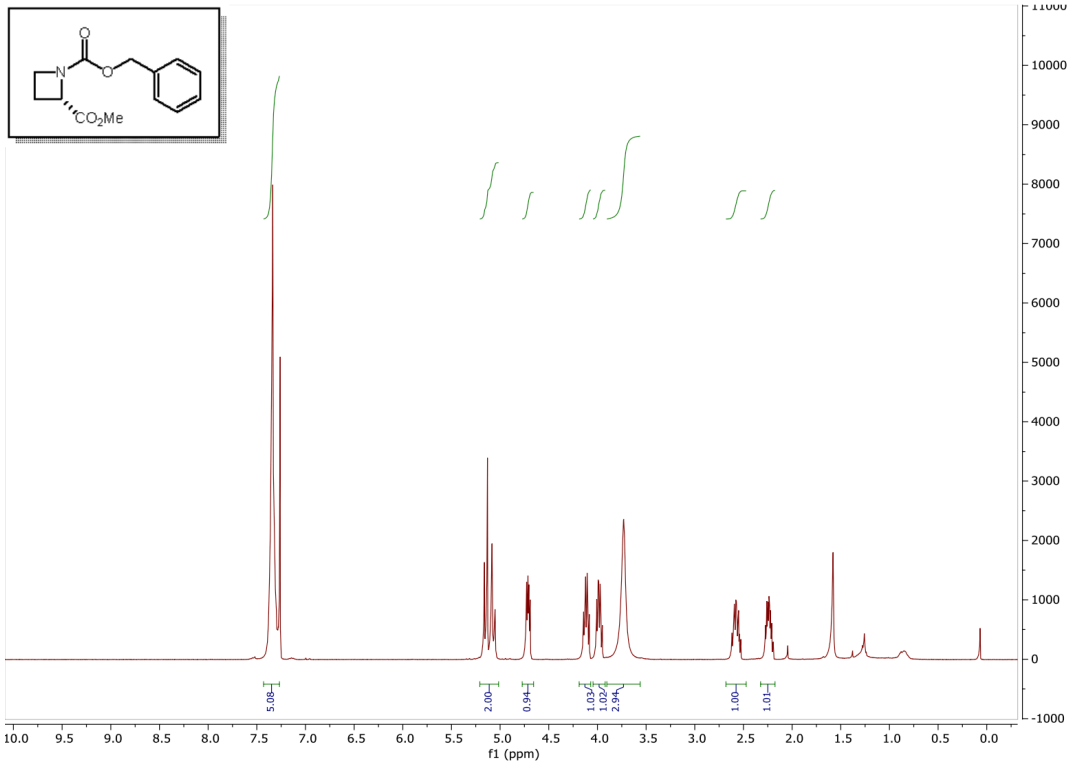


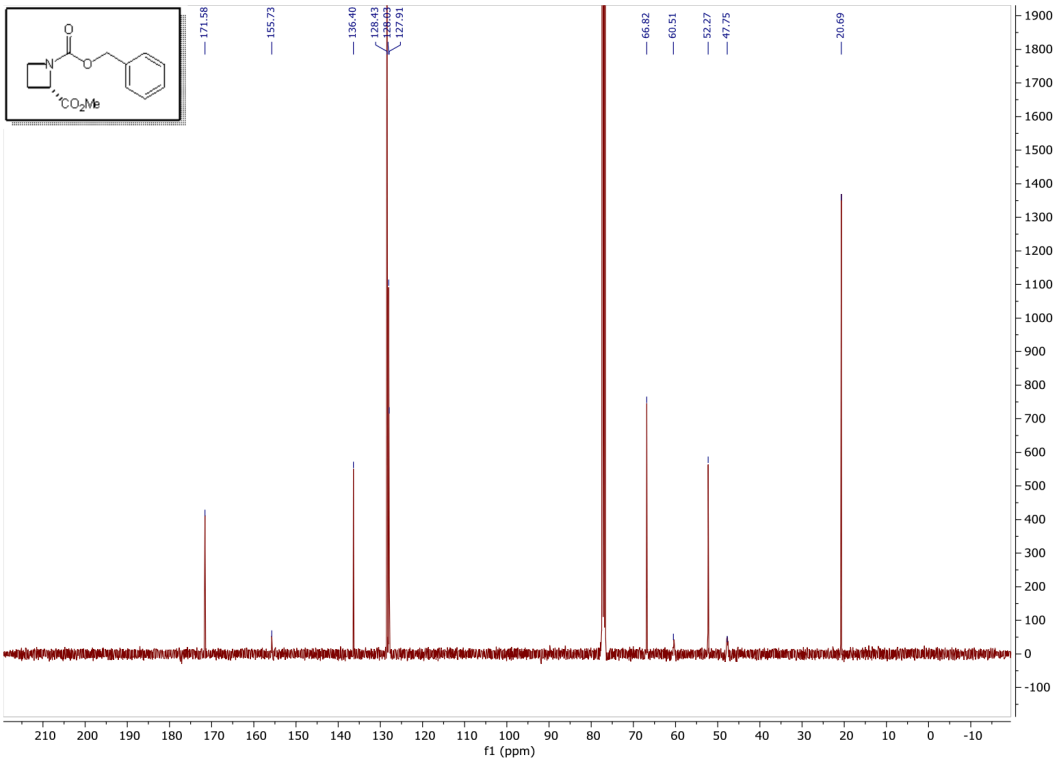
(N.B. methylene overlaps with HOD peak in 1H NMR, skewing integration)


171.20
155.70
136.40
128.40
127.99
127.89
66.79
66.78
60.69
47.62
21.86
20.74
10.22
CO2Pr
f1 (ppm)


CO2iPr
4.99
3.08
1.01
1.02
1.01
1.03
1.00
6.40
f1 (ppm)


170.63
155.70
136.41
128.40
127.97
127.86
68.84
66.75
60.85
47.62
21.65
21.60
20.68
CO2iPr
f1 (ppm)


CO2Me
5.08
2.00
0.94
1.03
1.02
2.94
1.00
1.01
f1 (ppm)


171.58
155.73
136.40
128.43
128.03
127.91
66.82
60.51
52.27
47.75
20.69
CO2Me
f1 (ppm)

A.11. References

1. Still, W.C.; Kahn, M.; Mitra, A. Rapid chromatographic technique for preparative separations with moderate resolution. *J. Org. Chem.* **1978**, *43*, 2923.
2. Sambrook, J.; Russel, D. Transformation of *E. coli* by Electroporation. *Cold Spring Harb. Protoc.* **2006**. Doi: 10.1101/pdb.prot3933.
3. 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.* **2013**, *2*, 83.
4. Gibson, D.G.; Young, L.; Chuang, R.-Y.; Venter, J.C.; Hutchinson, C.A.; Smith, H.O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* **2009**, *6*, 343.
5. Coehlo, P.S.; Wang, Z.J.; Ener, M.E.; Baril, S.A.; Kannan, A.; Arnold, F.H.; Brustad, E.M. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **2013**, *9*, 485.
6. Boddy, A.J.; Affron, D.P.; Cordier, C.J.; Rivers, E.L.; Spivey, A.C.; Bull, J.A. Rapid Assembly of Saturated Nitrogen Heterocycles in One-Pot: Diazo-Heterocyle "Stitching" by N–H Insertion and Cyclization. *Angew. Chem. Int. Ed.* **2019**, *58*, 1458.
7. Moore, E.G.; Xu, J.; Jocher, C.J.; Corneillie, T.M.; Raymond, K.N. Eu(III) complexes of functionalized octadentate 1-hydroxypyridin-2-ones: stability, bioconjugation, and luminescence resonance energy transfer studies. *Inorg. Chem.* **2010**, *49*, 9928.
8. Marlin, F.J. Stereoselective synthesis of orthogonally protected 2,3-disubstituted morpholines using a base-catalysed cascade reaction. *Tetrahedron Letters*, **2017**, *58*, 3078.
9. Searle, N.E. Ethyl Diazoacetate. *Org. Synth.* **1956**, *36*, 25.
10. Carminati, D.M. Intrieri, D.; Caselli, A.; Le Gac, S.; Boitrel, B.; Toma, L.; Legnani, L.; Gallo, E. Designing 'Totem' C_2-Symmetrical Iron Porphyrin Catalysts for Stereoselective Cyclopropanations. *Chem. -Eur. J.* **2016**, *22*, 13599.
11. Myhre, P.C.; Maxey, C.T.; Bebout, D.C.; Swedberg, S.H.; Petersen, B.L. Precursors to carbon-13-labeled reactive intermediates: preparation and NMR characterization of two doubly labeled isomers of methyl cyclopropene-3-carboxylate. *J. Org. Chem.* **1990**, *55*, 3417.
12. Hutchings-Goetz, L.S.; Yang, C.; Fyfe, J.W.B.; Snaddon, T.N. Enantioselective Syntheses of Strychnos and Chelidonium Alkaloids through Regio- and Stereocontrolled Cooperative Catalysis. *Angew. Chem. Int. Ed.* **2020**, *59*, 17556.
13. Tang, Y.; Dong, X.; Vennerstrom, J.L. The Reaction of Carbonyldiimidazole with Alcohols to Form Carbamates and *N*-Alkylimidazoles. *Synthesis*, **2004**, *15*, 2540.
14. McBurney, R.T.; Eisenschmidt, A.; Slawin, A.M.Z.; Walton, J.C. Rapid and selective *spiro*-cyclisations of *O*-centered radicals onto aromatic acceptors. *Chem. Sci.* **2013**, *4*, 2028.
15. Klock, C.; Herrera, Z.; Albertelli, M.; Khosla, C. Discovery of Potent and Specific Dihydroisoxazole Inhibitors of Human Transglutaminase 2. *J. Med. Chem.* **2014**, *57*, 9042

Chapter III

SURVEYING THE SEQUENCE-FUNCTION SPACE OF PROTOGLOBINS FOR NOVEL CARBENE TRANSFER ACTIVITIES

R.G.L participated in the design and execution of the research, including molecular biology, enzymatic reactions, and chemical synthesis.

ABSTRACT

This chapter describes efforts toward the discovery and evolution of new-to-nature carbene transfer activities in underexplored hemoprotein scaffolds. In particular, protoglobins, a family of archaeal globins, were studied for their ability to perform ring expansions of azetidines to generate proline analogs. Initially, protoglobins were screened for activity in which a one-carbon ring expansion is performed on *N*-benzylazetidine with ethyl diazoacetate as a carbene precursor. Though we were unable to find ring expansion activity with this system, we did discover a serendipitous side reaction that generates a quaternary ammonium species via protonation of an intermediate carbene-derived ylide. This activity was used to explore the protoglobin scaffold by generating a sequence-function dataset which tied changes in activity of this *N*-insertion reaction to mutations introduced by error-prone mutagenesis. Finally, several other heterocyclic compounds were explored as substrates for ring expansion activity; however, no desired activities were observed upon screening the protoglobins. In the course of this work several sites were identified in the protoglobin active site which proved to be impactful for other non-native activities. Additionally, in investigations of this biocatalytic system, we found that the behavior of heterologously expressed protoglobins in *Escherichia coli* (*E. coli*) is atypical, with certain variants expressing in the absence of the inductant isopropyl-β-D-1-thiogalactopyranoside (IPTG).

3.1. Introduction

Pyrrolidines are a privileged heterocyclic moiety in fine chemical synthesis. Of the 320 small molecule drugs approved by the FDA from 2013–2023, 13% contain a pyrrolidine ring.[1] Nearly half of these compounds feature a 2-carboxylate substitution, bearing a resemblance to the canonical amino acid proline (**Figure 3-1A**). When considering the synthesis of proline analogs, selectivity toward the *(S)*-configuration at the α-amino position is desirable, as these products would mimic the biological amino acid's natural chirality. Such proline analogs have garnered interest for their potential use as noncanonical amino acids in small molecule and peptide products.[2] The addition of substitutions in the pyrrolidine backbone of prolines can greatly alter the physicochemical properties of peptide compounds in which they have been incorporated.[3] For instance, introduction of a fluorine atom at the 4-position of proline (**FPro**) can influence the basicity of the ring amine, accelerate amide bond rotation, and modulate the preference for the *s-trans* and *s-cis* states.[4] Additionally, silaproline (**SiPro**) has been shown to greatly enhance bioavailability of a peptide product by improving chain lipophilicity.[5-6] Furthermore, it has been demonstrated that proline substitutions drive pyrrolidine puckering in the context of protein folding.[7-8] This effect has been utilized to modify protein dynamics and stability through global incorporation in protein sequences.[9-10]

Despite the value of proline analogs in biochemical contexts, there is a dearth of general routes for their enantioselective synthesis. As an example, to date, there has only been one reported method for the production of silaproline in a stereoselective fashion.[11] Recently, we developed a biocatalytic reaction for the one-carbon ring expansion of aziridines to furnish enantioenriched azetidine compounds using the engineered cytochrome P450 variant P411-AzetS (**Chapter 2**).[12] This work served to demonstrate that enzymatic carbene transfer is capable of catalyzing transformations of ammonium ylides with unprecedented stereocontrol when compared with small-molecule catalysts. In contrast to the ylide-derived ring expansion of aziridines, for which there was no literature precedent, carbene-based ring expansions of azetidine compounds with small molecule catalysts are well studied. Notably, when a tertiary azetidine is treated with a diazoacetate species to facilitate ylide formation, the resulting ring-expansion products are analogs of the

canonical amino acid proline. West and coworkers previously described a copper-catalyzed variation of this reaction, in which *N*-substituted azetidine compounds are expanded by one carbon with diazoacetates or diazoacetophenones toward the synthesis of highly substituted prolines.[13] This method struggles to effect stereocontrol over the diradical intermediates generated in the course of the [1,2]-Stevens rearrangement, resulting in poor stereoenrichment.

We envisioned that biocatalytic ring expansion could be extended to the synthesis of prolines in an enantioselective manner (**Figure 3-1C**). Not only would this method serve to deliver high-value pyrrolidine derivatives from achiral starting materials, it would also show that enzymatic control of ammonium ylides can be generalized to non-aziridinium systems. The discovery and engineering of a toolbox of ring expansion-competent enzymes would establish biocatalysis as a privileged reaction mode for the synthesis of heterocyclic compounds with high stereo- and chemoselectivity. Furthermore, as P411-AzetS can be utilized to produce azetidine compounds, the development of a downstream enzyme, which can further process these strained rings, would provide an opportunity to synthesize highly elaborated alkaloid compounds in a sequential manner via a biosynthetic cascade (**Figure 3-1D**). Keasling and coworkers recently reported the introduction of iridium-containing cytochromes P450 into *Saccharomyces cerevisiae* for the *in vivo* cyclopropanation of terpenoids[14] and endogenously synthesized styrene.[15] The sequential expansion of nitrogen heterocycles would be the first demonstration of multiple heterologously expressed carbene transfer enzymes working in sequence. Excited by this prospect, we sought to advance our existing methodology by engineering hemoproteins for the one-carbon ring expansion of azetidines to deliver substituted pyrrolidine species.

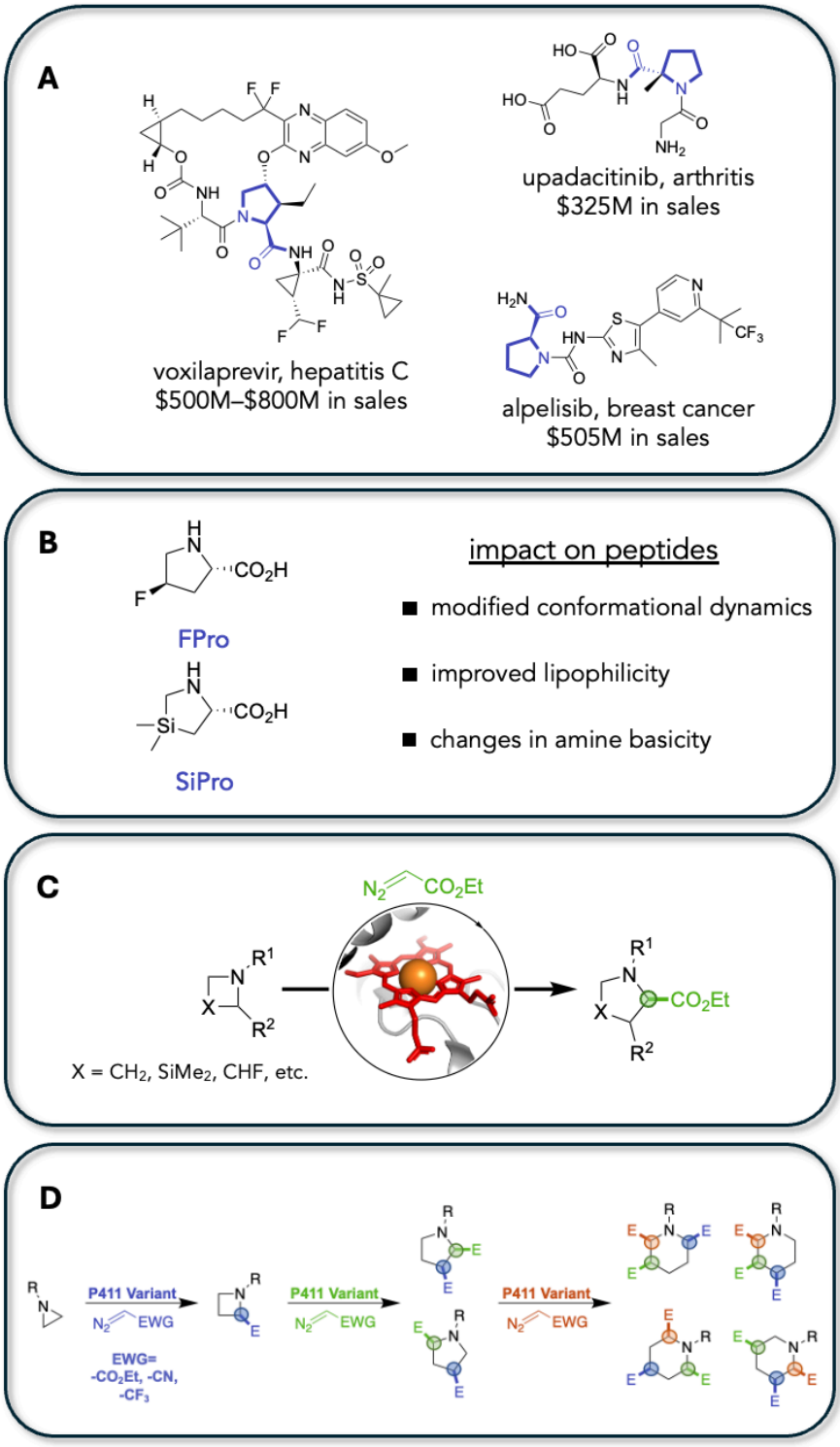


Figure 3-1. (A) Proline is a privileged heterocycle in the design of therapeutic small molecules. **(B)** Backbone modifications to the proline scaffold can improve the folding and pharmacokinetic properties of the peptides in which they are incorporated. **(C)** Enzymatic, enantioselective ring expansion of azetidines would all for the synthesis of proline

homologs. **(D)** Assembly of a multistep P411 pathway would provide access to functionalized pyrrolidine and piperidine cores from simple starting materials.

Here, we report our efforts to identify and engineer a hemoprotein capable of facilitating the ring expansion of *N*-benzylazetidine **1** with ethyl diazoacetate (EDA) as a carbene source (**Figure 3-2A**). While assaying whole-cell *Escherichia coli* (*E. coli*) suspensions harboring heterologously expressed hemoproteins, we observed what seemed to be ring expansion activity with several engineered cytochromes P411 and protoglobins. After electing to proceed with the highest performing protoglobin variant, we initiated a directed evolution campaign to improve reaction yield and stereoselectivity. To bolster our engineering efforts from an early stage, we collected a sequence-function dataset for a mutant library of this protoglobin scaffold using next-generation sequencing (NGS) and high-throughput screening methods. However, in the course of generating this dataset, we found that the signal previously believed to correspond to the desired proline product **2** actually corresponded to the azetidinium **3**. Nevertheless, the sequence-function dataset which we generated for this reaction proved to be of value in the determination of residues in the protoglobin scaffold which are highly impactful in the evolution of non-natural activities.

3.2. Results and Discussion

3.2.1. Initial Activity Discovery

To commence our search, we screened a compilation of cytochromes P411 which had previously been engineered for carbene and nitrene chemistries.[16] Whole-cell reactions were analyzed using liquid chromatography-mass spectrometry (LC-MS) to detect and quantify product formation. To our surprise, nearly every cytochrome P411 variant displayed activity for a product which had m/z 234 corresponding to the [M+H]+ ion and eluted at the same time as **2** (**Figure B-3** of **Appendix B**). Of the nearly 180 enzymes screened, approximately one third demonstrated greater than trace yield (**Figure 3-2B**).

Contemporaneously, researchers in our lab were beginning to investigate globins originating from thermophilic archaea for their ability to catalyze carbene transfer reactions. These proteins, termed protoglobins, are believed to natively bind gaseous species based on *in vitro* experimentation.[17] Protoglobins present several promising characteristics for engineering hemoprotein derived new-to-nature chemistries relative to cytochromes P411. Firstly, protoglobins are highly thermostable proteins.[18] Therefore, they can putatively accept more destabilizing mutations which may benefit the desired activity before dropping below the threshold of instability.[19] After nearly a decade of directed evolution having been performed on the cytochrome $P450_{BM3}$ scaffold, late-stage P411 variants exhibit poor thermostabilities, with T_{50}s as low as 40 °C, making them poor candidates for further directed evolution efforts. Additionally, protoglobins are relatively small proteins; at a length of 200 amino acids, they are less than half the size of the heme domain of $P450_{BM3}$. This makes them a simpler system to study as well as better candidates for heterologous expression in *E. coli*.[20]

Protoglobins were initially tested in our lab for their ability to catalyze carbene-derived cyclopropanations, eventually demonstrating that engineered protoglobins could exhibit yields and stereoselectivities rivaling those of P411s.[21] Emboldened by these results and the fact that putative ring expansion appeared to be facile for P411s, we elected to examine protoglobins for their ability to catalyze ylide-based ring expansions. A panel of approximately 60 protoglobin variants, comprised of mutants of 12 homologs, was screened. Gratifyingly, we found that nearly every protoglobin was capable of generating the same product as the previously assayed P411s (**Figure 3-2C**). The top performing mutant (which we term PgC10-G1, for its location in the screened 96-well plate) originated from *Aeropyrum pernix* protoglobin (*Ape*Pgb) and contained the following mutations: Y60G, V89A, and F145N (**Figure B-1** of **Appendix B**).

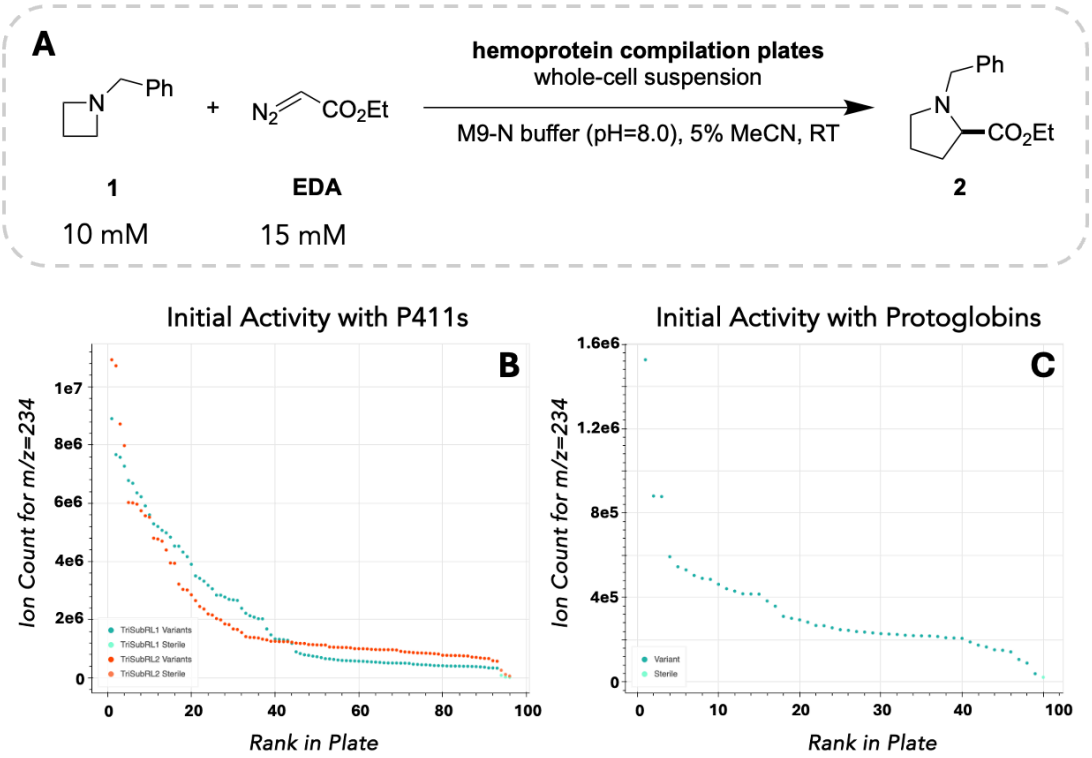


Figure 3-2. (A) Model conditions under which hemoproteins were screened for ring expansion activity. **(B)** Retention-of-function plot for two 96-well plates of cytochromes P411 for the production of prolines. Under the acidic LC-MS screening conditions, product **2** is protonated and exhibits m/z=234. **(C)** Retention-of-function plot for a compilation of protoglobins for the production of prolines.

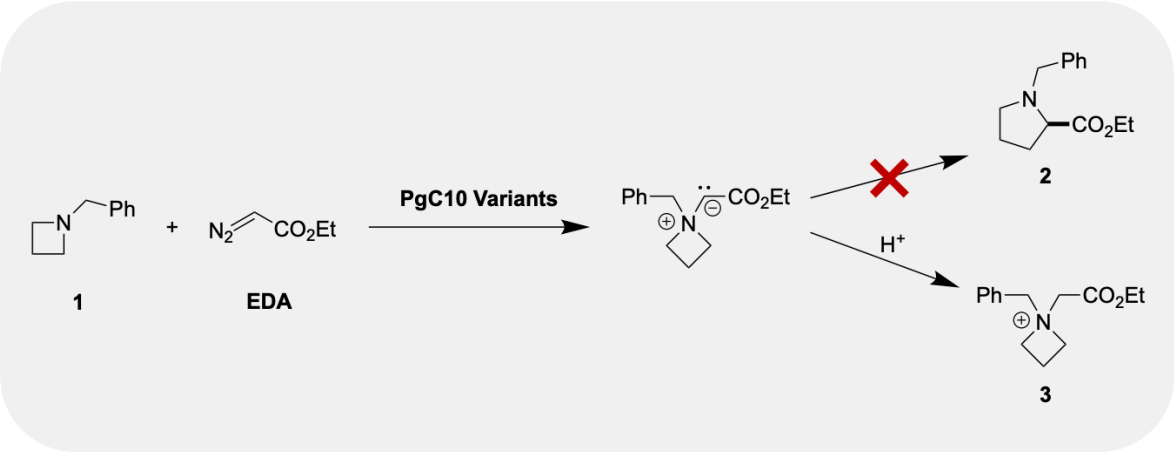
3.2.2. Azetidinium Formation is Favored Over Ring Expansion

Upon discovering what we believed was ring expansion chemistry with PgC10-G1, we proceeded to engineer this activity over two rounds of evolution to yield the variant PgC10-G3. The details of this directed evolution campaign are described in **section 3.2.3**. In testing of PgC10-G3, we soon came to find some peculiarities with the biocatalytic system. First, as our intent in this work was to establish a biocatalytic methodology for the synthesis of (*S*)-prolines, we were intent on determining the stereoselectivity of the PgC10 lineage. We developed a chiral gas chromatography (GC) method for separation of the enantiomers of proline **2**. To our surprise, despite repeated attempts, the reaction product was recalcitrant to extraction into solvents compatible with GC columns. While we made this finding early

in our evolutionary campaign, we ascribed this phenomenon to challenges associated with isolating tertiary amines from buffered systems.[22]

After arriving at PgC10-G3 we found that further evolution of this enzymatic activity was quite challenging. In the screening of eight libraries of single-site mutants we found that no variants improved product yield. With our suspicions raised, we reasoned that proline **2** could be acting as an inhibitor within the protoglobin active site. Complex amine products are well known to inhibit heme-binding proteins under physiological conditions.[23] To interrogate this possibility, we assessed the capability of PgC10-G3 to ring-expand azetidine **1** when the reaction is exogenously supplemented with authentic product **2**.[24] To our dismay, it became apparent after analysis of these spiked reactions that PgC10-G3 was not forming the desired ring-expanded species (**Figure B-5** of **Appendix B**). In the LC-MS traces of these reactions, it is apparent that the generated biocatalytic product is chemically distinct from **2**, albeit with the same observed m/z of 234. It seemed conceivable that under the protic conditions of aqueous buffer, ylide protonation to generate azetidinium **3** could be the dominant reaction pathway (**Scheme 3-1**). This cationic product would be expected to present the observed difficulties with extraction into organic solvents. While quantitative data on the half-lives of ammonium ylides are sparse in the literature, there is precedent for pendant carbonyl functionalities to stabilize these species, preventing rapid decomposition.[25] We speculated that the proximal ester moiety present in the intermediate zwitterion could extend its lifetime sufficiently for it to encounter and be quenched by an aqueous proton.

Scheme 3-1. Protonation of the intermediate ylide leads azetidinium **3** as the dominant product



3.2.3. Mapping the Protoglobin Fitness Landscape

Although we were disappointed with this result, we believed that the data generated in the course of evolving PgC10-G1 remained a valuable resource. As a note to the reader, for the remainder of this section we will refer to the formation of product **3**; however, throughout the collection of these data we were under the impression that **2** was being generated. At the outset of our investigation, the protoglobin scaffold was relatively understudied. In their early studies, Knight and coworkers queried mutations at positions 59, 73, 93, and 145 of *Ape*Pgb for improvements in carbene transfer activity as these residues had previously been confirmed to regulate the gas binding in the homologous *Methanosarcina acetivorans* protoglobin (*Ma*Pgb).[21,26] Concurrent with our studies, Porter *et al.* were engaged in a directed evolution campaign of *Ape*Pgb to functionalize diazirine species for the generation of iron-carbenoids at room temperature.[27]

Based upon the ubiquity of formation of **3** seen across hemoproteins under these conditions, we hypothesized that the observed activity would be difficult to eliminate. Interestingly, this activity is not observed in reactions with free heme (**Figure B-2** of **Appendix B**). As a result, we judged that this reaction provided an ideal system to explore the protoglobin backbone by generating a large collection of sequence-function data. To efficiently construct this dataset, we first developed an ultra-high-throughput LC-MS

method for detection of **3** using multiple injections in a single experimental run (MISER).[28] Importantly, Wittmann and coworkers had recently developed evSeq, a method for sequencing spatially separated variants in a protein library via next-generation sequencing (NGS), to collect the sequence of every screened enzyme.[29] In conjunction, the application of MISER and evSeq to mutants of PgC10-G1 enabled us to assess the sequence and activity of over 2000 protoglobin variants in under one month (**Figure 3-3A**). We were excited at the possibility of using this mapping of the protoglobin fitness landscape to accelerate this directed evolution campaign and future efforts through the training of machine learning algorithms.

As evSeq is designed for use with Illumina MiSeq instrumentation, the maximum possible sequencing length is 300 bp with 150-bp paired-end reads.[30] Thus, to fully investigate the 600-bp protoglobin sequence, we split the sequence into three 'tiles' to separately mutagenize and characterize in the native context of the rest of the protein. Each tile was a length of approximately 200 bp and broadly split PgC10-G1 into three regions: a variable region (so named for its variability among distinct protoglobins),[26] an active site region, and an interface region (so named for its presence at the interface of the protoglobin homodimer) (**Figure 3-3B**). We introduced mutations to each tile separately through the use of error-prone PCR and assessed the activities of the variants in each library (**Figure 3-4, Panels A–C**).

These data led to several findings relevant to the engineering of *Ape*Pgb. Within the variable region, we found that several residues between position 11 and 20 can positively impact activity. This is of note as this region seems to have no interaction with the heme cofactor based on the crystal structure of the homologous *Ma*Pgb. The amino acids in this range potentially influence protein expression and flexibility in the non-native *E. coli* cytosol. Another region of interest was presented by the residues found at the base of the helical bundle at the homodimer interface (positions 139–141). Previously, residues 145 and 149 were believed to be key to non-native activity as they are proximal to the heme ligand. However, these data indicate that disruptions along the whole homodimer interface can greatly impact activity.

Using these data, we assessed the activity of single-site mutagenesis (SSM) libraries of sites which demonstrated improved activity in the tile dataset. In two rounds of evolution with these sites we arrived at PgC10-G3 (PgC10-G1 C45G W59C C102D Q113R). PgC10-G3 catalyzes the formation of **3** approximately 6-fold more than PgC10-G1 (**Figure 3-4D**).

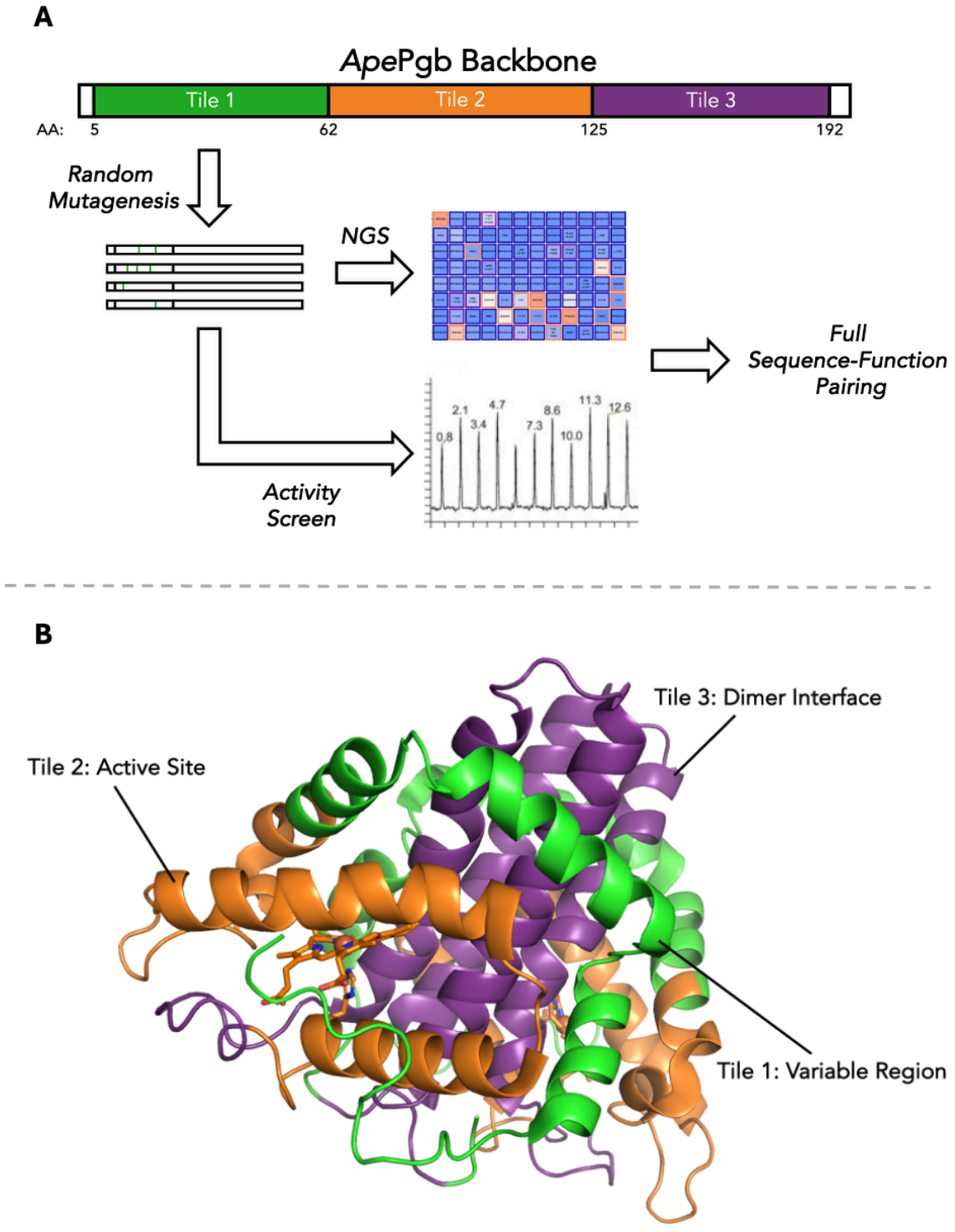


Figure 3-3. **(A)** Workflow for generating an extensive sequence-function dataset for a new-to-nature biocatalytic reaction. **(B)** The three "tiles" which are separately mutagenized correspond to functionally distinct regions of *Ape*Pgb (PDB: 2VEE).

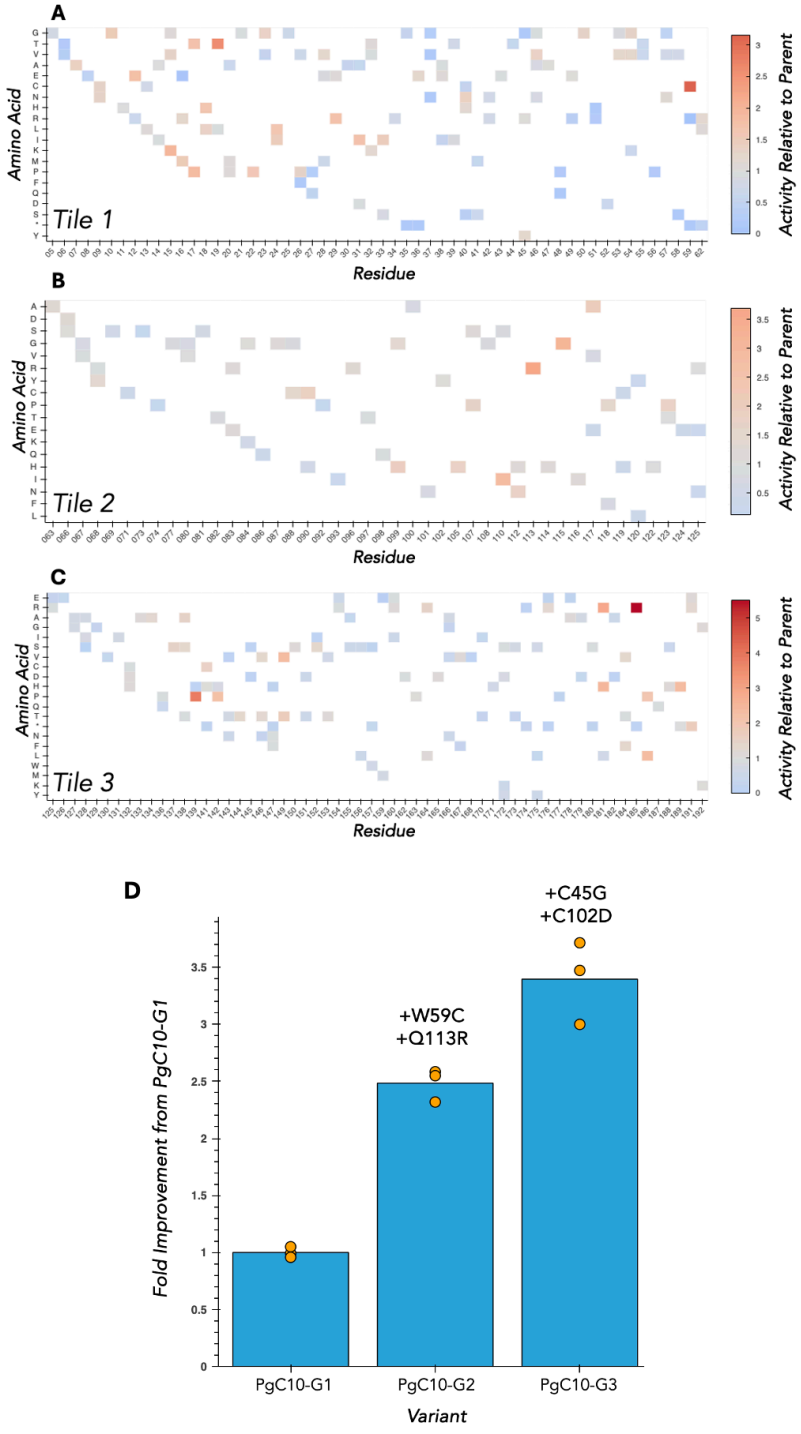


Figure 3-4. Activity data collected for Tile 1 **(A)**, Tile 2 **(B)**, and Tile 3 **(C)**. Approximately 700 random multi-mutants were assayed for each library. Data plotted here are represented as the average fitness relative to PgC10-G1 observed for each single amino acid

substitution irrespective of simultaneously observed mutations. **(D)** Directed evolution of the PgC10 lineage.

3.2.4. Learnings about Protoglobin Biochemistry

As described in **section 3.2.2**, we could not further improve yield of **3** with further evolution of PgC10-G3. As we could not access the desired proline with this biocatalytic system, we elected to halt our studies. Nevertheless, this work highlights the value of collecting complete sequence-function datasets at the outset of a directed evolution campaign. With a greater understanding of the reaction-pertinent regions of *Ape*Pgb we were able to rapidly improve *N*-insertion activity over two consecutive rounds of evolution. Furthermore, the findings of our tiling campaign have proven fruitful for several other directed-evolution efforts upon the protoglobin scaffold. Mutations at positions 118, 123, 185 (among others) which demonstrated hits for azetidinium formation were found to be hits for other carbene, and even nitrene, transfer activities performed by protoglobins.[31-33] As NGS becomes a more accessible technology[34] we anticipate that the collection of such sequence-function paired data will become commonplace in protein engineering.

In the course of attempting to evolve PgC10-G3 for further improvement in activity, we uncovered several emergent properties of evolved protoglobins. Perhaps most relevant to future engineering campaigns is the fact that, upon introduction of the mutation C45G, leaky expression (expression observed prior to addition of an inductant) of *Ape*Pgb in *E. coli* cultures was greatly enhanced. Leaky expression in T7 RNA polymerase-based (RNAP) systems is well known.[35] However, this phenomenon appears to be particularly pronounced in expressions of PgC10-G3. We observed that cultures of *E. coli* harboring the PgC10-G3 gene yielded approximately 1.5 times as much heme-loaded protein when left overnight without the addition of inductant IPTG as compared to cultures induced in a typical manner (**Figure B-12** of **Appendix B**). Although we were excited at the possibility of leveraging this leaky expression for the production of enzymes in a cost-effecting manner,[36] we also believe that it presents challenges for the directed evolution of protoglobins. Given the variant-specific nature of this heightened leaky expression, variable protein expression levels in variants arrayed in a 96-well plate may lead to increased noisiness in library screening. Furthermore, such basal expression can lower the

stability of protein production strains of *E. coli*.[37] Sicinski and coworkers have found similar results with variants demonstrating auto-expression often exhibiting loss of cell density in overnight expressions, and even complete cell lysis in some cases (Personal Communication). We suggest that leaky expression should be closely monitored in future engineering efforts with protoglobins.

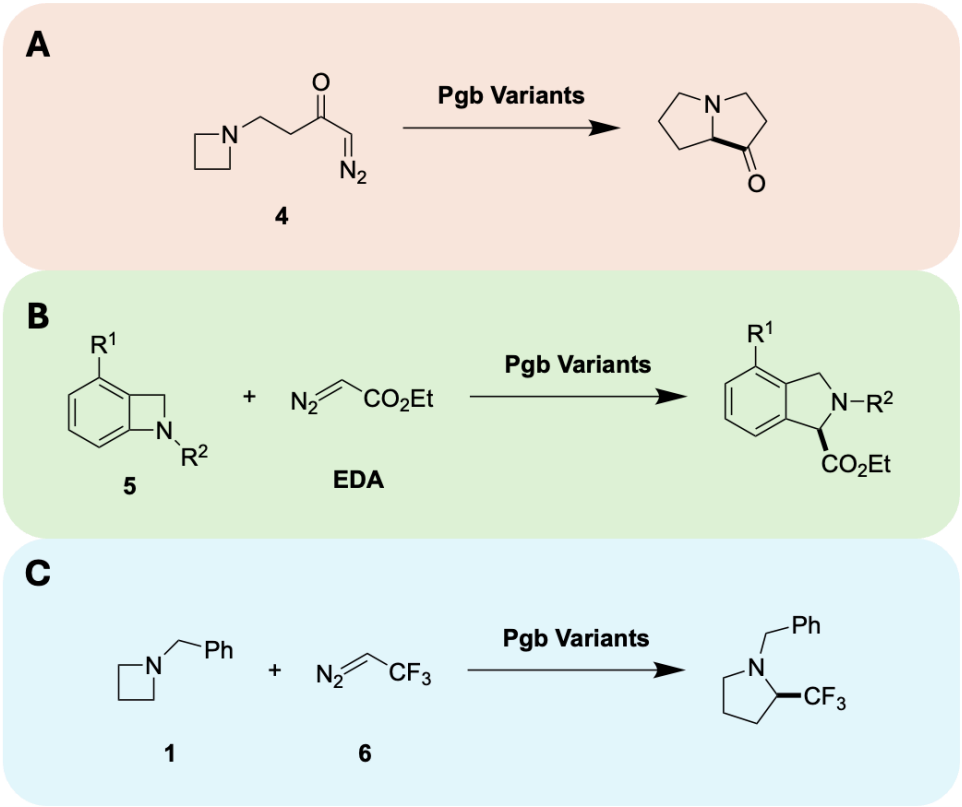
Finally, as we did not observe ring expansion with azetidine **1**, we were eager to test other heterocyclic systems for their ability to engage in ring expansion. By evolving a hemoprotein with a substrate more poised for ring expansion we would conceivably be able to access proline forming enzymes via a substrate walk.[38] Inspired by the work of West and Naidu,[39] we first tested azetidine **4** in hopes that intramolecular cyclization would purvey improve reaction kinetics towards forming pyrrolizidine products (**Scheme 3-2A**). We found no enzymatic activity with this reaction system. Furthermore, we found no ring-expanded products in testing of substrates primed to generate ylide intermediates with more ring strain[40] (benzazetidine **5**) or lower basicity at the ylide position[41] (diazo **6**) (**Scheme 3-2**, **Panels B**–**C**).

Given the observed generation of azetidinium **3**, one could envision that this enzymatic activity could be used to alkylate asymmetric tertiary amines to generate enantioenriched chiral-at-nitrogen quaternary ammonium species. Such chiral cations have garnered recent interest as phase transfer and bifunctional catalysts.[42] These organocatalysts are often limited to cinchona alkaloid-based salts as they have a rigid framework which cannot undergo pyramidal inversion, the fluxional event which makes stereoselective synthesis of ammonium species challenging.[43] With the stereocontrol imparted by an enzyme active site, the formal carbene-based alkylation of tertiary amines displayed by hemoproteins here could be a powerful technique for accessing unconstrained chiral ammonium species, a long-standing challenge in organic synthesis.[44] We intend to explore the capability of engineered hemoproteins to enantioselectively generate chiral ammonium compounds. Furthermore, it remains unclear if azetidinium formation is facilitated by dual-function enzymatic catalysis,[45] in which ylide formation and protonation are both enzyme-driven. An investigation of the mechanism of this system would be an important entry towards our

understanding of the lifetimes of nitrogen ylide species and their subsequent transformations.

Though these results were deeply discouraging, they underline a pressing challenge in the evolution of new-to-nature chemistries in biocatalysis: there is no guarantee that activity for a desired reaction can be found with the existing sequence diversity at hand. Even with our previous evolution of a P411-based system for aziridine expansion, all variants which demonstrated initial activity originated from a single engineering campaign by Brandenberg *et al.*[12,46] Had this previous work not been performed, then we would have likely dismissed the possibility of enzymatic aziridine expansion. We address the challenge of expanding enzyme promiscuity to improve the likelihood of discovering novel biocatalytic activity in **Chapter 5**.

Scheme 3-2. Additional ring-expansion reactions screened against an assortment of protoglobin variants.



Chapter III Bibliography

1. Marshall, C.M.; Federice, J.G.; Bell, C.N.; Cox, P.B.; Njardarson, J.T. An Update on the Nitrogen Heterocycle Compositions and Properties of U.S. FDA-Approved Pharmaceuticals (2013-2023). *J. Med. Chem.* **2024**, *67*, 11622.
2. Kubyshkin, V.; Rubini, M. Proline Analogues. *Chem. Rev.* **2024**, *124*, 8130.
3. Kubyshkin, V.; Mykhailiuk, P.K. Proline Analogues in Drug Design: Current Trends and Future Prospects. *J. Med. Chem.* **2024**, *67*, 20022.
4. Kubyshkin, V.; Davis, R.; Budisa, N. Biochemistry of fluoroprolines: the prospect of making fluorine a bioelement. *Beilstein J. Org. Chem.* **2021**, *17*, 439.
5. Cavalier, F.; Vivet, B.; Martinez, J.; Aubry, A.; Didierjean, C.; Vicherat, A.; Marraud, M. Influence of Silaproline on Peptide Conformation and Bioactivity. *J. Am. Chem. Soc.* **2002**, *124*, 2917.
6. Pujals, S.; Fernández-Carneado, J.; Kogan, M.J.; Martinez, J.; Cavelier, F.; Giralt, E. Replacement of a Proline with Silaproline Causes a 20-Fold Increase in the Cellular Uptake of a Pro-Rich Peptide. *J. Am. Chem. Soc.* **2006**, *128*, 8479.
7. DeRider, M.L.; Wilkens, S.J.; Waddell, M.J.; Bretscher, L.E.; Weinhold, F.; Raines, R.T.; Markley, J.L. Collagen Stability: Insights from NMR Spectroscopic and Hybrid Density Functional Investigations of the Effect of Electronegative Substituents on Prolyl Ring Conformations. *J. Am. Chem. Soc.* **2002**, *124*, 2497.
8. Shoulders, M.D.; Satyshur, K.A.; Forest, K.T.; Raines, R.T. Stereoelectronic and steric effects in side chains preorganize a protein main chain. *Proc. Nat. Acad. Sci.* **2010**, *107*, 559.
9. Holzberger, B.; Marx, A. Replacing 32 Proline Residues by a Noncanonical Amino Acid Results in a Highly Active DNA Polymerase. *J. Am. Chem. Soc.* **2010**, *132*, 15708.
10. Wright, M.M.; Rajewski, B.H.; Gerrein, T.A.; Xu, Z.; Smith, L.J.; Horne, W.S.; Del Valle, J.R. Stabilization of a miniprotein fold by an unpuckered proline analogue. *Commun. Chem.* **2025**, *8*, 76.
11. Remond, E.; Martin, C.; Martinez, J.; Cavalier, F. Silaproline, a Silicon-Containing Proline Surrogate. *Top. Heterocycl. Chem.* **2017**, *48*, 27.
12. Miller, D.C.; Lal, R.G.; Marchetti, L.A.; Arnold, F.H. Biocatalytic One-Carbon Ring Expansion of Aziridines to Azetidines via a Highly Enantioselective [1,2]-Stevens Rearrangement. *J. Am. Chem. Soc.* **2022**, *144*, 4739.
13. Bott, T.M.; Vanecko, J.A.; West, F.G. One-Carbon Ring Expansion of Azetidines via Ammonium Ylide [1,2]-Shifts: A Simple Route to Substituted Pyrrolidines. *J. Org. Chem.* **2009**, *74*, 2832.
14. Huang, J.; Liu, Z.; Bloomer, B.J.; Clark, D.S.; Mukhopadhyay, A.; Keasling, J.D.; Hartwig, J.F. Unnatural biosynthesis by an engineered microorganism with heterologously expressed natural enzymes and an artificial metalloenzyme. *Nat. Chem.* **2021**, *13*, 1186.
15. Huang, J.; Quest, A.; Cruz-Morales, P.; Deng, K.; Pereira, J.H.; Van Cura, D.; Kakumanu, R.; Baidoo, E.E.K.; Dan, Q.; Chen, Y.; Petzold, C.J.; Northen, T.R.; Adams, P.D.; Clark, D.S.; Balskus, E.P.; Hartwig, J.F.; Mukhopadhyay, A.; Keasling, J.D. Complete integration of carbene-transfer chemistry into biosynthesis. *Nature* **2023**, *617*, 403.

16. a) Coelho, P.S.; Wang, Z.J.; Ener, M.E.; Baril, S.A.; Kannan, A.; Arnold, F.H.; Brustad, E.M. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **2013**, *9*, 485. b) Prier, C.K.; Hyster, T.K.; Farwell, C.C.; Huang, A.; Arnold, F.H. Asymmetric Enzymatic Synthesis of Allylic Amines: A Sigmatropic Rearrangement Strategy. *Angew. Chem. Int. Ed.* **2016**, *55*, 4711. c) Wackelin, D.J.; Mao, R.; Sicinski, K.M.; Zhao, Y.; Das, A.; Chen, K.; Arnold, F.H. Enzymatic Assembly of Diverse Lactone Structures: An Intramolecular C–H Functionalization Strategy. *J. Am. Chem. Soc.* **2024**, *146*, 1580.
17. Pesce, A.; Bolognesi, M.; Nardini, M. Chapter Three – Protoglobin: Structure and Ligand-Binding Properties. *Adv. Microb. Physiol.* **2013**, *63*, 79.
18. Freitas, T.A.K.; Hou, S.; Dioum, E.M.; Saito, J.A.; Newhouse, J.; Gonzalez, G.; Gilles-Gonzalez, M.-A.; Alam, M. Ancestral hemoglobins in *Archaea*. *Proc. Nat. Acad. Sci*. **2004**, *101*, 6675.
19. Bloom, J.D.; Labthavikul, S.T.; Otey, C.R.; Arnold, F.H. Protein stability promotes evolvability. *Proc. Nat. Acad. Sci*. **2006**, *103*, 5869.
20. Frances, D.M.; Page, R. Strategies to Optimize Protein Expression in *E. coli*. *Curr. Protoc. Protein Sci.* **2010**, *61*, 5.24.1.
21. Knight, A.M.; Kan, S.B.J.; Lewis, R.D.; Brandenberg, O.F.; Chen, K.; Arnold, F.H. Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes. *ACS Cent. Sci.* **2018**, *4*, 372.
22. Jess, A.; Wasserscheid, P. *Chemical Technology: From Principles to Product*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2020.
23. Khojasteh, S.C.; Prabhu, S.; Kenny, J.R.; Halladay, J.S.; Lu, A.Y.H. *Eur. J. Drug Metab. Pharmacokinet.* **2011**, *36*, 1.
24. Engel, P.C.; Chen, S.-S. A Product-Inhibition Study of Bovine Liver Glutamate Dehydrogenase. *Biochem. J.* **1975**, *151*, 305.
25. a) Jemison, R.W.; Mageswaran, S.; Ollis, W.D.; Sutherland, I.O.; Thebtaranonth, Y. Base-catalysed Rearrangements involving ylide intermediates. Part 8. The preparation and some reactions of stable ammonium ylides. *J. Chem. Soc., Perkin Trans 1* **1981**, *0*, 1154. b) Herchl, R.; Stiftinger, M.; Waser, M. Identification of the best-suited leaving group for the diastereoselective synthesis of glycidic amides from stabilised ammonium ylides and aldehydes. *Org. Biomol. Chem.* **2011**, *9*, 7023.
26. a) Nardini, M.; Pesce, A.; Thijs, L.; Saito, J.A.; Dewilde, S.; Alam, M.; Ascenzi, P.; Coletta, M.; Ciaccio, C.; Moens, L.; Bolgnesi, M. Archael protoglobin structure indicates new ligand diffusion paths and modulation of haem-reactivity. *EMBO Rep.* **2008**, *9*, 157. b) Tilleman, L.; Abbruzzetti, S.; Ciaccio, C.; De Sanctis, G.; Nardini, M.; Pesce, A.; Desmet, F.; Moens, L.; Van Doorslaer, S.; Bruno, S.; Bolognesi, M.; Ascenzi, P.; Coletta, M.; Viappini, C.; Dewilde, S. Structural Bases for the Regulation of CO Binding in the Archaeal Protoglobin from *Methanosarcina acetivorans*. *PLoS ONE* **2015**, *10*, e0125959.
27. a) Porter, N.J.; Danelius, E.; Gonen, T.; Arnold, F.H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **2022**, *144*, 8892. b) Danelius, E.; Porter, N.J.; Unger, J.; Arnold, F.H.; Gonen, T. MicroED Structure of a Protoglobin Reactive Carbene Intermediate. *J. Am. Chem. Soc.* **2023**, *145*, 7159.

28. Equitz, T.R.; Rodriguez-Cruz, S.E. High-throughput analysis of controlled substances: Combining multiple injections in a single experiment run (MISER) and liquid chromatography-mass spectrometry (LC-MS). *Forensic Chem.* **2017**, *5*, 8.
29. 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.* **2022**, *11*, 1313.
30. Ravi, R.K.; Walton, K.; Khosroheidari, M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Disease Gene Identification*; DiStefano, J.K., Ed.; Humana Press, 2018; 223.
31. Hanley, D.; Li, Z.-Q.; Gao, S.; Virgil, S.C.; Arnold, F.H.; Alfonzo, E. Stereospecific Enzymatic Conversion of Boronic Acids to Amines. *J. Am. Chem. Soc.* **2024**, *146*, 19160.
32. Alfonzo, E.; Hanley, D.; Li, Z.-Q.; Sicinski, K.M.; Gao, S.; Arnold, F.H. Biocatalytic Synthesis of α-Amino Esters Nitrene C–H Insertion. *J. Am. Chem. Soc.* **2024**, *146*, 27267.
33. Kennemur, J.L.; Long, Y.; Ko, C.J.; Das, A.; Arnold, F.H. Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]alkanes. *ChemRxiv* **2025**. DOI: 10.26434/chemrxiv-2025-m7201.
34. a) Nekrutenko, A.; Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **2012**, *13*, 667. b) Okumus, B.; Pedraza, J.M.; Bakshi, S. Next-generation quantitative and synthetic biology: High-sensitivity, high-accuracy, and digital approaches. *Front. Bioeng. Biotechnol.* **2023**, *11*. DOI: 10.3389/fbioe.2023.1146729.
35. a) Studier, F.W. Protein production by auto-induction in high-density shaking cultures. *Protein Expression Purif.* **2005**, *41*, 207. b) Terpe, K. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **2006**, *72*, 211.
36. Ferreira, R.; Azzoni, A.R.; Freitas, S. Techno-economic analysis of the industrial production of a low-cost enzyme using *E. coli*: the case of recombinant β-glucosidase. *Biotechnol. Biofuels* **2018**, *11*, 81.
37. Schuster, L.A.; Reisch, C.R. Plasmids for Controlled and Tunable High-Level Expression in *E. coli*. *Appl. Environ. Microbiol.* **2022**, *88*, e00939-22.
38. Savile, C.K.; Janey, J.M.; Mundorff, E.C.; Moore, J.C.; Tam, S.; Jarvis, W.R.; Colbeck, J.C.; Krebber, A.; Fleitz, F.J.; Brands, J.; Devine, P.N.; Huisman, G.W.; Hughes, G.J. Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **2010**, *329*, 305.
39. a) West, F.G.; Naidu, B.N. Piperidines via Ammonium Ylide [1,2]-Shifts: A Concise, Enantioselective Route to (–)-Epilupinine from Proline Ester. *J. Am. Chem. Soc.* **1994**, *116*, 8420. b) Naidu, B.N.; West, F.G. A Short, Enantioselective Synthesis of (–)-Epilupinine from Proline via a Spirocyclic Ammonium Ylide. *Tetrahedron* **1997**, *53*, 16565. c) Vanecko, J.A.; West, F.G. Ring Expansion of Azetidinium Ylides: Rapid Access to the Pyrrolizidine Alkaloids Turneforcidine and Platynecine. *Org. Lett.* **2005**, *7*, 2949.
40. a) Wojciechowski, K. Aza-*ortho*-xylylenes in Organic Synthesis. *Eur. J. Org. Chem.* **2001**, *19*, 3587. b) He, G.; Lu, G.; Guo, Z.; Liu, P.; Chen, G. Benzazetidine

synthesis via palladium-catalysed intramolecular C–H amination. *Nat. Chem.* **2016**, *8*, 1131.

41. Pathak, D.; Deuri, S.; Phukan, P. Nucleophilicity and CO_2 fixation ability of phosphorous, nitrogen and sulfur ylides: insights on stereoelectronic factors from DFT study. *J. Chem. Sci.* **2021**, *133*, 127.
42. a) Oh, J.; Park, J.; Nahm, K. Counter-rotatable dual cinchona quinuclidinium salts and their phase transfer catalysis in enantioselective alkylation of glycine imines. *Chem. Commun.* **2021**, *55*, 6816. b) Novacek, J.; Waser, M. Syntheses and Applications of (Thio)Urea-Containing Chiral Quaternary Ammonium Salt Catalysts. *Eur. J. Org. Chem.* **2014**, *4*, 802. c) Fei, C.; Han, X.-L.; Yu, Y.; Wu, Y.; Lu, Z.; Li, Z.; Luo, J.; Deng, L. Catalytic Asymmetric Reactions of Alkyl Alkynyl Ketimines: A Versatile Approach to Chiral α-Trialkyl Amines. *J. Am. Chem. Soc.* **Article ASAP** DOI: 10.1021/jacs.5c06976.
43. Yang, X.; Deng, Y.; Ling, D.; Li, T.; Chen, L.; Jin, Z. Recent Progress in Chiral Quaternary Ammonium Salt-Promoted Asymmetric Nucleophilic Additions. *ACS Catal.* **2025**, *15*, 1973.
44. Gupta, S.S.; Sharma, U. Construction of stereogenic nitrogen compounds: emerging chemistry. *Trends in Chemistry* **2024**, *6*, 705.
45. Liu, Z.; Calvó-Tusell, C.; Zhou, A.Z.; Chen, K.; Garcia-Borrás, M.; Arnold, F.H. Dual-function enzyme catalysis for enantioselective carbon–nitrogen bond formation. *Nat. Chem.* **2021**, *13*, 1166.
46. Brandenberg, O.F.; Prier, C.K.; Chen, K.; Knight, A.M.; Wu, Z.; Arnold, F.H. Stereoselective Enzymatic Synthesis of Heteroatom-Substituted Cyclopropanes. *ACS Catal.* **2018**, *8*, 2629.

Appendix B

SUPPLEMENTARY INFORMATION FOR CHAPTER III

B.1. General Information and Protocols

B.1.1. Safety Statement

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure. Other than that, no unexpected or unusually high safety concerns were raised with these methods. Safety notes for individual synthetic procedures will be documented alongside the procedure.

B.1.2. General Information

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure to chemicals. Reagents and solvents were obtained commercially (Sigma-Aldrich, Alfa Aesar, VWR, Fischer, Matrix Scientific, Oakwood Chemical, TCI America, and other suppliers) and used without prior purification unless otherwise stated. Organic solutions were concentrated under reduced pressure on an IKA RV 10 rotary evaporator. Thin-layer chromatography (TLC) was performed on commercial Millipore Silica Gel 60 plates containing the F254 fluorescent indicator. Visualization of the developed chromatographs was performed by irradiation with UV light, or treating with and appropriate TLC staining solution (e.g., Ceric Ammonium Molybdate, $KMnO_4$, or Bromocresol Green) followed by heating if necessary. Chromatographic purification was accomplished by flash chromatography on Silacycle F60 silica gel according to the method of Still[1] or using a Biotage Isolera One instrument.

B.1.3. Spectral Data

All NMR spectra were obtained at the Caltech Liquid NMR Facility. For all cyclopropane compounds, ^{1}H NMR were recorded on a Bruker Prodigy 400 MHz instrument (400 MHz and 101 MHz). For intermediates, ^{1}H spectra were also recorded using a Varian 300 MHz spectrometer (300 MHz), a Varian 500 MHz spectrometer (500 MHz), and a Varian 600 MHz spectrometer (600 MHz). ^{1}H spectra are referred to residual $CDCl_3$ solvent signals referenced at δ 7.26 ppm. Data for ^{1}H NMR are reported as follows: chemical shift (δ ppm), integration, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentad, sext = sextet, hept = heptet, m = multiplet, br s = broad singlet), and coupling constant (Hz).

B.1.4. Mutagenesis of Relevant Protein Sequences

Site saturation mutagenesis (SSM) experiments were performed using primers bearing degenerate codons (NDT, VHG, TGG) as per the "22 codon trick" using a modified QuikChange™ protocol.[2] The PCR conditions were as follows (final concentrations): 2 ng/μL plasmid template, Phusion HF Buffer 1x, 0.2 mM dNTPs each, 0.5 μM of forward primers, 0.5 μM reverse primer, and 0.02 U/μL of Phusion polymerase. Upon completion of PCRs, the remaining template was digested with *Dpn*I.

Error-prone PCR experiments for the generation of randomly mutated tile variants were performed using primers designed to flank each tile and engage with evSeq (**Table B-1**).[3] A standard error-prone PCR protocol was utilized.[4] The PCR conditions were as follows (final concentrations): 1 ng/μL plasmid template, Standard *Taq* Reaction Buffer 1x, 0.4 mM dNTPs each, 0.2 μM of forward primer, 0.2 μM of reverse primer, 200-500 μM $MnCl_2$, 0.08 U/μL of *Taq* polymerase. Upon completion of PCRs, the remaining template was digested with *Dpn*I.

B.1.5. Cloning and Plasmid Isolation

Electrocompetent *Escherichia coli* (*E. coli*) cells were prepared following the protocol of Sambrook and Russell.[5] Phusion polymerase and *Dpn*I were purchased from New England Biolabs (NEB, Ipswich, MA). Gel purification was performed with a Zymoclean Gel DNA

Recovery Kit (Zymo Research Corp, Irvine, CA). The purified PCR products from **section B.1.4** were then assembled using the Gibson assembly protocol.[6]

The assembly products obtained were used to transform electrocompetent *E. cloni*® EXPRESS BL21(DE3) cells (Lucigen, Middleton, WI) with a MicroPulser Electroporator (Bio-Rad, Hercules, CA). Luria-Bertani medium (LB; 0.6 mL) was added to electroporated cells and they were incubated at 37 ºC with shaking at 220 rpm for 45 minutes before being plated on LB agar plates supplemented with 100 μg/mL ampicillin (LB-amp agar plates). Plates were incubated at 37 ºC overnight. Single colonies from these plates were used to inoculate flask cultures, prepare glycerol stocks, and isolate plasmids for sequencing. Plasmids were isolated using a QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany), and the genes were sequence-verified (Laragen, Inc.).

B.1.6. Generation of Tile Libraries and Protein Expression in 96-Well Plates

96-well deep-well plates are shaken in an INFORS HT Multitron Shaker in all instances. Single colonies from LB-agar plates were picked using sterilized toothpicks, which were used to inoculate 400 μL of LB containing 100 μg/mL of ampicillin (LB-amp) in 2 mL 96-well deep-well plates. The plates are incubated at 37 ºC and 220 rpm overnight. Eight 96-well plates were picked for each Tile library (**Table B-1**). Prior to expression, all variants were sequenced by evSeq (**Table B-5**). For expression cultures, 50 μL of these precultures are used to inoculate 900 μL of Terrific Broth (Research Products International) with 100 μg/mL of ampicillin (HB-amp) per well in 96-well deep-well plates. The remaining overnight culture plates are sealed are stored in a 4 ºC refrigerator until needed again. The expression cultures are initially incubated at 37 ºC and 225 rpm for 2.5 hours, at which point they are allowed to cool on ice for 30 minutes. Expression of proteins was induced with isopropyl-β-D-thiogalactoside (IPTG) and cellular heme production was increased with 5-aminolevulinic acid (ALA). An induction mixture containing IPTG and ALA in HB-amp (50 μL) was added to each well such that the final concentrations of IPTG and ALA were 0.5 mM and 1.0 mM respectively. The total culture volumes were 1 mL. The plates were then incubated at 22 ºC and 225 rpm overnight. SSM libraries were treated in a similar fashion, although only one 96-well plate was assayed for each investigated position.

Table B-1. Map for the arraying of variants in error-prone Tile libraries.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Sterile	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant
B	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant
C	Mutant	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant
D	Mutant	Mutant	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant
E	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant	Mutant	Mutant
F	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant	Mutant
G	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Parent	Mutant	Mutant	Mutant
H	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Mutant	Sterile

B.1.7. 96-Well Plate Library Reactions and Screening

Expression cultures containing *E. coli* expressing hemoproteins of interest were centrifuged at 5000 × g for 5 minutes at 4 ºC. The supernatant was discarded, and nitrogen-free M9 minimal media (M9-N, 380 µL) was added to each well. The plates were then put into a vinyl Coy anaerobic chamber (0 – 30 ppm O_2) and the pellets are resuspended. To each well were added 20 µL of a MeCN solution with 200 mM of the *N*-benzylazetidine **1** substrate and 300 mM of ethyl diazoacetate (EDA). The final reaction volume was 400 µL, and the final concentrations of the desired azetidine and EDA were 10 mM and 15 mM, respectively. The plates were then sealed carefully with a foil cover and shaken at room temperature for 16 hours in the Coy chamber. Once complete, plates were worked up for LC-MS analysis by adding 400 µL of a HPLC-grade acetonitrile (MeCN). The diluted plates were covered with foil and the solution allowed to mix by shaking for 30 minutes. The plates were then centrifuged (5000 × g for 5 minutes at 4°C) to pellet cell debris. Afterwards, 200 uL of the 1:1 H_2O:MeCN solutions were assayed by LC-MS via Multiple Injections in a Single Experimental Run (MISER).[7]

For SSM screening, if wells were identified which showed improved activity over control wells containing the parent variant, these were subjected to sequence identification and validation of activity. The corresponding wells, as well as a parent control well, in the overnight culture plate were streaked out on an LB-Amp plate. Single colonies from these

plates were then subjected to the small-scale protein expression conditions below (**B.1.8**), followed by the small-scale biocatalytic reaction protocol described below (**B.1.9**).

B.1.8. Small-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 6 mL of LB-amp in a 15 mL plastic culture tube. Cultures are incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 incubator. Two mL of these overnight cultures were used to inoculate 100 mL of HB-amp (1% v/v starter culture in expression culture) in 250-mL Erlenmeyer flasks. The remainder of the overnight culture was subjected to sequence identification (for new variants) and verification (for parent control wells). The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaker, at which point they were held on ice for 30 minutes. Protein expression was then induced by direct addition of 100 µL of stock solutions containing 500 mM IPTG and 1.0 mM ALA such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16 hours in an Innova 42 shaker.

B.1.9. Small-Scale Biocatalytic Reactions for Activity Validation

The corresponding 100 mL expression cultures were pelleted (5000 × g for 5 minutes at 4 ºC) and resuspended in 6 mL of M9-N buffer. 380 µL portions of whole cell suspension (WCS) were prepared in GC vials such that the final cell density corresponded to OD_{600}=30. The whole cell suspensions were put into a vinyl Coy anaerobic chamber, at which point 10 µL of a 400 mM solution of azetidine **1** in MeCN followed by 10 µL of a 600 mM solution of EDA in MeCN were added such that the final reaction concentrations were 10 mM of the azetidine substrate, and 15 mM of EDA. The GC vials were tightly capped with screwcaps with a septum, were brought out of the Coy chamber, and were allowed to shake at RT for 16 hours. Once complete, 300 uL of the reactions were transferred to a 1.7 mL Eppendorf tube and 300 µL of MeCN were added. The layers were thoroughly mixed, and the sample was centrifuged (14000 × g for 10 minutes at RT) to

separate cell debris. Afterwards, an aliquot of the worked up solution was subjected to LC-MS analysis.

B.2. Relevant DNA and Protein Sequences

Table B-2. The amino acid sequences of wild-type *Ape*Pgb and the previously engineered variant PgC10-G1. Colors in the PgC10-G1 sequence indicate which residues are mutated in Tile 1 (green), Tile 2 (orange), or Tile 3 (purple).

Protein Variant	Amino Acid Sequence
wild-type *Aeropyrum pernix* Protoglobin (*Ape*Pgb)	MTPSDIPGYDYGRVEKSPITDLEFDLLKKTVMLGEKDVM YLKKAGDVLKDQVDEILDLLVGWRASNEHLIYYFSNPDT GEPIKEYLERVRARFGAWILDTTSRDYNREWLDYQYEVG LRHHRSKKGVTDGVRTVPHIPLRYLIAQIYPLTATIKPFLA KKGGSPEDIEGMYNAWFKSVVLQVAIWSHPYTKENDW
Aeropyrum pernix Protoglobin Y60G V89A F145N (PgC10-G1)	MTPSDIPGYDYGRVEKSPITDLEFDLLKKTVMLGEKDV MYLKKACDVLKDQVDEILDLWGGWVASNEHLIYYFS NPDTGEPIKEYLERARARFGAWILDTTCRDYNREWLD YQYEVGLRHHRSKKGVTDGVRTVPHIPLRYLIANIYPI TATIKPFLAKKGGSPEDIEGMYNAWFKSVVLQVAIWSH PYTKENDW

DNA Sequence of PgC10-G1. Colors in the PgC10-G1 sequence indicate which residues are mutated in Tile 1 (green), Tile 2 (orange), or Tile 3 (purple).

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACC
CATCACGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAA
AGGACGTAATGTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGAT
GAGATCCTTGACTTGTGGGGTGGTTGGGTAGCATCAAATGAGCATTTGATTTA
TTACTTCTCCAATCCGGATACAGGAGAGCCTATTAAGGAATACCTGGAACGTGC
GCGCGCTCGCTTTGGAGCCTGGATTCTGGACACTACCTGCCGCGACTATAACC
GTGAATGGTTAGACTACCAGTACGAAGTTGGGCTTCGTCATCACCGTTCAAAG
AAAGGGGTCACAGACGGAGTACGCACCGTGCCCCATATCCCACTTCGTTATCT
TATCGCAAATATCTATCCTATCACCGCCACTATCAAGCCATTTTTGGCTAAGAAA
GGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGGTTCAAGTCTG
TAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATGACTGG

Table B-3. Evolutionary trajectory of protoglobin variants involved in this study.

Protein Variant	**Mutations Relative to PgC10-G1** (New Mutations in Bold)
PgC10-G1	None
PgC10-G2	**W59C Q113R**
PgC10-G3	W59C Q113R **C45G C102D**

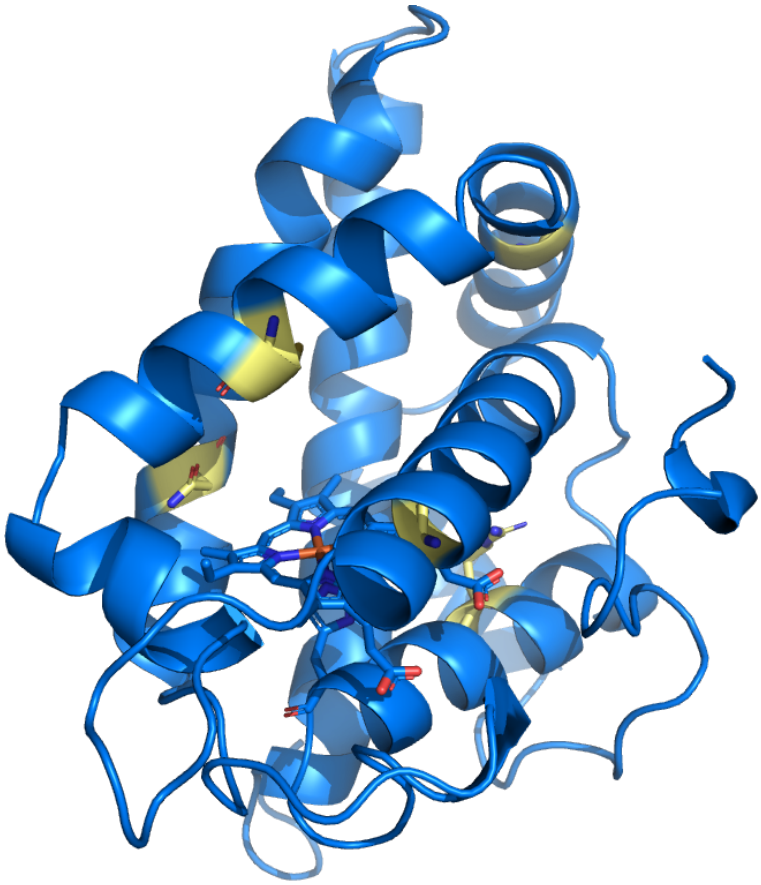


Figure B-1. Homology model of PgC10-G3 with all mutations relative to wild-type *Ape*Pgb shown in yellow.

Table B-4. Primers used in this work to generate enzyme variants and sequence them via evSeq. All primers were ordered from IDT (Coralville, IA).

Primer Name	Direction	Sequence	Description
Tile1_Frag_F	Forward	5'-CCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGACTCCCTCG-3'	Tile 1 Fragment Mutagenesis
Tile1_Frag_R	Reverse	5'-TGTATCCGGATTGGAGAAGTAATAAATCAAATGCTCATTTGATGCTAC-3'	Tile 1 Fragment Mutagenesis
Tile1_BB_F	Forward	5'-GTAGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATACAGG-3'	Tile 1 Backbone Amplification
Tile1_BB_R	Reverse	5'-CGAGGGAGTCATATGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGGG-3'	Tile 1 Backbone Amplification
Tile1_Seq_F	Forward	5'-CACCCAAGACCACTCTCCGGCTTTAAGAAGGAGATATACATATGACTCCCTCG -3'	Tile 1 Sequencing Adaptor
Tile1_Seq_R	Reverse	5'-CGGTGTGCGAAGTAGGTGCGAAGTAATAAATCAAATGCTCATTTGATGCTAC -3'	Tile 1 Sequencing Adaptor
Tile2_Frag_F	Forward	5'-AGATCCTTGACTTGTGGGGTGGTTGG-3'	Tile 2 Fragment Mutagenesis
Tile2_Frag_R	Reverse	5'-GGTGCGTACTCCGTCTGTGACCCC-3'	Tile 2 Fragment Mutagenesis
Tile2_BB_F	Forward	5'-GGGGTCACAGACGGAGTACGCACC-3'	Tile 2 Backbone Amplification
Tile2_BB_R	Reverse	5'-CCAACCACCCCACAAGTCAAGGATCTCATC-3'	Tile 2 Backbone Amplification
Tile2_Seq_F	Forward	5'-CACCCAAGACCACTCTCCGGGACTTGTGGGGTGGTTGG-3'	Tile 2 Sequencing Adaptor
Tile2_Seq_R	Reverse	5'-CGGTGTGCGAAGTAGGTGCCGTACTCCGTCTGTGACCCC-3'	Tile 2 Sequencing Adaptor

Tile3_Frag_F	Forward	5'-CGAAGTTGGGCTTCGTCATCACCGTTCAAAG-3'	Tile 3 Fragment Mutagenesis
Tile3_Frag_R	Reverse	5'-GGTGGTGCTCGAGCCAGTCATTCTC -3'	Tile 3 Fragment Mutagenesis
Tile3_BB_F	Forward	5'-GAGAATGACTGGCTCGAGCACCACC-3'	Tile 3 Backbone Amplification
Tile3_BB_R	Reverse	5'-CTTTGAACGGTGATGACGAAGCCCAAC-3'	Tile 3 Backbone Amplification
Tile3_Seq_F	Forward	5'-CACCCAAGACCACTCTCCGGGGCTTCGTCATCACCGTTCAAAG-3'	Tile 3 Sequencing Adaptor
Tile3_Seq_R	Reverse	5'-CGGTGTGCGAAGTAGGTGCGGTGCTCGAGCCAGTCATTC-3'	Tile 3 Sequencing Adaptor

B.3. Control Experiments

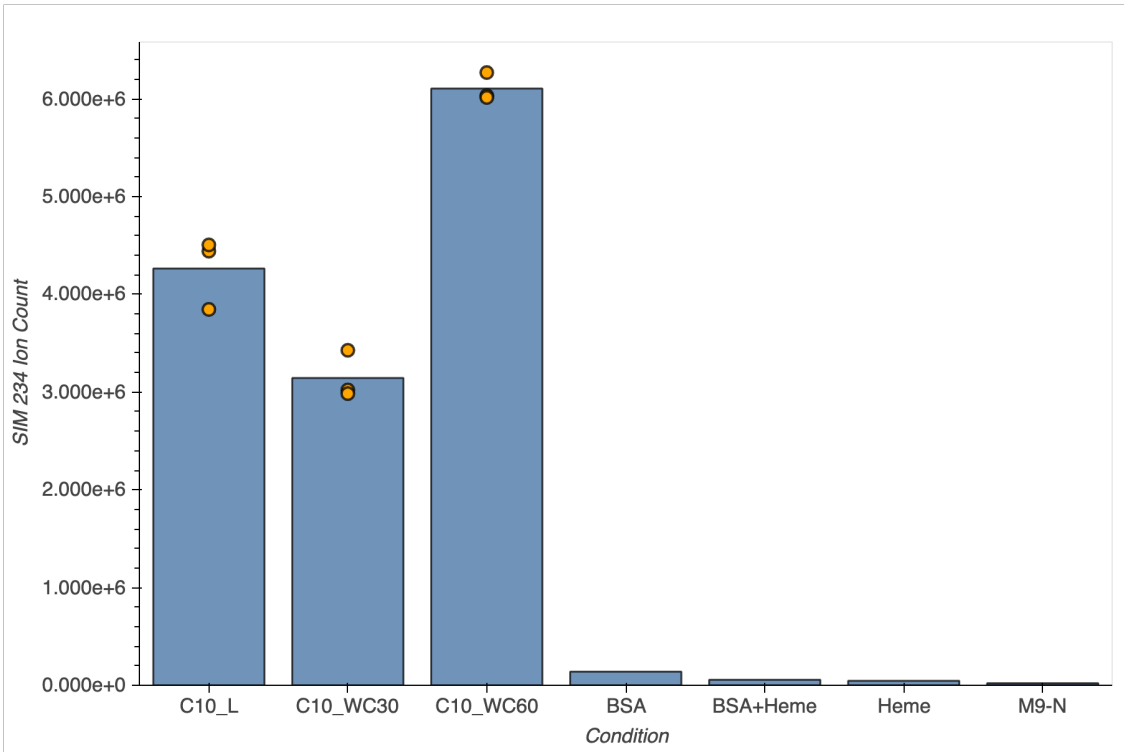
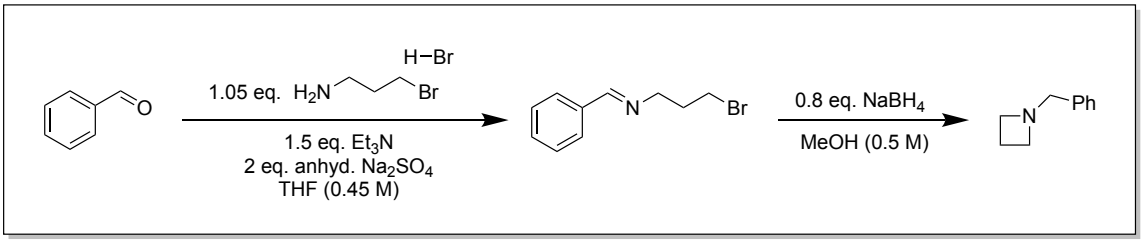
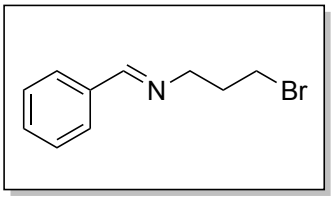


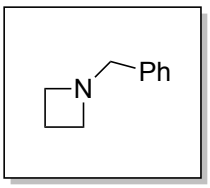
Figure B-2. Control experiments performed to assess potential background reactivity. C10_L refers to reactions in clarified lysates of *E. coli* expressing PgC10 corresponding to a cell density of OD_{600}=30. C10_WC30 and C10_WC60 refer to reactions in whole cell suspensions with OD_{600}=30 and OD_{600}=60 respectively. Conditions with BSA were performed with a concentration of 1 mg/mL protein and conditions with heme were performed with 10 μM loading. Reactions were performed on 4-µmol scale (**B.1.9**).

B.4. Preparation of Substrates



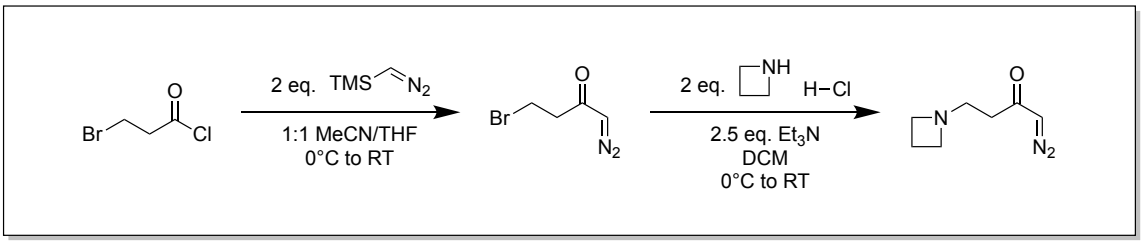


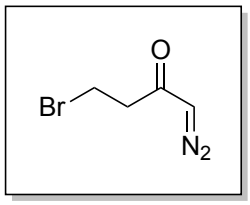
(*E*)-*N*-(3-bromopropyl)-1-phenylmethanimine. To an oven-dried round-bottom flask equipped with a stir bar: benzaldehyde (8.2 mL, 80 mmol, 1.0 equiv.) was dissolved in THF (180 mL). Anhydrous sodium sulfate (22.7 g, 160 mmol, 2 equiv.) and 3-bromopropylamine hydrobromide (18.4 g, 85 mmol, 1.05 equiv.) were added to the flask in a single portion. Triethylamine (16.7 mL, 120 mmol, 1.5 equiv.) was added dropwise while the reaction was agitated. The reaction was allowed to proceed with stirring at room temperature overnight. Once judged complete by TLC, the reaction was filtered over celite and the resulting precipitates were rinsed with diethyl ether. The organics were decanted from the drying agent and the volatiles were stripped under vacuum, yielding a pale yellow oil. The crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



***N*-benzylazetidine**. To a round-bottom flask equipped with a stir bar: the crude imine from the previous step (14.4 g, 80 mmol, 1 equiv.) was dissolved in methanol (160 mL). Sodium borohydride was added portion-wise to control the evolution of hydrogen gas and prevent overflow due to foaming. The reaction is allowed to proceed with stirring for one hour, after which the reaction vessel is heated to 55°C. The reaction was allowed to stir vigorously until judged complete by TLC analysis. Once complete, the reaction was

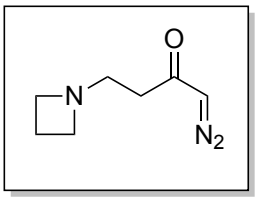
concentrated to *ca.* 5 mL under reduced pressure and then diluted in 50 mL Et_2O. This solution was transferred to a separatory funnel and washed with saturated sodium bicarbonate solution (2x 50 mL) and brine (50 mL) prior to drying over anhydrous sodium sulfate. Once dry, the organics were decanted from the drying agent and the volatiles were stripped under vacuum yielding a yellow oil. The crude product was purified by vacuum distillation (0.4 mmHg) (gradient from 100% hexanes to 20% EtOAc in hexanes) to afford 2.6 g (22% overall yield) of the titled compound as a colorless oil: bp 37°-40°C/0.4 mmHg (lit. 65°-70°C/4 mmHg).[8] ^{1}H NMR (600 MHz, $CDCl_3$) δ 7.7 – 7.4 (m, 5H), 3.58 (s, 2H), 3.23 (t, J = 8.1 Hz, 4H), 2.1 (quintet, J = 8.1 Hz, 2H). The spectrum obtained is in accord with prior literature reports.[9]



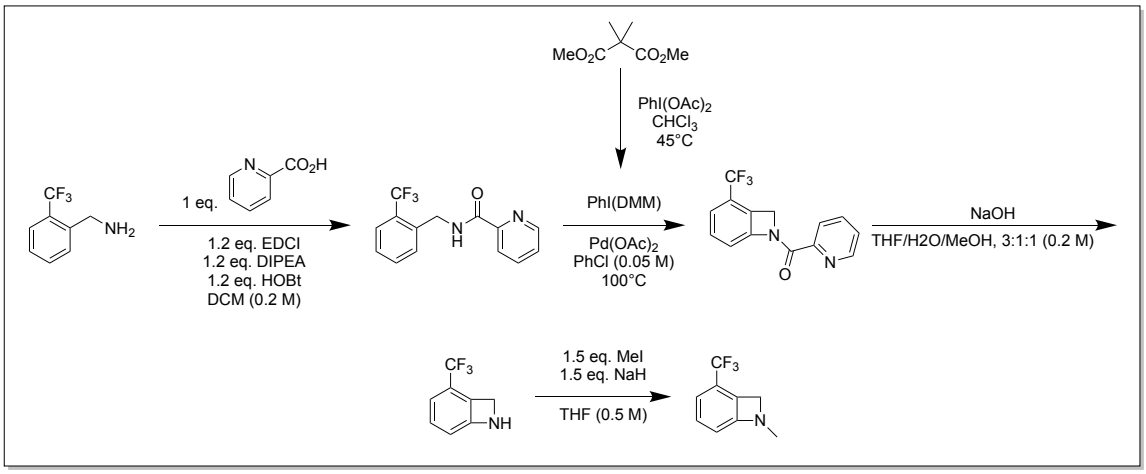


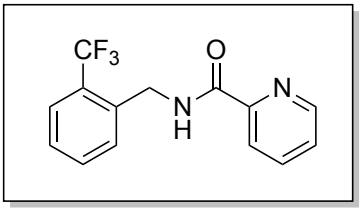
4-bromo-1-diazobutan-2-one. Synthesized using the procedure of Das.[10] A solution of trimethylsilyldiazomethane (5 mL, 2 M in Et_2O, 10 mmol, 2 equiv.) is added to a 1:1 mixture of MeCN and THF (10 mL) in a round bottom flask equipped with a stir bar at 0°C. Separately, a solution of 3-bromopropionyl chloride (860 mg, 5 mmol, 1 equiv.) was dissolved another 1:1 mixture of MeCN and THF (10 mL). The acid chloride solution is added dropwise to the chilled diazo-containing flask. After stirring for 2 hours at 0°C, the reaction was allowed to ambiently warm to RT and proceed until judged complete by TLC. The reaction is quenched with saturated sodium carbonate solution (10 mL). The resulting mixture is transferred to a separation funnel and partitioned into 20 mL Et_2O. The aqueous layer was washed twice more with Et_2O (20 mL). The combined organics were washed with brine and dried over anhydrous sodium sulfate. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 35% EtOAc in hexanes)

to afford 71 mg (8% yield) of the titled compound as a bright yellow oil. ^{1}H NMR (400 MHz, CDCl3) δ 5.31 (s, 1H), 3.65 – 3.57 (m, 2H), 2.89 (t, J = 6.7 Hz, 2H). The spectrum obtained is in accord with prior literature reports.[10]

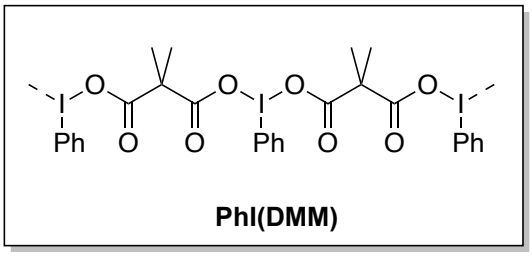


4-(azetidine-1-yl)-1-diazobutan-2-one. The diazo product (71 mg, 0.4 mmol, 1 equiv.) from the previous step was dissolved in DCM (2 mL) in a dram vial equipped with stir bar. The vial was then cooled to 0°C. A solution of azetidine hydrochloride (83 mg, 0.8 mmol, 2 equiv.) was prepared in DCM (2 mL) and added dropwise to the reaction vessel. Triethylamine (0.14 mL, 1 mmol, 2.5 equiv.) was then added dropwise. This solution was allowed to stir at 0°C until judged done by TLC. The reaction was quenched with saturated sodium bicarbonate solution (5 mL). This mixture was transferred to a separatory funnel and the layers separated. The aqueous phase was washed with DCM (3x 10 mL) after which the combined organics were rinsed with brine (10 mL). The organic layers were then dried over sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% DCM to 5% MeOH in DCM) to afford 45 mg (73% yield) of the titled compound as an orange solid. ^{1}H NMR (400 MHz, CDCl3) δ 5.63 (s, 1H), 3.58 (q, J = 7.3 Hz, 4H), 3.01 (q, J = 7.3 Hz, 2H), 1.42 (t, J = 7.3 Hz, 6H), 1.33 (t, J = 7.3 Hz, 3H).

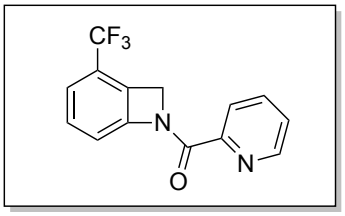




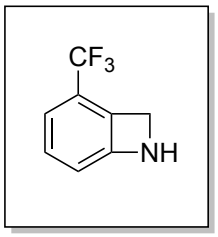
***N*-(2-(trifluoromethyl)benzyl)picolinamide.** Synthesized using the procedure of He and coworkers.[11] An oven-dried round bottom flask was charged with a stir bar and DCM (75 mL). To this flask were picolinic acid 2-(trifluoromethyl)-benzylamine (2.6 g, 15 mmol, 1 equiv.), picolinic acid (1.85 g, 15 mmol, 1 equiv.), diisopropylethylamine (3.1 mL, 18 mmol, 1.2 equiv.), HOBt (2.4 g, 18 mmol, 1.2 equiv.), and EDCI (3.45 g, 18 mmol, 1.2 equiv.). The reaction was allowed to stir overnight at room temperature. The following day, the reaction was partitioned into water (75 mL). The mixture was transferred to a separatory funnel and the phases were separated. The organic layers were washed twice with water (75 mL) and rinsed with brine. The resulting organic layer was allowed to dry over anhydrous sodium sulfate. Once dried, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from 100% hexanes to 60% EtOAc in hexanes) to afford 2.1 g (50% yield) of the titled compound as a feathery white solid. ^{1}H NMR (400 MHz, CDCl3) δ 8.48 (ddd, J = 4.8, 1.8, 0.9 Hz, 1H), 8.39 (s, 1H), 8.16 (dt, J = 7.8, 1.1 Hz, 1H), 7.80 (td, J = 7.7, 1.7 Hz, 1H), 7.59 (dd, J = 14.3, 7.8 Hz, 2H), 7.45 (t, J = 7.6 Hz, 1H), 7.41 – 7.27 (m, 2H), 4.80 (d, J = 6.4 Hz, 2H). The spectrum obtained is in accord with prior literature reports.[11]



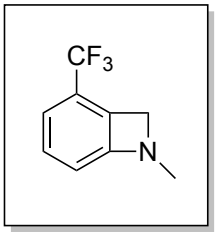
Phenyliodonium dimethylmalonate [PhI(DMM)]. PhI(DMM) is an oligomeric complex of dimethylmalonate with hypervalent iodobenzene described by He *et al.*[1] (Diacetoxyiodo)benzene (3.2 g, 10 mmol, 1 equiv.) and 2,2-dimethylmalonate (1.32 g, 10 mmol, 1 equiv.) were dissolved in chloroform (50 mL) in a round bottom flask equipped with stir bar. The solution was brought to 45°C and agitated for 10 minutes, after which the volatiles were stripped under reduced pressure. This process is repeated twice (2x 50 mL $CHCl_3$), yielding 3.4 g of a sticky white solid (97% yield). ^{1}H NMR (400 MHz, CDCl3) δ 13.25 (d, J = 19.2 Hz, 0H), 8.20 (s, 6H), 7.27 (tt, J = 4.4, 2.2 Hz, 10H), 6.51 (s, 1H). The spectrum obtained is in accord with prior literature reports.[11]



Pyridin-2-yl(2-(trifluoromethyl)-7-azabicyclo[4.2.0]octa-1,3,5-trien-7-yl)methanone. The picolinamide (840 mg, 3 mmol, 1 equiv.) and hypervalent iodine species (2 g, 6 mmol, 2 equiv.) generated in the previous steps were dissolved in chlorobenzene (60 mL). To this solution was added palladium diacetate (67 mg, 0.3 mmol, 10 mol%). The suspension was brought to 100°C and proceeded until judged complete by TLC. The reaction was diluted with EtOAc (150 mL) and partitioned with saturated sodium bicarbonate solution (100 mL). In a separatory funnel the organic layers were rinsed twice with saturated sodium bicarbonate solution (2x 50 mL) and brine (50 mL). The combined organic layers were dried over anhydrous sodium sulfate. The organics were decanted from the drying agent and the volatiles were stripped under vacuum, yielding a dark red residue. After purification by silica gel column chromatography (gradient from 100% hexanes to 30% EtOAc in hexanes), 690 mg (82% yield) of a crystalline white solid was isolated. ^{1}H NMR (400 MHz, CDCl3) δ 8.60 (ddd, J = 4.7, 1.8, 0.9 Hz, 2H), 8.21 – 8.08 (m, 3H), 7.90 – 7.72 (m, 3H), 7.51 (s, 1H), 7.52 – 7.39 (m, 2H), 7.43 – 7.37 (m, 2H), 7.41 – 7.26 (m, 3H), 7.26 – 7.10 (m, 3H), 5.77 (s, 4H), 5.20 (s, 1H), 4.76 – 4.65 (m, 1H), 1.23 – 1.11 (m, 3H). The spectrum obtained is in accord with prior literature reports.[11]

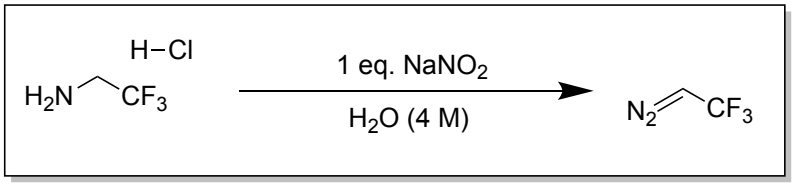


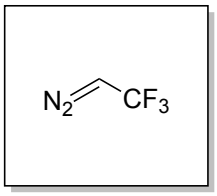
2-(trifluoromethyl)-7-azabicyclo[4.2.0]octa-1,3,5-triene. The protected benzazetidine (690 mg, 2.5 mmol, 1 equiv.) was dissolved in a mixture of THF, MeOH, and H_2O (12.5 mL, *ca.* 3:1:1 respectively). Crushed sodium hydroxide pellets were added portion-wise (2x 120 mg, 6 mmol) until all of the starting material was consumed. The volatiles were stripped under reduced pressure, leaving the reaction components in aqueous solution. The aqueous phase was diluted with H_2O (25 mL) and partitioned into DCM (25 mL). The mixture was transferred to a separatory funnel and the aqueous layer was extracted with DCM (2x 25 mL) prior to drying the organics over anhydrous sodium sulfate. Once dry, the organics were decanted and the volatiles stripped under vacuum. The crude product was purified by silica gel column chromatography (gradient from 100% hexanes to 40% EtOAc in hexanes) to afford 240 mg (56% yield) of the titled compound as an orange oil. ^{1}H NMR (400 MHz, CDCl3) δ 10.32 (s, 1H), 8.73 (s, 1H), 8.50 (ddd, J = 4.8, 1.7, 0.9 Hz, 1H), 8.14 (ddt, J = 26.5, 7.8, 1.1 Hz, 1H), 7.79 (td, J = 7.7, 1.7 Hz, 1H), 7.44 – 7.36 (m, 1H), 7.40 – 7.18 (m, 4H), 7.16 – 7.07 (m, 2H), 7.07 – 7.02 (m, 2H), 6.79 (d, J = 7.7 Hz, 2H), 4.75 – 4.65 (m, 6H). The spectrum obtained is in accord with prior literature reports.[11]



7-methyl-2-(trifluoromethyl)-7-azabicyclo[4.2.0]octa-1,3,5-triene. The material isolated in the previous step (240 mg, 1.4 mmol, 1 equiv.) was dissolved in anhydrous THF in an oven-dried vial. Sodium hydride (60% dispersion, 125 mg, 2 mmol, 1.5 equiv.) and a stir bar were added to the reaction vessel and the mixture was allowed to stir for *ca.* 1 hour. Subsequently, iodomethane (0.13 mL, 2 mmol, 1.5 equiv.) was added dropwise. The reaction proceeded with stirring until judged complete by TLC. Once done, the reaction was quenched with water (3 mL). The aqueous phase was extracted with DCM (3x 6 mL) and the combined organics dried over anhydrous sodium sulfate. Once dry, the volatiles were stripped under vacuum and the product purified by silica gel column chromatography (gradient from 100% hexanes to 40% EtOAc in hexanes) to afford 94 mg (36% yield) of

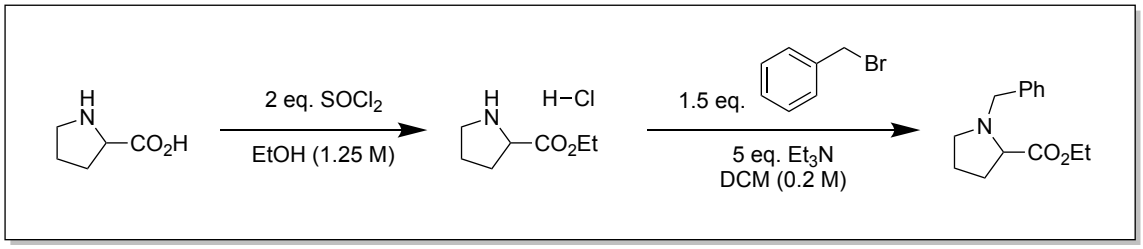
the titled compound as an orange residue. ^{1}H NMR (400 MHz, CDCl3) δ 10.32 (s, 2H), 8.74 (s, 2H), 8.50 (ddd, J = 4.7, 1.8, 0.9 Hz, 2H), 8.15 – 8.08 (m, 2H), 7.79 (td, J = 7.8, 1.7 Hz, 2H), 7.54 – 7.46 (m, 1H), 7.45 – 7.35 (m, 3H), 7.30 – 7.19 (m, 4H), 7.19 (s, 8H), 7.17 – 6.99 (m, 6H), 6.92 (d, J = 7.6 Hz, 1H), 6.81 (dd, J = 15.6, 7.9 Hz, 1H), 5.35 – 5.21 (m, 2H), 5.19 – 5.12 (m, 1H), 4.93 (s, 0H), 4.71 – 4.61 (m, 6H), 4.56 (dd, J = 21.6, 13.0 Hz, 2H), 3.98 – 3.84 (m, 2H), 3.38 (s, 2H), 3.37 – 3.17 (m, 1H), 2.19 (s, 1H), 2.00 (t, J = 7.2 Hz, 1H), 1.19 (s, 1H).

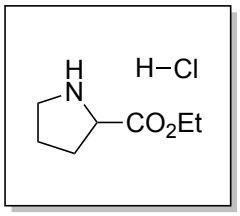




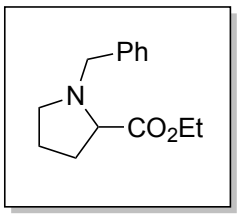
2,2,2-trifluorodiazoethane. Following the protocol of Zhang.[12] Trifluoroethylamine hydrochloride (1.9 g, 14 mmol, 1 equiv.) and sodium nitrite (970 mg, 14 mmol, 1 equiv.) were dissolved in separate aliquots of water (1.5 mL and 2 mL respectively). The following reaction-trap setup was prepared: 1) the reaction vessel is charged with the ethylamine solution, sealed, and attached to an argon line and a separate empty vial via PTFE tubing and needle attachment on either end, 2) the empty vial is attached by tubing to a trap vial charged with 7 mL EtOH kept on ice. The reaction is commenced with slow addition of the sodium nitrite solution (0.9 mL/hr). As the reaction proceeds, the diazo product is trapped in the ethanol phase in the trap vial as evidenced by the yellowing of the solution. After the nitrite addition is complete the reaction is stirred for one additional hour, at which point the trap vial is removed and stored at -20°C as a solution. The product concentration was measured by ^{19}F NMR with 1 M fluorobenzene as an internal standard. ^{19}F NMR: δ =-54 ppm.

B.5. Preparation of Authentic Standards



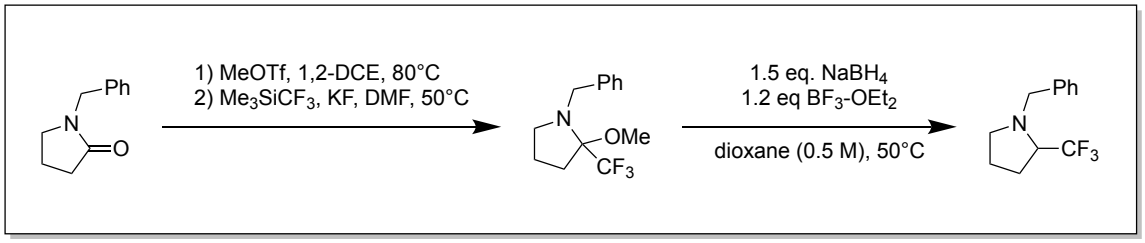


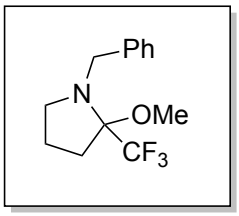
DL-ethyl prolinate hydrochloride. To an oven-dried 8-mL dram vial equipped with a stir bar, septum, and drying tube: DL-proline (460 mg, 4 mmol, 1 equiv.) was suspended in ethanol (3.2 mL). The reaction was chilled in an ice bath, and thionyl chloride (0.6 mL, 8 mmol, 2 equiv.) was added dropwise. The reaction was allowed to ambiently warm to RT with stirring until judged complete by TLC. Once done, the solution was concentrated directly from this dram vial under vacuum to remove volatiles, and the crude product was carried forward with no further purification. For the purposes of the subsequent reaction, the yield was presumed to be quantitative.



DL-ethyl benzylprolinate. The crude prolinate (710 mg, 4 mmol, 1 equiv.) synthesized above is suspended in methylene chloride (25 mL) and the dram vial is equipped with a stir bar, a septum, and a drying tube. The vial is chilled in an ice bath prior to the addition of benzyl bromide (0.71 mL, 6 mmol, 1.5 equiv.). Once addition is complete, triethylamine (2.8 mL, 20 mmol, 5 equiv.) is added dropwise. The reaction is allowed to warm to RT and stirring is continued until the reaction is judged complete by TLC analysis. Once complete, the reaction is transferred to a separatory funnel and quenched with 25 mL of saturated sodium bicarbonate solution. The layers are thoroughly mixed, and the aqueous layer is removed. The organic layer is washed with water (1x 25 mL) and brine (1x 20 mL) prior to drying over sodium sulfate. Once dry, the organics are

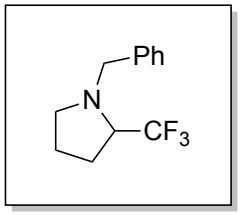
decanted and the volatiles are removed under vacuum. The crude product is purified by silica gel column chromatography (gradient from DCM to 5% MeOH in DCM) to afford 147 mg (63% yield) of the titled compound as an orange oil. ^{1}H NMR (400 MHz, CDCl3) δ 7.40 – 7.31 (m, 3H), 7.35 – 7.29 (m, 1H), 7.33 – 7.20 (m, 1H), 4.15 (qd, J = 7.1, 1.6 Hz, 2H), 3.96 (d, J = 12.9 Hz, 1H), 3.60 (d, J = 12.9 Hz, 1H), 3.27 (t, J = 7.1 Hz, 1H), 3.08 (ddd, J = 10.5, 7.8, 3.2 Hz, 1H), 2.44 (t, J = 10.2 Hz, 1H), 2.23 – 2.08 (m, 1H), 2.05 – 1.74 (m, 3H), 1.27 (t, J = 7.1 Hz, 3H).





(±)-1-benzyl-2-methoxy-2-(trifluoromethyl)pyrrolidine. Synthesized according to the method of Levin *et al.*[13] 1-benzylpyrrolidinone (700 mg, 4 mmol, 1 equiv.) was dissolved in 1,2-dichloroethane (4 mL) in an oven-dried 40 mL dram vial equipped with a stir bar. Methyl triflate (0.94 mL, 4.8 mmol, 1.2 equiv.) was added with stirring and the reaction was brought to 80°C. The reaction was allowed to stir for 15 minutes, after which the reaction was cooled to room temperature and concentrated under reduced pressure. The resulting residue was dissolved in DMF (7 mL) with the addition of trimethyl(trifluoromethyl)silane (1 mL, 6 mmol, 1.5 equiv.) and potassium fluoride (290 mg, 5 mmol, 1.25 equiv.). This solution was stirred for 1 hour at 50°C. After cooling, the reaction is quenched with saturated sodium bicarbonate solution (2 mL) and then diluted with water (10 mL). In a separatory funnel the aqueous layer is washed with a 1:1 mixture of hexanes and Et_2O (3x 10 mL). The combined organics are dried over anhydrous sodium sulfate. Once dry, the organics are decanted and the volatiles stripped under reduced pressure. The crude product is purified by triethylamine-neutralized silica gel column chromatography (gradient from hexanes to 10% EtOAc in hexanes) to afford 294 mg (28%

yield) of the titled compound a yellow oil. ^{1}H 1H NMR (400 MHz, CDCl3) δ 7.31 – 7.13 (m, 8H), 4.16 (d, J = 14.1 Hz, 2H), 3.94 (s, 1H), 3.59 (d, J = 14.1 Hz, 2H), 3.00 (t, J = 9.4 Hz, 1H), 2.88 (td, J = 8.1, 3.8 Hz, 2H), 2.70 (q, J = 8.0 Hz, 2H), 2.41 (dddd, J = 12.5, 9.6, 7.7, 4.1 Hz, 3H), 2.26 (s, 2H), 2.00 – 1.62 (m, 5H). The spectrum obtained is in accord with prior literature reports.[13]



(±)-1-benzyl-2-(trifluoromethyl)pyrrolidine. The *α*-methoxyamine synthesized above (294 mg, 1.1 mmol, 1 equiv.) was dissolved in dioxane (2.3 mL) in a 20 mL dram vial equipped with a stir bar. Sodium borohydride (64 mg, 1.7 mmol, 1.5 equiv.) and boron trifluoride diethyl etherate (0.17 mL, 1.4 mmol, 1.2 equiv.) were added in succession to the reaction vial. The resulting suspension was stirred at 50°C for 2 hours, cooled to room temperature and quenched with saturated sodium bicarbonate solution (3 mL). The mixture was diluted with water (5 mL) and extracted with a 1:1 mixture of hexanes and Et_2O (3x 10 mL). After drying over sodium sulfate, the organics are decanted and the volatiles are removed under vacuum. The crude product is purified by silica gel column chromatography (gradient from hexanes to 33% EtOAc in hexanes) to afford 179 mg (69% yield) of the titled compound as a colorless oil. ^{1}H NMR (400 MHz, CDCl3) δ 7.31 – 7.17 (m, 2H), 7.21 – 7.10 (m, 1H), 4.10 (d, J = 13.3 Hz, 1H), 3.52 (d, J = 13.3 Hz, 1H), 3.25 – 3.11 (m, 1H), 2.88 (ddd, J = 8.7, 6.4, 1.9 Hz, 1H), 2.30 (td, J = 9.7, 6.0 Hz, 1H), 1.95 – 1.82 (m, 2H), 1.82 – 1.68 (m, 1H), 1.72 – 1.60 (m, 1H). The spectrum obtained is in accord with prior literature reports.[13]

B.6. Evidence for Azetidinium Formation

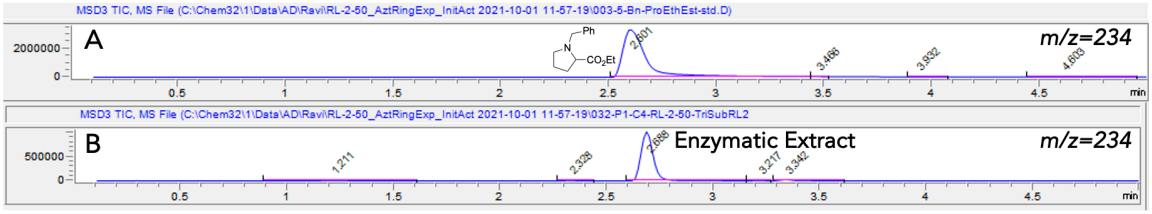


Figure B-3. Initial Activity Determination (A) LC-MS trace for the authentic standard of target compound ethyl benzylprolinate after addition to a suspension of *E. coli*. The sample was worked up as if it were a biocatalytic reaction (**B.1.9**). **(B)** LC-MS trace of analytes from biocatalytic reaction of *N*-benzylazetidine and ethyl diazoacetate with PgC10-G1. The observed differences in peak shape and retention time were attributed to variations in product concentration and matrix composition at the time of this experiment.

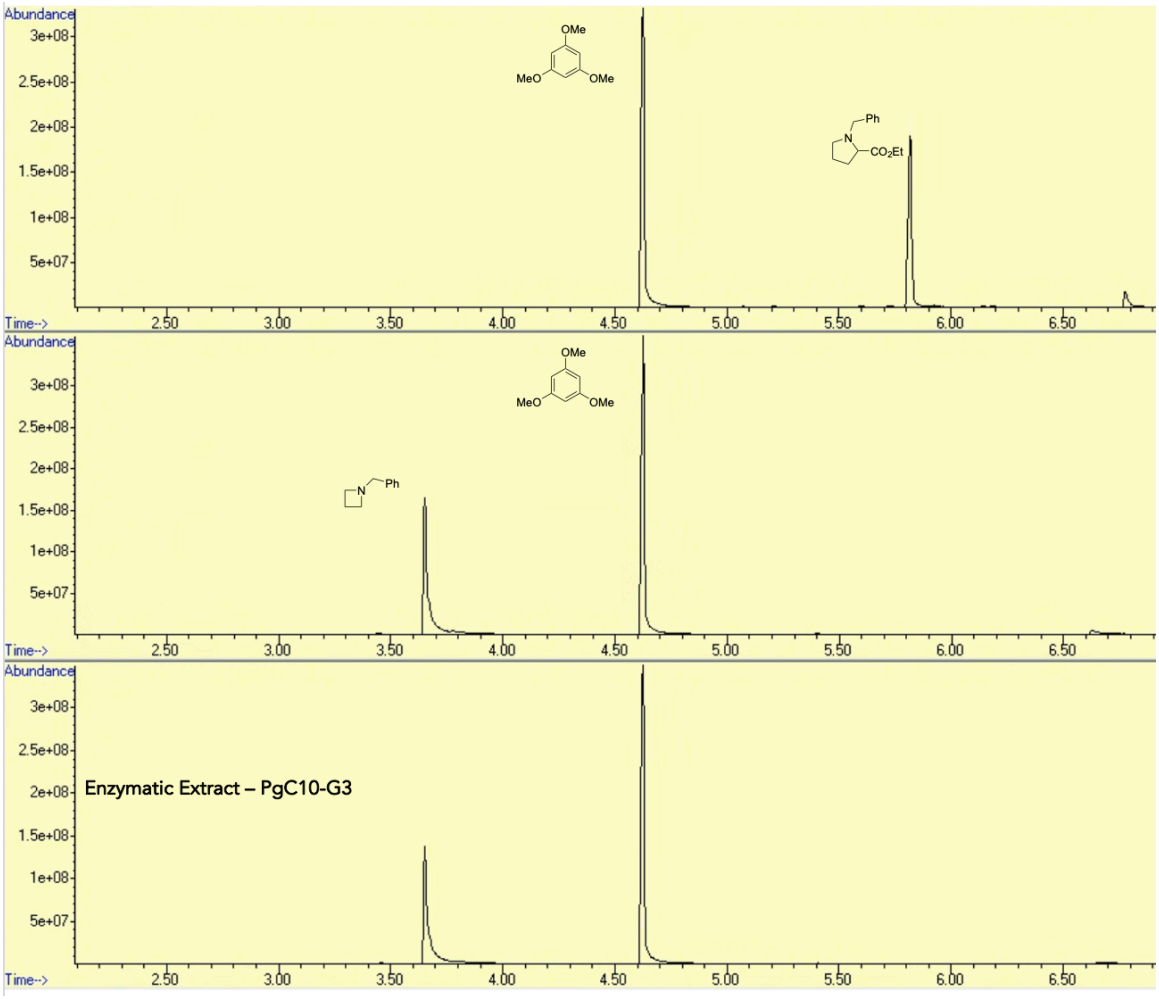


Figure B-4. (A) GC-MS chromatogram of suspension of uninduced *E. coli* spiked with ethyl benzylprolinate with ethyl acetate containing 1 mM 1,3,5-trimethoxybenzene. **(B)** GC-MS chromatogram of suspension of uninduced *E. coli* spiked with *N*-benzylazetidine extracted with ethyl acetate containing 1 mM 1,3,5-trimethoxybenzene. **(C)** GC-MS chromatogram of a biocatalytic whole-cell reaction extracted with ethyl acetate containing 1 mM 1,3,5-trimethoxybenzene. None of the expected product is observed despite the presence of a peak in the LC-MS trace for the same reaction with m/z=234. This indicates that the product that is formed does not partition into organic phases.

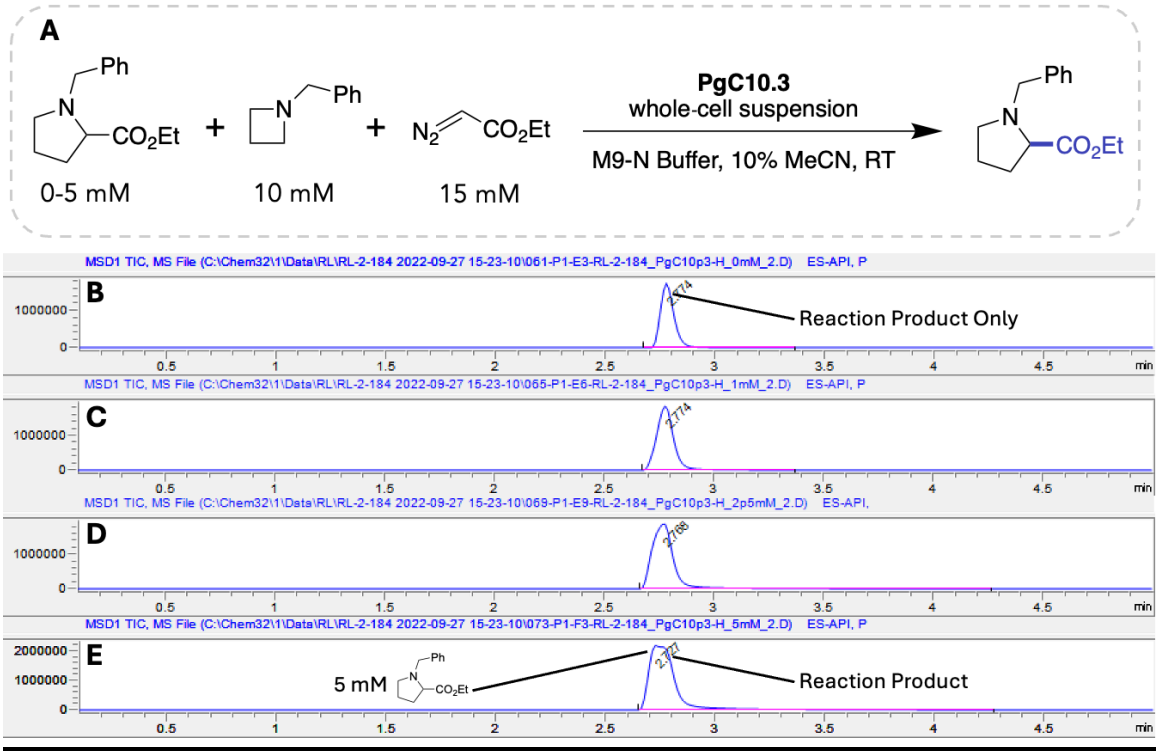


Figure B-5. (A) Reaction conditions for the investigation of product inhibition in reaction of *N*-benzylazetidine with PgC10-G3. Conditions without the expected product **(B)** alongside 1 mM **(C)**, 2.5 mM **(D)**, and 5 mM **(E)** ethyl benzylprolinate were tested. At higher concentrations of added prolinate, it becomes apparent from peak distortion that the enzymatic conditions do not deliver the desired ring-expansion product.

B.7. Sequence-Function Data Collected from Protoglobin Tiles

B.7.1. Protoglobin Tile Design

Protoglobin 'tiles' in this work refer to distinct, contiguous regions of the *Ape*Pgb sequence. These regions are defined by lengths which can be fully sequenced by 150-bp end-paired Illumina sequencing, such that individual variants containing mutations within a single tile can be sequenced by evSeq.[3] Tiles could then be mutagenized by error-prone PCR and the activity of each sequenced variant could be assessed to generate a sequence-function dataset. For each tile, first an optimal error rate was determined by testing libraries generated from PCRs containing varied quantities of $MnCl_2$.[4] Subsequently, eight 96-well plates of variants were generated at the optimized error rate for each tile and sequenced. Finally, the activity for each variant was assessed and the formation of azetidinium product was normalized against simultaneously assayed parent sequences. For each tile we show: the number of mutations observed within the tile at each position, the count of each possible mutation seen in multi-site mutants (only for Tile 1), the average activity for each observed mutation observed in multi-site mutants (only for Tile 1), the count of all single-site mutants, and the average activity for each of these single-mutants.

Tile 1:

Positions: 5-62

Error-prone PCR Conditions: 600 μM $MnCl_2$

Error Rate: 0.882 mutations per variant

Mutant Counts: 368 unique variants, 219 of these mutants contain a single amino acid mutation

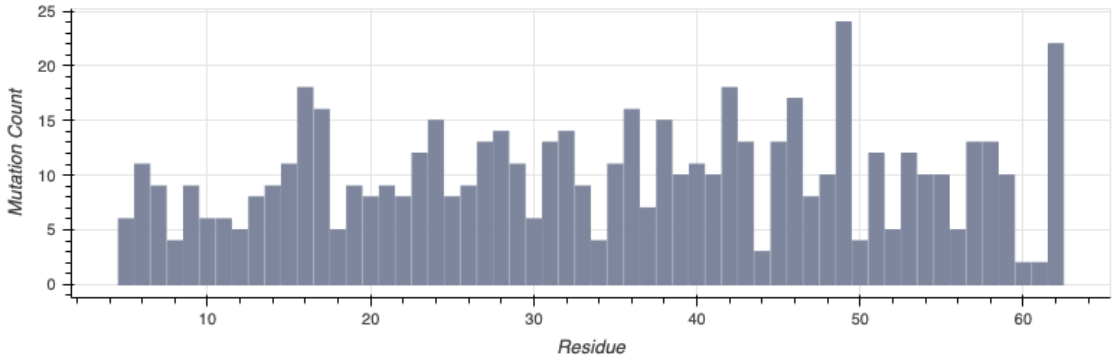


Figure B-6. Counts for the number of variants which contained mutations at each position in Tile 1.

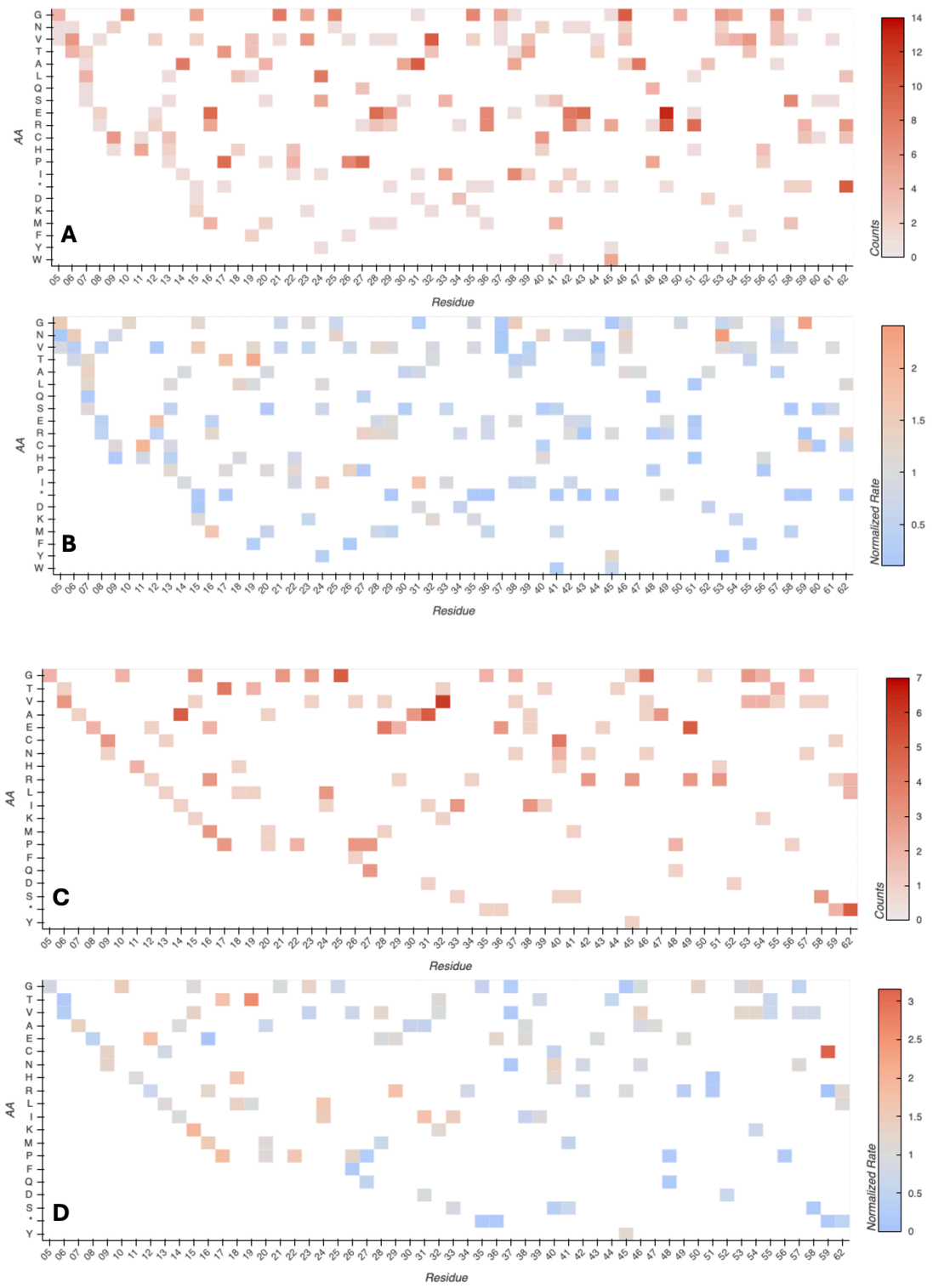


Figure B-7. (A) Counts for the number of instances that each amino acid mutation in Tile 1 is observed in multi-mutants of this library. **(B)** Average activity observed for each amino acid mutation observed in multi-mutants of Tile 1. **(C)** Counts for the number of instances

that each amino acid mutation in Tile 1 is observed as a single mutant. **(D)** Average activity observed for each amino acid mutation observed in single mutants of Tile 1.

Tile 2:

Positions: 62-125

Error-prone PCR Conditions: 400 μM $MnCl_2$

Error Rate: 0.452 mutations per variant

Mutant Counts: 130 unique variants, 90 of these mutants contain a single amino acid mutation

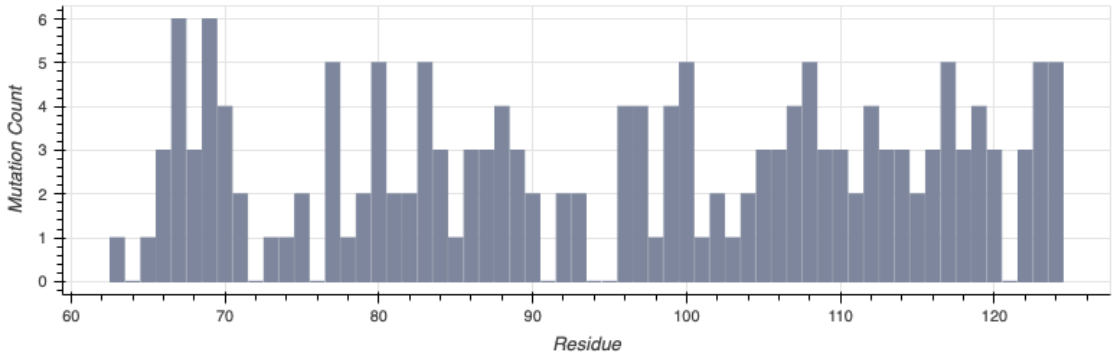


Figure B-8. Counts for the number of variants which contained mutations at each position in Tile 2.

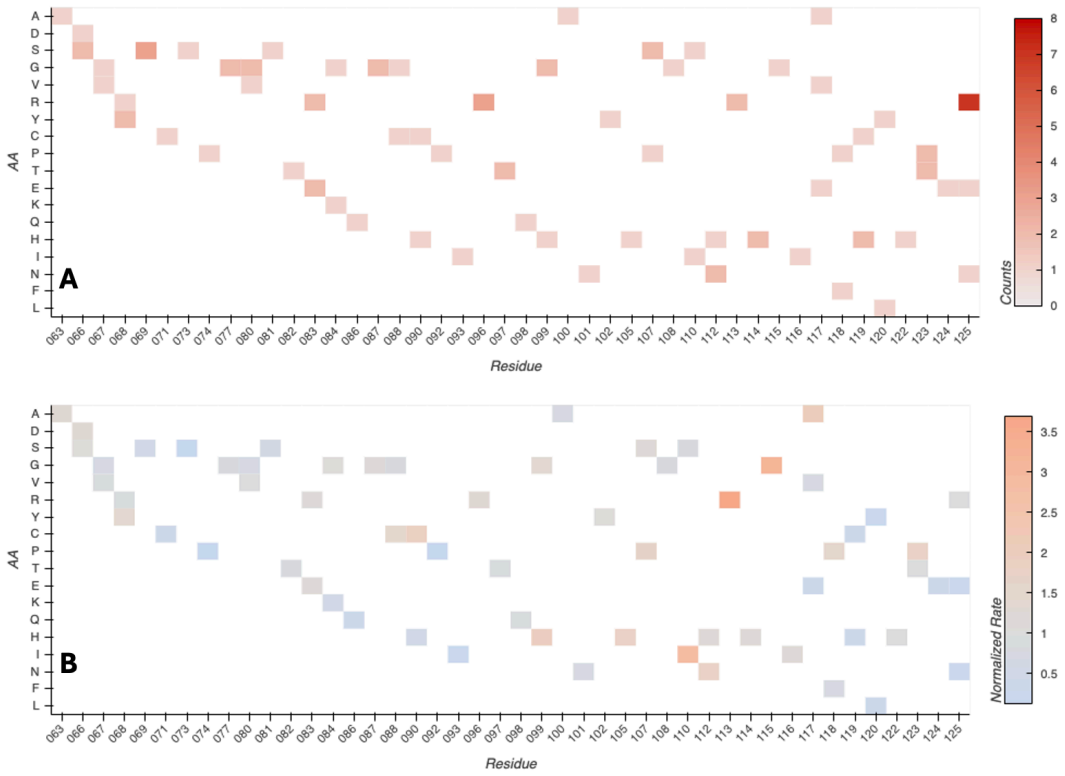


Figure B-9. (A) Counts for the number of instances that each amino acid mutation in Tile 2 is observed as a single mutant. **(B)** Average activity observed for each amino acid mutation observed in single mutants of Tile 2.

Tile 3:

Positions: 125-192

Error-prone PCR Conditions: 500 μM $MnCl_2$

Error Rate: 0.805 mutations per variant

Mutant Counts: 359 unique variants, 217 of these mutants contain a single amino acid mutation

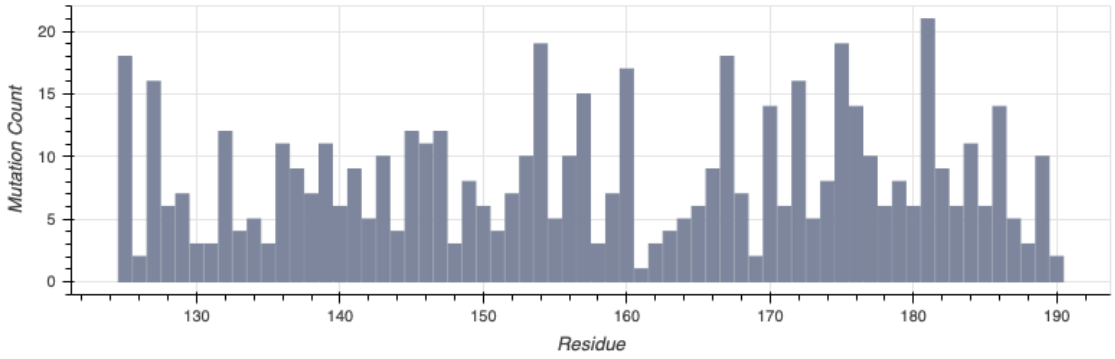


Figure B-10. Counts for the number of variants which contained mutations at each position in Tile 3.

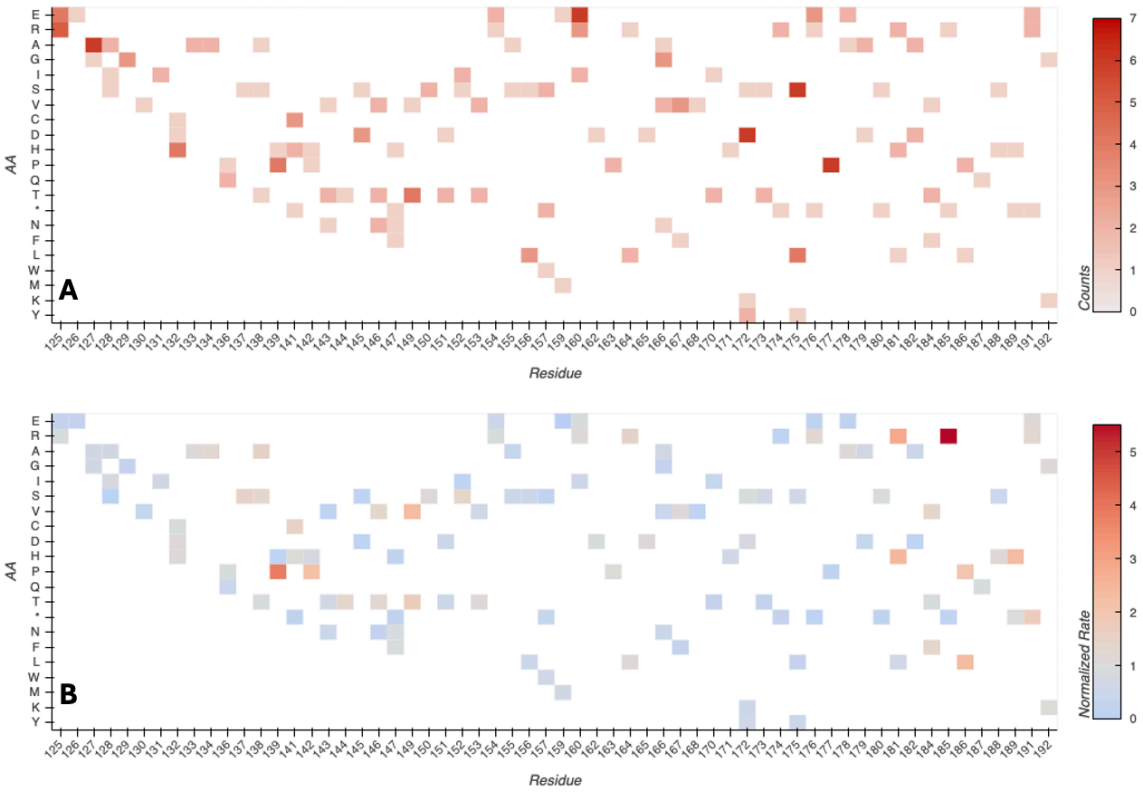


Figure B-11. (A) Counts for the number of instances that each amino acid mutation in Tile 3 is observed as a single mutant. **(B)** Average activity observed for each amino acid mutation observed in single mutants of Tile 3.

B.8. Expression Tests for Protoglobin Variants

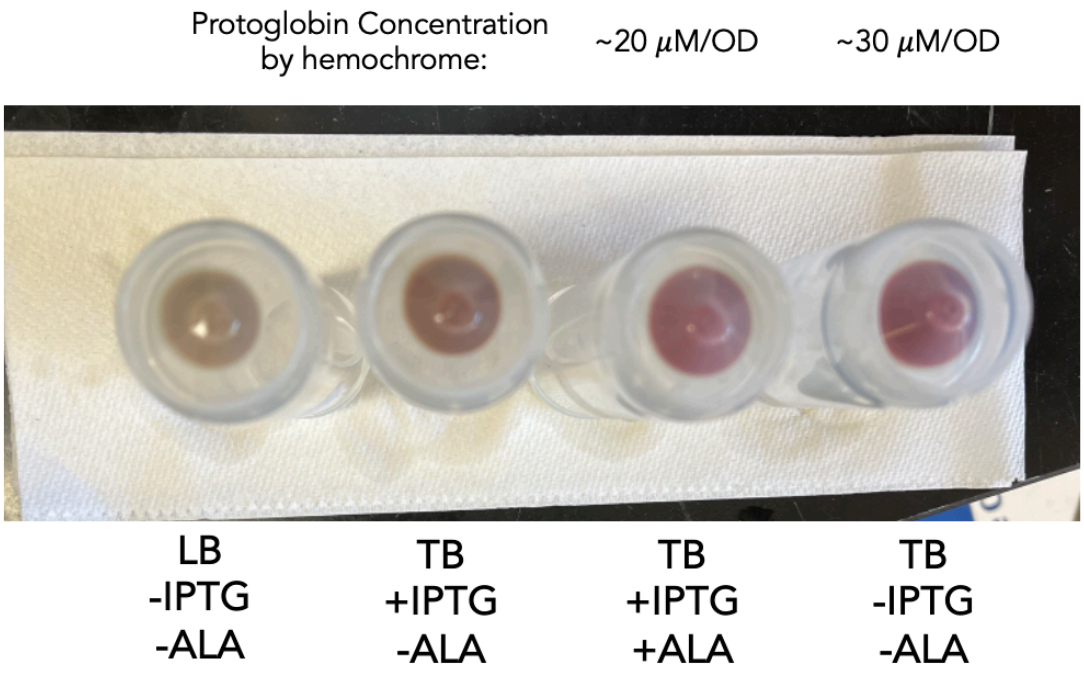
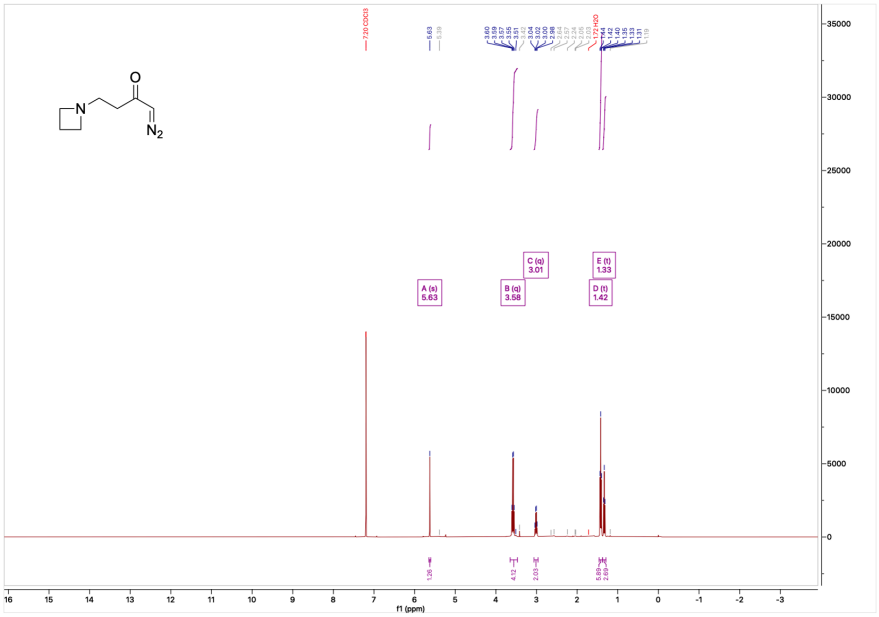


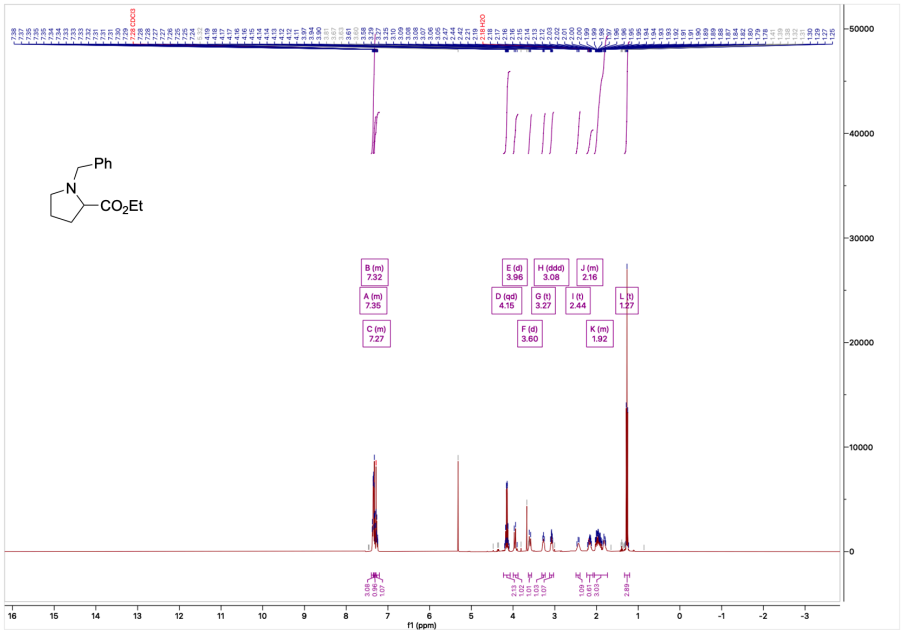
Figure B-12. PgC10-G3 displays high levels of leaky expression. In the absence of IPTG for induction PgC10-G3 is expressed in approximately 1.5 fold greater quantity than with IPTG.

B.9. ^{1}H, ^{13}C, and ^{19}F NMR Spectra of Enzymatic Reaction Products

^{1}H NMR for 4-(azetidin-1-yl)-1-diazobutan-2-one



^{1}H NMR for N-benzylproline ethyl ester



B.10. References

1. Still, W.C.; Kahn, M.; Mitra, A. Rapid Chromatographic technique for preparative separations with moderate resolution. *J. Org. Chem.* **1978**, *43*, 2923.
2. Coehlo, P.S.; Wang, Z.J.; Ener, M.E.; Baril, S.A.; Kannan, A.; Arnold, F.H. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **2013**, *9*, 485.
3. 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.* **2022**, *11*, 1313.
4. Cirino, P.C.; Mayer, K.M.; Umeno, D. Generating Mutant Libraries Using Error-Prone PCR. *Directed Evolution Library Creation*; Arnold, F.H., Georgiou, G., Eds.; Humana Press, 2003; pp 3-9.
5. Sambrook, J.; Russel, D. Transformation of *E. coli* by Electroporation. *Cold Spring Harb. Protoc.* **2009**, *6*, 343.
6. Gibson, D.G.; Young, L.; Chuang, R.-Y.; Venter, J.C.; Hutchinson, C.A.; Smith, H.O. Enzymatic assembly of DNA molecules up to several hunder kilobases. *Nat. Methods.* **2009**, *6*, 343.
7. Ewuitz, T.R.; Rodriguez-Cruz, S.E. High-throughput analysis of controlled substances: Combining multiple injections in a single experiment run (MISER) and liquid chromatography-mass spectrometry (LC-MS). *Forensic Chem.* **2017**, *5*, 8.
8. Sammes, P.G.; Smith, S. Preparation of azetidines from 1,3-aminopropanols. *J. Chem. Soc. Perkin Trans. 1* **1984**, 2415.
9. Lai, G. A FACILE AND EFFICIENT SYNTHESIS OF *N*-BENZYLAZETIDINE. *Synthetic Communications* **2001**, *31*, 565.
10. Das, A.; Wang, D.; Belhomme, M.-C.; Szabó, K.J. Copper-Catalyzed Cross-Coupling of Allylboronic Acids with α-Diazoketones. *Org. Lett.* **2015**, *17*, 4754.
11. He, G.; Lu, G.; Liu, P.; Chen, G. Benzazetidine synthesis via palladium-catalysed intramolecular C–H amination. *Nat. Chem.* **2016**, *8*, 1131.
12. Zhang, J.; Huang, X.; Zhang, R.K.; Arnold, F.H. Enantiodivergent α-Amino C–H Fluoroalkylation Catalyzed by Engineered Cytochrome P450s. *J. Am. Chem. Soc.* **2019**, *141*, 9798.
13. Levin, V.V.; Kozlov, M.A.; Dilman, A.D.; Belyakov, P.A.; Struchkova, M.I.; Tartakovsky, V.A. Trifluoromethylation of the amide group. *Russ. Chem. Bull., Int. Ed.* **2009**, *58*, 484.

Chapter IV

ACTIVE LEARNING-ASSISTED DIRECTED EVOLUTION

Material from this chapter appears in: "Yang, J. †; Lal, R.G. †; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.*; Arnold, F.H.* Active learning-assisted directed evolution. *Nat. Commun.* **2025**, *16*, 714."

R.G.L participated in the conceptualization of the project, as well as execution of the research, including enzymatic reactions, substrate scope studies, and synthetic applications. R.G.L participated in writing and reviewing the manuscript.

ABSTRACT

Directed evolution (DE) is a powerful tool to optimize protein fitness for a specific application. However, DE can be inefficient when mutations exhibit non-additive, or epistatic, behavior. Here, we present Active Learning-assisted Directed Evolution (ALDE), an iterative machine learning-assisted DE workflow that leverages uncertainty quantification to explore the search space of proteins more efficiently than current DE methods. We apply ALDE to an engineering landscape that is challenging for DE: optimization of five epistatic residues in the active site of an enzyme. In three rounds of wet-lab experimentation, we improve the yield of a desired product of a non-native cyclopropanation reaction from 12% to 93%. We also perform computational simulations on existing protein sequence-fitness datasets to support our argument that ALDE can be more effective than DE. Overall, ALDE is a practical and broadly applicable strategy to unlock improved protein engineering outcomes.

4.1. Introduction

Protein engineering is an optimization problem, where the goal is to find the amino acid sequence that maximizes "fitness," a quantitative measurement of the efficacy or functionality for a desired application, from chemical synthesis to bioremediation and therapeutics.[1] Protein fitness optimization can be thought of as navigating a protein fitness landscape, a mapping of amino acid sequences to fitness values, to find higher-fitness variants.[2] However, since protein sequence space is vast, as a protein of length N can take on 20^N distinct sequences and functional proteins are vanishingly rare, finding an optimal sequence is hard. Because functional proteins are surrounded by other functional proteins one mutation away,[3] protein engineers often use directed evolution (DE) to optimize protein fitness.[4,5]

In its simplest form, DE involves accumulating beneficial mutations by searching through sequences near one that exhibits some level of desired function for variants that exhibit enhanced performance on a target fitness metric (**Figure 4-1A**). This approach can be thought of as greedy hill climbing optimization across the protein fitness landscape (**Figure 4-1B**). DE is limited because screening for performance can only explore a small, local region of sequence space. Additionally, taking one mutational step at a time can cause the experiment to become stuck at a local optimum, especially on rugged protein fitness landscapes where mutation effects exhibit epistasis.[6] Machine learning (ML) techniques offer a pathway to circumvent these obstacles, providing strategies to more efficiently navigate these complex landscapes.[7–11]

While supervised ML has been used to propose ideal combinations of mutations–such as in ML-assisted DE (MLDE)[12,13]–these approaches are often limited to small design spaces as they do not take advantage of the fundamentally iterative manner in which protein engineering can take place in real-world applications. By contrast, active learning is an ML paradigm that gathers data iteratively using a supervised model which is, in turn, updated as new data are acquired (**Figure 4-1C**). By leveraging uncertainty quantification to choose which variants should be tested at each step, active learning has the potential to unlock improved engineering outcomes (**Figure 4-1D**).[14–18] Approaches related to active learning have been used in the wet lab to optimize artificial metalloenzymes, nucleases, and other

proteins.[19–23] Past work has also explored the use of Bayesian optimization (BO), a particular class of active learning algorithms, to experimentally improve the thermostability of protein chimeras[24,25] and to optimize proteins with one to several mutations.[14,26,27] However, few studies have explored the utility of active learning methods in comparison to DE, especially where epistatic effects are prevalent.[20,28] In addition, understanding of the practical role of uncertainty quantification in the context of deep learning[29–31] and high-dimensional[32] representations learned from protein language models[33,34] is limited.

To address the limitations of existing methods, we introduce Active Learning-Assisted Directed Evolution (ALDE), a computationally assisted workflow for protein engineering that employs batch Bayesian optimization. ALDE alternates between collecting sequence-fitness data using a wet-lab assay and training an ML model to prioritize new sequences to screen in the wet lab (**Figure 4-1C**); it resembles existing wet-lab mutagenesis and screening workflows for DE and is generally applicable to any protein engineering objective. In this study, we use ALDE to find the ideal combination of five mutations in the active site of a biocatalyst based on a protoglobin from *Pyrobaculum arsenaticum* (*Par*Pgb) for performing a non-native cyclopropanation reaction with high yield and stereoselectivity. We chose this model system because the residues of interest are in close structural proximity and there is evidence of negative epistasis, which hinders DE. After performing three rounds of ALDE (exploring only ~0.01% of the design space), the optimal variant has 99% total yield and 14:1 selectivity for the desired diastereomer of the cyclopropane product. The mutations present in the final variant are not expected from the initial screen of single mutations at these positions, demonstrating that the consideration of epistasis through ML-based modeling is important. We solidify our argument that ALDE is more effective than DE by computationally simulating ALDE on two combinatorially complete protein fitness landscapes. We also provide an extensive analysis of the effects of protein sequence encodings, models, acquisition functions, and uncertainty quantification for protein fitness optimization, to determine best practices for real-world engineering campaigns. In short, we find that frequentist uncertainty quantification works more consistently than typical Bayesian approaches, and incorporating deep learning does not always boost performance. Ultimately, we demonstrate that ALDE is a practical and

effective tool for navigating protein fitness landscapes and provide experimental and computational tools (https://github.com/jsunn-y/ALDE) so that the method is easy to use and broadly applicable.

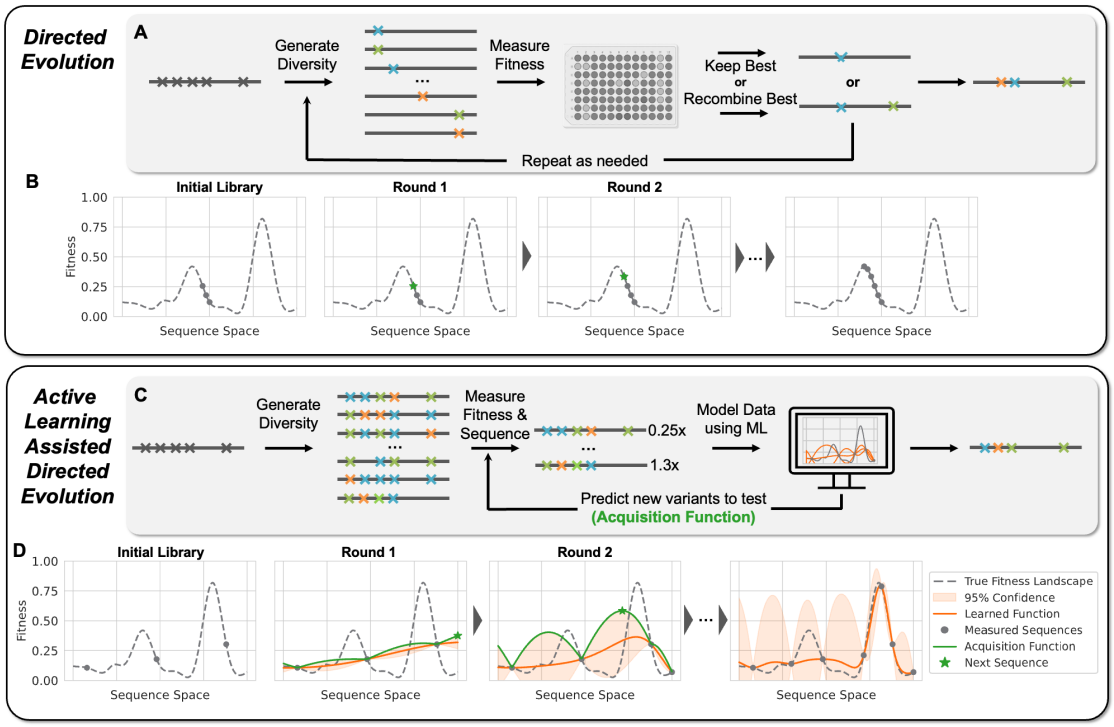


Figure 4-1. Conceptual differences between DE and ALDE. (A) A common workflow for DE, where a starting protein is mutated and fitnesses of variants are measured (screened). The best variant is used as the starting point for the next round of mutation and screening, until desired fitness is achieved. **(B)** Conceptualization of DE as greedy hill climbing optimization on a hypothetical protein fitness landscape. **(C)** Workflow for ALDE. An initial training library is generated, where k residues are mutated simultaneously (for example k=5). A small subset of this library is randomly picked, after which the variants are sequenced and their fitnesses are screened. A supervised ML model with uncertainty quantification is trained to learn a mapping from sequence to fitness. An acquisition function is used to propose new variants to test, balancing exploration (high uncertainty) and exploitation (high predicted fitness). The process is repeated until desired fitness is achieved. **(D)** Conceptualization of active learning on a hypothetical protein fitness landscape. Active learning is often more effective than DE for finding optimal combinations of mutations. In these conceptualizations, a single sequence is queried in each round, but in practical settings, active learning operates in batch where multiple sequences are tested in each round.

4.2. Results

4.2.1. Practical implementation of ALDE

Broadly, ALDE alternates between library synthesis/screening in the wet lab to collect sequence-fitness labels and computationally training an ML model to learn a mapping from sequence to fitness in order to suggest a new batch of sequences to test (**Figure 4-1C**), resembling batch BO. Before beginning ALDE, a combinatorial design space on *k* residues is defined, corresponding to 20^k possible variants. The choice of *k* will vary depending on the system, as larger values of *k* can consider a greater extent of epistatic effects (allowing for better possible outcomes) but will likely require collecting more data to find an optimal variant. First, those *k* residues are simultaneously mutated, and an initial round of sequence-fitness data is collected in the wet lab. ALDE is compatible with low-N, batch protein engineering settings where tens to hundreds of sequences are screened in each round. The collected sequence-fitness data are then used to computationally train a supervised ML model that can predict sequence from fitness. Different ways to encode protein sequence numerically and different types of models which can provide uncertainty quantification are analyzed in this study. Afterward, an acquisition function is applied to the trained model to rank all sequences in the design space, from most to least likely to have high fitness. Several acquisition functions are evaluated in this study, to balance *exploration* of new areas of protein space with *exploitation* of variants that are predicted to have high fitness (**Figure 4-1D**). The computational component of ALDE can be performed using the codebase at https://github.com/jsunn-y/ALDE. For the next round of ALDE, the top N variants from the ranking are then assayed in the wet-lab to provide additional sequence-fitness data, and the cycle is repeated until fitness is sufficiently optimized.

4.2.2. The active site of ParPgb is a challenging design space for standard DE

To initiate wet lab studies with ALDE, we identified a target enzymatic activity on a protein design space that would be difficult to engineer with simple DE methods. Enzyme-catalyzed carbene transfer reactions have the potential to be useful in many synthetic

chemistry applications, and thus we decided to focus on the cyclopropanation of 4-vinylanisole (**1a**) using ethyl diazoacetate (**EDA**) as a carbene precursor to afford the 1,2-disubstituted cyclopropanes *trans*-**2a** and *cis*-**2a** (**Figure 4-2A**). Enzyme engineering for styrenyl cyclopropanation poses a stimulating challenge for evolution toward two properties, higher yield *and* improved selectivity toward one of the diastereomers of the cyclopropane product. While this non-native chemistry has been demonstrated with cytochromes P411,[35] we decided to engineer this activity in a protoglobin. Protoglobins are archaeal hemoproteins, which are attractive engineering targets due to their high thermostability (T_{50} ~ 60°C), small size (~200 amino acids),[36] and ability to perform novel carbene and nitrene transfer chemistries.[37–40] After screening a diverse set of protoglobins, including wild-types and engineered homologs, for cyclopropanation activity (**Figure C-31** of **Appendix C**), we decided to proceed with *Par*Pgb W59L Y60Q (ParLQ) as a starting point (parent) for ALDE. Because our goal was to arrive at a variant with high yield and high selectivity for the *cis*-product, we defined the objective to be explicitly optimized as the difference between the yield of *cis*-**2a** and the yield of *trans*-**2**a. The ParLQ variant demonstrates only moderate cyclopropanation yield (~40% yield) and stereoselectivity (3:1 preferring *trans*-**2a**) under screening conditions, and no known protein variant of *Par*Pgb has high fitness for our objective.

Based on previous engineering studies using protoglobin scaffolds, we selected five active-site residues (W56, Y57, L59, Q60, and F89; WYLQF) positioned above the distal face of the heme cofactor, which display epistatic effects and are known to impact non-native activity (**Figure 4-2B**).[38,39] Single-site saturation mutagenesis (SSM) was performed at these sites, and variants were screened by gas chromatography for their cyclopropanation products. None of the screened mutants demonstrated a significant, desirable shift in the value of the objective (**Figure 4-2C**) or related metrics such as *cis* yield and *cis*/*trans* selectivity (**Figures C-32–C-46** of **Appendix C**). Given these data, a protein engineer might opt to perform a simple recombination of all positive variants to exploit the typically additive character of mutations.[41] However, in our recombination studies of the single-site mutants with the highest fold-change in *cis* yield (DAYFW), the objective (DGMDW), or the selectivity (DGMVW), respectively, we did not observe a variant which generated ***cis***-

2a with high yield and selectivity (**Figure 4-2D**). Overall, these findings suggest that our design problem is quite challenging for standard DE approaches.

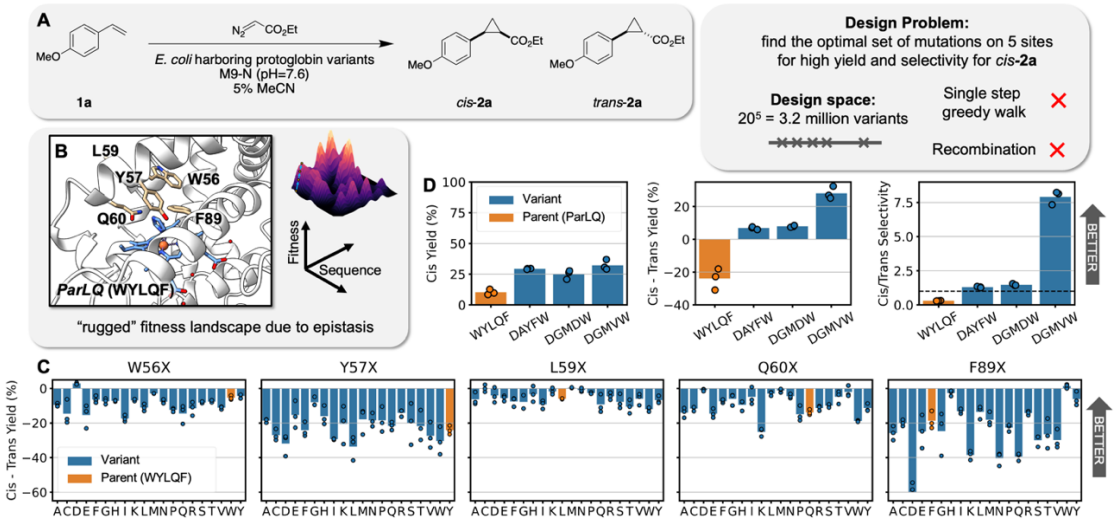


Figure 4-2. A challenging, epistatic protein design space: optimization of five active site residues in *Par*Pgb. (A) Our objective was to optimize an enzyme to catalyze the formation of the *cis* product of a cyclopropanatiom reaction with high yield and high selectivity, which we quantify in a single value as *cis* – *trans* Yield. **(B)** The parent protein ParLQ is two mutations (W59L and V60Q) away from the wild-type ParPgb sequence. Five residues in the active site of ParLQ which were likely to exhibit epistasis were targeted: W56, Y57, L59, Q60, and F89. **(C)** The single mutations from parent at the five targeted sites do not offer significant improvements to the objective of *cis* – *trans* Yield. Very few single-mutation variants have the desired selectivity (positive *cis* – *trans* Yield), and it would not be obvious which variant to take forward in a DE campaign. Parent yields vary between runs but consistently show moderate yield and selectivity for the t*rans* product. **(D)** Various recombinations of ideal single mutations are not effective proteins for the desired objective (*cis* – *trans* Yield), and related metrics such as *cis* Yield and *cis/trans* Selectivity. DAYFW, DGMDW, and DGMVW are the ideal combinations of single mutations naively predicted to have the highest *cis* Yield, *cis* – *trans* Yield (objective), and *cis*/*trans* Selectivity, respectively. Yields were measured in biological triplicate. Overall, these results suggest an optimization problem that is challenging for standard DE methods.

4.2.3. Using ALDE to efficiently optimize ParPgb for a non-native carbene transfer reaction

With the design space confined to five residues and a well-defined objective, we began an ALDE engineering campaign. First, we synthesized an initial library of ParLQ variants

which were mutated at all five positions under study (**Figure 4-3A**). Mutants in this library were generated through sequential rounds of PCR-based mutagenesis methods utilizing NNK degenerate codons. We elected to use random selection from this library because we did not know if any zero-shot predictors might enrich the starting library with useful variants.[8,13] In fact, retrospective analysis of the initial library revealed that our objective is not strongly correlated with conventional zero-shot predictors,[13,42] likely because the objective involves non-native chemistry, for which fitness is not sufficiently captured by evolutionary or stability-based metrics alone (**Figure C-71** of **Appendix C**). Four 96-well plates of these random variants were picked and sequenced using the LevSeq long-read pooled sequencing method (**Figures C-7–C-10** of **Appendix C**),[43] yielding 216 unique variants without stop codons. Screening revealed that nearly all of the variants had higher cyclopropanation activity than free-heme background activity, likely because *Par*LQ was moderately active to begin with, and its high thermostability allows it to tolerate multiple mutations. The majority of variants displaying improved cyclopropanation yield strongly favored formation of *trans*-**2a**; however, several of the randomly selected sequences were capable of forming *cis*-**2a** in much higher yield than any previously tested ParLQ variant (**Figure 4-3B**). Notably, the F89Y mutation was particularly important for inverting selectivity to favor the *cis*-**2a**, but only in the context of certain mutations at positions 56, 57, 59, and 60.

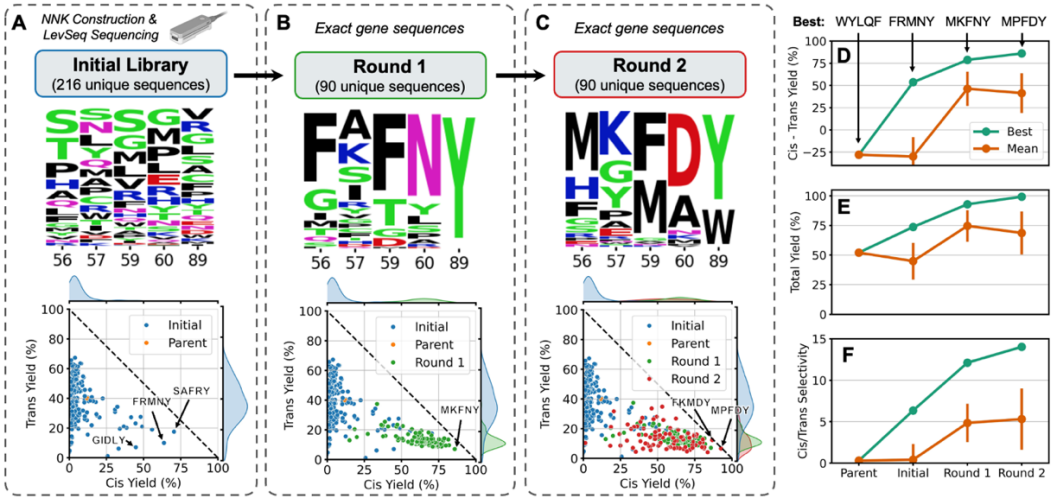


Figure 4-3. ALDE optimization trajectory on the *Par*Pgb active site. The optimization campaign started with **(A)** constructing an initial library with mutations at all five sites under study using NNK degenerate codons, randomly selecting 384 for screening for product formation, and mapping to sequences using LevSeq. This was followed by two rounds of ALDE– **(B)** Round 1 and **(C)** Round 2. In Round 1 and Round2, exact genes were ordered as ENFINIA DNA produced by Elegen Corp. and screened for product formation. For each round, we present the distribution of amino acids sampled at each site and the distribution of yields for the *cis* and *trans* products, with a few of the top-performing variants labeled. Overall improvement in **(D)** *cis* – *trans* Yield, **(E)** Total Yield, and **(F)** *cis*/*trans* Selectivity over several rounds of ALDE for the best variant in each round and the mean across variants in each round. The best variant in each round, defined by the obejctive of *cis* – *trans* Yield is labeled. Error bars indicate the standard deviation across variants in the round. Yields were measured in biological triplicate.

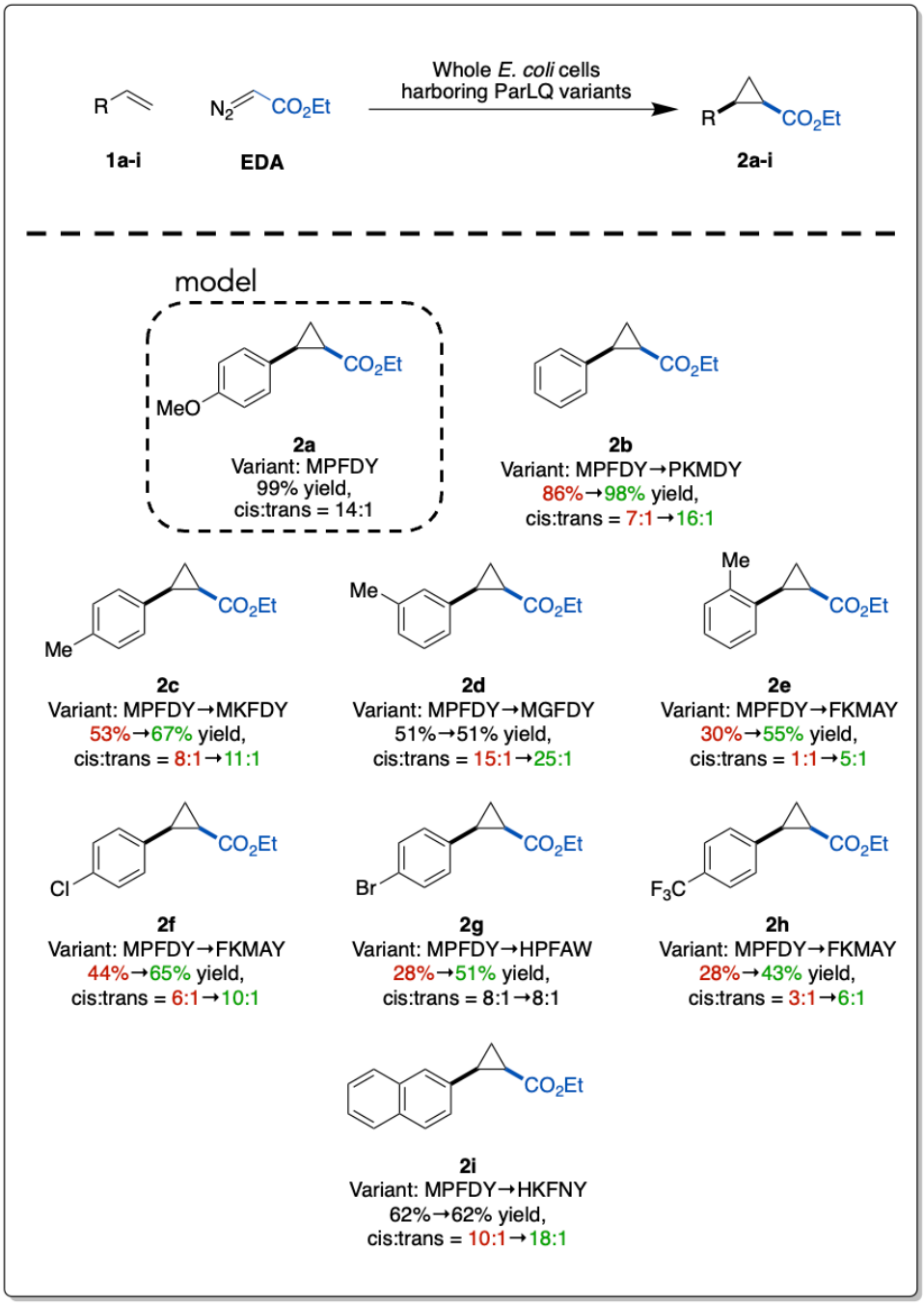
The ALDE computational package was used to train a predictive model on sequences and labels in the initial 216-member library and to suggest sequences for testing based on our acquisition function. Based on our extensive computational simulations (described in the following section), we decided to use the DNN ensemble with one-hot encoding of the five targeted residues for model training and Thompson sampling as the acquisition function. Genes encoding the top 90 amino acid sequences, optimized for expression in *E. coli*, were prepared by exact DNA synthesis for screening (Round 1, **Figure 4-3B**). Details regarding DNA sequence design are described in the included supplementary materials. Subsequent activity screening of these sequences in triplicate showed that nearly a third of Round 1 sequences met the objective better than the best variant in the initial, randomly selected set (**Figure 4-3B**). The best variant in the Round 1 library, MKFNY (W56M Y57K L59F

Q60N F89Y), demonstrated a total cyclopropanation yield of 93% and a *cis:trans* selectivity ratio of 12:1.

We then gave the newly collected data back to the ALDE computational algorithm for a second round of active learning. The top 90 predicted sequences were again synthesized and tested exactly as before (Round 2, **Figure 4-3C**). Interestingly, the model *explored* the sequence space more in this round, as reflected in the expanded mutational diversity present in Round 2 and the increased variance in the activities of these sequences (both reaction yield and diastereoselectivity) (**Figure 4-3, Panels D-F**). Impressively, the top-performing variant among these sequences (MPFDY) displayed a total cyclopropane yield of 99% and a 14:1 *cis:trans* selectivity ratio. None of the mutations in MPFDY obviously optimized the objective in the single-site mutagenesis studies (**Figure 4-2C**); they work together, however, to deliver an optimal variant. Furthermore, after screening the reaction products of all predicted variants with chiral gas chromatography methods, we found that all of these sequences were generally capable of generating ***cis*-2a** in high enantiopurity (**Figure C-70** of **Appendix C**).

Having concluded the ALDE-based evolutionary campaign with substrate **1a,** we sought to understand the substrate scope of the sequences explored in this project. We screened eight styrene derivatives (**1b**–**1i**) for cyclopropanation using the sequences from Round 2 of ALDE (**Scheme 4-1**). The variants show different yields for each of the substrates, even though some of these substrates differed from **1a** only by a single atom. Nevertheless, for every substrate, nearly all of the Round 2 variants were higher yielding and more selective for their respective *cis*- diastereomers than the parent protein, ParLQ (**Figure C-51–C-66** of **Appendix C**). Interestingly, the top-performing variants for each substrate differed in sequence from MPFDY, the top enzyme for **1a** cyclopropanation. For all the predicted variants in Round 1 and 2 of ALDE, sequences were confirmed with LevSeq, and the yield and selectivity of the top variant from each round was validated in vial format, showing good overall consistency (**Figure C-30B** of **Appendix C**).

Scheme 4-1. Substrate scope for *Par*Pgb variants explored during Round 2 of the wet-lab ALDE campaign. All substrates were tested on each of the variants. Total yields and selectivities are shown for the top-performing variant for each substrate as well as MPFDY. Improvements in yield and selectivity from MPFDY are indicated with a shift from red to green.



4.2.4. Computational simulations on combinatorial protein datasets support the utility of ALDE

The design choices used for the wet-lab ALDE campaign were determined by performing computational simulations on two combinatorial landscape datasets for GB1[44] and TrpB[45]. On these landscapes, fitnesses have been measured experimentally for nearly all of the 20^4 = 160,000 variants in a library where four amino acid residues were mutated to all possible amino acids. GB1 refers to the B1 domain of protein G, an immunoglobin binding protein where fitness is measured by binding affinity-based sequence enrichment. The fitness of TrpB, the b-subunit of tryptophan synthase, was measured by coupling growth to the rate of tryptophan formation. Our baseline was DE greedy walk, where one residue was mutated to all possible amino acids, the best mutation was fixed, and the process was repeated at each of the residues under study (**Figure 4-4A**). DE simulations were performed from all active variants as starting points, using all 24 possible orders to enumerate the residues under interest.

The ALDE simulation consisted of batch BO, as shown in **Figure 4-4B**. In each simulation, a random batch of 96 initial samples was selected, followed by four rounds of 96 samples each, with the surrogate model retrained and proposing new samples (via the acquisition function) in each round. This simulation setup was chosen to closely imitate a real wet-lab active learning campaign. The different parameters explored for ALDE, including encodings, models, and acquisition functions, are summarized in **Table 4-1**. We expanded the analysis beyond Gaussian process (GP) models, which are the typical surrogate models for BO, to deep kernel learning (DKL) models[29,31] and frequentist models based on boosting and deep neural network (DNN) ensembles. This was motivated by the rise of high dimensional encodings of protein sequences, such as those from protein language models (i.e. ESM2[33]), which have shown utility in certain property prediction tasks.[47,48] Visualizations of the acquisition functions (greedy, upper confidence bound (UCB), and Thompson sampling (TS)) on hypothetical models are given in **Figure 4-4C**, with more details in **Methods**.

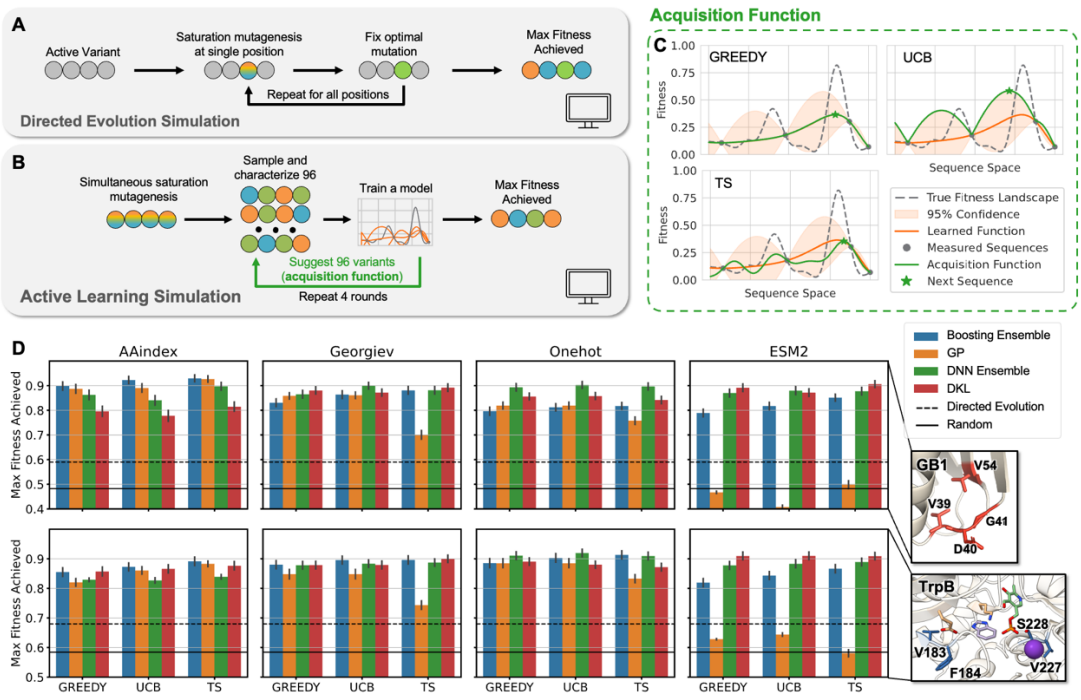


Figure 4-4. Performance of simulated ALDE campaigns on two combinatorially complete protein datasets, GB1 and TrpB. (A) Each DE simulation as a greedy single-step walk on four residues, where each residue is fixed to the optimal mutation until all four residues have been iterated across. DE simulations start from every variant that has some measurable function, with all 24 possible orderings of four residues simulated. **(B)** Each ALDE simulation starts from a random sample of 96 variants on the 4-site landscape, with four rounds of learning and proposing new sequences to test, each with 96 protein variants. **(C)** Hypothetical visualization of the three acquisition functions explored in this work: greedy, upper confidence bound (UCB), and Thompson sampling (TS). **(D)** ALDE for four encodings, four models, and three acquisition functions generally outperforms the average DE simulation and random sampling on the GB1 and TrpB datasets. Performance is quantified as the normalized maximum fitness achieved by the end of the ALDE campaign. Error bars indicate standard deviation across 70 random initializations.

The performance of each simulated ALDE campaign was quantified as the maximum fitness achieved at the end of the campaign, normalized to the variant with maximum fitness in the design space (**Figure 4-4D**). Full optimization trajectories at each iteration of the campaign are provided in **Figure C-72** (**Appendix C**). We conclude that active learning can significantly outperform the average performance of DE and random sampling, and results are consistent across the two different protein datasets. ALDE is competitive with similar methods[14,15] and also outperforms a single round of MLDE (**Figure C-73** of **Appendix C**). Higher dimensional encodings (Onehot and ESM2) generally work better with deep learning-based models (DNN Ensemble and DKL), while non-deep learning

models might learn better from low dimensional AAIndex and Georgiev parameters. The simulations further suggest that encodings from protein language models may not offer much benefit, which corroborates previous findings[13] but stands in contrast to other protein properties that can be predicted more effectively by transfer learning from protein language models.[21,47,48] We find that ESM2 encodings cannot be used by GPs, likely because they are too high dimensional. Other studies suggest that using the right lengthscale priors with GP can enable them to work in these settings;[49,50] while we did not observe this same effect for our application, further exploration may be interesting here. In our acquisition functions, samples in the batch were sampled independently of each other. We also explored batch expected improvement,[51] but this ran extremely slowly without noticeable improvement in performance. Overall, the frequentist ensemble models perform the most consistently across different encodings.

Table 4-1. Summary of encodings of protein sequences, models, and acquisition functions tested in this work.

Encoding	Dimension per Residue	Description
AAIndex	4	Continuous fixed amino acid descriptors
Georgiev[46]	19	Continuous fixed amino acid descriptors
Onehot	20	Categorical (which amino acid)
ESM2[33]	1280	Learned embedding from a protein language model (ESM2 with 650 million parameters)

Model	Bayesian?	Deep Learning?	Description
Boosting Ensemble	N	N	An ensemble of 5 boosting models
Gaussian Process (GP)	Y	N	A collection of continuous functions described by a posterior
DNN Ensemble	N	Y	An ensemble of 5 multilayer perceptrons (deep neural networks, DNNs)
Deep Kernel Learning (DKL)[29]	Y	Y	A GP on the last layer of a deep neural network

Acquisition Function	Deterministic?	Description
Greedy	Y	Acquires the maximum value of the mean from the posterior
Upper Confidence Bound (UCB)	Y	Acquires the maximum value of a certain confidence interval from the posterior (tuned by a hyperparameter)
Thompson Sampling (TS)	N	Acquires the maximum value of a random function sampled from the posterior

To better understand which models are the most advantageous, we assessed how well calibrated their uncertainties were (**Figure 4-5A**). For a calibrated model, an n% confidence interval should contain n% of true labels across different values of n, which can be evaluated and visualized based on a calibration curve. Hypothetical calibrated, underconfident, and overconfident models are visualized in **Figure 4-5B**, with their

associated calibration curves. The calibration curves for different encodings and models are given in **Figure C-74** (**Appendix C**). The area between a calibration curve and perfect calibration (dashed line) is defined as its miscalibration area, which should be low. Another way to measure uncertainty calibration is by measuring the Spearman correlation between uncertainty from the model (σ) and the mean absolute error from the model (MAE), which should be high.

Overall, the Boosting and DNN Ensembles have the lowest MAEs, which suggests that they are the most accurate models (**Figure 4-5A)**. DNN ensembles have the lowest miscalibration areas, suggesting that they are the most calibrated and best models overall. These results are generally consistent across encodings and datasets, with a few outliers. In general, calibrated uncertainty is desirable,[52,53] and it is thought that it is important to understand how calibration shifts when extrapolating beyond the training set.[54,55] However, in this study, we find that performance in ALDE simulations (by max fitness achieved) is not necessarily correlated to how calibrated the uncertainties are for each model. For example, DKL performs the best for the ESM2 encoding, but these models have the least calibrated uncertainties and the highest MAEs. The poor calibration of DKL models may result from some mode collapse where out-of-distribution inputs are mapped close to the training representations by the neural network.[56] Because calibration is measured on the entire combinatorial design space, it may not directly correspond to the ability to find an optimal variant.

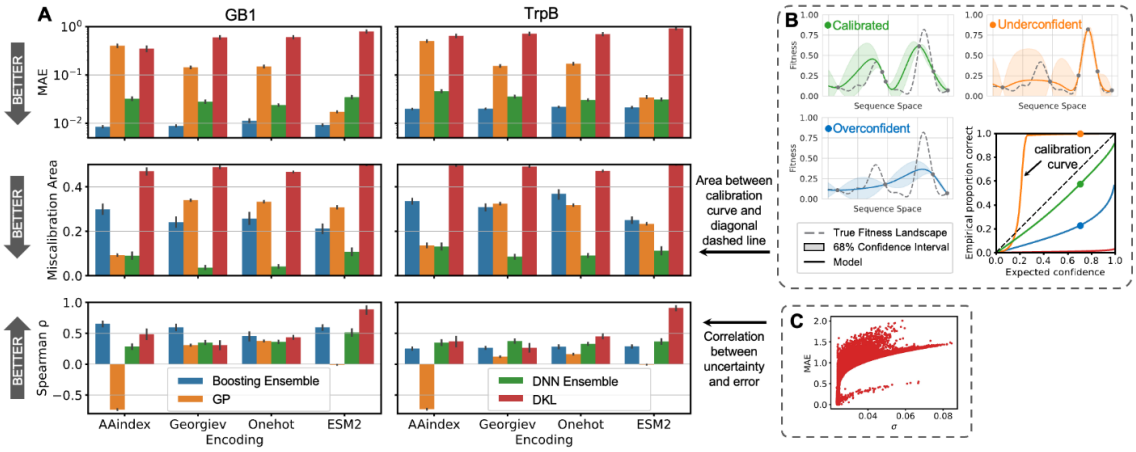


Figure 4-5. Analysis of uncertainty quantification on simulated ALDE campaigns. (A) Metrics used to evaluate how well calibrated each of the four models are for four encodings. Metrics for evaluation are the mean absolute error (MAE), the miscalibration area for the calibration curve, and the Spearman correlation between uncertainty and error. All metrics are calculated based on all measured points in the combinatorial design space. All results are based on the campaigns using UCB as the acquisition function, during the final round of the campaign. Error bars indicate standard deviation across 70 random initializations. **(B)** Visualizations of three hypothetical models with underconfident, calibrated, and overconfident uncertainties, and their respective calibration curves. **(C)** Visualization of how the Spearman correlation between uncertainty and error is calculated.

4.3. Discussion

Overall, ALDE is an effective method for navigating protein fitness landscapes, and it offers several advantages compared to DE. First, ALDE can unlock engineering outcomes not possible with simple DE. By considering multiple interacting positions, ALDE can search for combinations of mutations which may demonstrate desirable epistatic effects[57,58] and reduce the risk of getting trapped at a local optimum. We demonstrated the advantage of ALDE on *Par*Pgb as a particular wet-lab case study – though proof is not possible without testing every DE greedy single-step walk (which is experimentally intractable in the wet lab). Computational simulations of ALDE support this conclusion, as ALDE consistently outperforms MLDE and DE baselines. While ALDE and MLDE[12,13] are similar conceptually and practically, MLDE only uses greedy acquisition, whereas ALDE can consider model uncertainty, which is potentially useful for exploring larger design spaces. Compared to previous computational studies on BO for protein variants,[14,26] our

study examines a more comprehensive range of encodings, models, and acquisition functions, and it introduces analysis of the role of calibrated uncertainty quantification. Interestingly, we found that frequentist ensembles work the best in terms of performance and uncertainty quantification,[30,59] rather than Bayesian approaches such as typical GP models used in BO. Other ways to quantify uncertainty and improve overall performance could be explored in the future.[16,30] Overall, classical notions of uncertainty quantification seem to play a more limited role than expected in these real-world applications. In a related study, we show that ALDE can be combined with various zero-shot predictors and that our findings here still hold for 16 different protein-fitness landscapes, including those with fewer active variants.[60]

In the wet-lab engineering campaign, we were pleased to find that ALDE enabled access to a compilation of enzymes which, when considered together, demonstrate a broader substrate scope than that of a single enzyme. By contrast, DE is limited because it often "locks" one into high yield for only a single substrate or closely related ones, as the final variant is generally a single optimized sequence. Here, we observe an emergent advantage inherent to ALDE: since sequences that balance *exploration* and *exploitation* for a given task are proposed, they can be serendipitously proficient at related tasks.

ALDE is enabled by several recent advancements in biotechnology. For the initial library constructed using degenerate codons, high-throughput sequencing was necessary to identify the sequences of variants in each well. For this work we utilized LevSeq,[43,61] a method that leverages real-time nanopore sequencing. Furthermore, rapid and reliable access to directly synthesized DNA was instrumental to the speed with which evolution was performed. The ALDE workflow was significantly enhanced with (1) the delivery of exact genes in one week, which shortened time between rounds of evolution, (2) the high fidelity of the delivered gene products meant that no sequencing was required for Rounds 1 and 2 of ALDE, and (3) no over-screening was needed because the exact sequences were arrayed individually. Overall, the time and screening cost of the wet-lab engineering campaign with ALDE was lower than for a greedy walk strategy with DE. A total of six 96-well plates were screened before arriving at a final variant: four plates of random variants, and two plates of predicted sequences within three rounds. By comparison, a

greedy walk with DE would have required around five rounds of evolution–typically one mutation is accumulated in each round of a DE greedy walk–with increased screening in the later rounds, which would require greater experimental resources such as chemical reagents and analysis time. We expect that exact gene synthesis will be increasingly important for powering active learning workflows in protein engineering.

In this work, we illustrated ALDE's power for simultaneously increasing the activity and selectivity of an enzyme for a non-natural reaction, but ALDE is a general workflow that can be used for a broad range of protein engineering applications. Additionally, ALDE could be integrated into robotic systems for automated and efficient protein engineering workflows, and library design could utilize tools such as DeCOIL.[62] While we only engineered on five residues in this study, ALDE should naturally extend to even larger design spaces on more residues, as long as assay-labeled data is collected on variants with mutations spread across those residues. Determining the residues to target is an open challenge: these residues should tolerate mutations and have the potential to increase the fitness of interest. Initial domain knowledge, evolutionary conservation, or initial mutational screening may be useful here. Library design could also benefit from limiting the number of simultaneous mutations or using zero-shot scores.[13,19] Despite our *in silico* simulations using combinatorially complete datasets and wet-lab demonstration of ALDE, it remains an open question how generally our findings can be applied to the engineering of other enzymatic systems. Further work is needed to understand how the number of samples and/or rounds required to achieve successful engineering outcomes will increase (linearly or exponentially) with the number of sites explored simultaneously and how epistasis affects this. Future work here may also involve generative modeling if it is not possible to enumerate the acquisition function on the entire design space. Overall, accompanied by a user-friendly codebase, ALDE is a broadly applicable tool that can unlock more efficient and effective protein engineering.

4.4. Methods

4.4.1 Cloning of Random ParPgb Variants

4.4.1.1 Cloning for Single Site-Saturation Mutagenesis

Chemically competent *Escherichia coli* (*E. coli*) cells (T7 Express Competent *E. coli*) were purchased from New England Biolabs (NEB, Ipswich, MA). Phusion polymerase and *Dpn*I were purchased from NEB. SSM experiments were performed using primers bearing degenerate codons (NNK) using a modified QuikChange™ protocol.[63] The PCR conditions were (final concentrations): Phusion HF Buffer 1x, 0.2 mM dNTPs each, 0.5 μM of forward primers, 0.5 μM reverse primer, and 0.02 U/μL of Phusion polymerase. The standard Phusion PCR protocol was used. Upon completion of PCRs, the remaining template was digested with *Dpn*I. Gel purification was performed with a Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA). The purified PCR product was then assembled using the Gibson assembly protocol.[64]

4.4.1.2. Transformation of Single Site Mutants

96-well deep-well plates are shaken in an INFORS HT Multitron Shaker in all instances. The assembly products obtained were used to transform T7 Express Competent *E. coli* (High Efficency) cells (NEB, Ipswich, MA) following the recommended protocol. Upon heat-shock transformation, mixtures were recovered in 0.4 mL Luria-Bertani medium (LB) (Research Products Int.), after which the cells were incubated at 37 ºC with shaking at 220 rpm for 30 minutes before being plated on LB-agar plates with 100 μg/mL ampicillin (LB-amp agar plates). Single colonies from LB-agar plates were picked using sterilized toothpicks, which were used to individually inoculate 400 μL of LB containing 100 μg/mL of ampicillin (LB-amp) in 2 mL 96-well deep-well plates. The plates were incubated at 37 ºC and shaken at 220 rpm for 16-18 hours. The following morning 50 μL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 μL of 50% glycerol solution. These glycerol stocks were

stored at -80°C for future inoculation. Additionally, the sequences of protoglobin genes contained in every well were sequenced using the evSeq protocol.[61]

4.4.1.3. Cloning for Multisite-Saturation Mutagenesis

Mutations were simultaneously incorporated as with single SSM using the ParLQ_quadNNK primers (**Table C-4** of **Appendix C**). The library transformation was recovered in 0.4 mL LB. 50 µL of transformation mixture were used to inoculate 6 mL of LB-Amp in a 15 mL plastic culture tube. This culture was allowed to shake overnight at 37°C. The following morning, this library preculture was miniprepped using a QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany). This miniprep sample was used as the new template for mutagenesis with the primers for SSM of site 89. The Gibson products for the new five-site library were transformed using the recommended protocol into T7 Express Competent *E. coli*. Upon heat-shock transformation, mixtures were recovered in 0.4 mL Luria-Bertani medium (LB) (Research Products Int.), after which the cells were incubated at 37 ºC with shaking at 220 rpm for 30 minutes before being plated on LB-agar plates with 100 µg/mL ampicillin (LB-amp agar plates). Single colonies from LB-agar plates were picked using sterilized toothpicks, which were used to individually inoculate 400 µL of LB containing 100 µg/mL of ampicillin (LB-amp) in 2 mL 96-well deep-well plates across 4 separate plates. The plates were incubated at 37 ºC and shaken at 220 rpm for 16-18 hours. The following morning 50 µL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 µL of 50% glycerol solution. These glycerol stocks were stored at -80°C for future inoculation. Additionally, the sequences of protoglobin genes contained in every well were sequenced using LevSeq sequencing.[43]

4.4.2. Cloning of ParPgb Predicted Sequences

4.4.2.1. 96-Well Plate Gibson Protocol

Exact genes encoding ParLQ mutants predicted by Active Learning-Assisted Directed Evolution (ALDE) were synthesized and delivered by Elegen Corp. (San Carlos, CA). DNA fragments were received as dry residues in 96-well PCR plates in 2-4 μg quantities. These DNA samples were dissolved in 100 μL of double-distilled H_2O (ddH_2O), yielding concentrations between 20-40 ng/μL. 0.7 μL of these gene solutions were added to the wells of a 96-well PCR plate (Globe Scientific Inc., Mahwah, NJ). 1.0 μL of an aqueous solution containing 60 ng/μL of linearized pET-22b(+) backbone with overhangs designed for Gibson ligation with the ordered DNA sequences was added to each of the wells of this plate. Finally, to each well was added 5 μL of Gibson assembly mix. The 96-well plate was then incubated at 50°C for 60 minutes, after which the Gibson products were placed on ice. These Gibson products could then either be directly used for transformation or stored at -20°C for later use.

4.4.2.2. 96-Well Plate Transformation Protocol

To each well of the previously described Gibson assembly plate was added 5 μL of T7 Express Competent *E. coli*. The cell solutions were allowed to incubate on ice for 20 minutes, after which they were heat-shocked at 42°C for 10 seconds in a water bath. The cells were then recovered with the addition of 100 μL of LB. Without outgrowth at 37°C, 10 μL of each transformation mixture was used to inoculate the wells of a 2 mL 96-well deep-well plate in which the wells had been preloaded with 400 μL LB-Amp. This plate was incubated at 37 ºC and shaken at 220 rpm for 16-18 hours. The following morning the plate was removed from shaking and allowed to sit at room temperature for 8-10 hours. After this rest phase, 1 μL from each well was used to reinoculated yet another 96-well deep-well plate preloaded with 400 μL LB-Amp. This cell passage plate was incubated at 37 ºC and shaken at 220 rpm for 16-18 hours. The following morning 50 μL of preculture from each well was added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 μL of 50% glycerol solution. These glycerol stocks were

stored at -80°C for future inoculation. The sequences of transformants generated in this manner were confirmed by LevSeq long-read sequencing.

4.4.3. Protocols for the Screening of ParPgb Variants

4.4.3.1. 96-Well Plate Library Expression

The wells of a 2 mL 96-well deep-well plates were filled with 400 µL LB-Amp. Previously generated 96-well plate glycerol stocks were removed from -80°C storage and placed on dry ice. Multichannel pipet tips were used to scratch the frozen glycerol stock surface and used to inoculate the aforementioned deep-well plate. These pre-expression cultures were incubated at 37 ºC and shaken at 220 rpm for 16-18 hours. For expression cultures, the following morning 50 µL of these precultures were used to inoculate 900 µL of Terrific Broth (TB) (Research Products Int.) with 100 µg/mL of ampicillin (TB-amp) per well in 96-well deep-well plates. These expression cultures were initially incubated at 37 ºC and 220 rpm for 2.5 hours, at which point they were allowed to sit at room temperature for 30 minutes. Expression of proteins was induced with isopropyl-β-D-thiogalactoside (IPTG) and cellular heme production was increased with 5-aminolevulinic acid (ALA). An induction mixture containing IPTG and ALA in TB-amp (50 µL) was added to each well such that the final concentrations of IPTG and ALA were 0.5 mM and 1.0 mM, respectively. The total culture volumes were 1 mL. The plates were then incubated at 22 ºC and 220 rpm overnight.

4.4.3.2. 96-Well Plate Library Reactions and Screening

Expression cultures containing *E. coli* expressing hemoproteins of interest were centrifuged at 4000 × g for 10 minutes at 4 ºC. The supernatant was discarded, and nitrogen-free M9 minimal medium (M9-N, 380 µL) was added to each well. The pellets were resuspended in this medium via shaking at room temperature for 30 minutes. The plates were then transferred into a vinyl Coy anaerobic chamber (0 – 30 ppm O_2). To each well was added 20 µL of a MeCN solution with 200 mM of the desired styrene substrate

and 300 mM of ethyl diazoacetate (EDA). The final reaction volume was 400 μL, and the final concentrations of the styrene and EDA were 10 mM and 15 mM, respectively. The plates were then sealed carefully with a foil cover and shaken at room temperature for 16 hours in the Coy chamber. Once complete, plates were worked up for processing by adding 600 μL of a 1:1 solution of ethyl acetate:cyclohexane containing 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). A silicone sealing mat (AWSM1003S, ArcticWhite) was used to cover the plate and the two layers were thoroughly mixed by rapid inversion of the plate. The plate was then centrifuged (5000 × g for 5 minutes at room temperature) to separate the phases. Afterwards, a 200 μL aliquot of the organic layer was transferred to a GC vial insert in a GC vial, and the samples were analyzed by GC-FID.

4.4.4. Machine Learning Details

The initial training data for the ParPgb campaign was obtained by merging sequencing data with screening yield data. Measured yields were averaged for sequences with the same amino acid combination and normalized to the yield of the *cis* product formation of the parent variants (WYLQF) on each plate. These normalized values were used for model training and acquiring new points, which followed the same protocol as the computational simulations on GB1 and TrpB. For the wet-lab campaign, we trained the model with onehot encodings, the DNN ensemble with 5 models and bootstrapping using 90% of the available training data for each model, and Thompson sampling as the acquisition function. These design choices correspond to the most consistent strategy based on the computational simulations. Detailed instructions on how to reproduce our results and run ALDE for other engineering campaigns are provided at https://github.com/jsunn-y/ALDE.

Most Bayesian optimization algorithms consist of two main components: (1) a probabilistic surrogate model of the objective function and (2) an acquisition function. The surrogate model predicts the objective function values at unobserved inputs, while the acquisition function quantifies the potential benefit of evaluating any given batch of inputs based on these predictions. In each iteration of the Bayesian optimization loop, a new batch of inputs

is selected by maximizing the acquisition function. After evaluating the objective function at these new inputs, the surrogate model is updated, and the process repeats. Below, we describe in detail the probabilistic models and acquisition functions explored in this work, which were implemented using BoTorch[65] and GPyTorch.[66]

4.4.4.1. Probabilistic Models for Bayesian Optimization

Let $\boldsymbol{X}$ denote the input space (i.e., the space of feasible protein sequences) and let $f: \boldsymbol{X} \rightarrow \boldsymbol{R}$ denote the objective function (i.e., the metric we wish to optimize). In this work, we explore four classes of probabilistic surrogate models of the objective function: regular Gaussian processes (GP), deep kernel Gaussian processes (DKL), deep ensembles (DNN ensemble), and boosting ensembles.

Gaussian Processes. A Gaussian process model is defined in terms of a prior mean function $\mu_0: \boldsymbol{X} \rightarrow \boldsymbol{R}$ and a prior covariance function $\mathrm{K}_0: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \boldsymbol{R}$ and it encodes a Bayesian prior distribution over f. Given a dataset of n evaluations of the objective function, denoted as $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, one can derive the posterior distribution of f given $\mathcal{D}_n$. If these evaluations are corrupted by i.i.d. additive Gaussian noise, i.e., $y_i = f(x_i) + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Gaussian with mean zero and variance σ^2, the posterior is again a Gaussian process characterized by a posterior mean function $\mu_n: \boldsymbol{X} \rightarrow \boldsymbol{R}$ and a posterior covariance function $\mathrm{K}_n: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \boldsymbol{R}$. These functions can be computed in closed form in terms of the prior mean and covariance functions as well as the data using the classical Gaussian process regression formulas. [67] The noise variance σ^2 and other hyperparameters of the model (such as the lengthscale parameters) can be estimated by maximizing the log marginal likelihood.

Deep Kernel Learning. Gaussian process models with classical covariance functions, such as the Matern or squared exponential covariance functions, are known to perform poorly in high-dimensional input spaces.[32] To address this limitation, Wilson et al. (2015) proposed *deep kernel learning*.[29] Succinctly, this approach uses a covariance function of the form $\mathrm{K}(x, x') = k\big(\phi_w(x), \phi_w(x')\big)$, where k is a regular covariance function (e.g., squared exponential) and ϕ_w is a deep neural network with weights w. These weights are

treated like hyperparameters of the model, which can also be estimated by maximizing the log marginal likelihood.

Boosting Ensembles. Boosting models leverage a sequential training strategy where each new model is trained to correct the errors of the previously combined models.[68] The final prediction is often a weighted sum of the predictions made by earlier models, where the weights reflect each model's accuracy. Unlike methods such as bagging, which train models independently and in parallel, boosting specifically designs each new model to address the weaknesses of the existing ensemble, thereby creating a strong predictive model from a sequence of weaker ones. While boosting does not inherently offer a probabilistic interpretation like Bayesian methods, it is highly effective for reducing bias and variance in predictive modeling tasks. Here, we train the boosting ensembles with bootstrapping; each ensemble consists of 5 models where 90% of the total training data is randomly seen during training.

Deep Ensembles. Deep neural network (DNN) ensemble models are constructed by training identical deep neural network architectures multiple times, each with different random initializations of the weight parameters. Here, we train the deep ensembles with bootstrapping; each ensemble consists of 5 models where 90% of the total training data is randomly seen during training. These independently trained networks are then collectively used as if they were samples from a Bayesian posterior distribution over the objective function f. Unlike Gaussian processes, deep ensembles lack a proper Bayesian interpretation. However, Izmailov and Wilson argue it is possible to see these models as a form of approximate Bayesian inference.[59] We adopt this view in our work.

4.4.4.2. Acquisition Functions for Bayesian Optimization

Expected Improvement. The expected improvement (EI) acquisition function is given by $\alpha_n(x) = E_n[\{f(x) - f_n^*\}^+]$, where $f_n^* = \max_{i=1,\ldots,n} f(x_i)$ and the expectation is computed with respect to the posterior distribution given $\mathcal{D}_n$.[51] For Gaussian posterior distributions and noise-free observations (where f_n^* is a constant rather than a random variable), the EI can be expressed in a closed form using the posterior mean and variance. In scenarios where

these conditions do not hold, computing the EI often requires approximate calculation, typically through Monte Carlo sampling techniques. When extending the EI to the batch setting, the acquisition function becomes $\alpha_n(X) = E_n\left[\{\max_{x\in X} f(x) - f_n^*\}^+\right]$, where $X = (x_1, \ldots, x_q) \in \boldsymbol{X}^q$ is a batch of q inputs (qEI). Maximizing the batched EI poses significant computational challenges due to the requirement to optimize over $\boldsymbol{X}^q$. However, by exploiting the submodularity of the acquisition function, an efficient approximation can be achieved through a greedy optimization strategy, selecting each input in the batch sequentially. In this study, we tested qEI, but it ran slowly without noticeable improvement, so it was not included in the final results.

Upper Confidence Bound. The upper confidence bound (UCB) acquisition function is defined by $\alpha_n(x) = \mu_n(x) + \beta_n^{1/2}\sigma_n(x)$, where $\mu_n(x)$ and $\sigma_n(x)$ are the posterior mean and standard deviation, respectively, and β_n is a parameter that controls the exploration-exploitation trade-off. In our experiments, we set $\beta_n = 4$. While there are sophisticated batch extensions of the UCB acquisition function available in the literature,[69] our approach utilizes a straightforward heuristic. Specifically, we form batches by selecting the q inputs that yield the highest values of $\alpha_n(x)$, evaluated across all discrete x in the design space. The Greedy acquisition function can be thought of as a specific case of UCB with $\beta_n = 0$ so the acquisition function becomes $\alpha_n(x) = \mu_n(x)$. For the frequentist ensemble models, we evaluate $\mu_n(x)$ and $\sigma_n(x)$ as the mean and standard deviation of all models in the ensemble, respectively.

Thompson Sampling. Thompson Sampling (TS) is a randomized selection strategy where the next input to evaluate is obtained by drawing a sample (function) from the posterior distribution of f and selecting the point that maximizes this sample. For the GP and DKL models, we approximate samples from the posterior using 1000 random Fourier features.[70] For the frequentist ensemble models, the random function sample is drawn as one of the models in the ensemble. In the batch setting, each input in the batch is obtained as an independent sample. Unlike the EI and UCB, TS is inherently stochastic as opposed to deterministic; however, we note that since our ensembles have five models each, TS is less stochastic in this setting.

4.4.5. Data Availability

All experimental and simulation data that support the findings of this study are available at https://github.com/jsunn-y/ALDE and https://zenodo.org/records/12196802

4.4.6. Code Availability

All code that accompanies this study is available at https://github.com/jsunn-y/ALDE under the MIT license.

Chapter IV Bibliography

1. Reisenbauer, J.C.; Sicinski, K.M.; Arnold, F.H. Catalyzing the future: recent advances in chemical synthesis using enzymes. *Curr. Opin. Chem. Biol.* **2024**, *83*, 102536.
2. Romero, P.A.; Arnold, F.H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866.
3. Smith, J.M. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225*, 563.
4. Packer, M.S.; Liu, D.R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **2015**, *16*, 379.
5. Wang, Y.; Xue, P.; Cao, M.; Yu, T.; Lane, S.T.; Zhao, H. Directed Evolution: Methodologies and Applications. *Chem. Rev.* **2021**, *121*, 12384.
6. Miton, C.M.; Buda, K.; Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 160.
7. Yang, K.K.; Wu, Z.; Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687.
8. Yang, J.; Li, F.-Z.; Arnold, F.H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **2024**, *10*, 226.
9. Wittmann, B.J.; Johnston, K.E.; Wu, Z.; Arnold, F.H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 11.
10. Freschlin, C.R.; Fahlberg, S.A.; Romero, P.A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **2022**, *75*, 102713.
11. Aghazadeh, A.; Nisonoff, H.; Ocal, O.; Brookes, D.H.; Huang, Y.; Koyluoglu, O.O.; Listgarten, J.; Ramchandran, K. Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nat. Commun.* **2021**, *12*, 5225.
12. Wu, Z.; Kan, S.B.J.; Lewis, R.D.; Wittmann, B.J.; Arnold, F.H. Machine learning-assisted directed evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **2019**, *116*, 8852.
13. Wittmann, B.J.; Yue, Y.; Arnold, F.H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **2021**, *12*, 1026.
14. Qiu, Y.; Hu, J.; Wei, G.-W. Cluster learning-assisted directed evolution. *Nat. Comput. Sci.* **2021**, *1*, 809.
15. Qiu, Y.; Wei, G.-W. CLADE 2.0: Evolution-Driven Cluster Learning-Assisted Directed Evolution. *J. Chem. Inf. Model.* **2022**, *62*, 4629.
16. Hie, B.; Bryson, B.D.; Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Syst.* **2020**, *11*, 461.
17. Greenman, K.P.; Amini, A.P.; Yang, K.K. Benchmarking uncertainty quantification for protein engineering. *PLoS Comput. Biol.* **2025**, *21*, e1012639.
18. Hie, B.L.; Yang, K.K. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* **2022**, *72*, 145.
19. Vornholt, T.; Mutný, M.; Schmidt, G.W.; Schellhaas, C.; Tachibana, R.; Panke, S.; Ward, T.R.; Krause, A.; Jeschek, M. Enhanced Sequence-Activity Mapping

and Evolution of Artificial Metalloenzymes by Active Learning. *ACS Cent. Sci.* **2024**, *10*, 1357.
20. Thomas, N.; Belanger, D.; Xu, C.; Lee, H.; Hirano, K.; Iwai, K.; Polic, V.; Nyberg, K.D.; Hoff, K.; Frenz, L.; Emrich, C.A.; Kim, J.W.; Chavarha, M.; Ramanan, A.; Agresti, J.J.; Colwell, L.J. Engineering highly active and diverse nuclease enzymes by combining machine learning and ultra-high-throughput screening. *bioRxiv* **2024**, https://doi.org/10.1101/2024.03.21.585615.
21. Jiang, K.; Yan, Z.; Di Bernardo, M.; Sgrizzi, S.R.; Villiger, L.; Kayabolen, A.; Kim, B.; Carscadden, J.K.; Hiraizumi, M.; Nishimasu, H.; Gootenberg, J.S.; Abudayyeh, O.O. Rapid protein evolution by few-shot learning with a protein language model. *bioRxiv* **2024**, https://doi.org/10.1101/2024.07.17.604015
22. Landwehr, G.M.; Bogart, J.W.; Magalhaes, C.; Hammarlund, E.G.; Karim, A.S.; Jewett, M.C. Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nat. Commun.* **2025**, *16*, 865.
23. Ding, K.; Chin, M.; Zhao, Y.; Huang, W.; Mai, B.K.; Wang, H.; Liu, P.; Yang, Y.; Luo, Y. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nat. Commun.* **2024**, *15*, 6392.
24. Romero, P.A.; Krause, A.; Arnold, F.H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci.* **2012**, *110*, E193.
25. Rapp, J.T.; Bremer, B.J.; Romero, P.A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **2024**, *1*, 97.
26. Hu, R.; Fu, L.; Chen, Y.; Chen, J.; Qiao, Y.; Si, T. Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Brief. Bioinform.* **2023**, *24*, 1.
27. Bal, M.I.; Sessa, P.G.; Mutný, M.; Krause, A. Optimistic Games for Combinatorial Bayesian Optimization with Application to Protein Design. *arXiv* **2025**, https://doi.org/10.48550/arXiv.2409.18582
28. Gantz, M.; Mathis, S.V.; Nintzel, F.E.H.; Zurek, P.J.; Knaus, T.; Patel, E.; Boros, D.; Weberling, F.-M.; Kenneth, M.R.A.; Klein, O.J.; Medcalf, E.J.; Moss, J.; Herger, M.; Kaminski, T.S.; Mutti, F.G.; Lio, P.; Hollfelder, F. Microdroplet screening rapidly profiles a biocatalyst to enable its AI-assisted engineering. *bioRxiv* **2024**, https://doi.org/10.1101/2024.04.08.588565.
29. Wilson, A.G.; Hu, Z.; Salakhutdinov, R.; Xing, E.P. Deep Kernel Learning. *PMLR* **2016**, *51*, 370.
30. Abe, T.; Buchanan, E.K.; Pleiss, G.; Zemel, R.; Cunningham, J.P. Deep ensembles work, but are they necessary? *NeurIPS* **2022**, *36*, 33646.
31. Bowden, J.; Song, J.; Chen, Y.; Yue, Y.; Desautels, T.A. Deep Kernel Bayesian Optimization. *Conf. UAI* **2021**.
32. Eriksson, D.; Pearce, M.; Gardner, J.; Turner, R.D.; Poloczek, M. Scalable Global Optimization via Local Bayesian Optimization. *NeurIPS* **2019**, *33*, 5496.
33. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. **2023**, *379*, 1123.

34. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Towards Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mech. Intell.* **2022**, *44*, 7112.
35. Coelho, P.S.; Brustad, E.M.; Kannan, A.; Arnold, F.H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* **2013**, *339*, 307.
36. Pesce, A.; Bolognesi, M.; Nardini, M. Protoglobin: Structure and Ligand-Binding Properties. *Adv. Microb. Physiol.* **2013**, *63*, 79.
37. Knight, A.M.; Kan, S.B.J.; Lewis, R.D.; Brandenberg, O.F.; Chen, K.; Arnold, F.H. Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes. *ACS Cent. Sci.* **2018**, *4*, 372.
38. Porter, N.J.; Danelius, E.; Gonen, T.; Arnold, F.H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **2022**, *144*, 8892.
39. Gao, S.; Das, A.; Alfonzo, E.; Sicinski, K.M.; Rieger, D.; Arnold, F.H. Enzymatic Nitrogen Incorporation Using Hydroxlamine. *J. Am. Chem. Soc.* **2023**, *145*, 20196.
40. Hanley, D.; Li, Z.-Q.; Gao, S.; Virgil, S.C.; Arnold, F.H.; Alfonzo, E. Stereospecific Enzymatic Conversion of Boronic Acids to Amines. *J. Am. Chem. Soc.* **2024**, *146*, 19160.
41. Park, Y.; Metzger, B.P.H.; Thornton, J.W. Epistatic drift causes gradual decay of predictability in protein evolution. *Science* **2022**, *376*, 823.
42. 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.* **2017**, *35*, 128.
43. Long, Y.; Mora, A.; Li, F.-Z.; Gürsoy, E.; Johnston, K.E.; Arnold, F.H. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *ACS Synth. Biol.* **2025**, *14*, 230.
44. 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* **2016**, *5*, e16965.
45. Johnston, K.E.; Almhjell, P.J.; Watkins-Dulaney, E.J.; Liu, G.; Porter, N.J.; Yang, J.; Arnold, F.H. A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proc. Natl. Acad. Sci.* **2024**, *121*, e2400439121.
46. Georgiev, A.G. Interpretable numerical descriptors of amino acid space. *J. Comput. Biol.* **2009**, *16*, 703.
47. Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y.S. Evaluating protein transfer learning with TAPE. *NeurIPS* **2019**, *33*, 9689.
48. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **2021**, *118*, e2016239118.
49. Michael, R.; Kæstel-Hansen, J.; Groth, P.M.; Bartels, S.; Salomon, J.; Tian, P.; Hatzakis, N.S.; Boomsma, W. A systematic analysis of regression models for protein engineering. *PLoS Comput. Biol.* **2024**, *20*, e1012061.

50. Hvarfner, C.; Hellsten, E.O.; Nardi, L. Vanilla Bayesian Optimization Performs Great in High Dimensions. *PMLR* **2024**, *235*, 20793.
51. Letham, B.; Karrer, B.; Ottoni, G.; Bakshy, E. Constrained Bayesian Optimization with Noisy Experiments. *Bayesian Anal.* **2019**, *14*, 195.
52. Luo, Y.; Liu, Y.; Peng, J. Calibrated geometric deep learning improves kinase-drug binding predictions. *Nat. Mach. Intell.* **2023**, *5*, 1390.
53. Stanton, S.; Maddox, W.; Wilson, A.G. Bayesian Optimization with Conformal Prediction Sets. *PMLR* **2023**, *206*, 959.
54. Fannjiang, C.; Bates, S.; Angelopoulos, A.N.; Listgarten, J.; Jordan, M.I. Conformal prediction under feedback covariate shift for biomolecular design. *Proc. Natl. Acad. Sci.* **2022**, *119*, e2204569119.
55. Fannjiang, C.; Listgarten, J. Is Novelty Predictable? *Cold Spring Harb. Perspect.* **2024**, *16*, a041469.
56. Ober, S.W.; Rasmussen, C.E.; van der Wilk, M. The promises and pitfalls of deep kernel learning. *PMLR* **2021**, *161*, 1206.
57. Fröhlich, C.; Bunzel, H.A.; Buda, K.; Mulholland, A.J.; van der Kamp, M.W.; Johnsen, P.J.; Leiros, H.-K.S.; Tokuriki, N. Epistasis rises from shifting the rate-limiting step during enzyme evolution of a β-lactamase. *Nat. Catal.* **2024**, *7*, 499.
58. Hollmann, F.; Sanchis, J.; Reetz, M.T. Learning from Protein Engineering by Deconvolution of Multi-Mutational Variants. *Angew. Chem. Int. Ed.* **2024**, *63*, e202404880.
59. Wilson, A.G.; Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *NeurIPS* **2020**, *34*, 4697.
60. Li, F.-Z.; Yang, J.; Johnston, K.E.; Gürsoy, E.; Yue, Y.; Arnold, F.H. Evaluation of Machine Learning-Assisted Directed Evolution Across Diverse Combinatorial Landscapes. *bioRxiv* **2024**, https://doi.org/10.1101/2024.10.24.619774.
61. 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.* **2022**, *11*, 1313.
62. Yang, J.; Ducharme, J.; Johnston, K.E.; Li, F.-Z.; Yue, Y.; Arnold, F.H. DeCOIL: Optimization of Degenerate Codon Libraries for Machine Learning-Assisted Protein Engineering. *ACS Synth. Biol.* **2023**, *12*, 2444.
63. Nov, Y. When Second Best Is Good Enough: Another Probabilistic Look at Saturation Mutagenesis. *Appl. Environ. Microbiol.* **2012**, *78*, 258.
64. Gibson, D.G.; Young, L.; Chuang, R.-Y.; Venter, J.C.; Hutchison III, C.A.; Smith, H.O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **2009**, *6*, 343.
65. Balandat, M.; Karrer, B.; Jiang, D.R.; Daulton, S.; Letham, B.; Wilson, A.G.; Bakshy, E. BOTORCH: a framework for efficient monte-carlo Bayesian optimization. *NeurIPS* **2020**, *34*, 21524.
66. Gardner, J.R.; Pleiss, G.; Bindel, D.; Weinberger, K.Q.; Wilson, A.G. GPyTorch: blackbox matrix-matrix Gaussian process interference with GPU acceleration. *NeurIPS* **2018**, *32*, 7587.
67. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press, 2018.

68. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining* **2016**, *22*, 785.
69. Desautels, T.; Krause, A.; Burdick, J.W. Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization. *JMLR* **2014**, *15*, 4053.
70. Rahimi, A.; Recht, B. Random features for large-scale kernel machines. *NeurIPS* **2007**, *21*, 1177.

Appendix C

SUPPLEMENTARY INFORMATION FOR CHAPTER IV

C.1. General Information

C.1.1. Safety Statement

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure. Other than that, no unexpected or unusually high safety concerns were raised with these methods. Safety notes for individual synthetic procedures will be documented alongside the procedure.

C.1.2. General Information

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure to chemicals. Reagents and solvents were obtained commercially (Sigma-Aldrich, Alfa Aesar, VWR, Fisher Scientific, Matrix Scientific, Oakwood Chemical, TCI America, and other suppliers) and used without prior purification unless otherwise stated. Organic solutions were concentrated under reduced pressure on an IKA RV 10 rotary evaporator. Thin-layer chromatography (TLC) was performed on commercial Millipore Silica Gel 60 plates containing the F254 fluorescent indicator. Visualization of the developed chromatographs was performed by irradiation with UV light or by treatment with an appropriate TLC staining solution (e.g., Ceric Ammonium Molybdate, $KMnO_4$, or Bromocresol Green) followed by heating if necessary. Chromatographic purification was accomplished by flash chromatography using a Biotage Isolera One instrument.

C.1.3. Spectral Information

All NMR spectra were obtained at the Caltech Liquid NMR Facility. For all cyclopropane compounds, ^{1}H NMR were recorded on a Bruker Prodigy 400 MHz instrument (400 MHz

and 101 MHz). For intermediates, ^{1}H spectra were also recorded using a Varian 300 MHz spectrometer (300 MHz), a Varian 500 MHz spectrometer (500 MHz), and a Varian 600 MHz spectrometer (600 MHz). ^{1}H spectra are referred to residual $CDCl_3$ solvent signals referenced at δ 7.26 ppm. Data for ^{1}H NMR are reported as follows: chemical shift (δ ppm), integration, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentad, sext = sextet, hept = heptet, m = multiplet, br s = broad singlet), and coupling constant (Hz).

C.1.4. Gas Chromatography Data

GC chromatography (GC) was performed on an Agilent Technologies 7820A GC system equipped with a split-mode capillary injection system and flame-ionization detector. For achiral analyses, an Agilent J&W HP-5 Column was used as the stationary phase. For chiral analyses, the specific stationary phase is provided along with the chiral traces.

C.2. Cloning and Sequence Design

C.2.1. Relevant Protoglobin Sequences

Table C-1. The amino acid sequences of wild-type *Par*Pgb and the previously engineered variant, ParLQ. Residues in each sequence highlighted in yellow are the sites which were investigated in this work for wet-lab experimentation: W56, Y57, L59, Q60, and F89 in ParLQ.

Protein Variant	Amino Acid Sequence
wild-type *Pyrobaculum arsenenticum* Protoglobin (*Par*Pgb)	MAVPGYDFGKVPDAPISDADFESLKKTVM WGEEDEKYRKMACEALKGQVEDILDLWY GWVGSNQHLIYYFGDKSGRPIPQYLEAVRK RFGLWIIDTLCKPLDRQWLNYMYEIGLRHH RTKKGKTDGVDTVEHIPLRYMIAFIAPIGLTI KPILEKSGHPPEAVERMWAAWVKLVVLQV AIWSYPYAKTGEW
Pyrobaculum arsenenticum Protoglobin W59L V60Q (ParLQ)	MAVPGYDFGKVPDAPISDADFESLKKTVM WGEEDEKYRKMACEALKGQVEDILDLWY GLQGSNQHLIYYFGDKSGRPIPQYLEAVRK RFGLWIIDTLCKPLDRQWLNYMYEIGLRHH RTKKGKTDGVDTVEHIPLRYMIAFIAPIGLTI KPILEKSGHPPEAVERMWAAWVKLVVLQV AIWSYPYAKTGEW

DNA Sequence of ParLQ

ATGGCGGTTCCCGGCTACGATTTTGGCAAAGTCCCGGATGCCCCAATCTCAGA
CGCGGATTTTGAGAGTTTAAAAAAAACCGTGATGTGGGGTGAGGAAGATGAG
AAATATCGCAAAATGGCTTGCGAAGCCTTAAAGGGTCAAGTAGAAGATATTTT
AGATTTGTGGTACGGCCTGCAGGGAAGCAATCAACACCTTATCTACTACTTCG
GTGATAAGAGTGGTCGTCCAATTCCGCAATACCTGGAAGCGGTCCGCAAGCGT
TTCGGGTTGTGGATCATTGATACATTGTGTAAGCCACTGGACCGCCAGTGGTTG
AATTACATGTACGAAATTGGCCTTCGCCATCACCGTACCAAGAAAGGGAAGAC
AGATGGCGTAGATACTGTTGAACATATCCCATTACGCTACATGATTGCTTTCATC
GCTCCCATCGGTCTGACTATTAAGCCGATCTTGGAAAAATCGGGACATCCGCC
AGAGGCCGTGGAGCGTATGTGGGCAGCATGGGTTAAGTTGGTGGTGTTACAG
GTAGCTATCTGGTCGTACCCCTATGCAAAGACGGGCGAATGGCTCGAGCACCA
CCACCACCACCAC

Table C-2. Codons utilized when ALDE-predicted mutations were incorporated into the ParLQ DNA sequence. Codons which are generally recognized as the most prevalently found in the *E. coli* genome were selected.

Amino Acid	Codon
A	GCC
C	TGC
D	GAT
E	GAA
F	TTT
G	GGC
H	CAT
I	ATT
K	AAA
L	TTA
M	ATG
N	AAC
P	CCG
Q	CAG
R	CGT
S	AGC
T	ACC
V	GTG
W	TGG
Y	TAT

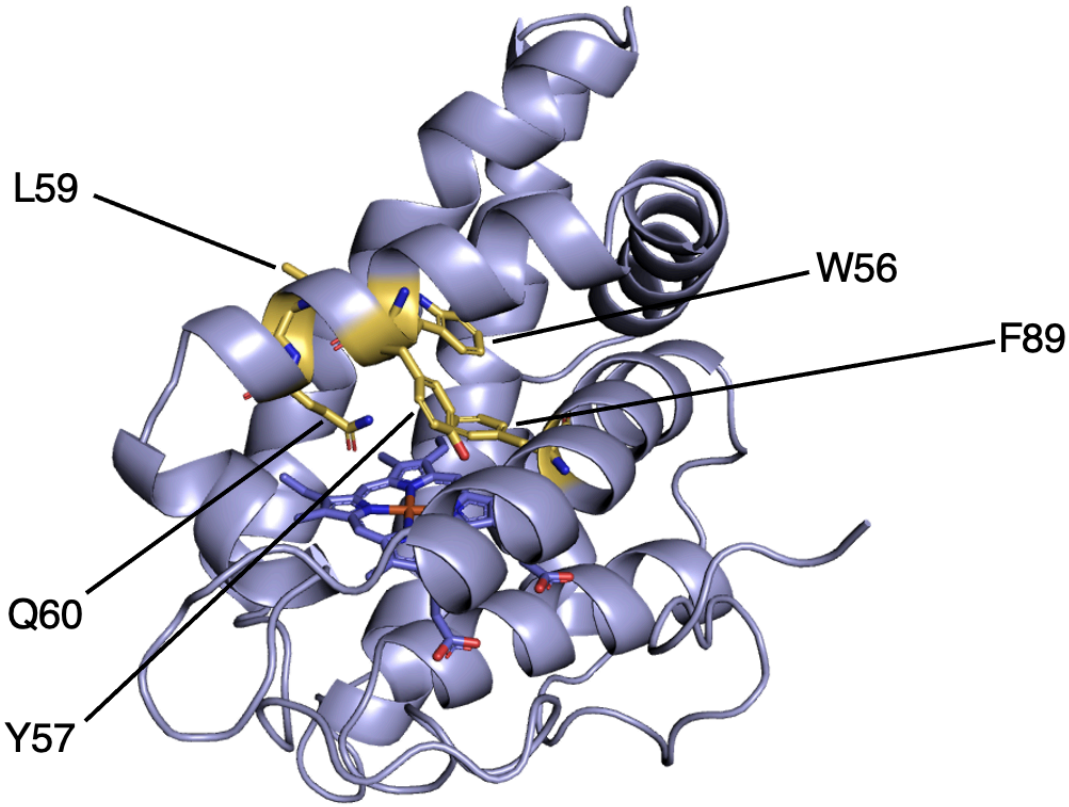


Figure C-1. Homology model of ParLQ with docked heme. Structural modeling was performed with AlphaFold3.[1] Residues which were mutated in this study are illustrated in yellow.

C.2.2. Nomenclature for Variant Naming

Single-site mutants are named using the standard nomenclature: (original AA)(site)(new AA). The mutant W56F refers to a variant of ParLQ mutated from tryptophan to phenylalanine. Five-site multi-mutants are named as a string of the five amino acids to which the positions of interest have been mutated. The variant MPFDY refers to a variant of ParLQ bearing the mutations W56M, Y57P, L59F, Q60D, and F89Y. In this nomenclature, ParLQ is named WLYQF.

C.2.3. Primer Design for Site-Saturation Mutagenesis (SSM)

General Cloning Primers

Table C-3. Primers 007 and primers 008 are used to generate amplicons of linearized pET-22b(+)[2] backbone. Primers 005 and 006 were used with primers internal to the protoglobin gene (**Table C-4**) to generate mutant protoglobin genes. All primers were ordered from IDT (Coralville, IA).

Primer Name	**Direction**	**Sequence**	**Description**
005	Forward	5'-GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG-3'	Upstream of N-term, anneals with 007
006	Reverse	5'-GCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAG-3'	Downstream of C-term, anneals with 008
007	Reverse	5'-CATATGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTC-3'	Upstream of N-term, anneals with 005
008	Forward	5'-CTCGAGCACCACCACCACCACCACTGAGATCCGGC-3'	Downstream of C-term, anneals with 006

Cloning Primers for Site-Saturation Mutagenesis

Table C-4. Primers containing degenerate codons at sites of interest for SSM. Primers were used in conjunction with either primer 005 or 006 (**Table S3**) to generate mutant protoglobin genes. All primers were ordered from IDT (Coralville, IA).

Primer Name	**Direction**	**Sequence**	**Description**
ParLQ_56X_F	Forward	5'-GGTCAAGTAGAAGATATTTTAGATTTGNNKTACGGCCTGCAGGGAAGCAATC-3'	SSM for site 56
ParLQ_56X_R	Reverse	5'-CAAATCTAAAATATCTTCTACTTGACCCTTTAAGGCTTC-3'	SSM for site 56
ParLQ_57X_F	Forward	5'-CAAGTAGAAGATATTTTAGATTTGTGGNNKGGCCTGCAGGGAAGCAATCAAC-3'	SSM for site 57
ParLQ_57X_R	Reverse	5'-CCACAAATCTAAAATATCTTCTACTTGACCCTTTAAGGC-3'	SSM for site 57
ParLQ_59X_F	Forward	5'-TATTTTAGATTTGTGGTACGGCNNKCAGGGAAGCAATCAACACCTTATCTACTAC-3'	SSM for site 59
ParLQ_59X_R	Reverse	5'-GCCGTACCACAAATCTAAAATATCTTCTACTTGACCC-3'	SSM for site 59
ParLQ_60X_F	Forward	5'-TTTGTGGTACGGCCTGNNKGGAAGCAATCAACACCTTATCTACTACTTCGG-3'	SSM for site 60
ParLQ_60X_R	Reverse	5'-CAGGCCGTACCACAAATCTAAAATATCTTCTACTTG-3'	SM for site 60
ParLQ_89X_F	Forward	5'-GCGGTCCGCAAGCGTNNKGGGTTGTGGATC-3'	SSM for site 89

ParLQ_89X_R	Reverse	5'-CCMNNACGCTTGCGGACCGCTTCCAGG-3'	SSM for site 89
ParLQ_quad NNK_F	Forward	5'-CAAGTAGAAGATATTTTAGATTTG NNKNNKGGCNNKNNKGGAAGCAATCAACA CCTTATC-3'	Multisite SSM for sites 56, 57, 59, and 60
ParLQ_quad NNK_R	Reverse	5'-GATAAGGTGTTGATTGCTTCCMNNMNNGCC MNNMNNCAAATCTAAAATATCTTCTACTTG-3'	Multisite SSM for sites 56, 57, 59, and 60

C.3. Cloning Protocols and Results

C.3.1. Protocols for the Cloning of Random ParLQ Variants

C.3.1.1. Cloning for Single Site-Saturation Mutagenesis (SSM)

Chemically competent *Escherichia coli* (*E. coli*) cells (T7 Express Competent *E. coli*) were purchased from New England Biolabs (NEB, Ipswich, MA). Additionally, Phusion polymerase and *Dpn*I were purchased from NEB. SSM experiments were performed using primers bearing degenerate codons (NNK) using a modified QuikChange™ protocol (**Table S4**).[3] The following PCR amplicons were generated using a parent plasmid (pET-22b(+) harboring ParLQ):

Table C-5. Primer combinations for the generation of PCR amplicons for the construction of expression plasmids containing mutagenized ParLQ variants. '*site*' refers to the specific site which is being saturated.

Fragment Name	**Forward Primer**	**Reverse Primer**
SSM_Frag1	005	ParLQ_*site*_R
SSM_Frag2	ParLQ_*site*_F	006
pET_Backbone	008	007

The PCR conditions were as follows (final concentrations): Phusion HF Buffer 1x, 0.2 mM dNTPs each, 0.5 μM of forward primers, 0.5 μM reverse primer, and 0.02 U/μL of Phusion polymerase. The standard Phusion PCR protocol was used.[4] Upon completion of PCRs, the remaining template was digested with *Dpn*I. Gel purification was performed with a Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA). The purified PCR product was then assembled using the Gibson assembly protocol.[5]

C.3.1.2. Transformation of E. coli with the Genes Coding for Single-Site Mutants

96-Well deep-well plates were shaken in an INFORS HT Multitron Shaker in all instances. The assembly products obtained were used to transform T7 Express Competent *E. coli* (High Efficency) cells (NEB, Ipswich, MA) following the protocol recommended by the manufacturer. Upon heat-shock, freshly transformed *E. coli* cells were recovered in 0.4 mL Luria-Bertani medium (LB) (Research Products Int.) at 37 ºC with shaking at 220 rpm for 30 minutes before being plated on LB-agar plates with 100 μg/mL ampicillin (LB-Amp agar plates). Single colonies from LB-agar plates were picked using sterilized toothpicks to individually inoculate 400 μL of LB containing 100 μg/mL of ampicillin (LB-Amp) in 2-mL 96-well deep-well plates. The plates were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 μL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 μL of 50% glycerol solution. These glycerol stocks were stored at -80 °C. Additionally, the sequences of protoglobin genes contained in every well were sequenced using evSeq.[6]

C.3.1.3. Rearray of Single-Site Mutants:

Single-site mutagenesis mutants were rearrayed from the previously described randomly picked plate to reduce screening burden and to collect data in triplicate. The sequences of protoglobin genes contained in every well of the randomly picked plates were collected using evSeq,[6] yielding the following plate maps:

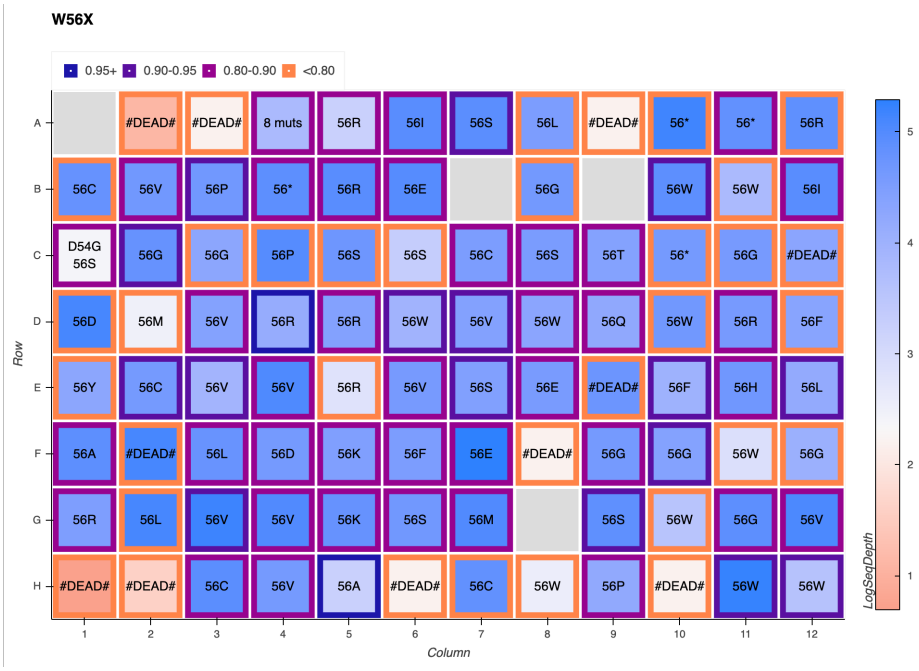


Figure C-2. evSeq data for single site-saturation mutagenesis of ParLQ at site W56. Any missing residues were found in a second round of cloning, sequencing, and picking.

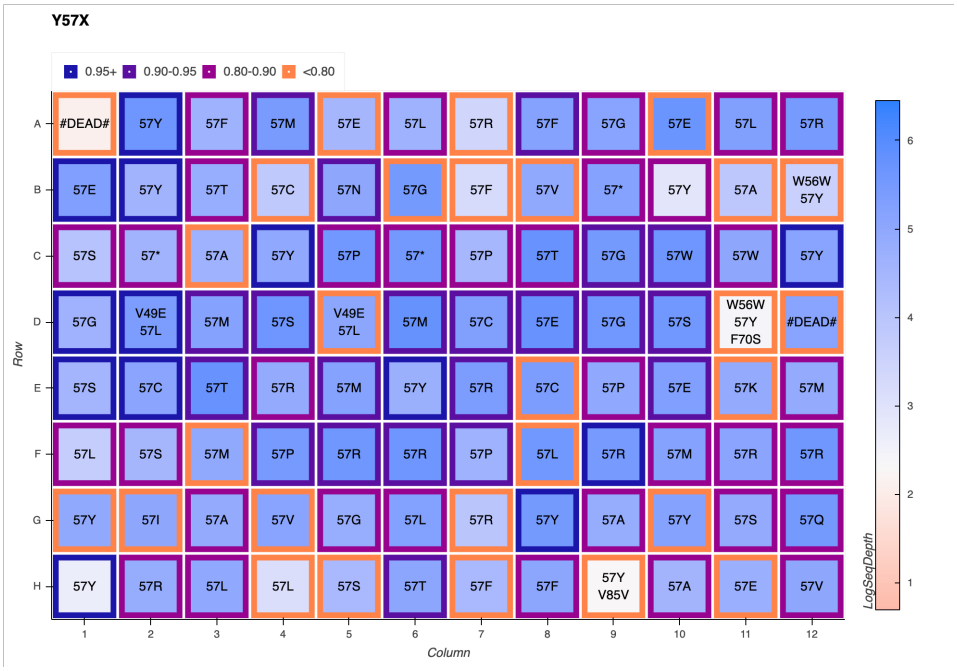


Figure C-3. evSeq data for single site-saturation mutagenesis of ParLQ at site Y57. Any missing residues were found in a second round of cloning, sequencing, and picking.

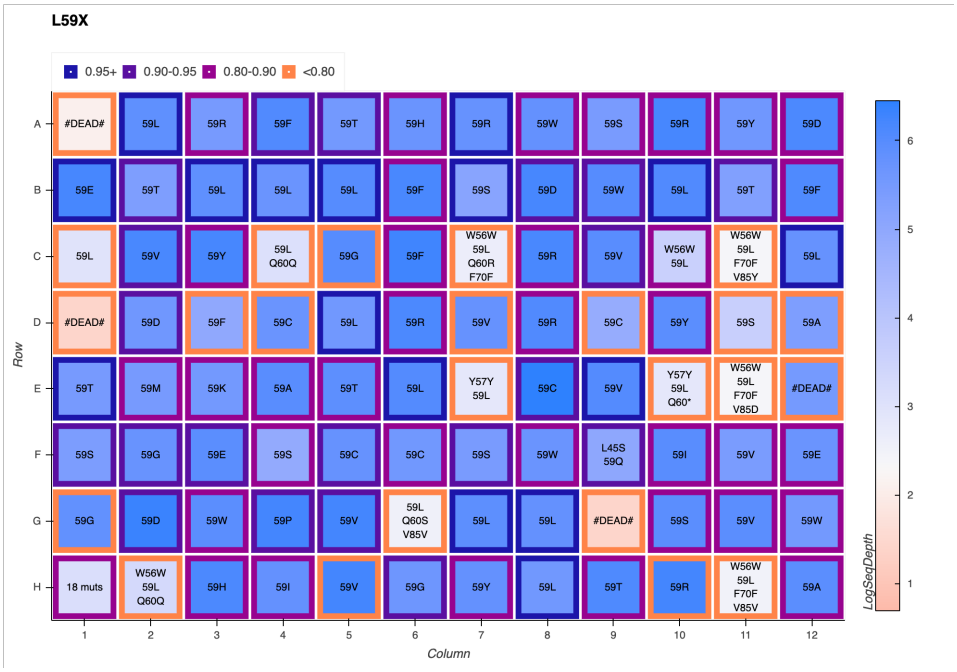


Figure C-4. evSeq data for single site-saturation mutagenesis of ParLQ at site L59. Any missing residues were found in a second round of cloning, sequencing, and picking.

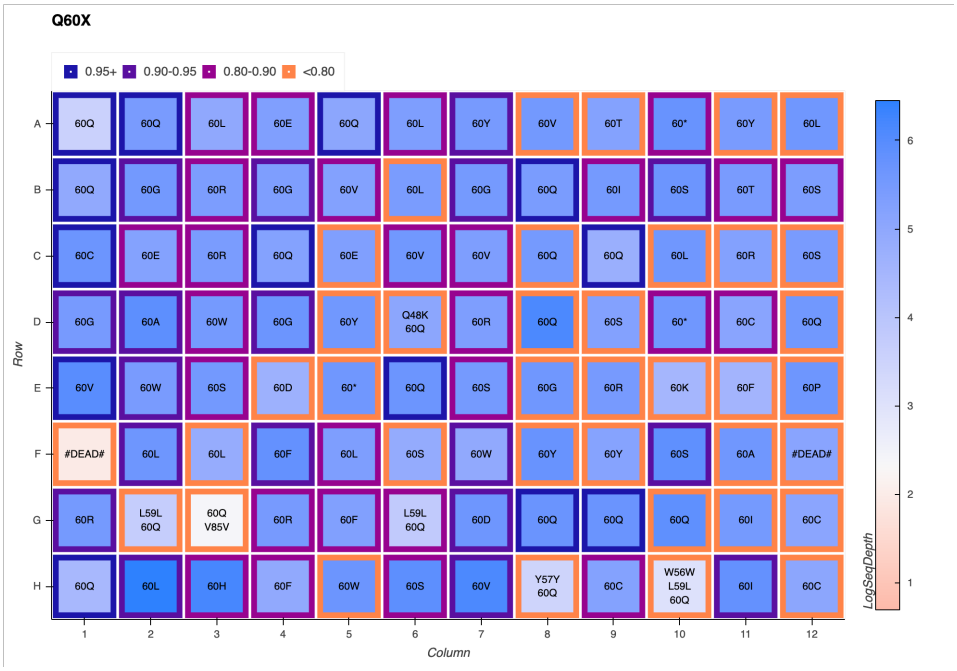


Figure C-5. evSeq data for single site-saturation mutagenesis of ParLQ at site Q60. Any missing residues were found in a second round of cloning, sequencing, and picking.

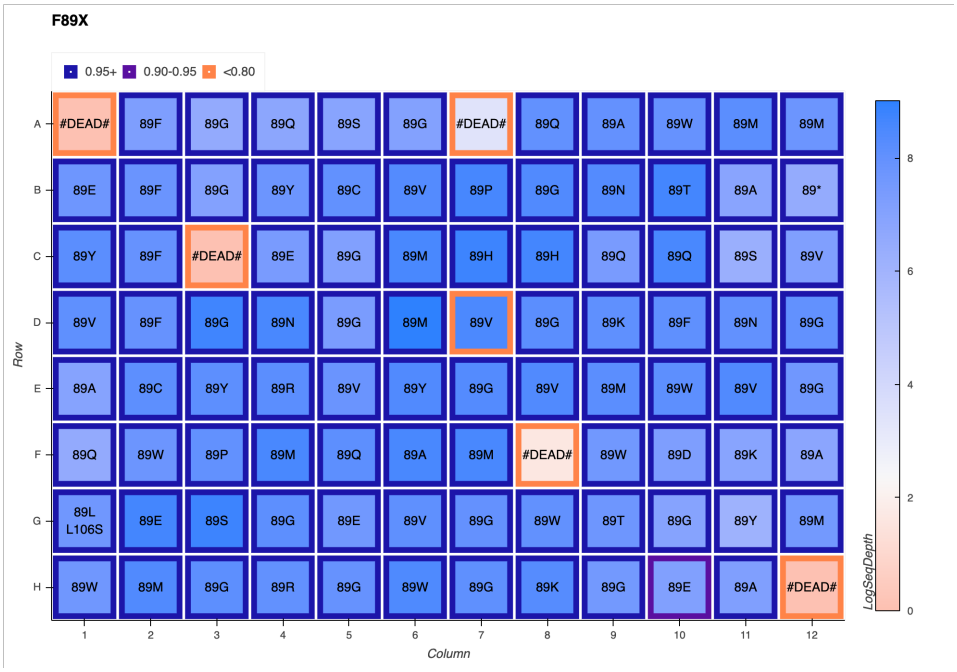


Figure C-6. evSeq data for single site-saturation mutagenesis of ParLQ at site F89. Any missing residues were found in a second round of cloning, sequencing, and picking.

Protoglobin mutants for each site were rearrayed from wells containing each attained single-site mutant. In cases where a mutation was found multiple times on a plate, the well for that mutation with the highest confidence and sequencing depth was selected for rearray plate inoculation. Libraries were arrayed in the following pattern:

Table C-6. Single site mutants were rearrayed in triplicate. If a mutation was not observed in the picked library, then the wells corresponding to that mutation are left sterile.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty
B	Empty	A	C	D	E	F	G	H	I	K	L	Empty
C	Empty	A	C	D	E	F	G	H	I	K	L	Empty
D	Empty	A	C	D	E	F	G	H	I	K	L	Empty
E	Empty	M	N	P	Q	R	S	T	V	W	Y	Empty
F	Empty	M	N	P	Q	R	S	T	V	W	Y	Empty
G	Empty	M	N	P	Q	R	S	T	V	W	Y	Empty
H	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty	Empty

The center 60 wells of a 2-mL 96-well deep-well plate were filled with 400 µL LB-Amp. Previously generated 96-well plates were removed from -80 °C storage and placed on dry ice. Pipet tips were used to scratch the frozen glycerol stock surface and used to inoculate the aforementioned deep-well plate according to **Table S6**. These overnight cultures were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 µL of overnight culture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 µL of 50% glycerol solution. These glycerol stocks were stored at -80°C for future inoculation.

C.3.1.4. Cloning for Multisite-Saturation Mutagenesis

Mutations were simultaneously incorporated as with single site-saturation mutagenesis using the ParLQ_quadNNK primers. Upon transformation, the 4-site library

transformation was recovered in 0.4 mL of LB medium at 37 °C and 220 rpm for 30 minutes. From this recovered transformation mixture, 50 µL was transferred into 6 mL of LB-Amp in a 15-mL culture tube. This culture was allowed to shake overnight at 37 °C and 220 rpm. The following morning, this library overnight culture was miniprepped using a QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany). The miniprepped plasmid DNA pool was used as the new template for mutagenesis with the primers for site-saturation mutagenesis of site 89X. T7 Express Competent *E. coli* were transformed with the Gibson products for the new five-site library using the recommended protocol. Upon heat-shock transformation, the freshly transformed *E. coli* cells were recovered in 0.4 mL LB medium at 37 ºC with shaking at 220 rpm for 30 minutes before being plated on LB-Amp agar plates. Single colonies from LB-agar plates were picked with sterilized toothpicks to individually inoculate 400 µL of LB-Amp in 2-mL 96-well deep-well plates across four separate plates. The plates were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 µL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 µL of 50% glycerol solution. These glycerol stocks were stored at -80°C for future inoculation. Additionally, the sequences of protoglobin genes contained in every well were sequenced using LevSeq sequencing.[7]

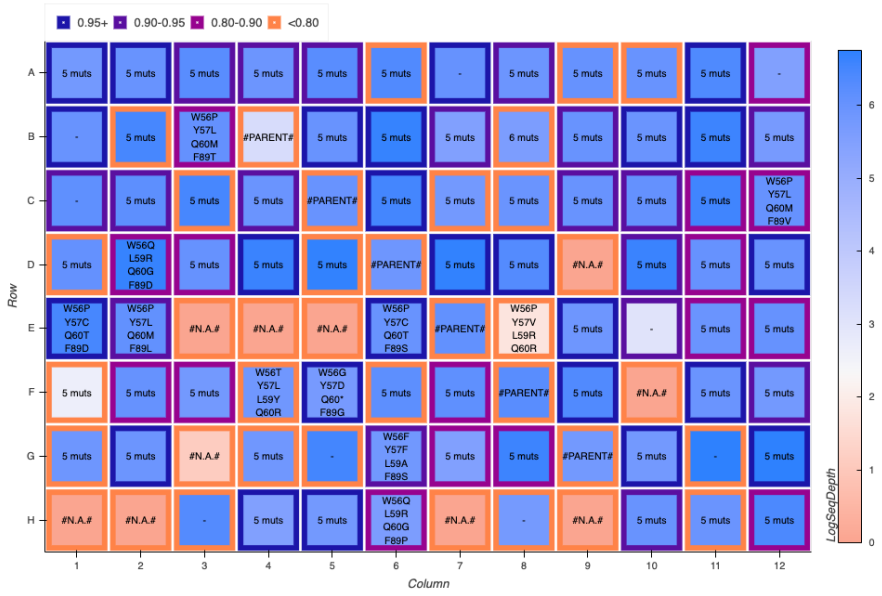


Figure C-7. LevSeq plate sequencing data for plate 1 of the randomly generated multi-mutant library.

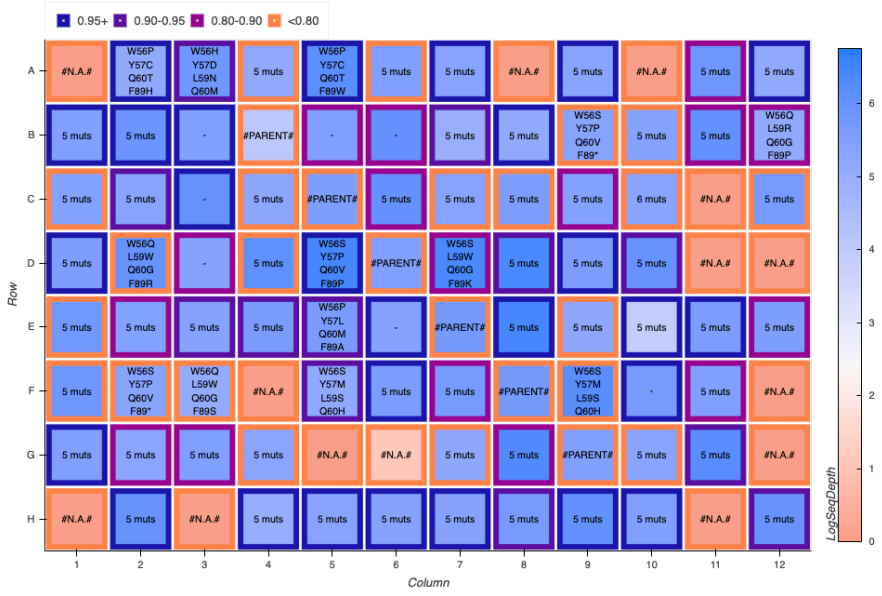


Figure C-8. LevSeq plate sequencing data for plate 2 of the randomly generated multi-mutant library.

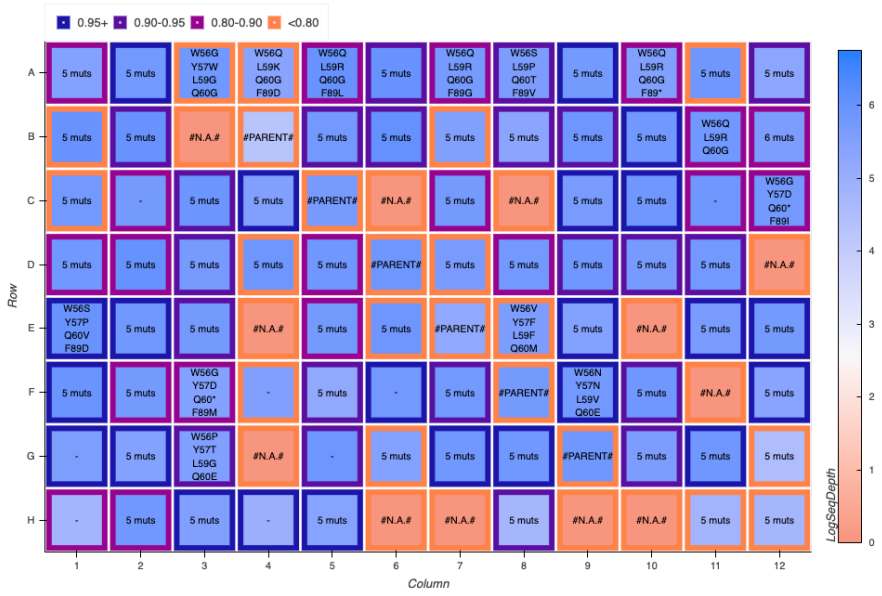


Figure C-9. LevSeq plate sequencing data for plate 3 of the randomly generated multi-mutant library.

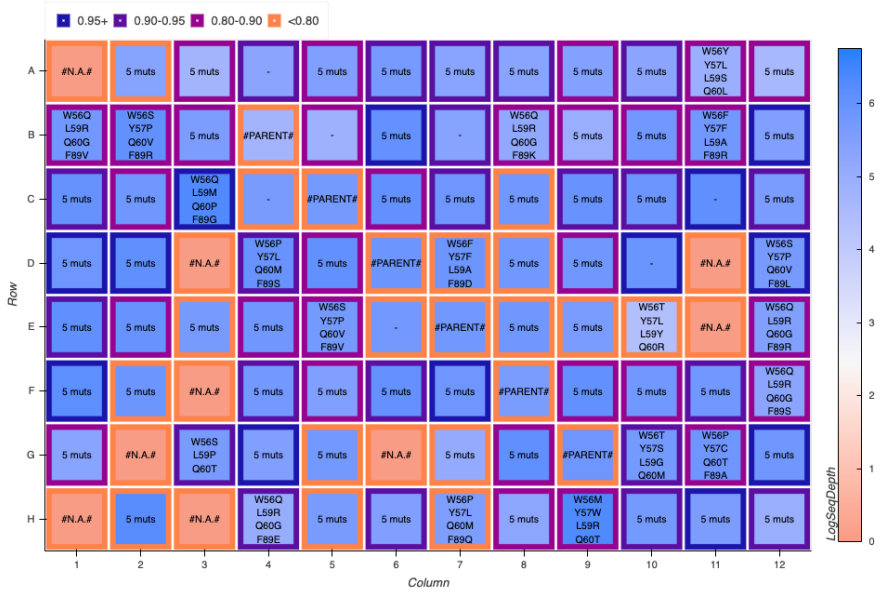


Figure C-10. LevSeq plate sequencing data for plate 4 of the randomly generated multi-mutant library.

C.3.2. Protocols for the Cloning of ALDE Predicted Sequences

C.3.2.1. 96-Well Plate Gibson Protocol

Exact genes encoding ParLQ mutants predicted by Active Learning-Assisted Directed Evolution (ALDE) were synthesized and delivered by Elegen Corp. (San Carlos, CA). DNA fragments were received as dry residues in 96-well PCR plates in 2–4-µg quantities. These DNA samples were dissolved in 100 µL of double-distilled H_2O (ddH_2O), yielding concentrations between 20–40 ng/µL. A 0.7-µL aliquot of these resuspended gene solutions were added to the wells of a 96-well PCR plate (Globe Scientific Inc., Mahwah, NJ). To each well of this plate was then added 1.0 µL of an aqueous solution containing 60 ng/µL of linearized pET-22b(+) backbone with overhangs designed for Gibson ligation with the DNA samples. Finally, to each well were added 5 µL of Gibson assembly mix.[5] The 96-well plate was then incubated at 50 °C for 60 minutes, after which the resulting Gibson products are placed on ice. These Gibson products could then either be directly used for transformation or stored at -20 °C for future use.

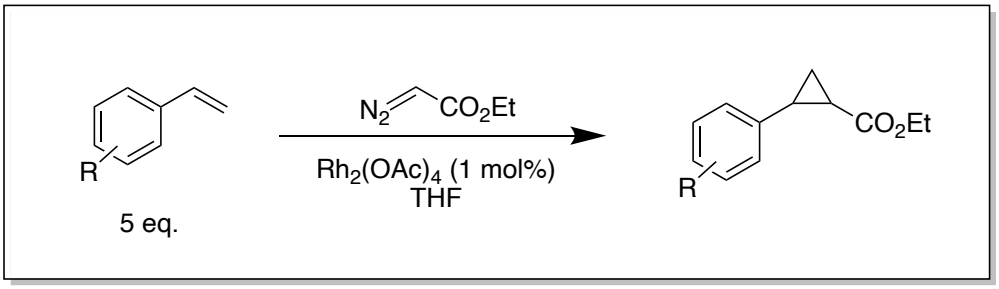
C.3.2.2. Transformation Protocol for 96-Well Plate Scale

To each well of the previously described Gibson assembly plate were added 5 µL of T7 Express Competent *E. coli*. The cell solutions were allowed to incubate on ice for 20 minutes, after which they were heat-shocked at 42 °C for 10 seconds in a water bath. The cells were then recovered with the addition of 100 µL of LB but without outgrowth at 37 °C. Immediately, 10 µL of each transformation mixture were used to inoculate the wells of a 2-mL 96-well deep-well plate in which the wells had been preloaded with 400 µL LB-Amp. This plate was incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning the plate was removed from the incubator and allowed to sit at room temperature for 8–10 hours. After this rest phase, 1 µL from each well was used to reinoculated yet another 96-well deep-well plate preloaded with 400 µL LB-Amp. This cell passage plate was incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 µL of preculture from each well were added to the wells of a 96-well flat-

bottom tissue culture plate (ThermoFisher) preloaded with 50 µL of 50% glycerol solution. These glycerol stocks were stored at -80°C for future use.

C.4. Preparation of Authentic Standards

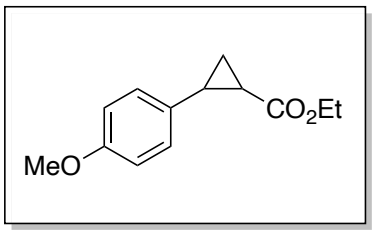
C.4.1. General Procedure for Cyclopropane Synthesis



The protocol for cyclopropane authentic standard synthesis was adapted from the procedure of Doyle and coworkers.[8] Rhodium acetate dimer (40 μmol, 18 mg) was added to a 40-mL dram vial equipped with a stir bar. The vial was sealed with a septum cap. A solution of olefin (20 mmol) was dissolved in 2 mL anhydrous THF and added to the sealed vial. Separately, a solution of ethyl diazoacetate (4 mmol, 0.4 mL) was dissolved in 1.2 mL of THF. The ethyl diazoacetate solution was added to the reaction over the course of two hours while the reaction was stirred at room temperature. The reaction was allowed to proceed overnight. The crude reaction mixture was filtered through a plug of activated alumina, and the filtrate was concentrated under reduced pressure. The resultant concentrate was loaded on a SNAP Ultra silica flash cartridge and separated using an Isolera flash purification system (Biotage, Charlotte, NC) with a hexane/ethyl acetate gradient from 0–10% ethyl acetate. Fractions containing the desired product were pooled and concentrated under reduced pressure. Only in certain cases could the *trans*- and *cis*-cyclopropane products be chromatographically separated.

C.4.2. Synthesis and Spectral Characterization of Cyclopropane Products

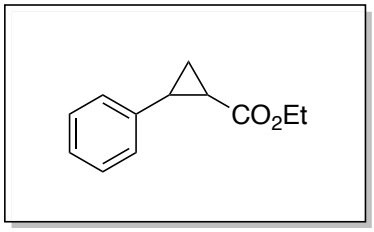
Ethyl 2-(4-methoxyphenyl)cyclopropane-1-carboxylate (2a)



Compound **2a** was synthesized with the general procedure for cyclopropane synthesis from 4-vinylanisole (**1a**) and EDA. The *trans*- and *cis*- products were separated by column chromatography, yielding 346 mg (39% yield) of the *trans*- diastereomer and 229 mg (26% yield) of the *cis*- diastereomer were isolated. Both *trans*- and *cis*-**2a** have been previously characterized in the literature.[9]

trans-***2a***: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.04 – 6.89 (m, 2H), 6.82 – 6.68 (m, 2H), 4.09 (q, *J* = 7.1 Hz, 2H), 3.71 (s, 3H), 2.41 (ddd, *J* = 9.2, 6.6, 4.2 Hz, 1H), 1.75 (ddd, *J* = 8.4, 5.2, 4.2 Hz, 1H), 1.53 – 1.44 (m, 2H), 1.25 – 1.14 (m, 4H).
cis-**2a**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.22 – 7.14 (m, 2H), 6.84 – 6.76 (m, 2H), 3.89 (q, *J* = 7.1 Hz, 2H), 3.77 (s, 3H), 2.52 (td, *J* = 8.8, 7.3 Hz, 1H), 2.08 – 1.98 (m, 1H), 1.65 (ddd, *J* = 7.5, 5.6, 5.0 Hz, 1H), 1.35 – 1.23 (m, 1H), 1.01 (t, *J* = 7.1 Hz, 3H).

Ethyl 2-phenylcyclopropane-1-carboxylate (2b)

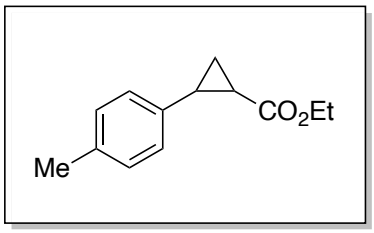


Compound **2b** was synthesized with the general procedure for cyclopropane synthesis from styrene (**1b**) and EDA. After separation by column chromatography, 124 mg (16% yield) of an 8:1 mixture of *trans*-/*cis*- isomers and 97 mg (13% yield) of pure *cis*-**2b** were isolated. Both *trans*- and *cis*-**2b** have been previously characterized in the literature.[10,11]

trans-**2b**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.34 – 7.21 (m, 2H), 7.24 – 7.16 (m, 1H), 7.14 – 7.06 (m, 2H), 4.17 (q, *J* = 7.1 Hz, 2H), 2.62 – 2.47 (m, 1H), 1.90 (ddd, *J* = 8.4, 5.3, 4.2 Hz, 1H), 1.60 (ddd, *J* = 9.2, 5.3, 4.5 Hz, 1H), 1.38 – 1.23 (m, 4H).

cis-**2b**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.29 (s, 2H), 7.33 – 7.17 (m, 2H), 3.90 (q, *J* = 7.1 Hz, 2H), 2.61 (td, *J* = 9.0, 7.5 Hz, 1H), 2.10 (ddd, *J* = 9.4, 7.8, 5.6 Hz, 1H), 1.74 (dt, *J* = 7.5, 5.4 Hz, 1H), 1.40 – 1.24 (m, 2H), 0.99 (t, *J* = 7.1 Hz, 3H).

Ethyl 2-(p-tolyl)cyclopropane-1-carboxylate (2c)

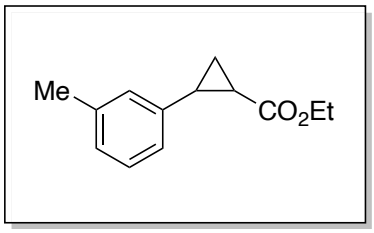


Compound **2c** was synthesized with the general procedure for cyclopropane synthesis from 1-methyl-4-vinylbenzene (**1c**) and EDA. After separation by column chromatography, 254 mg (31% yield) of a 1.92:1 mixture of *trans*-/*cis*- isomers (determined by NMR integration ratios) of **2c** was isolated. Both *trans*- and *cis*-**2c** have been previously characterized in the literature.[12] Reported *trans*- and *cis*- NMR peak assignments are assigned from the same NMR spectrum.

trans-**2c**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.12 – 6.95 (m, 5H), 4.23 – 4.12 (m, 2H), 2.49 (ddd, *J* = 10.0, 6.3, 4.0 Hz, 1H), 2.31 (s, 3H), 1.86 (dddd, *J* = 8.3, 5.1, 4.2, 0.8 Hz, 1H), 1.63 – 1.51 (m, 2H), 1.38 – 1.23 (m, 5H).

cis-**2c**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.19 – 7.09 (m, 4H), 3.89 (q, *J* = 7.1 Hz, 2H), 2.54 (q, *J* = 7.9, 7.3 Hz, 1H), 2.30 (s, 3H), 2.05 (ddd, *J* = 9.4, 7.8, 5.6 Hz, 1H), 1.68 (dt, *J* = 7.2, 5.0 Hz, 1H), 1.41 – 1.31 (m, 1H), 1.06 – 0.96 (m, 3H).

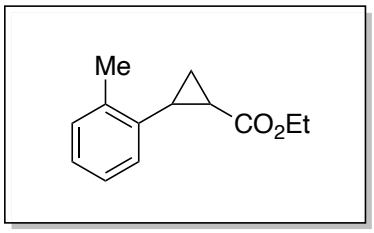
Ethyl 2-(m-tolyl)cyclopropane-1-carboxylate (2d)



Compound **2d** was synthesized with the general procedure for cyclopropane synthesis from 1-methyl-3-vinylbenzene (**1d**) and EDA. After separation by column chromatography, 242 mg (30% yield) of a 1.87:1 mixture of *trans*-/*cis*- isomers (determined by NMR integration ratios) of **2d** was isolated. Both *trans*- and *cis*-**2d** have been previously characterized in the literature.[12] Reported *trans*- and *cis*- NMR peak assignments are assigned from the same NMR spectrum.

trans-**2d:** ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.04 – 6.98 (m, 2H), 6.94 – 6.87 (m, 2H), 4.17 (q, J = 7.1 Hz, 2H), 2.48 (ddd, J = 9.2, 6.5, 4.2 Hz, 1H), 2.33 – 2.32 (m, 3H), 1.89 (ddd, J = 8.4, 5.3, 4.2 Hz, 1H), 1.58 (ddd, J = 9.2, 5.3, 4.5 Hz, 2H), 1.36 – 1.19 (m, 5H).
cis-**2d**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.19 – 7.04 (m, 5H), 3.89 (q, J = 7.1 Hz, 2H), 2.55 (q, J = 8.6 Hz, 1H), 2.31 (d, J = 0.7 Hz, 3H), 2.06 (ddd, J = 9.3, 7.8, 5.6 Hz, 1H), 1.69 (ddd, J = 7.5, 5.6, 5.0 Hz, 1H), 0.99 (t, J = 7.1 Hz, 3H).

Ethyl 2-(o-tolyl)cyclopropane-1-carboxylate (2e)

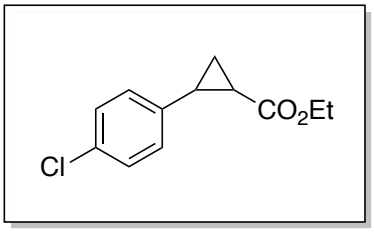


Compound **2e** was synthesized with the general procedure for cyclopropane synthesis from 1-methyl-2-vinylbenzene (**1e**) and EDA. The *trans*- and *cis*- products were separated by column chromatography, yielding 118 mg (14% yield) of the *trans*- diastereomer and 81

mg (10% yield) of the *cis*- diastereomer were isolated. Both *trans*- and *cis*-**2e** have been previously characterized in the literature.[12]

trans-**2e**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.24 – 7.06 (m, 3H), 7.03 – 6.95 (m, 1H), 4.30 – 4.12 (m, 2H), 2.56 – 2.42 (m, 1H), 2.39 (s, 3H), 1.79 (dtd, *J* = 8.3, 4.8, 0.8 Hz, 1H), 1.57 (ddd, *J* = 9.3, 5.1, 4.3 Hz, 1H), 1.30 (td, *J* = 7.1, 0.8 Hz, 4H).
cis-**2e**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.25 – 7.16 (m, 1H), 7.21 – 7.07 (m, 3H), 3.85 (q, *J* = 7.1 Hz, 2H), 2.45 (q, *J* = 8.4 Hz, 1H), 2.16 (ddd, *J* = 9.1, 7.9, 5.4 Hz, 1H), 1.75 (dt, *J* = 7.6, 5.2 Hz, 1H), 1.61 – 1.55 (m, 1H), 1.40 – 1.26 (m, 1H), 0.93 (t, *J* = 7.1 Hz, 3H).

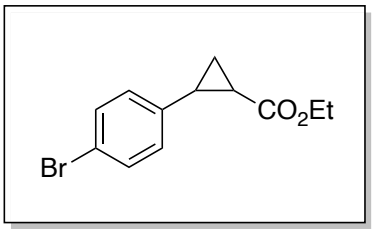
Ethyl 2-(4-chlorophenyl)cyclopropane-1-carboxylate (2f)



Compound **2f** was synthesized with the general procedure for cyclopropane synthesis from 1-chloro-4-vinylbenzene (**1f**) and EDA. After separation by column chromatography, 226 mg (25% yield) of a 2:1 mixture of *trans*-/*cis*- isomers (determined by NMR integration ratios) of **2f** was isolated. Both *trans*- and *cis*-**2f** have been previously characterized in the literature.[13,14]

trans-**2f**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.23 – 7.18 (m, 2H), 7.09 – 7.00 (m, 2H), 4.18 (q, *J* = 7.0 Hz, 2H), 2.52 – 2.44 (m, 1H), 1.87 (ddd, *J* = 8.5, 5.3, 4.1 Hz, 1H), 1.61 (ddd, *J* = 9.2, 5.4, 4.6 Hz, 2H), 1.29 (t, *J* = 7.1 Hz, 3H).
cis-**2f**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.27 – 7.22 (m, 4H), 3.90 (q, *J* = 7.1 Hz, 2H), 2.57 – 2.50 (m, 1H), 2.08 (ddd, *J* = 9.2, 7.8, 5.6 Hz, 1H), 1.67 (dt, *J* = 7.5, 5.4 Hz, 1H), 1.39 – 1.31 (m, 1H), 1.02 (t, *J* = 7.1 Hz, 3H).

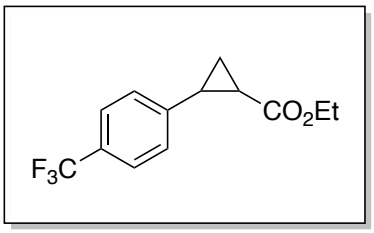
Ethyl 2-(4-bromophenyl)cyclopropane-1-carboxylate (2g)



Compound **2g** was synthesized with the general procedure for cyclopropane synthesis from 1-bromo-4-vinylbenzene (**1g**) and EDA. The *trans*- and *cis*- products were separated by column chromatography, yielding 334 mg (31% yield) of the *trans*- diastereomer and 297 mg (28% yield) of the *cis*- diastereomer were isolated. Both *trans*- and *cis*-**2g** have been previously characterized in the literature.[9,15]

trans-**2g**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.43 – 7.35 (m, 2H), 7.00 – 6.93 (m, 2H), 4.17 (q, *J* = 7.1 Hz, 2H), 2.47 (ddd, *J* = 9.4, 6.5, 4.1 Hz, 1H), 1.91 – 1.82 (m, 1H), 1.60 (dt, *J* = 9.5, 5.0 Hz, 1H), 1.32 – 1.22 (m, 4H).
cis-**2g**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.47 – 7.34 (m, 2H), 7.14 (d, *J* = 8.2 Hz, 2H), 3.90 (q, *J* = 7.1 Hz, 2H), 2.50 (q, *J* = 8.5 Hz, 1H), 2.08 (ddd, *J* = 8.9, 7.9, 5.6 Hz, 1H), 1.67 (dt, *J* = 7.5, 5.4 Hz, 1H), 1.57 (d, *J* = 1.8 Hz, 1H), 1.35 (dt, *J* = 8.4, 4.2 Hz, 1H), 1.34 – 1.22 (m, 0H), 1.02 (t, *J* = 7.1 Hz, 3H).

Ethyl 2-(4-(trifluoromethyl)phenyl)cyclopropane-1-carboxylate (2h)

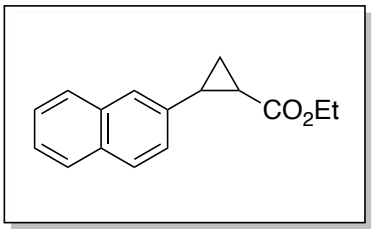


Compound **2h** was synthesized with the general procedure for cyclopropane synthesis from 1-(trifluoromethyl)-4-vinylbenzene (**1h**) and EDA, with the exception that the reaction was on a 1-mmol basis for EDA with all corresponding quantities being adjusted accordingly. After separation by column chromatography, 89 mg (30% yield) of a 3:1 mixture of *trans*-

/*cis*- isomers (determined by NMR integration ratios) of **2h** was isolated. Both *trans*- and *cis*-**2h** have been previously characterized in the literature.[12]

trans-**2h**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.57 – 7.49 (m, 2H), 7.22 – 7.18 (m, 1H), 4.18 (q, *J* = 7.2 Hz, 2H), 2.58 – 2.51 (m, 1H), 1.94 (ddd, *J* = 8.5, 5.4, 4.2 Hz, 1H), 1.66 (ddd, *J* = 9.2, 5.4, 4.7 Hz, 1H), 1.36 – 1.31 (m, 1H), 1.29 (t, *J* = 7.1 Hz, 3H).
cis-**2h**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.44 (d, *J* = 4.1 Hz, 4H), 7.31 (dq, *J* = 7.5, 0.8 Hz, 2H), 3.82 (qd, *J* = 7.1, 2.1 Hz, 2H), 2.52 (d, *J* = 7.9 Hz, 1H), 2.07 (ddd, *J* = 9.3, 7.9, 5.7 Hz, 1H), 1.67 (dt, *J* = 7.5, 5.4 Hz, 1H), 1.33 (ddd, *J* = 8.7, 7.9, 5.2 Hz, 1H), 0.92 (t, *J* = 7.1 Hz, 3H).

Ethyl 2-(naphthalen-2-yl)cyclopropane-1-carboxylate (2i)



Compound **2i** was synthesized with the general procedure for cyclopropane synthesis from 2-vinylnapthylene (**1i**) and EDA. After separation by column chromatography, 467 mg (48% yield) of a 1.82:1 mixture of *trans*-/*cis*- isomers (determined by NMR integration ratios) of **2h** was isolated. Both *trans*- and *cis*-**2h** have been previously characterized in the literature.[12]

trans-**2i**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.85 – 7.56 (m, 7H), 4.23 – 4.16 (m, 2H), 2.71 – 2.65 (m, 1H), 2.01 (ddd, *J* = 8.4, 5.3, 4.2 Hz, 1H), 1.68 (ddd, *J* = 9.2, 5.3, 4.6 Hz, 1H), 1.43 (ddd, *J* = 8.4, 6.4, 4.6 Hz, 2H), 1.30 (t, *J* = 7.1 Hz, 3H).
cis-**2i**: ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.05 – 7.89 (m, 3H), 7.69 – 7.59 (m, 2H), 7.21 (dd, *J* = 8.5, 1.8 Hz, 2H), 3.83 (q, *J* = 7.1 Hz, 2H), 2.73 (dd, *J* = 8.0, 1.7 Hz, 1H), 2.16 (ddd, *J* = 9.3, 7.8, 5.6 Hz, 1H), 1.86 (dt, *J* = 7.5, 5.3 Hz, 1H), 1.47 – 1.44 (m, 1H).

C.5. Preparation of Calibration Curves for Analytical Yield Determination

All reactions were assayed by gas chromatography equipped with flame ionization detection (GC-FID). To quantify cyclopropane yields, calibration curves were made with the synthesized authentic standards. Calibration curves were constructed by one of two methods.

Calibration Curve Construction – Method A

Method A was used in cases where the two diastereomers of the cyclopropane could be separated by column chromatography (compounds **2a**, **2b**, **2e**, and **2g**). For these compounds, a separate 10 mM stock of cyclopropane was made for each diastereomer in a solution of 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). This stock solution was diluted in the same of ethyl acetate:cyclohexane solution containing 1.0 mM standard to final concentrations of 0.5–10 mM cyclopropane. Samples were then analyzed by GC-FID.

Calibration Curve Construction – Method B

Method A was used in cases where the two diastereomers of the cyclopropane were isolated as a mixture after column chromatography (compounds **2c**, **2d**, **2f**, **2h**, and **2i**). For these compounds, a single 20 mM stock of cyclopropane, containing both diastereomers, was prepared in a 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). This stock solution was diluted in the same of ethyl acetate:cyclohexane solution containing 1.0 mM standard to final concentrations of 0.5–20 mM cyclopropane. Concentrations for the separate diastereomers were computed according to the molar ratios determined by NMR, and it was assumed that the presence of both diastereomers in each sample would not affect analytical signal. Samples were then analyzed by GC-FID.

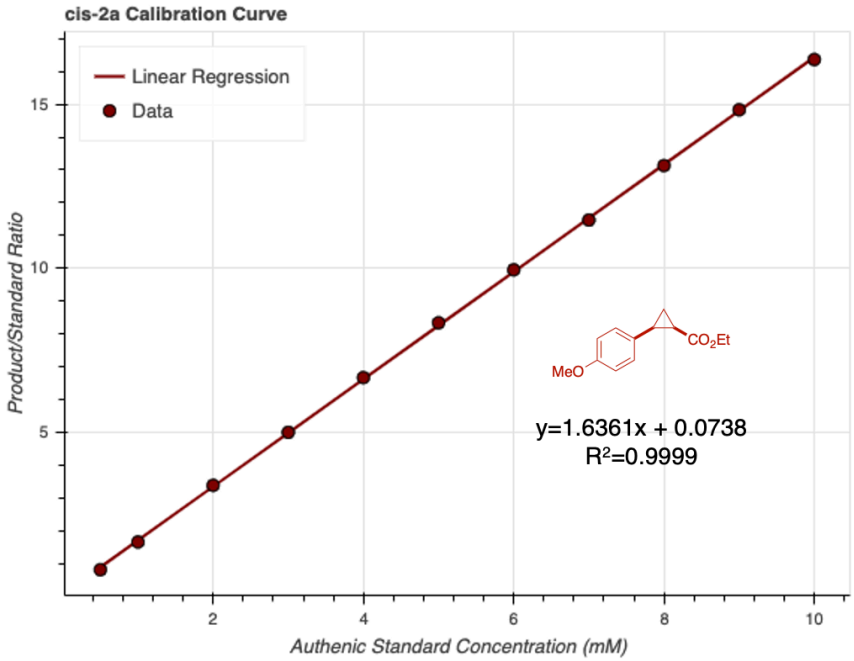


Figure C-11. Achiral GC-FID calibration curve for *cis*-**2a**. Samples were generated using method A.

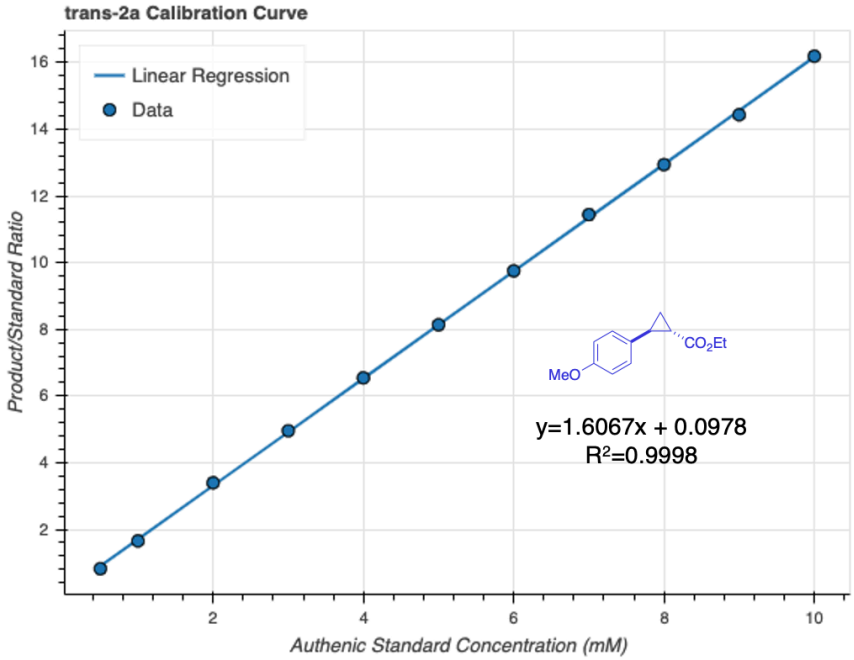


Figure C-12. Achiral GC-FID calibration curve for *trans*-**2a**. Samples were generated using method A.

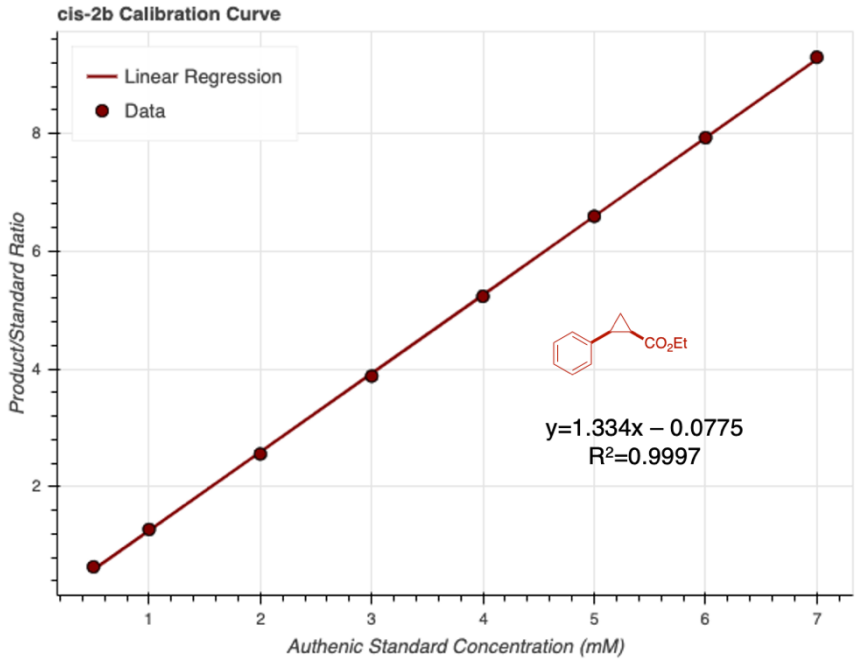


Figure C-13. Achiral GC-FID calibration curve for *cis*-**2b**. Samples were generated using method A.

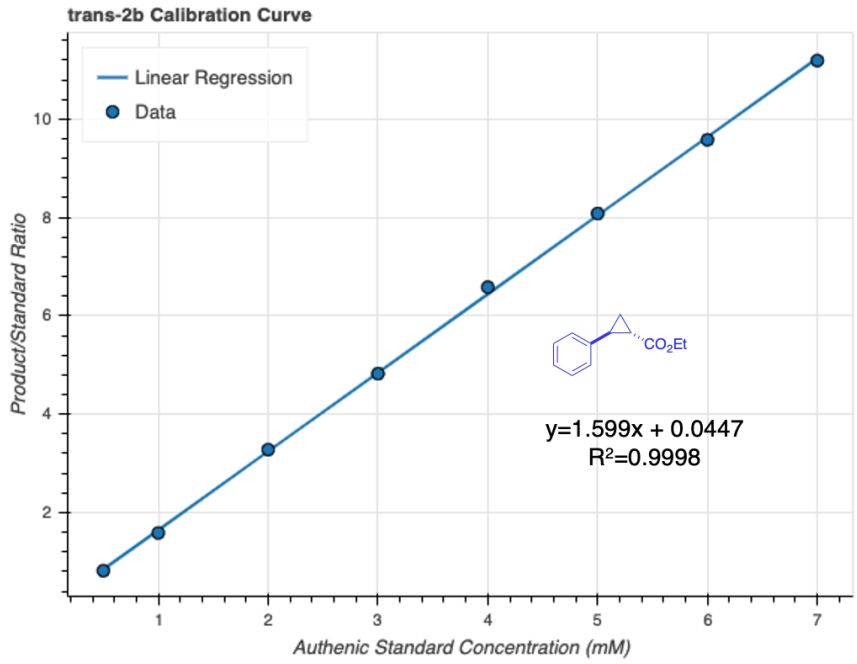


Figure C-14. Achiral GC-FID calibration curve for *trans*-**2b**. Samples were generated using method A.

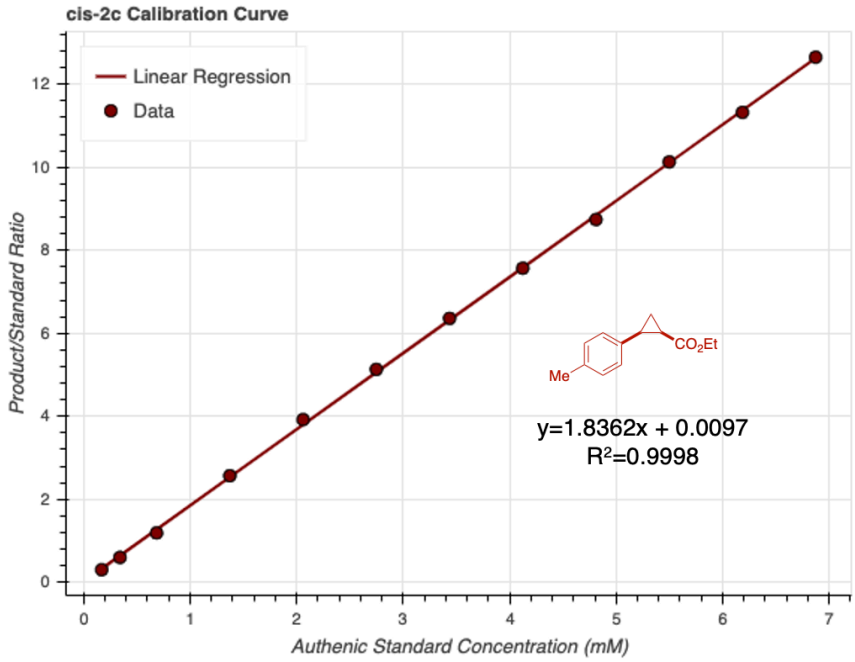


Figure C-15. Achiral GC-FID calibration curve for *cis*-**2c**. Samples were generated using method B.

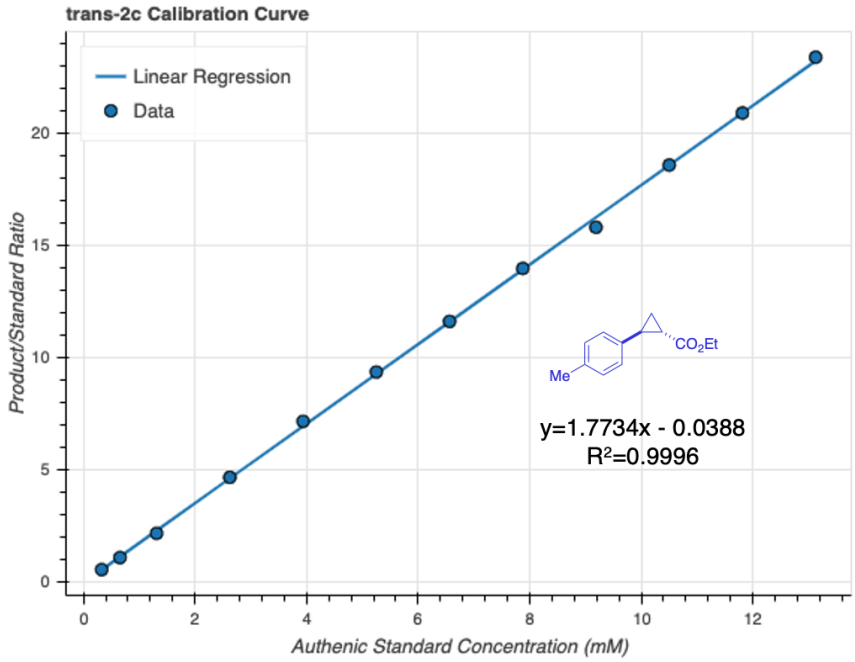


Figure C-16. Achiral GC-FID calibration curve for *trans*-**2c**. Samples were generated using method B.

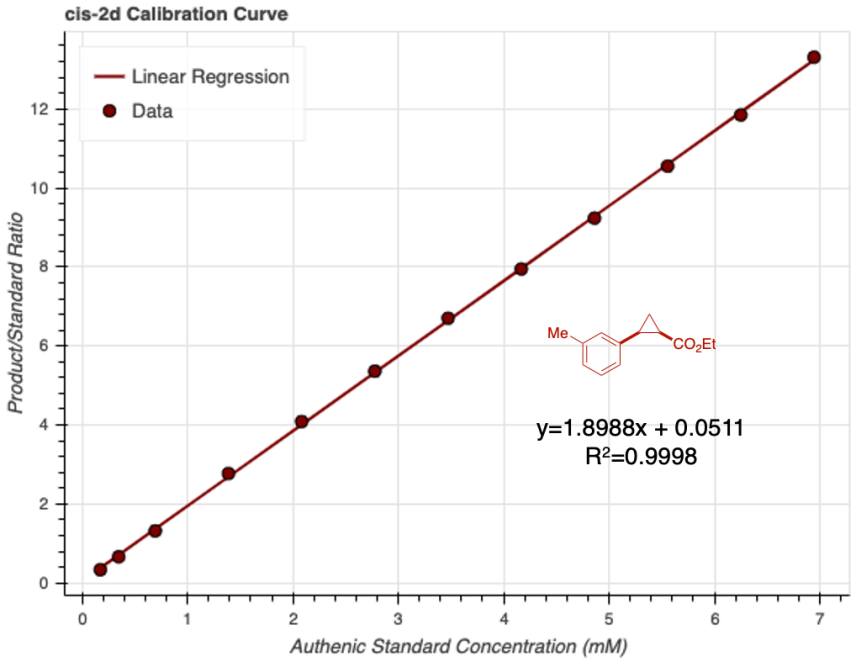


Figure C-17. Achiral GC-FID calibration curve for *cis*-**2d**. Samples were generated using method B.

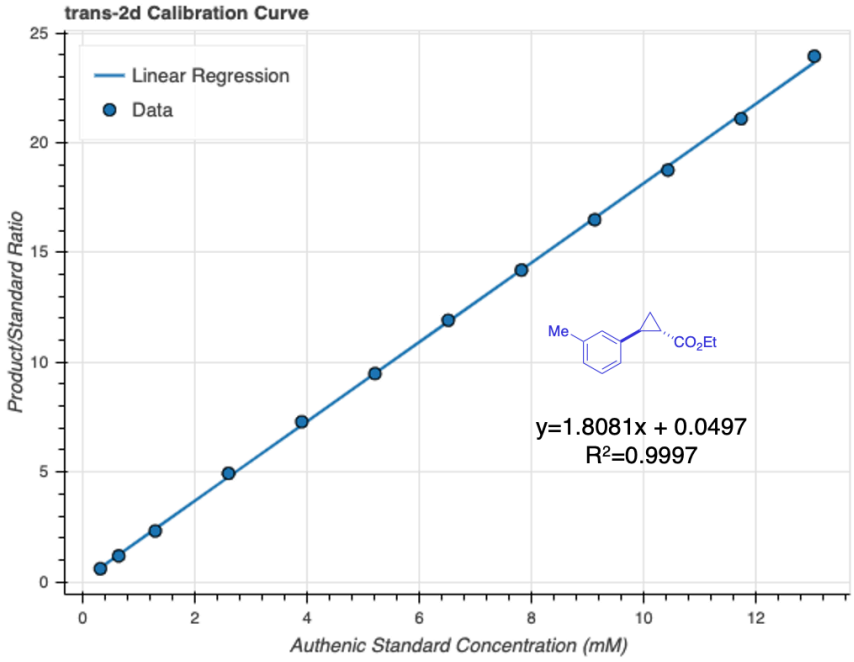


Figure C-18. Achiral GC-FID calibration curve for *trans*-**2d**. Samples were generated using method B.

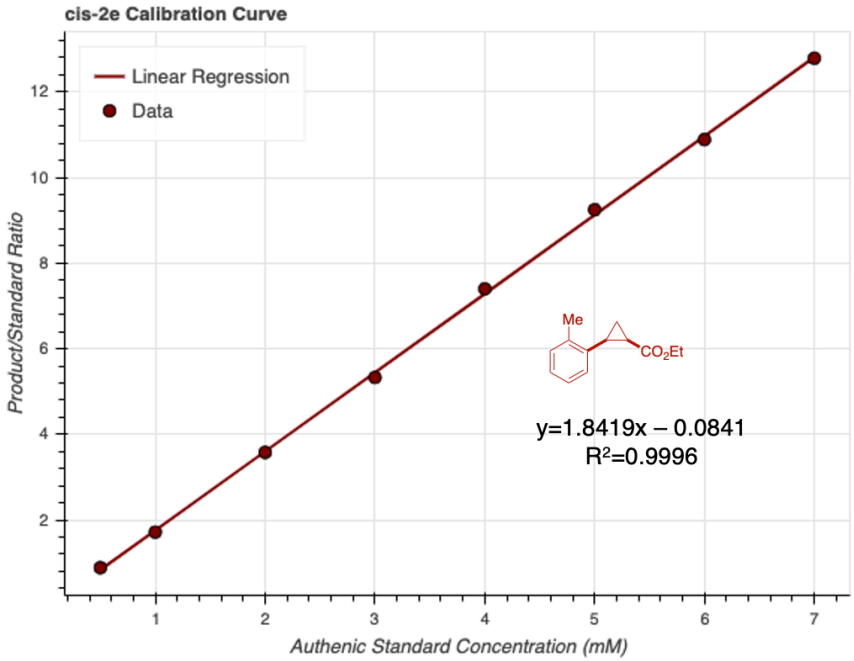


Figure C-19. Achiral GC-FID calibration curve for *cis*-**2e**. Samples were generated using method A.

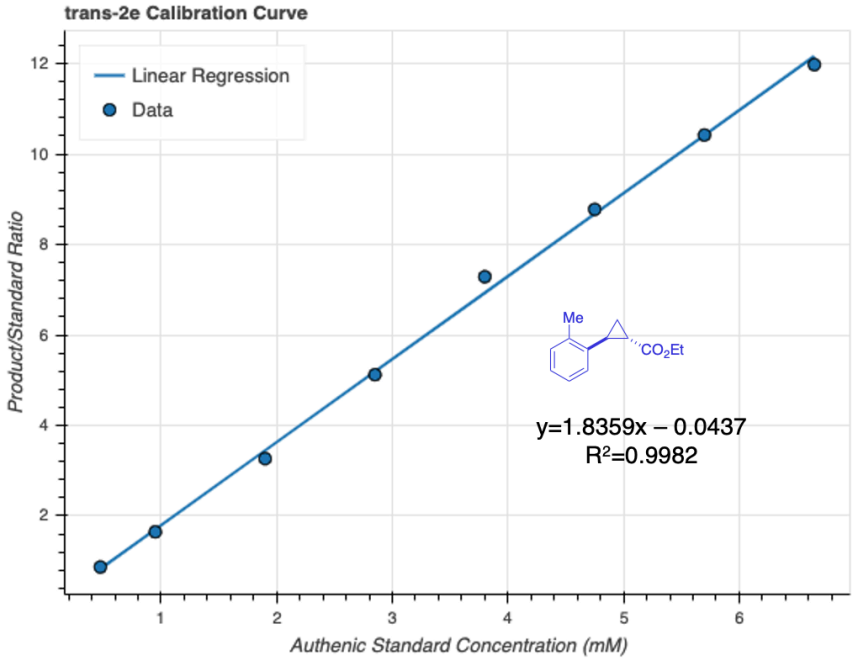


Figure C-20. Achiral GC-FID calibration curve for *trans*-**2e**. Samples were generated using method A.

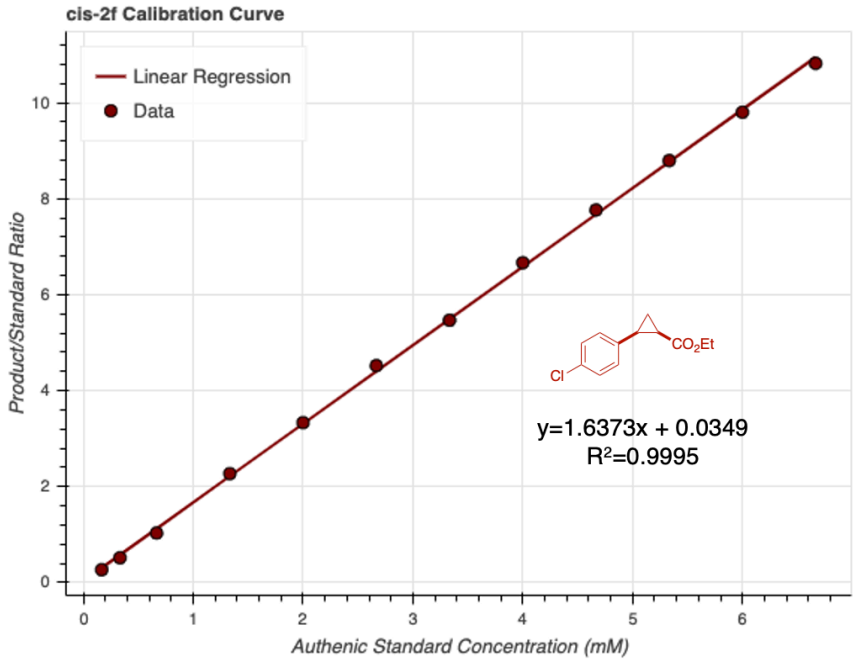


Figure C-21. Achiral GC-FID calibration curve for *cis*-**2f**. Samples were generated using method B.

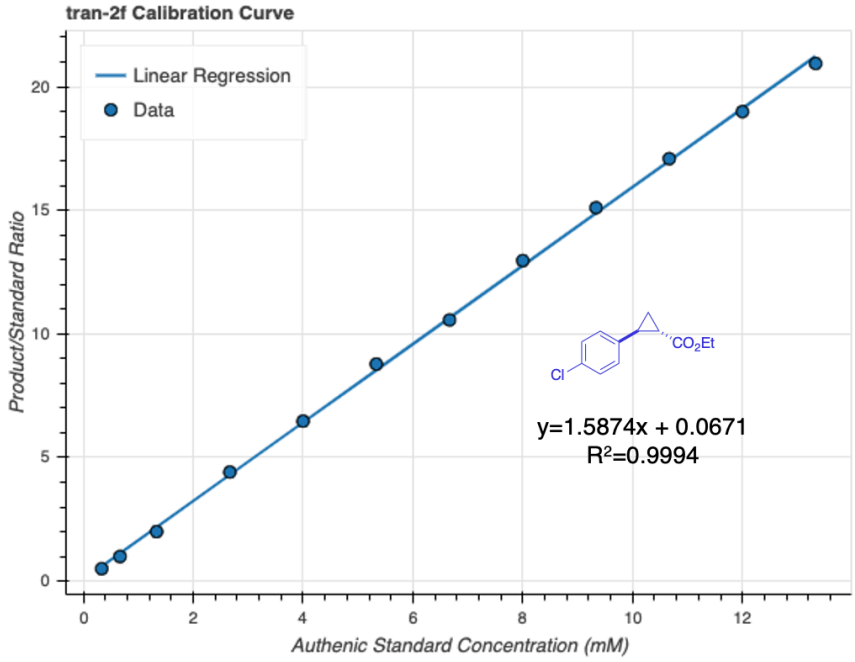


Figure C-22. Achiral GC-FID calibration curve for *trans*-**2f**. Samples were generated using method B.

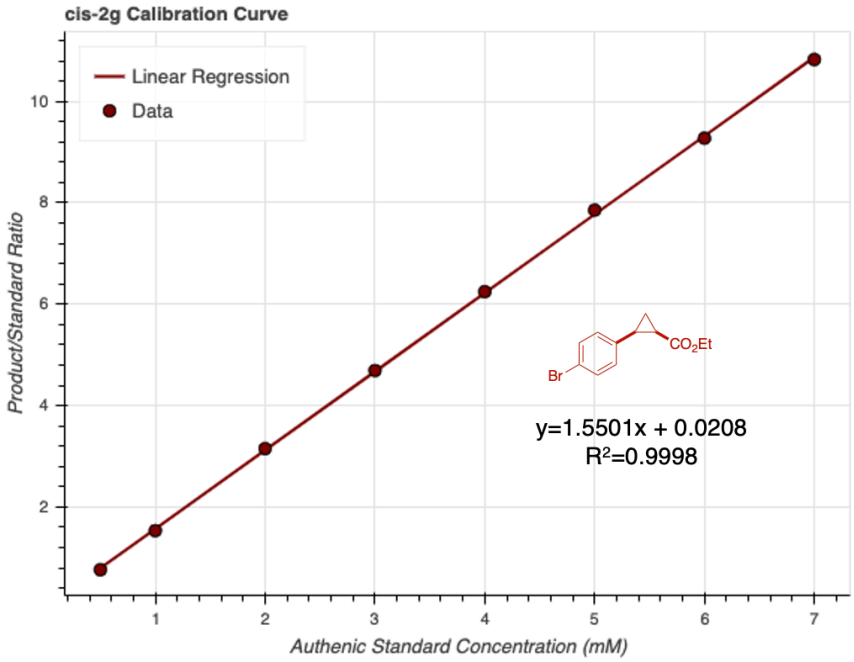


Figure C-23. Achiral GC-FID calibration curve for *cis*-**2g**. Samples were generated using method A.

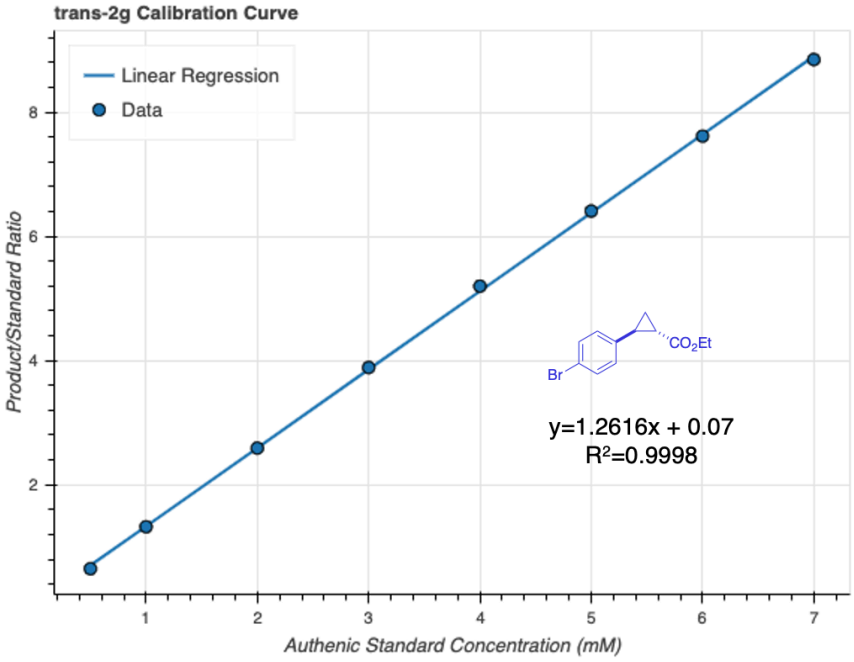


Figure C-24. Achiral GC-FID calibration curve for *trans*-**2g**. Samples were generated using method A.

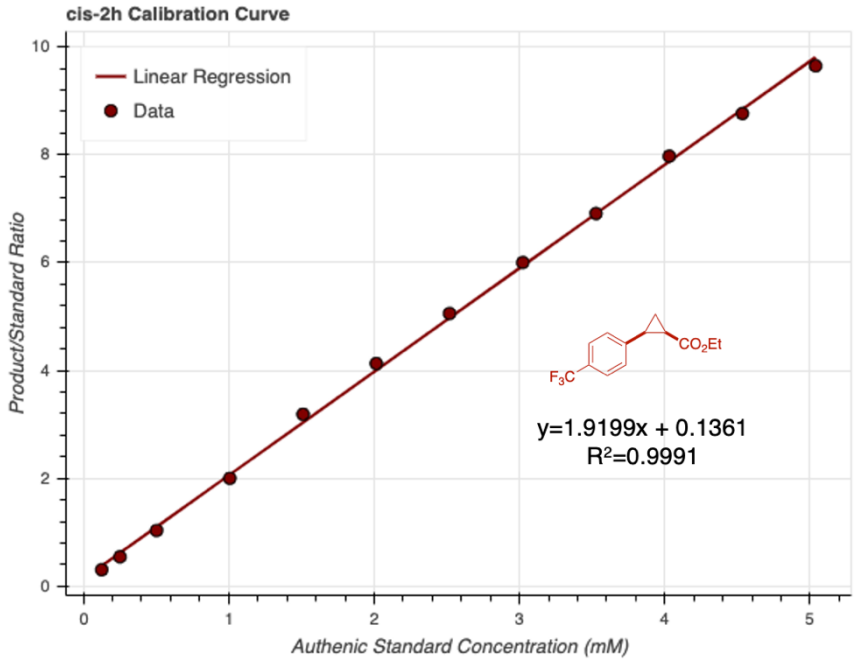


Figure C-25. Achiral GC-FID calibration curve for *cis*-**2h**. Samples were generated using method B.

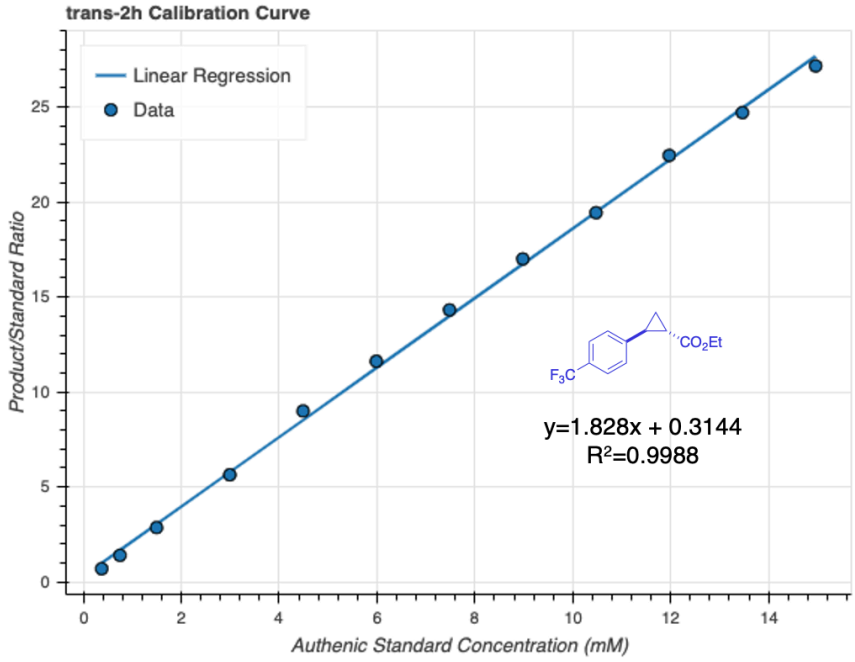


Figure C-26. Achiral GC-FID calibration curve for *trans*-**2h**. Samples were generated using method B.

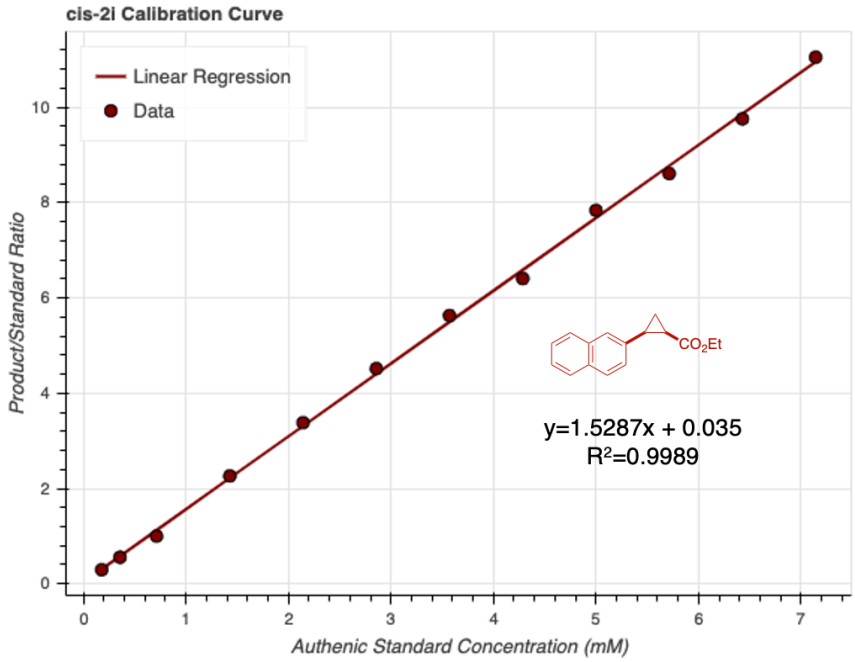


Figure C-27. Achiral GC-FID calibration curve for *cis*-**2i**. Samples were generated using method B.

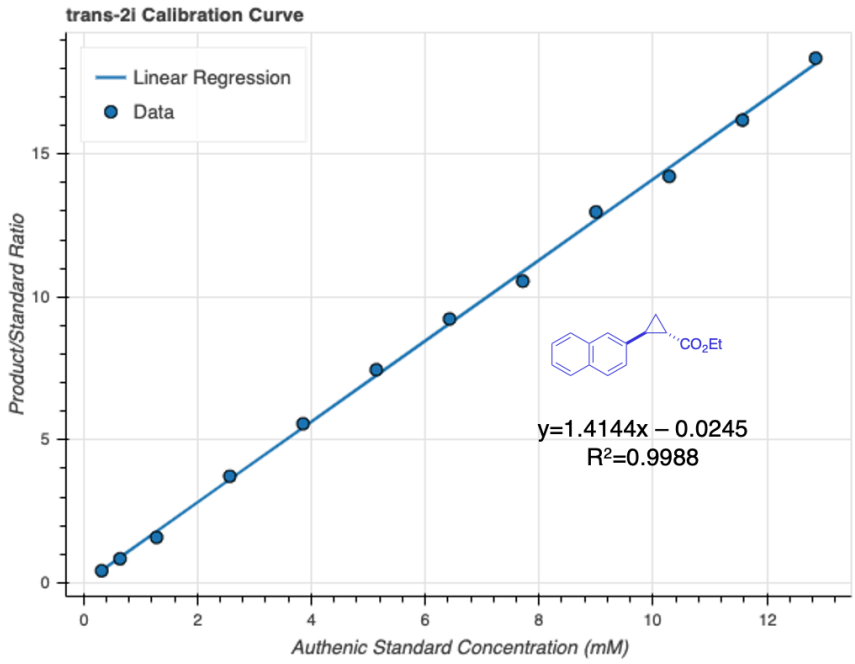


Figure C-28. Achiral GC-FID calibration curve for *trans*-**2i**. Samples were generated using method B.

C.6. Control and Validation Experiments

C.6.1. Protocols for Small-Scale Reaction Setup in GC Vials

C.6.1.1. Small-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 6 mL of LB-Amp in a 15-mL culture tube. Cultures were incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 shaking incubator. A 1-mL aliquot of each of these overnight cultures was used to inoculate 50 mL of Terrific Broth with 100 µg/mL of ampicillin (TB-Amp) (0.5% v/v starter culture in expression culture) in 125-mL unbaffled Erlenmeyer flasks. The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaking incubator, at which point they are moved to room temperature for 25 minutes. Protein expression was then induced by direct addition of 50 µL of stock solutions containing 500 mM isopropyl-β-D-thiogalactoside (IPTG) and 1.0 M 5-aminolevulinic acid (ALA) such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16–18 hours in an Innova 42 shaker.

C.6.1.2. Small-Scale Biocatalytic or Control Reactions

The corresponding 50-mL expression cultures were pelleted (4,000 × g for 15 minutes at 4 ºC) and resuspended in 5 mL of M9-N buffer. The optical density at 600 nm (OD_{600}) of this suspension was measured and adjusted to OD_{600} = 31.5 with the addition of more M9-N buffer. A 380 µL aliquot of the cell suspension was added to 2-mL GC vials (Agilent). For control reactions, compounds and additives were added to GC vials according to the conditions described in **Figure S29**. These whole-cell suspensions were transferred into a vinyl Coy anaerobic chamber, at which point 10 µL of a 400 mM solution of styrene in MeCN followed by 10 µL of a 600 mM solution of EDA in MeCN were added such that the final reaction concentrations were 10 mM of styrene and 15 mM of EDA. The GC vials were tightly capped with screwcaps with a septum and were allowed to shake at room

temperature for 16 hours. Once complete, the reactions were transferred to a 1.7-mL Eppendorf tube and 600 μL of a 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). The layers are vigorously mixed, and the samples were centrifuged (14,000 × g for 10 minutes at RT). Afterwards, an aliquot of the organic layer was subjected to GC analysis.

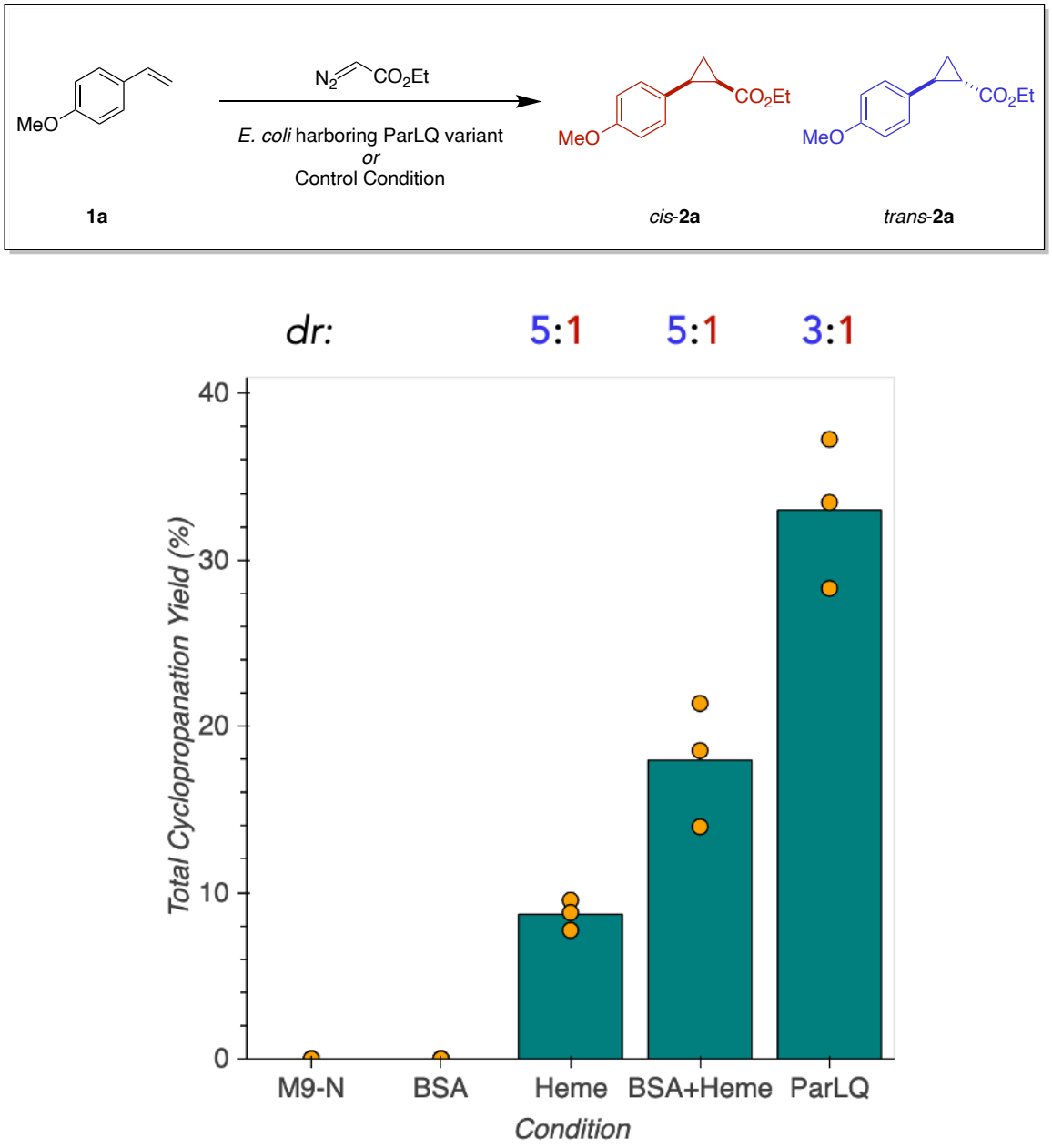


Figure C-29. Cyclopropanation yields and selectivities for control conditions relative to ParLQ. All reactions are run in 380 µL M9-N (pH=7.6) and 20 µL MeCN. BSA was loaded at 1 mg/mL and heme was loaded at 1 mM. Diastereoselectivities are shown above the plot as the trans:cis cyclopropane ratio.

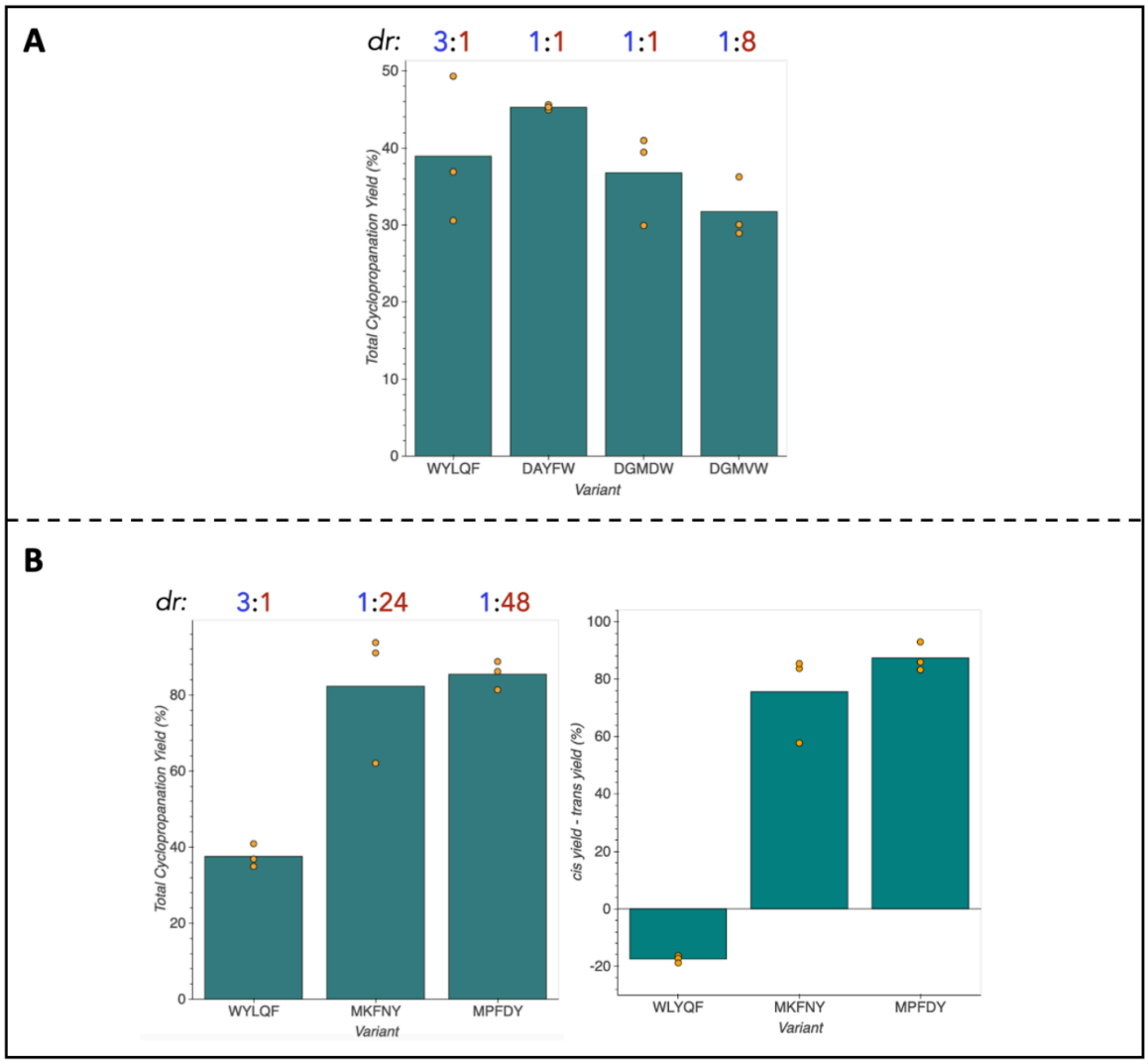
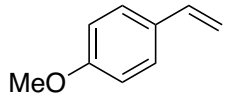
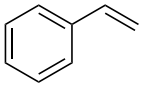
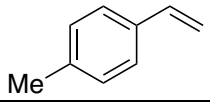
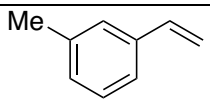
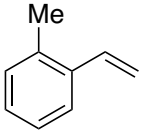
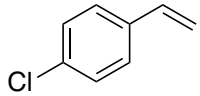
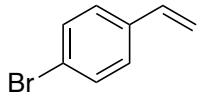
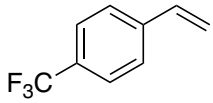
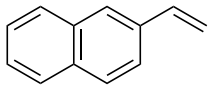


Figure C-30. (A) Cyclopropanation yields and selectivities for select 5-site multi-mutants of ParLQ. These variants are the result of the recombination of the single-site mutations which showed the most improvement for various possible acquisition functions. DAYFW recombines the top cyclopropane yielding mutations, DGMDW recombines the mutants which were the most selective for *cis*- formation, and DGMVW recombines the variants which had the highest objective function. Diastereoselectivities are shown above the plot as the trans:cis cyclopropane ratio. **(B)** Objective functions for the top performing variants found in each round of ALDE predictions in comparison to the parent ParLQ (WLYQF). Reactions were performed according to the method "Small-Scale Biocatalytic or Control Reactions."

C.7. Biocatalytic Cyclopropanation of Olefins

C.7.1. Substrates Utilized in Plate Screening

Table C-7. Table of styrenyl substrates investigated in this study.

Substrate	Name	Structure
1a	4-vinylanisole	
1b	Styrene	
1c	1-methyl-4-vinylbenzene	
1d	1-methyl-3-vinylbenzene	
1e	1-methyl-2-vinylbenzene	
1f	1-chloro-4-vinylbenzene	
1g	1-bromo-4-vinylbenzene	
1h	1-(trifluoromethyl)-4-vinylbenzene	
1i	2-vinylnaphthylene	

C.7.2. Protocols for the Screening of Protoglobin Variants in 96-well Plate Format

C.7.2.1. 96-Well Plate Library Expression

The wells of a 2-mL 96-well deep-well plate were filled with 400 µL LB-Amp. Previously generated 96-well plate glycerol stocks were removed from -80°C storage and placed on

dry ice. Multichannel pipet tips were used to scratch the frozen glycerol stock surface and used to inoculate the aforementioned deep-well plate. These overnight cultures were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. For expression cultures, the following morning 50 µL of the precultures were used to inoculate 900 µL TB-amp per well in 96-well deep-well plates. The expression cultures were initially incubated at 37 ºC and 220 rpm for 2.5 hours, at which point they were allowed to sit at room temperature for 30 minutes. Expression of proteins was induced with IPTG and cellular heme production was increased with ALA. An induction mixture containing IPTG and ALA in TB-amp (50 µL) was added to each well such that the final concentrations of IPTG and ALA were 0.5 mM and 1.0 mM respectively. The total culture volumes were 1 mL per well. The plates were then incubated at 22 ºC and 220 rpm overnight.

C.7.2.2. 96-Well Plate Library Reactions and Screening

Expression cultures containing *E. coli* expressing hemoproteins of interest were centrifuged at 4,000 × g for 10 minutes at 4 ºC. The supernatant was discarded, and nitrogen-free M9 minimal media (M9-N, 380 µL) was added to each well. The pellets were resuspended in this media via shaking at room temperature for 30 minutes. The plates were then pumped into a vinyl Coy anaerobic chamber (0–30 ppm O_2). To each well were added 20 µL of a MeCN solution with 200 mM of the desired styrene substrate and 300 mM of ethyl diazoacetate (EDA). The final reaction volume was 400 µL, and the final concentrations of the desired styrene and EDA were 10 mM and 15 mM, respectively. The plates were then sealed carefully with a foil cover and shaken at room temperature for 16 hours in the Coy chamber. Once complete, plates are worked up for processing by adding 600 µL of a 1:1 solution of ethyl acetate:cyclohexane containing 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). A silicone sealing mat (AWSM1003S, ArcticWhite) was used to cover each of the plates, and the two layers were thoroughly mixed by rapid inversion of the plates. The plates were then centrifuged (5,000 × g for 5 minutes at room temperature) to separate the phases. Afterwards, a 200-µL aliquot of the organic layer was transferred to a GC vial insert in a GC vial, and the samples were assayed by GC-FID.

C.7.2.3. Protoglobin Compilation Screening Data

Initial reaction screening was performed on a compilation of protoglobin sequences. This compilation plate was primarily composed of mutants of the protoglobin from *Aeropyrum pernix*. Reactions were prepared according to the protocols for the screening of protoglobin variants in 96-well plate format (page S41). Cyclopropanation yields for formation of *cis*-**2a** and *trans*-**2a** are presented.

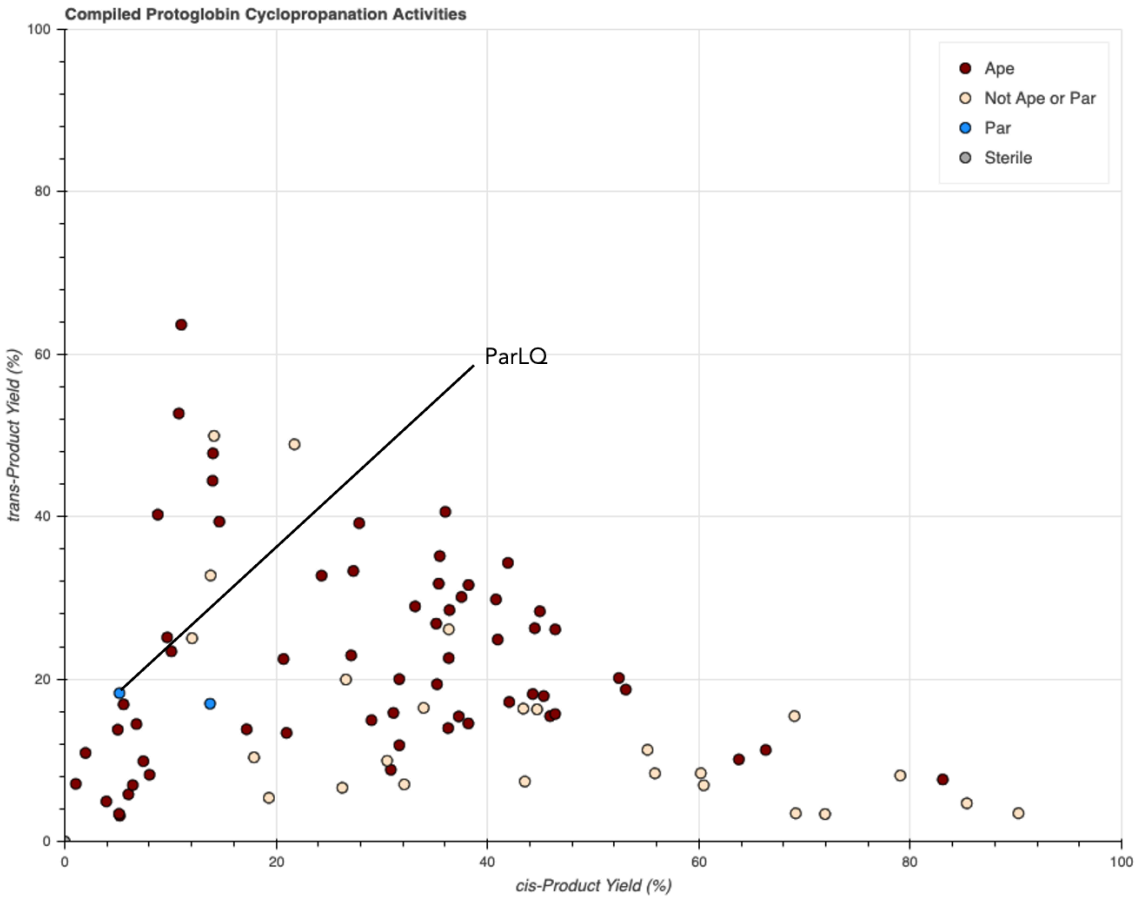


Figure C-31. Yields of various protoglobin homologs and mutants for the formation of *cis*- and *trans*-**2a**.

C.7.2.4. ParLQ Single-Site Mutant Screening Data

Reactions of single-site mutants were prepared according to the protocols for the screening of protoglobin variants in 96-well plate format (page S41). Total cyclopropanation activity for formation of **2a**, *cis*-**2a**:*trans*-**2a** diastereoselectivity, and resulting objective function data are presented for reactions in triplicate.

Activity of W56X Mutants

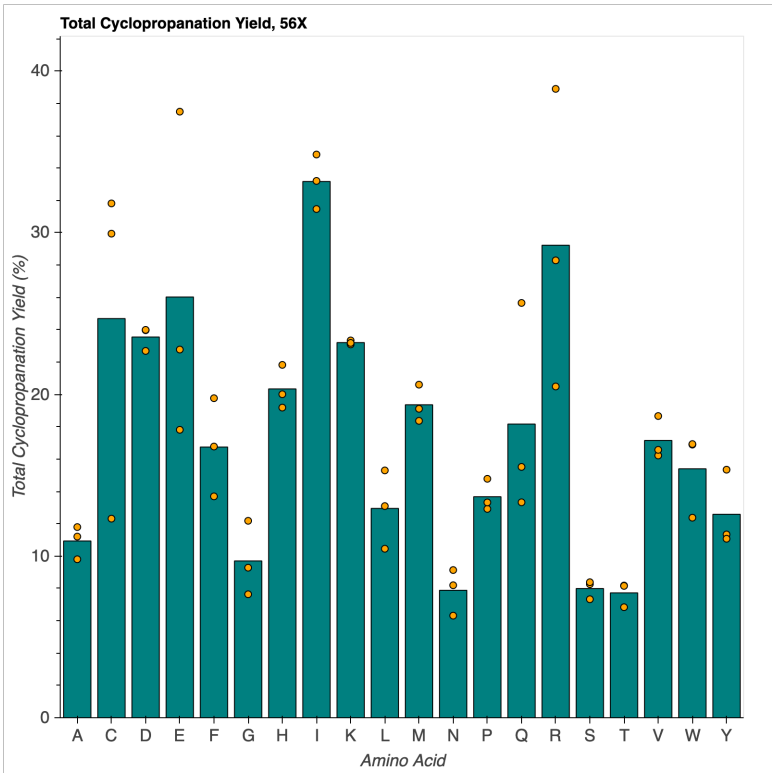


Figure C-32. Total cyclopropanation activity of single-site mutants at position 56X. The original amino acid at this position was W. Data for variants W56L and W56N were collected separately and the yields were normalized according to parent activities across both experiments. Each variant was tested in biological triplicate.

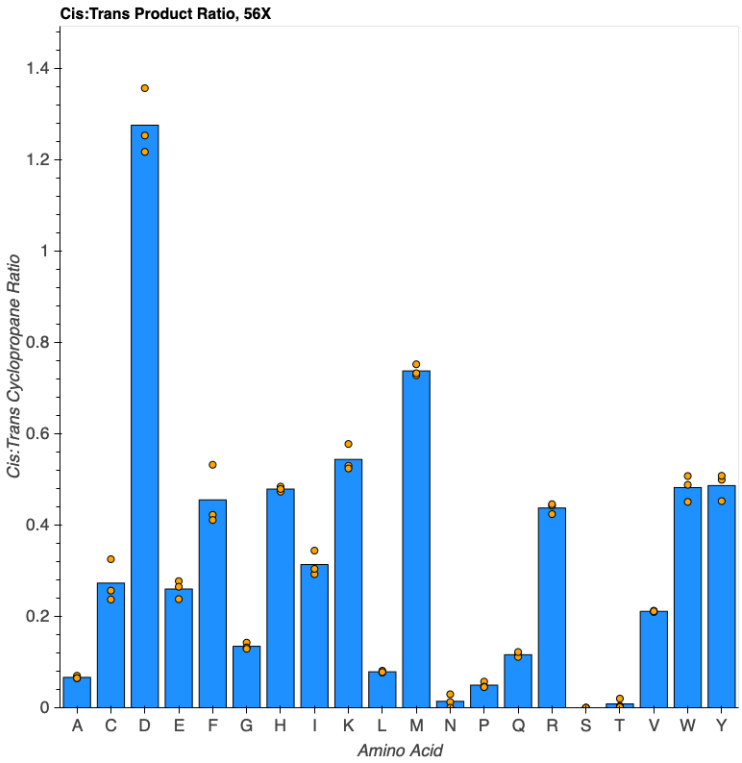


Figure C-33. Diastereoselectivity of single-site mutants at position 56X. The original amino acid at this position was W. Each variant was tested in biological triplicate.

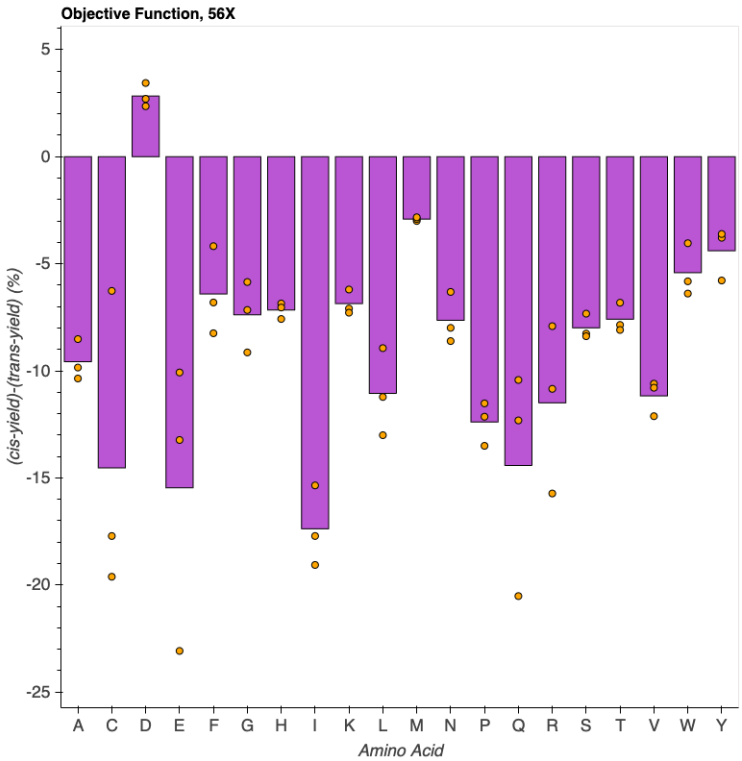


Figure C-34. Objective function observed for single-site mutants at position 56X. The original amino acid at this position was W. Each variant was tested in biological triplicate.

Activity of Y57X Mutants

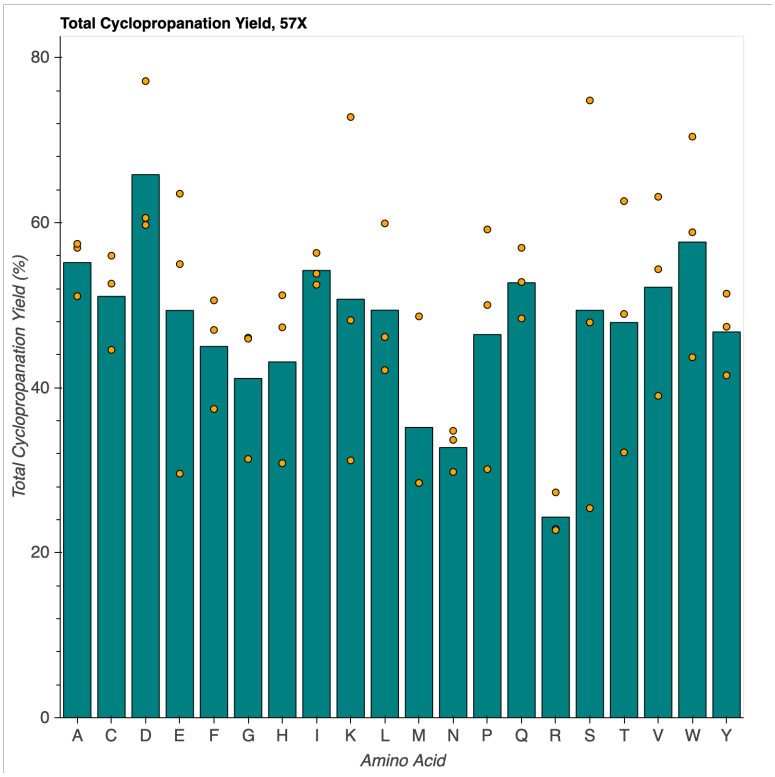


Figure C-35. Total cyclopropanation activity of single-site mutants at position 57X. The original amino acid at this position was Y. Data for variant Y57D were collected separately and the yields were normalized according to parent activities across both experiments. Each variant was tested in biological triplicate.

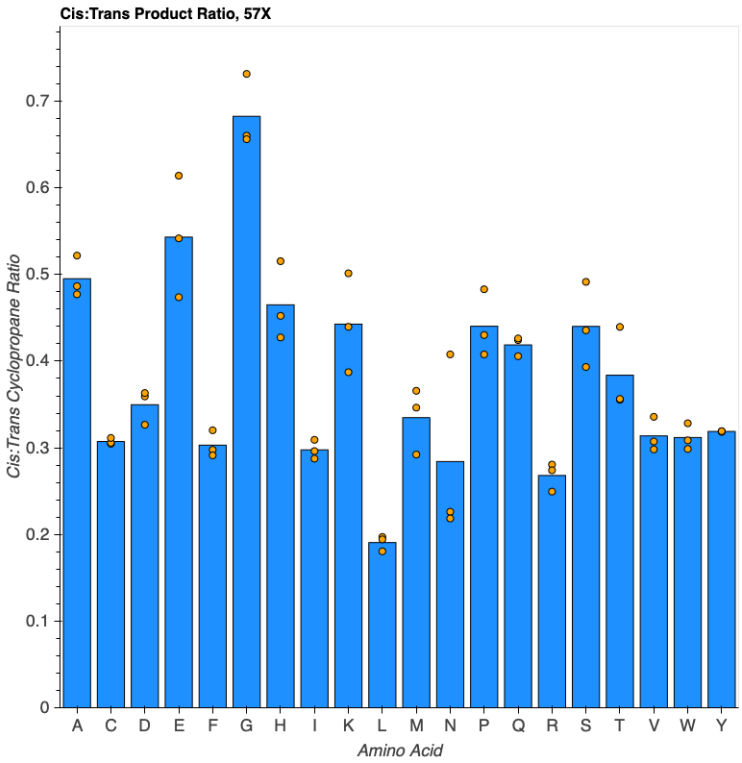


Figure C-36. Diastereoselectivity of single-site mutants at position 57X. The original amino acid at this position was Y. Each variant was tested in biological triplicate.

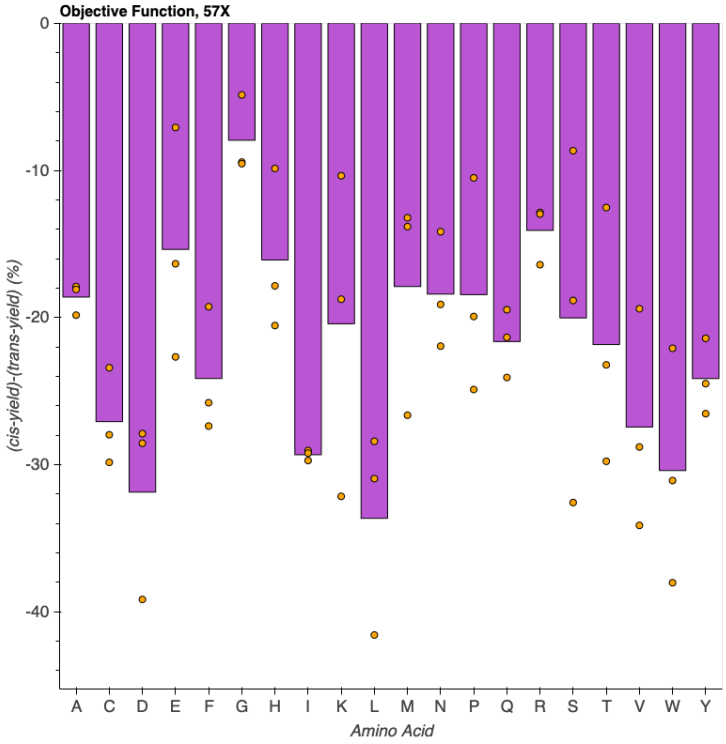


Figure C-37. Objective function observed for single-site mutants at position 57X. The original amino acid at this position was Y. Each variant was tested in biological triplicate.

Activity of L59X Mutants

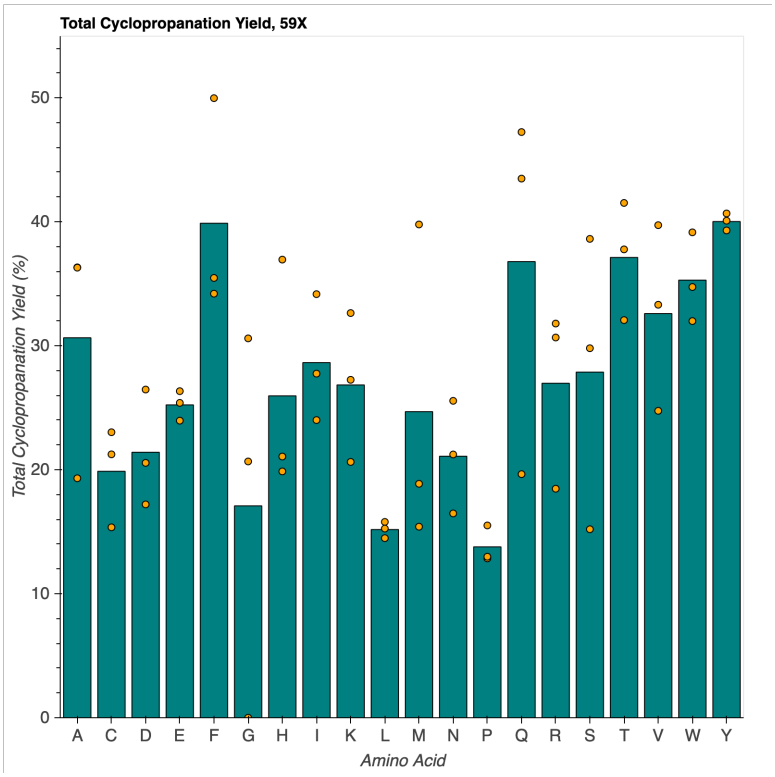


Figure C-38. Total cyclopropanation activity of single-site mutants at position 59X. The original amino acid at this position was L. Each variant was tested in biological triplicate.

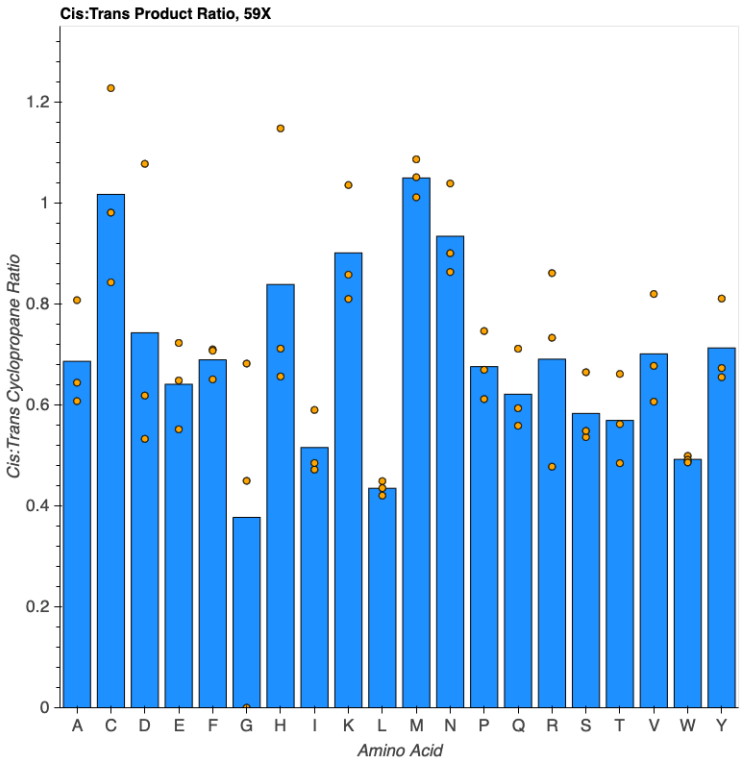


Figure C-39. Diastereoselectivity of single-site mutants at position 59X. The original amino acid at this position was L. Each variant was tested in biological triplicate.

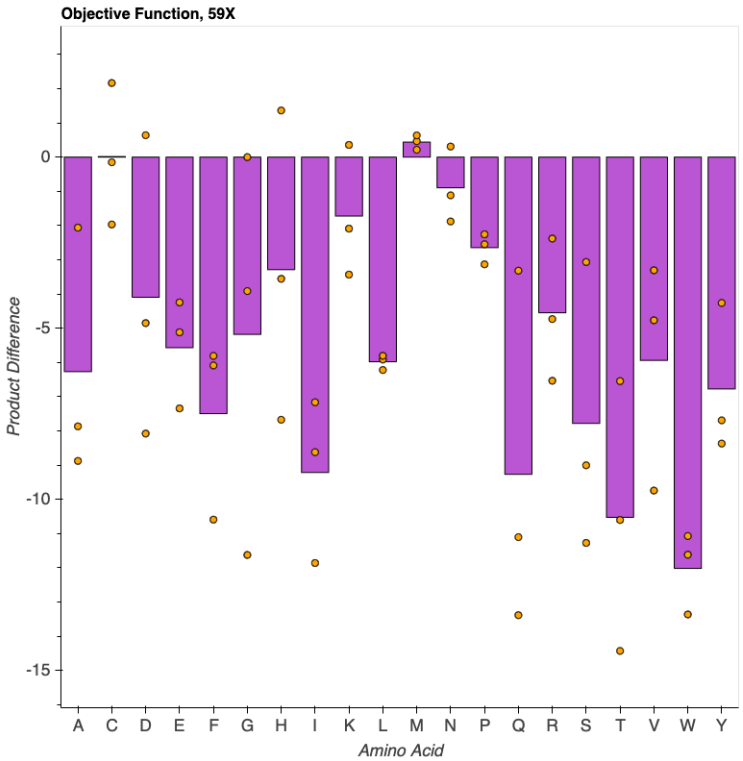


Figure C-40. Objective function observed for single-site mutants at position 59X. The original amino acid at this position was L. Each variant was tested in biological triplicate.

Activity of Q60X Mutants

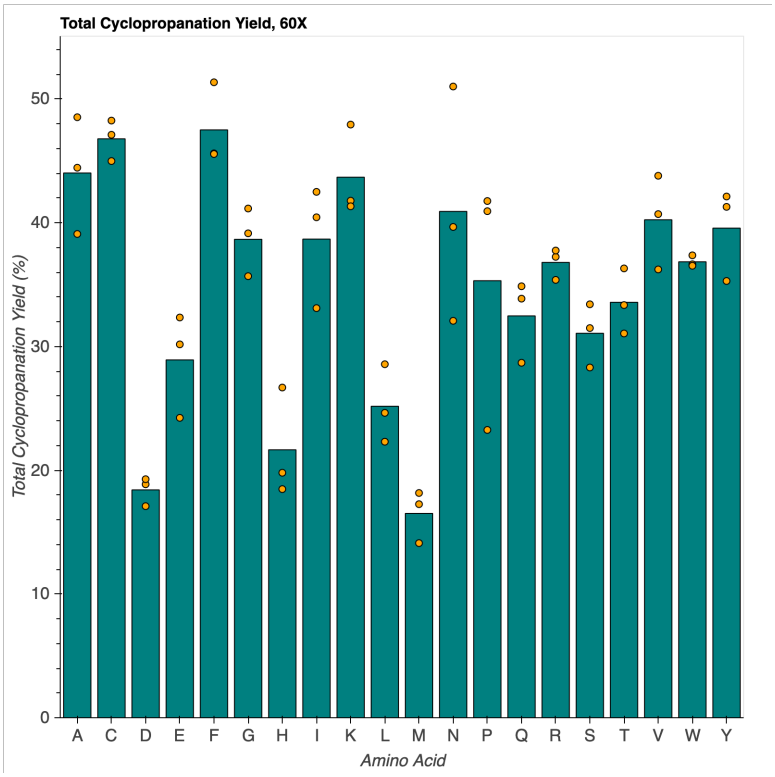


Figure C-41. Total cyclopropanation activity of single-site mutants at position 60X. The original amino acid at this position was Q. Data for variants Q60L, Q60M, and Q60N were collected separately and the yields were normalized according to parent activities across both experiments. Each variant was tested in biological triplicate.

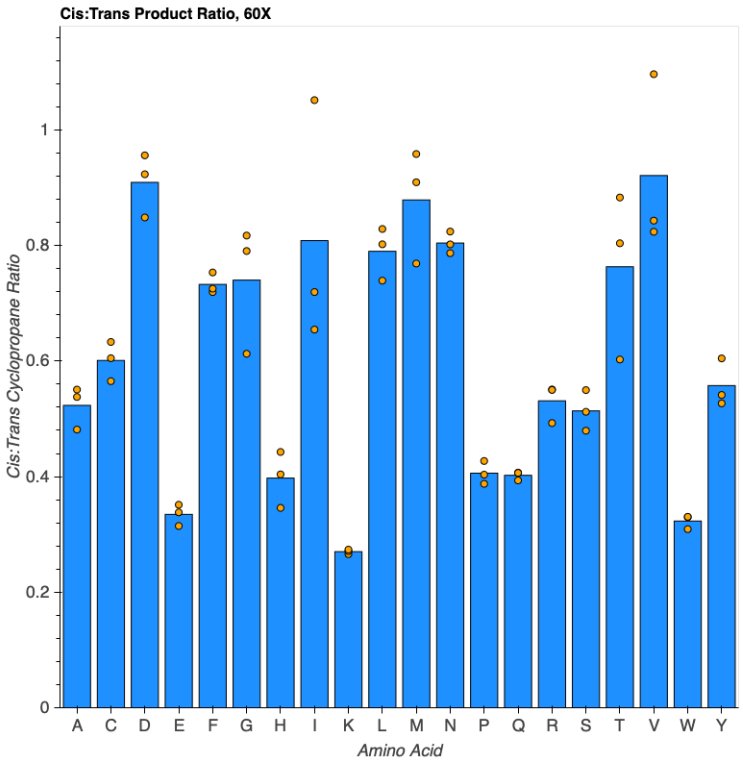


Figure C-42. Diastereoselectivity of single-site mutants at position 60X. The original amino acid at this position was Q. Each variant was tested in biological triplicate.

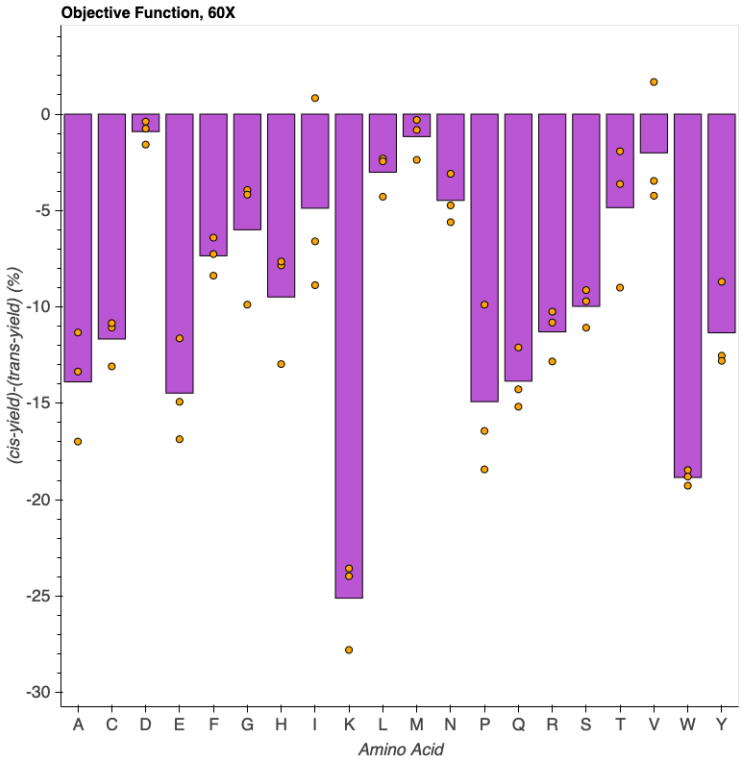


Figure C-43. Objective function observed for single-site mutants at position 60X. The original amino acid at this position was Q. Each variant was tested in biological triplicate.

Activity of F89X Mutants

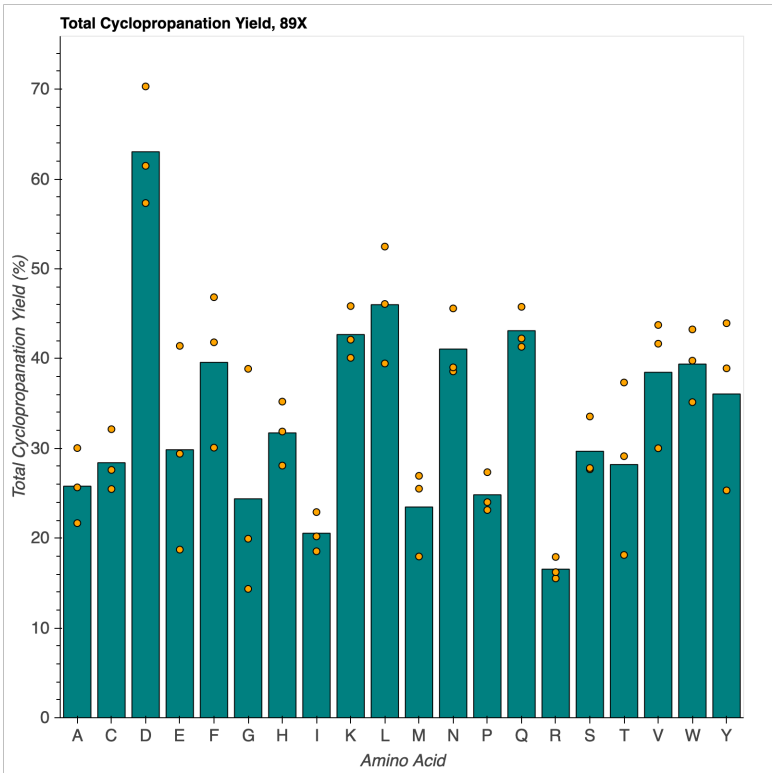


Figure C-44. Total cyclopropanation activity of single-site mutants at position 89X. The original amino acid at this position was Q. Data for variant F89I were collected separately and the yields were normalized according to parent activities across both experiments. Each variant was tested in biological triplicate.

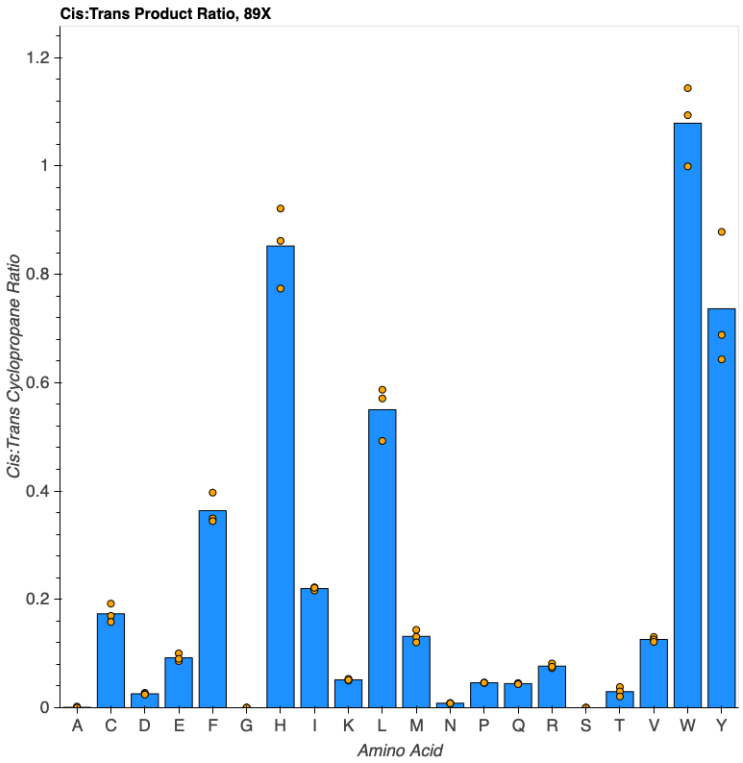


Figure C-45. Diastereoselectivity of single-site mutants at position 89X. The original amino acid at this position was Q. Each variant was tested in biological triplicate.

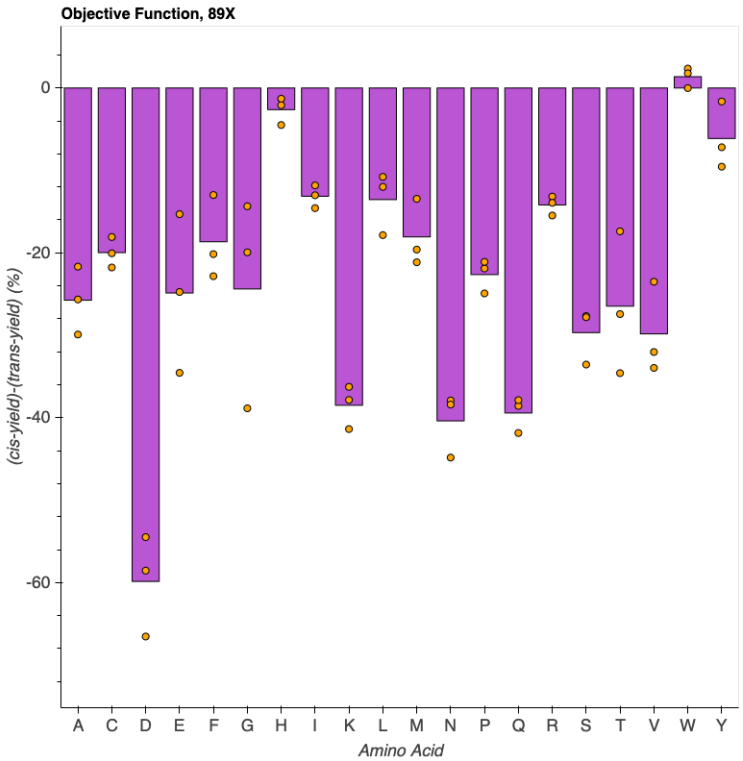


Figure C-46. Total cyclopropanation activity of single-site mutants at position 89X. The original amino acid at this position was Q. Each variant was tested in biological triplicate.

C.7.2.5. ParLQ 5-site Multi-Mutant Screening Data with the Model Substrate 1a

Reactions were prepared according to the protocols for the screening of protoglobin variants in 96-well plate format (page S41). Total cyclopropanation for formation of *cis*-**2a** and *trans*-**2a** are presented.

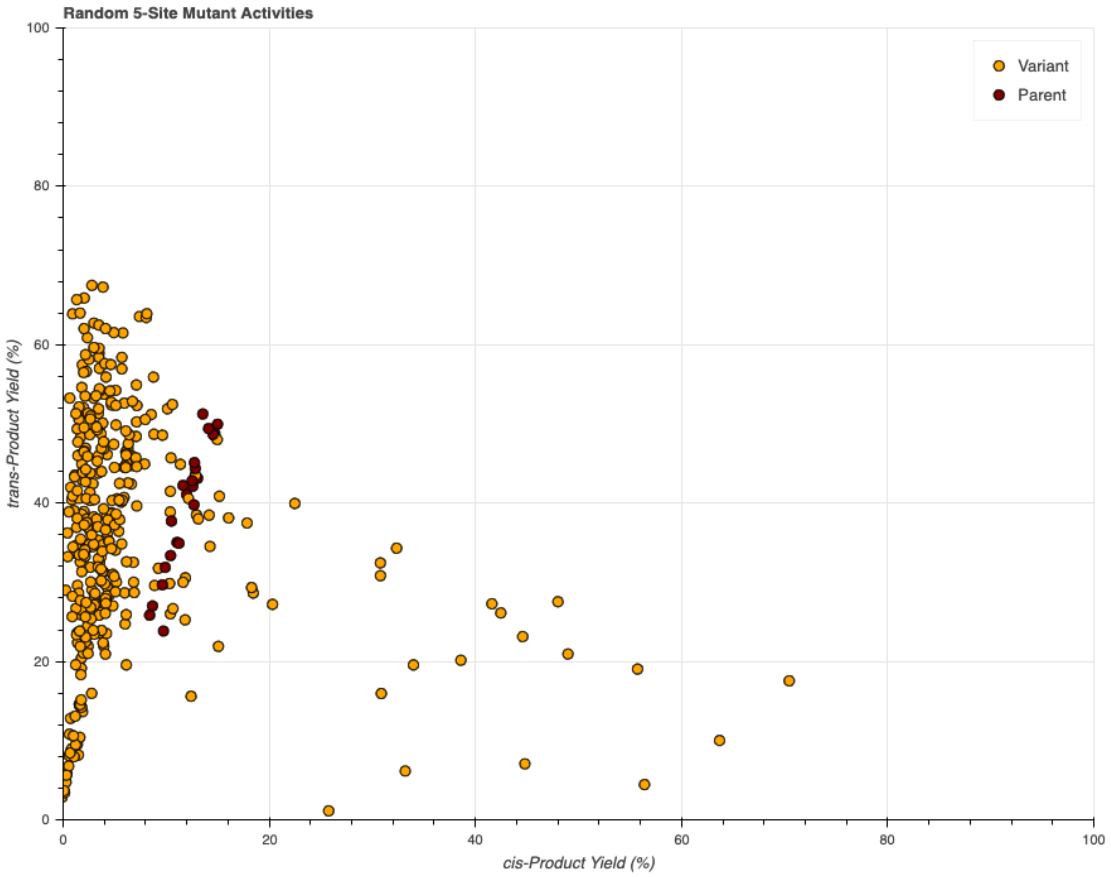


Figure C-47. Yields of various random ParLQ 5-site mutants for the formation of *cis*- and *trans*-**2a**. Only ~10% of variants in this random library had improved selectivity for *cis*-**2a.**

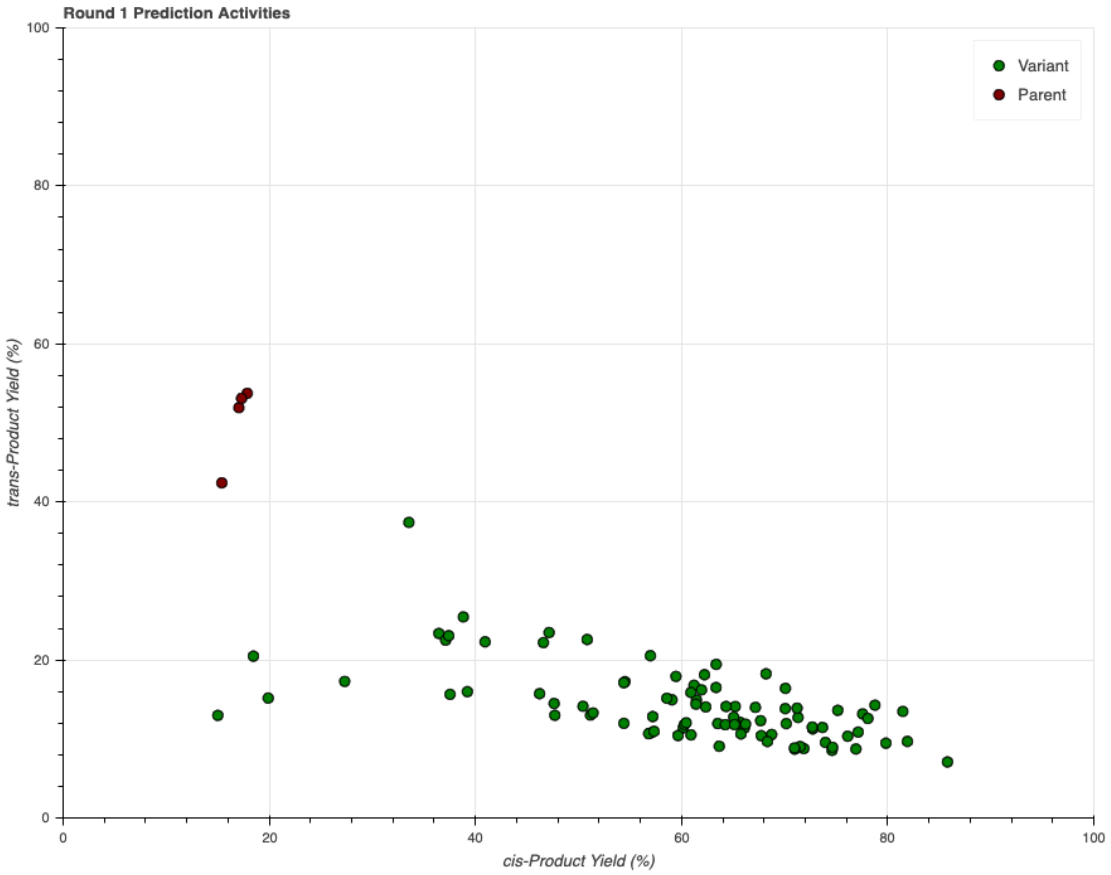


Figure C-48. Activities of the first round of mutants predicted by ALDE. Reactions with this library of variants were run in biological triplicate, and the average yields for the formation of *cis*- and *trans*-**2a** observed for each variant are shown.

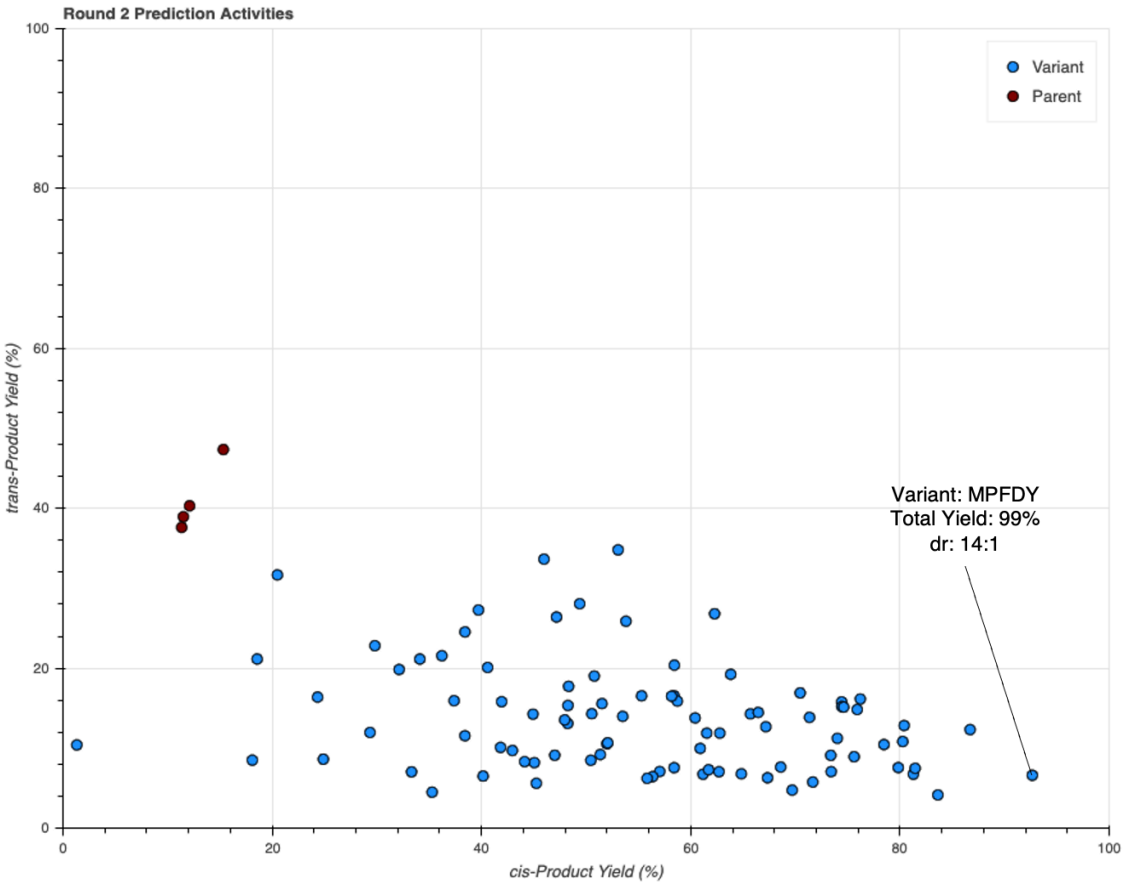
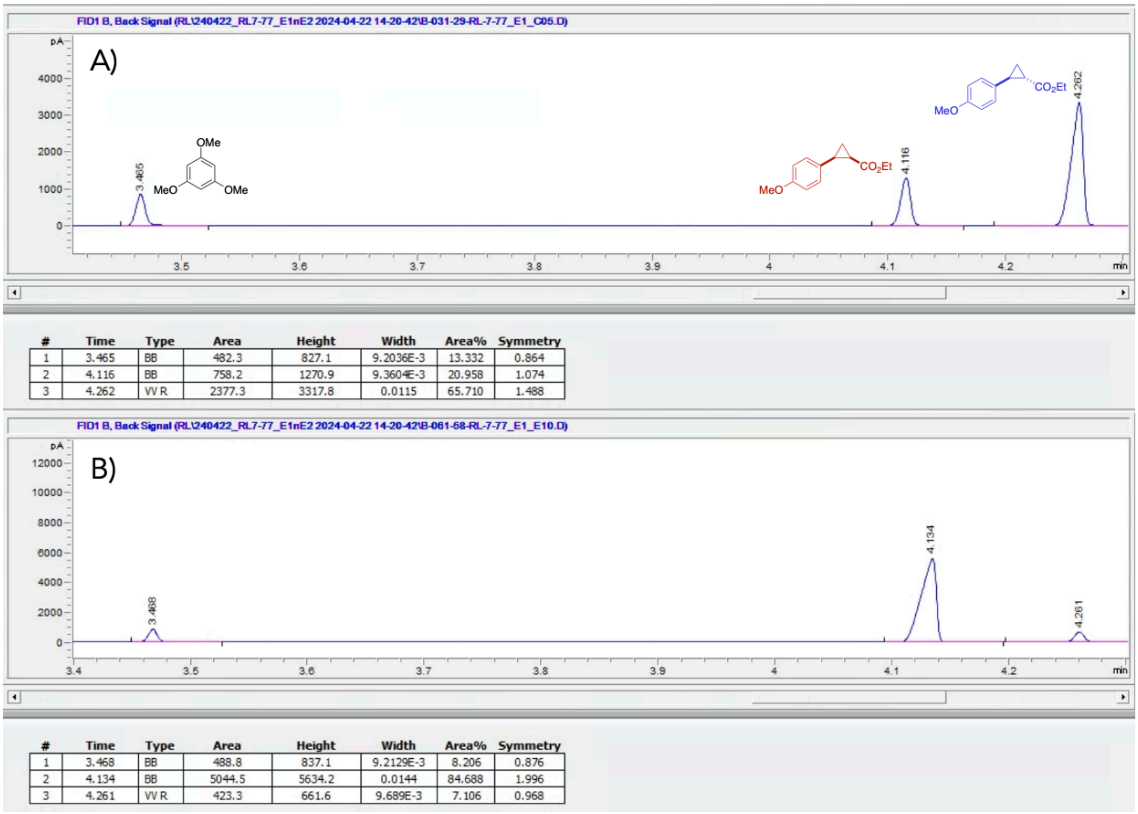


Figure C-49. Activities of the second round of mutants predicted by ALDE. Reactions with this library of variants were run in biological triplicate, and the average yields for the formation of *cis*- and *trans*-**2a** observed for each variant are shown. The highest-performing variant, MPFDY, demonstrated a yield of 99% with a 14:1 selectivity for *cis*-**2a**.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.465	BB	482.3	827.1	9.2036E-3	13.332	0.864
2	4.116	BB	758.2	1270.9	9.3604E-3	20.958	1.074
3	4.262	VV R	2377.3	3317.8	0.0115	65.710	1.488

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.468	BB	488.8	837.1	9.2129E-3	8.206	0.876
2	4.134	BB	5044.5	5634.2	0.0144	84.688	1.996
3	4.261	VV R	423.3	661.6	9.689E-3	7.106	0.968

Figure C-50. (A) Representative GC-FID trace for reactions of ParLQ with **1a** and EDA. **(B)** GC-FID trace for the top-performing predicted variant, MPFDY, in reactions with **1a** and EDA.

C.7.2.6. Reactions of ALDE Round 2 Predictions with Non-Model Substrates

Reactions were prepared according to the protocols for the screening of protoglobin variants in 96-well plate format (page S41). Total cyclopropanation for the formation of the *cis*- and *trans*- stereoisomers of each substrate are presented.

Substrate 1b – Styrene

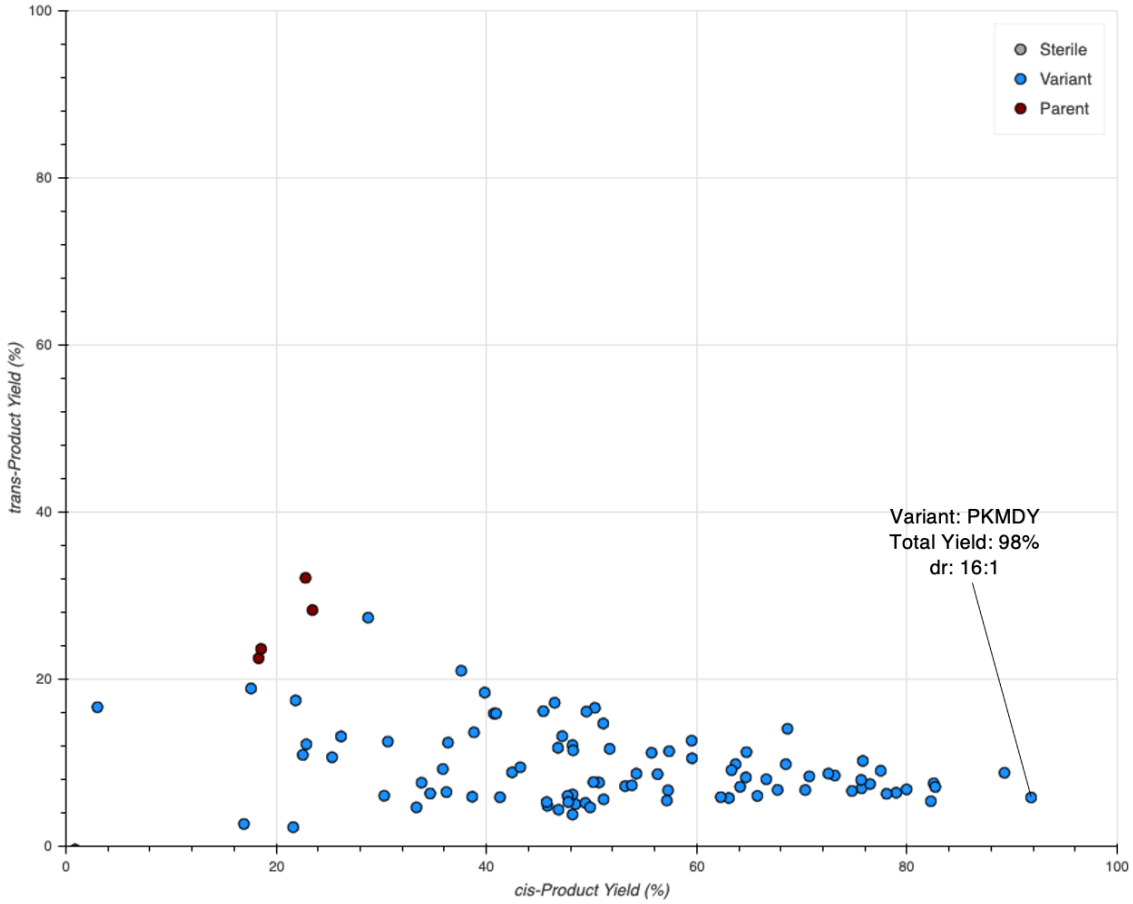
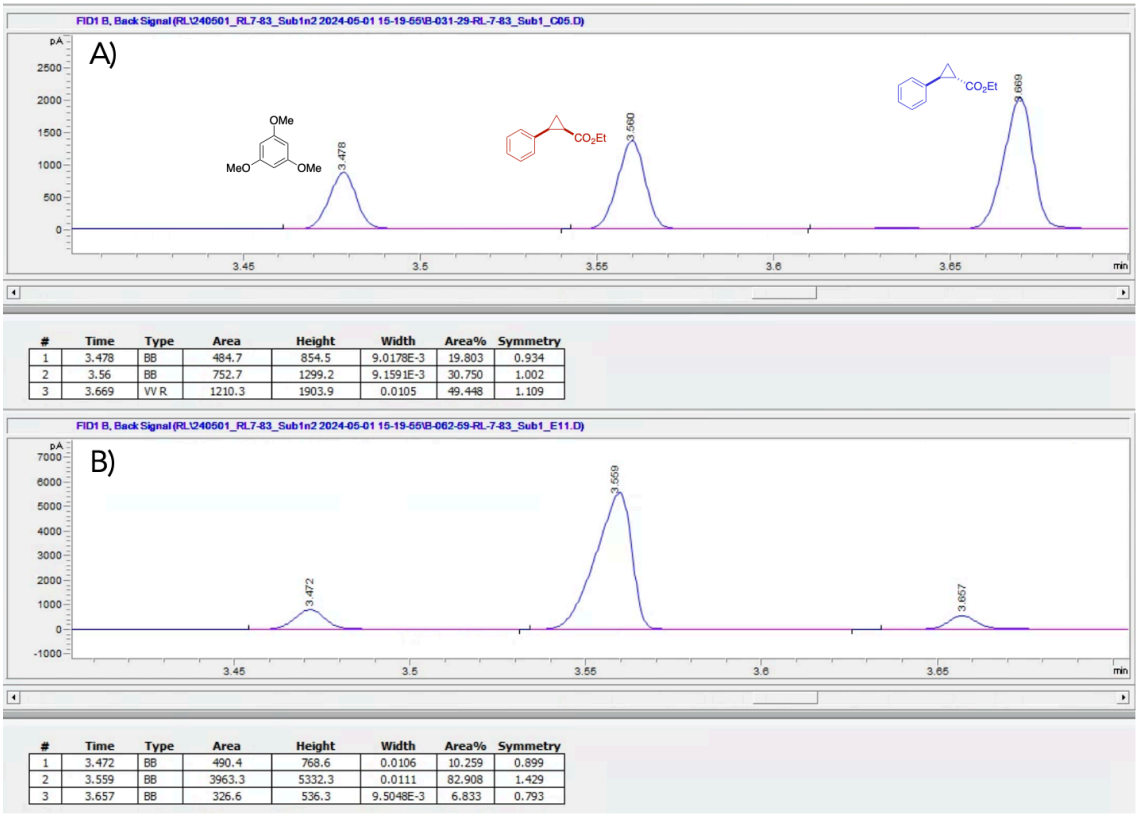


Figure C-51. Activities of the second round of mutants predicted by ALDE with substrate **1b**. Observed yields for the formation of *cis*- and *trans*-**2b** are shown. The highest-performing variant, PKMDY, demonstrated a yield of 98% with a 16:1 selectivity for *cis*-**2b**.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.478	BB	484.7	854.5	9.0178E-3	19.803	0.934
2	3.56	BB	752.7	1299.2	9.1591E-3	30.750	1.002
3	3.669	VV R	1210.3	1903.9	0.0105	49.448	1.109

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.472	BB	490.4	768.6	0.0106	10.259	0.899
2	3.559	BB	3963.3	5332.3	0.0111	82.908	1.429
3	3.657	BB	326.6	536.3	9.5048E-3	6.833	0.793

Figure C-52. (A) Representative GC-FID trace for reactions of ParLQ with **1b** and EDA. **(B)** GC-FID trace for the top-performing variant, PKMDY, in reactions with **1b** and EDA.

Substrate 1c - 1-methyl-4-vinylbenzene

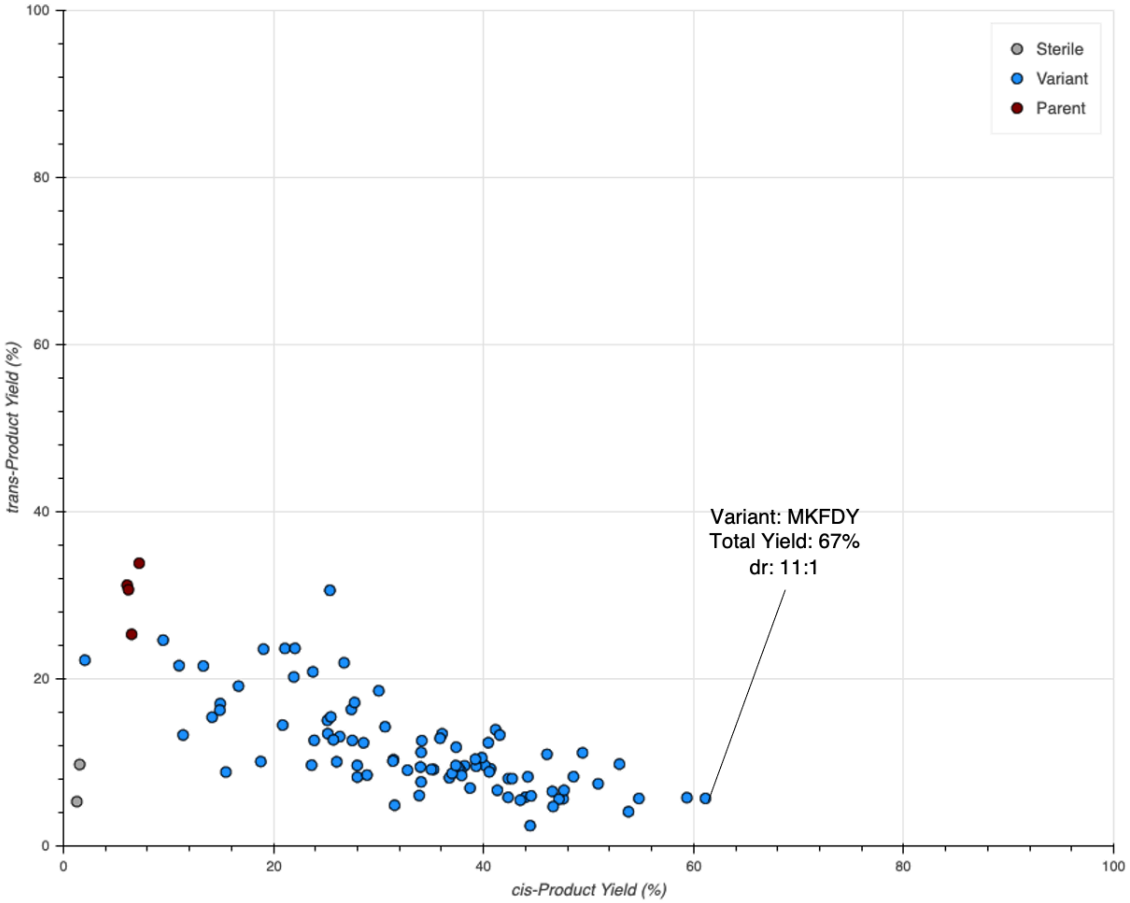


Figure C-53. Activities of the second round of mutants predicted by ALDE with substrate **1c**. Observed yields for the formation of *cis*- and *trans*-**2c** are shown. The highest-performing variant, MKFDY, demonstrated a yield of 67% with a 11:1 selectivity for *cis*-**2c**.

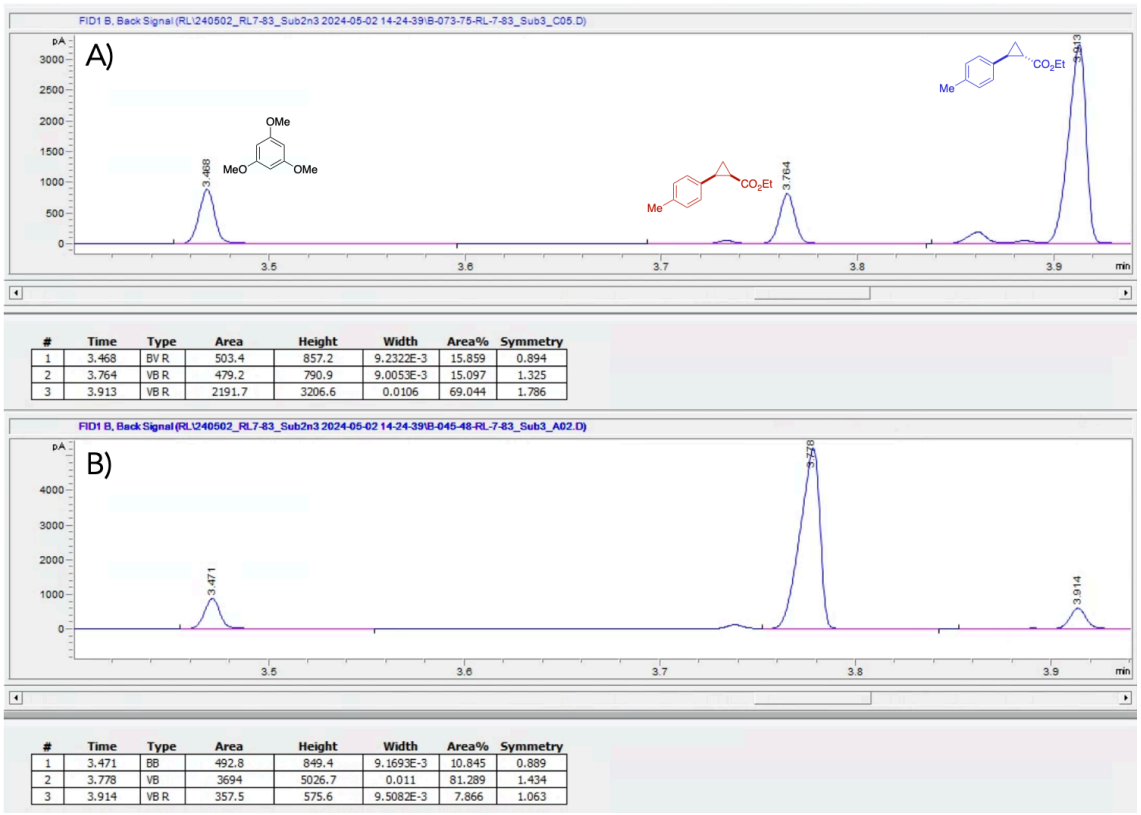


Figure C-54. (A) Representative GC-FID trace for reactions of ParLQ with **1c** and EDA. **(B)** GC-FID trace for the top-performing variant, MKFDY, in reactions with **1c** and EDA.

Substrate 1d - 1-methyl-3-vinylbenzene

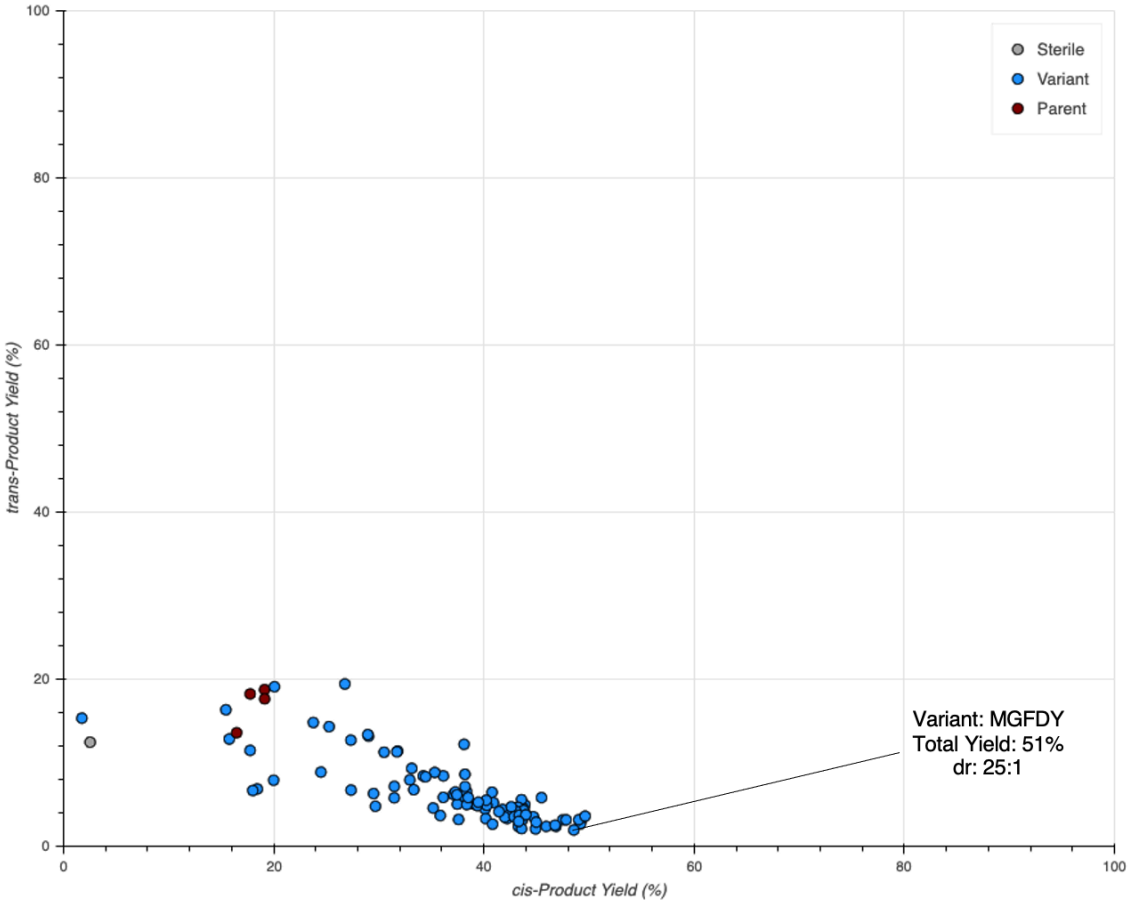
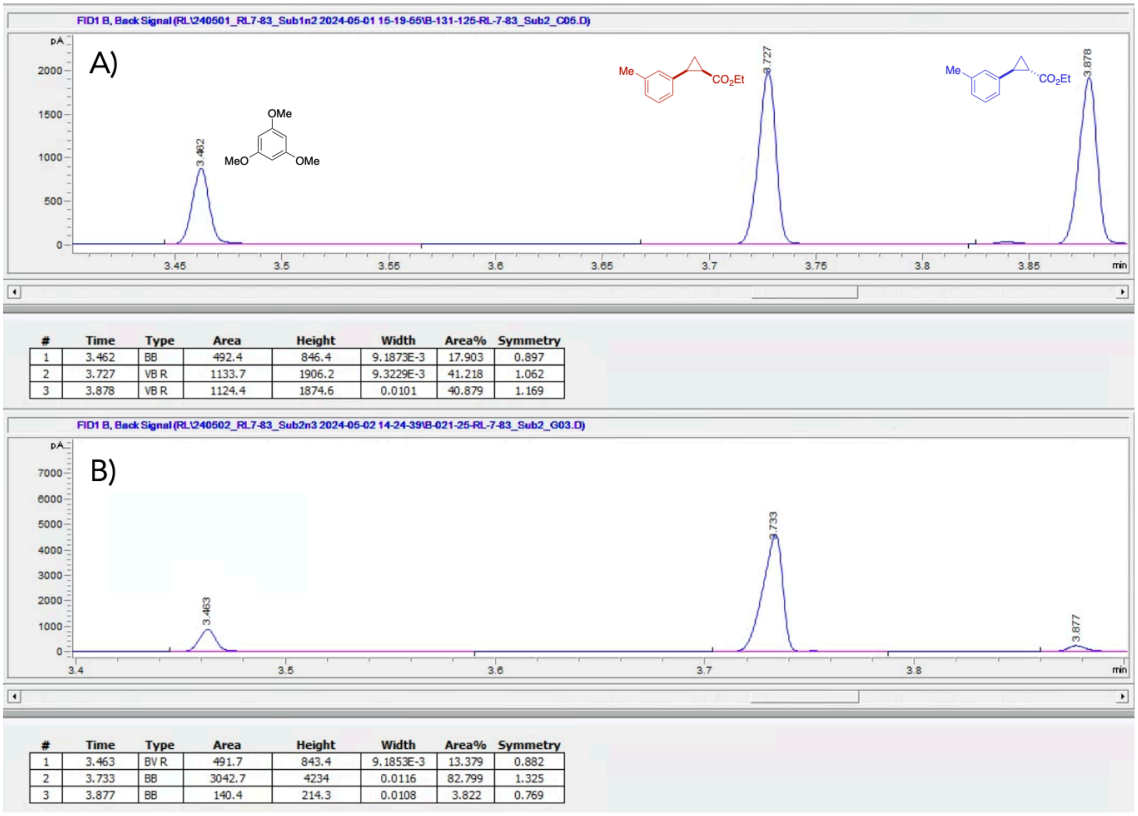


Figure C-55. Activities of the second round of mutants predicted by ALDE with substrate **1d**. Observed yields for the formation of *cis*- and *trans*-**2d** are shown. The highest-performing variant, MGFDY, demonstrated a yield of 51% with a 25:1 selectivity for *cis*-**2d**.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.462	BB	492.4	846.4	9.1873E-3	17.903	0.897
2	3.727	VB R	1133.7	1906.2	9.3229E-3	41.218	1.062
3	3.878	VB R	1124.4	1874.6	0.0101	40.879	1.169

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.463	BV R	491.7	843.4	9.1853E-3	13.379	0.882
2	3.733	BB	3042.7	4234	0.0116	82.799	1.325
3	3.877	BB	140.4	214.3	0.0108	3.822	0.769

Figure C-56. (A) Representative GC-FID trace for reactions of ParLQ with **1d** and EDA. **(B)** GC-FID trace for the top-performing variant, MGFDY, in reactions with **1d** and EDA.

Substrate 1e - 1-methyl-2-vinylbenzene

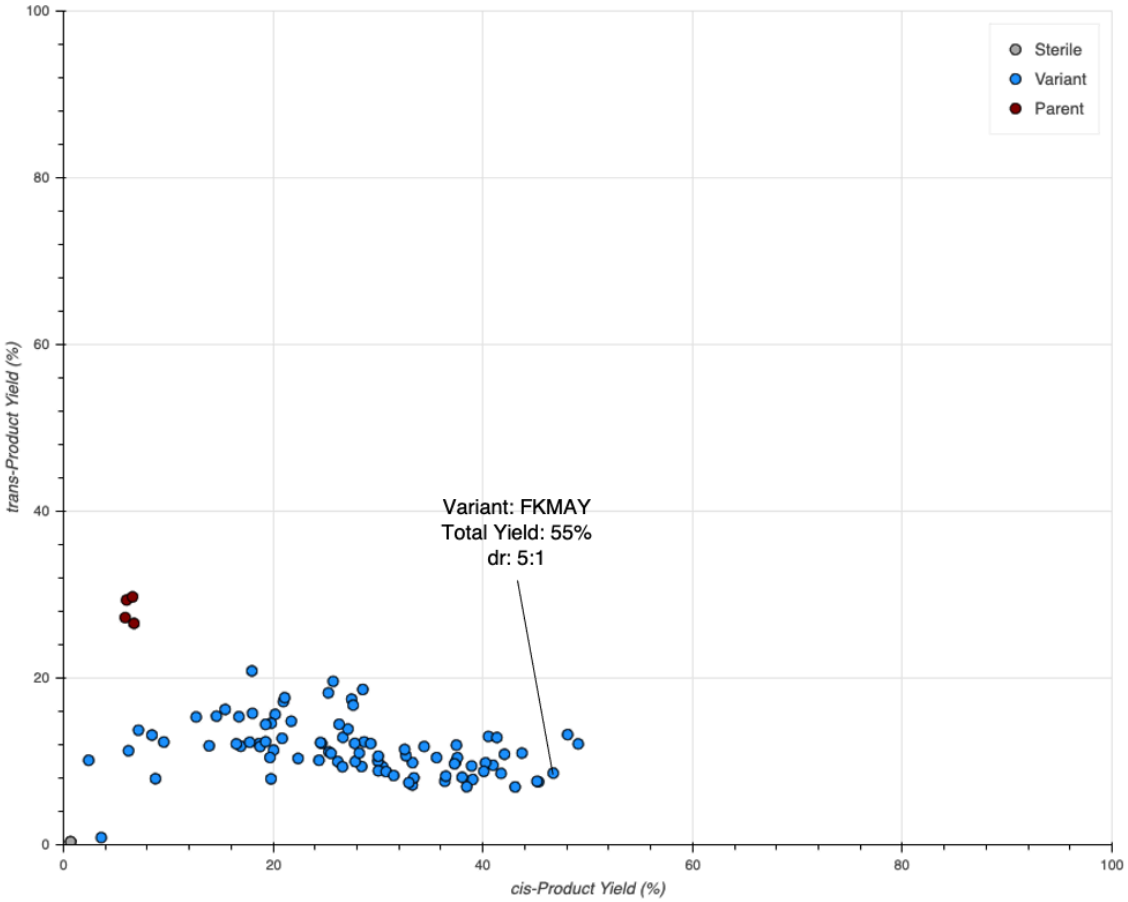


Figure C-57. Activities of the second round of mutants predicted by ALDE with substrate **1e**. Observed yields for the formation of *cis*- and *trans*-**2e** are shown. The highest-performing variant, FKMAY, demonstrated a yield of 55% with a 5:1 selectivity for *cis*-**2e**.

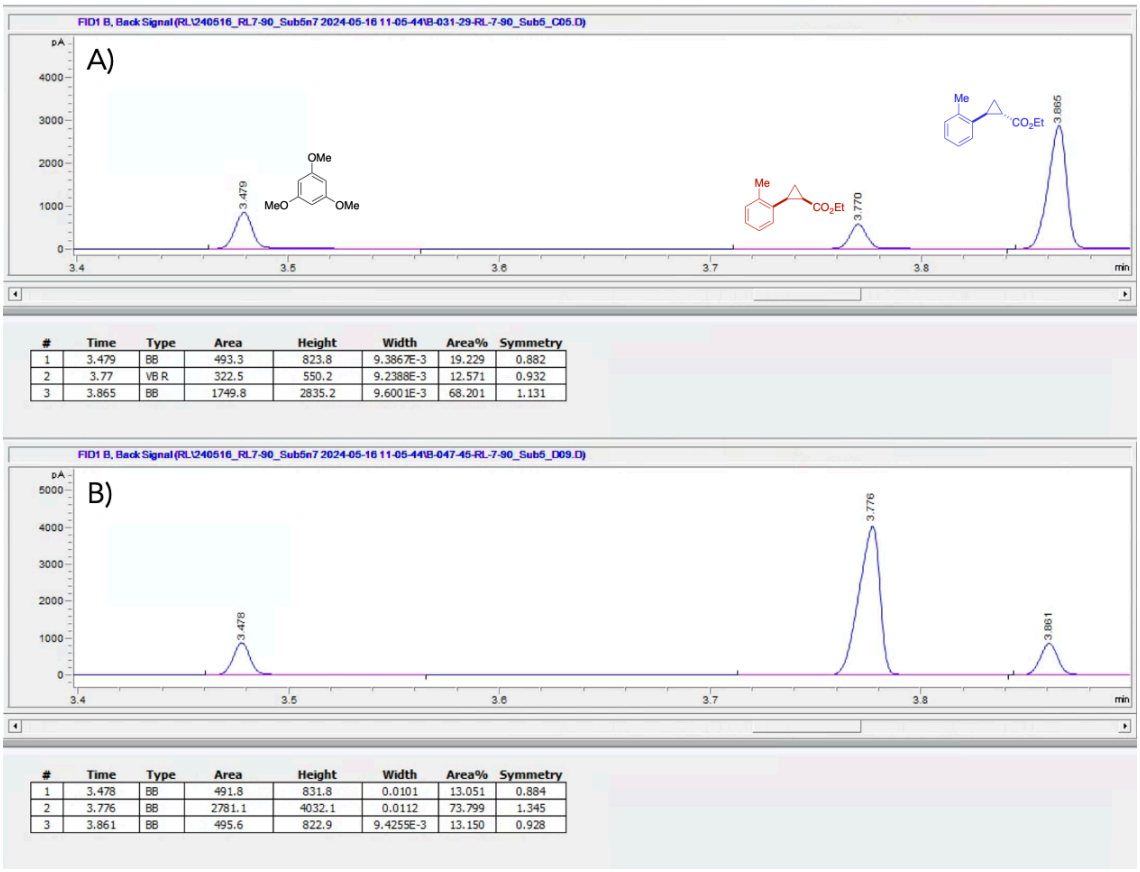


Figure C-58. (A) Representative GC-FID trace for reactions of ParLQ with **1e** and EDA. **(B)** GC-FID trace for the top-performing variant, FKMAY, in reactions with **1e** and EDA.

Substrate 1f - 1-chloro-4-vinylbenzene

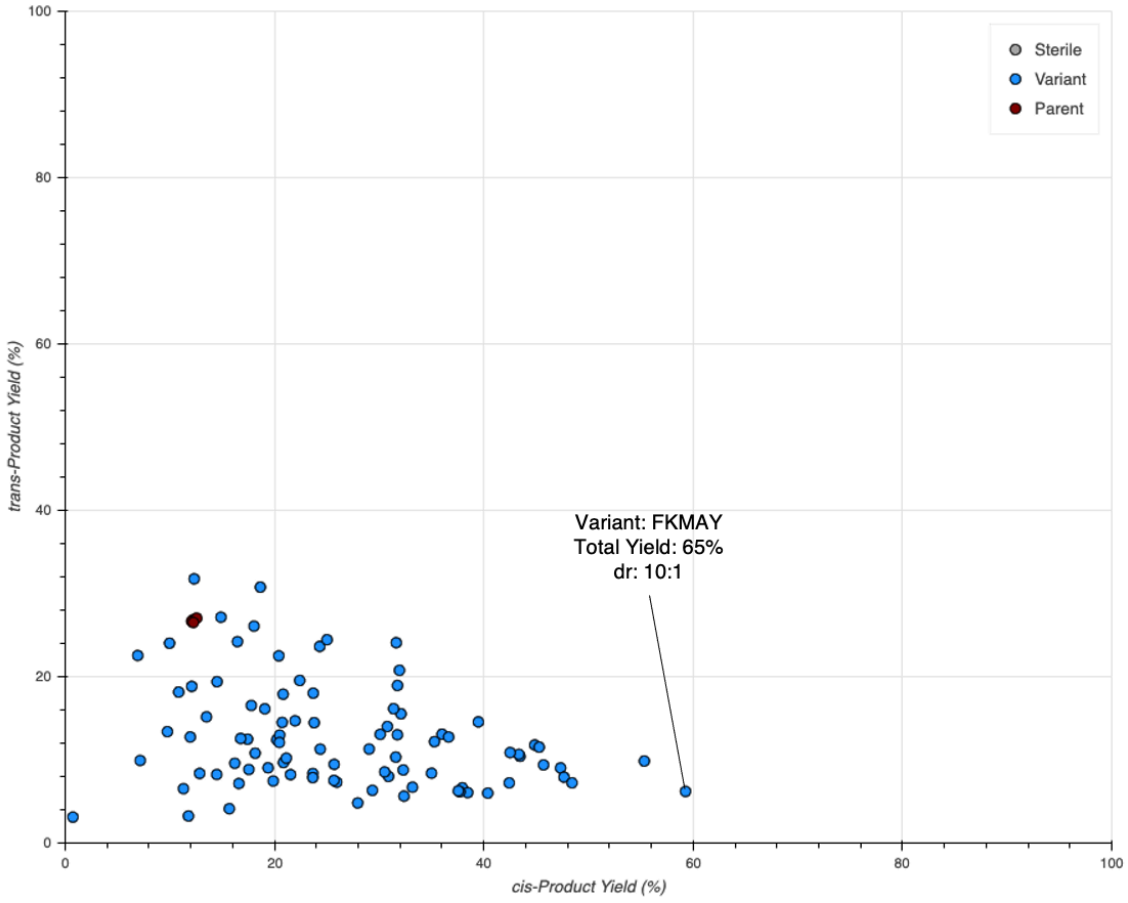


Figure C-59. Activities of the second round of mutants predicted by ALDE with substrate **1f**. Observed yields for the formation of *cis*- and *trans*-**2f** are shown. The highest-performing variant, FKMAY, demonstrated a yield of 65% with a 10:1 selectivity for *cis*-**2f**.

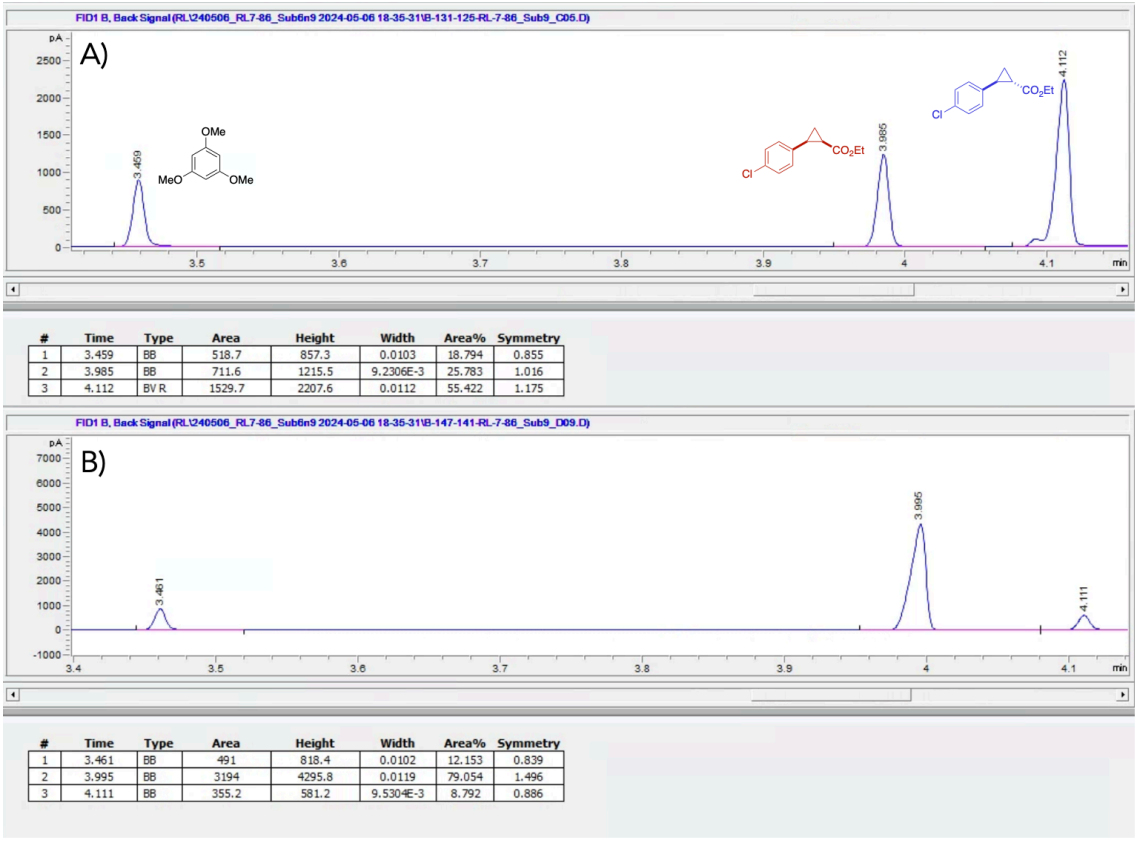


Figure C-60. (A) Representative GC-FID trace for reactions of ParLQ with **1f** and EDA. **(B)** GC-FID trace for the top-performing variant, FKMAY, in reactions with **1f** and EDA.

Substrate 1g - 1-bromo-4-vinylbenzene

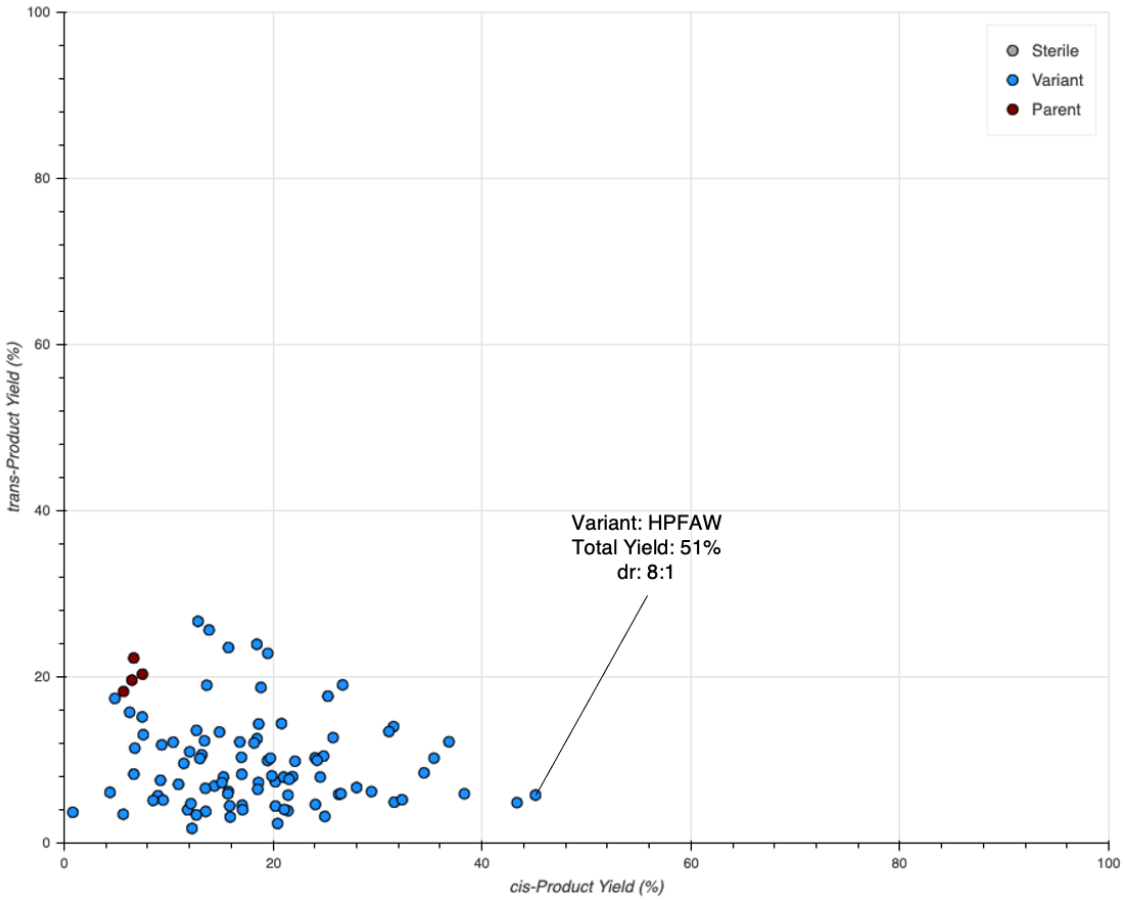


Figure C-61. Activities of the second round of mutants predicted by ALDE with substrate **1g**. Observed yields for the formation of *cis*- and *trans*-**2g** are shown. The highest-performing variant, HPFAW, demonstrated a yield of 51% with a 8:1 selectivity for *cis*-**2g**.

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.473	BV R	507.7	814.4	0.0105	27.889	0.82
2	4.232	BV	404.4	654.9	0.0104	22.214	0.928
3	4.37	VB R	908.4	1329.6	0.0103	49.897	1.035

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.464	BV R	503.9	812.9	9.5788E-3	15.585	0.784
2	4.224	BV R	2447.4	3309.6	0.0115	75.698	1.459
3	4.351	VV R	281.8	424.8	0.01	8.717	0.875

Figure C-62. (A) Representative GC-FID trace for reactions of ParLQ with **1g** and EDA. **(B)** GC-FID trace for the top-performing variant, HPFAW, in reactions with **1g** and EDA.

Substrate 1h - 1-(trifluoromethyl)-4-vinylbenzene

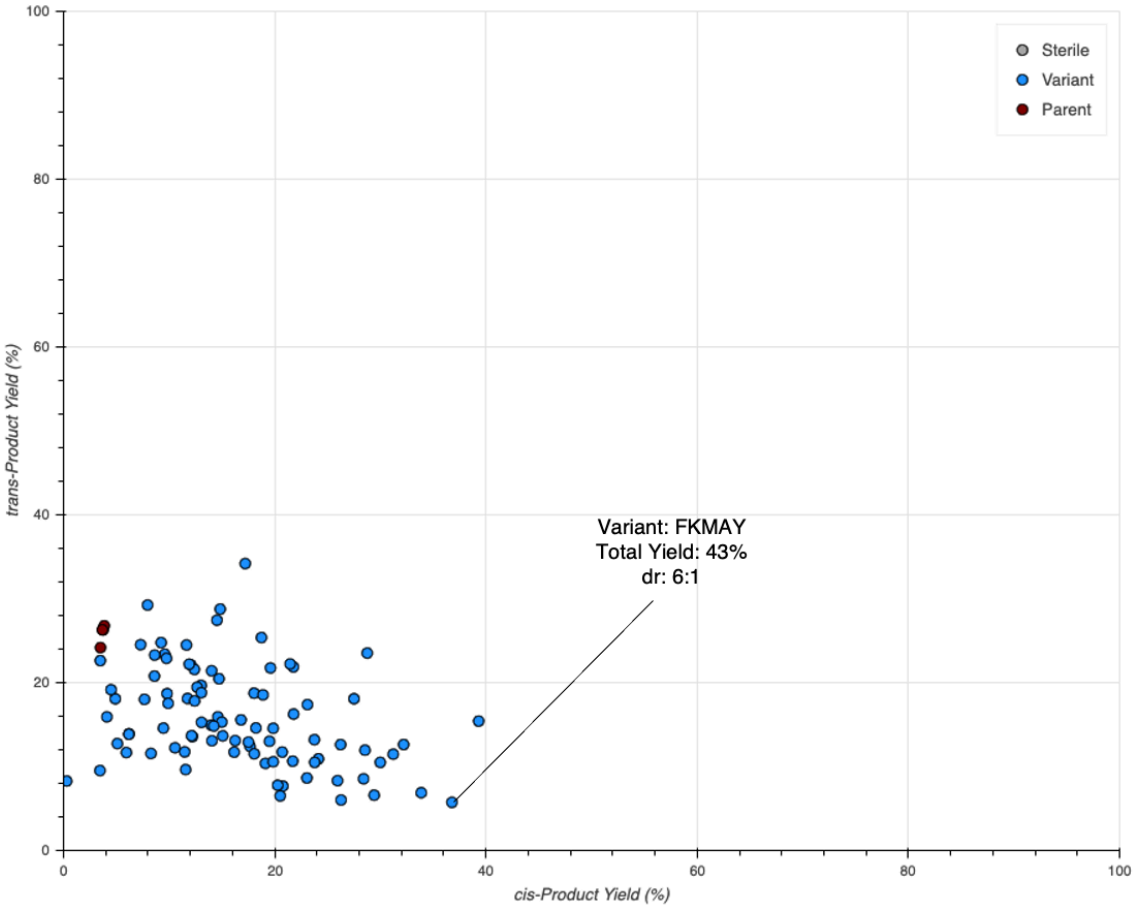
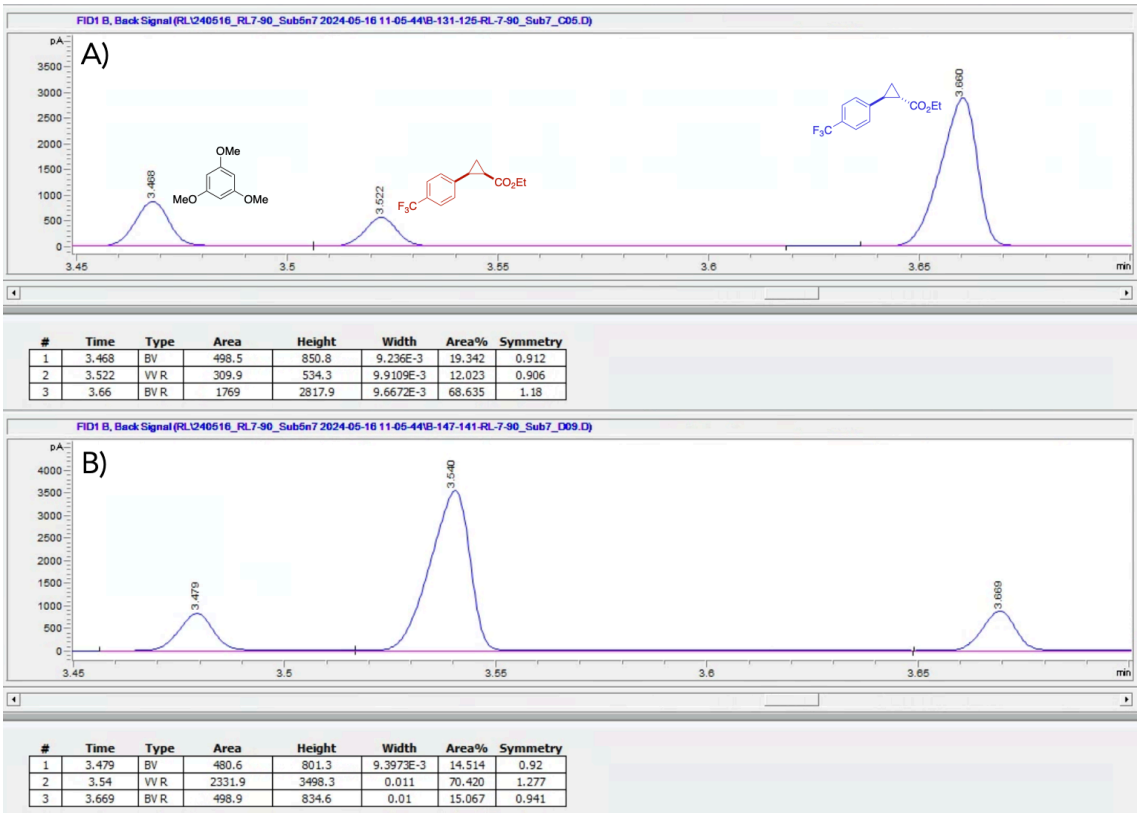


Figure C-63. Activities of the second round of mutants predicted by ALDE with substrate **1h**. Observed yields for the formation of *cis*- and *trans*-**2h** are shown. The highest-performing variant, FKMAY, demonstrated a yield of 43% with a 6:1 selectivity for *cis*-**2h**.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.468	BV	498.5	850.8	9.236E-3	19.342	0.912
2	3.522	VV R	309.9	534.3	9.9109E-3	12.023	0.906
3	3.66	BV R	1769	2817.9	9.6672E-3	68.635	1.18

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.479	BV	480.6	801.3	9.3973E-3	14.514	0.92
2	3.54	VV R	2331.9	3498.3	0.011	70.420	1.277
3	3.669	BV R	498.9	834.6	0.01	15.067	0.941

Figure C-64. (A) Representative GC-FID trace for reactions of ParLQ with **1h** and EDA. **(B)** GC-FID trace for the top-performing variant, FKMAY, in reactions with **1h** and EDA.

Substrate 1i – 2- vinylnaphthylene

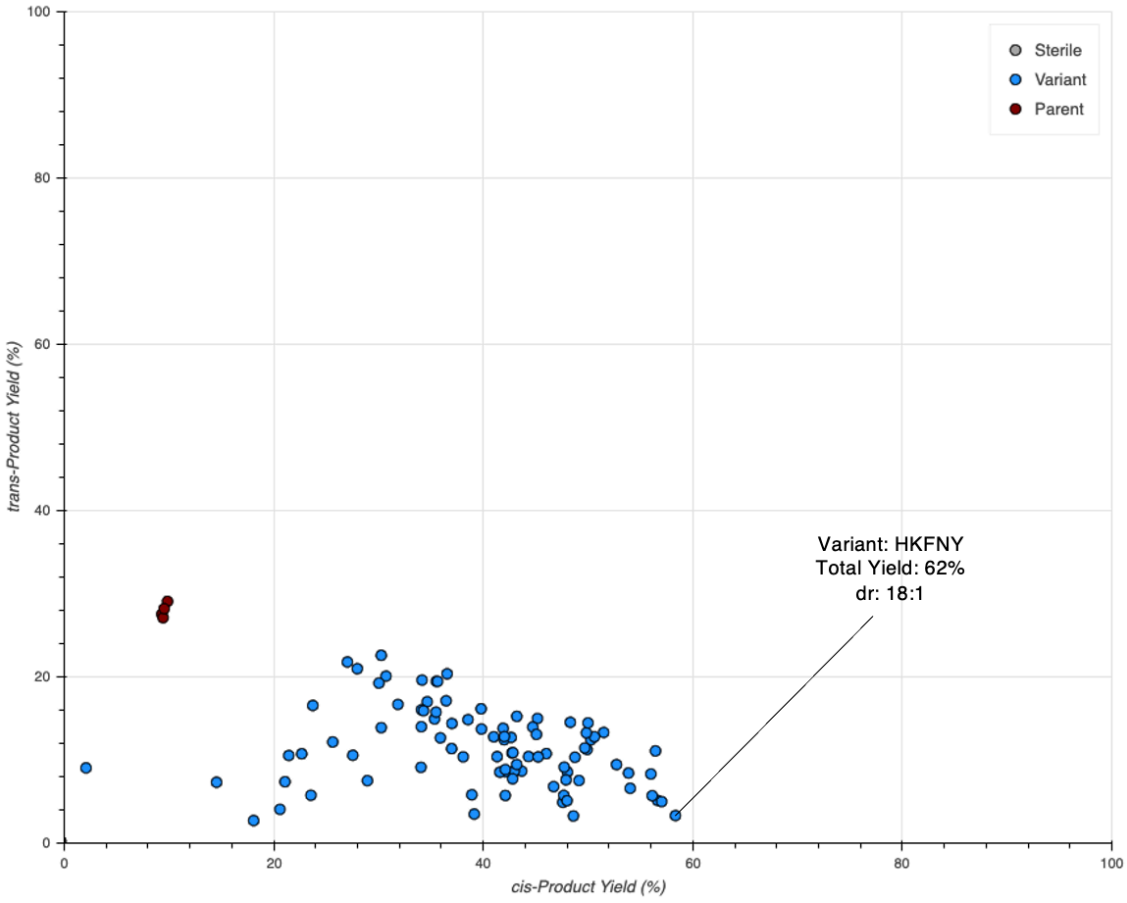


Figure C-65. Activities of the second round of mutants predicted by ALDE with substrate **1i**. Observed yields for the formation of *cis*- and *trans*-**2i** are shown. The highest-performing variant, HKFNY, demonstrated a yield of 62% with an 18:1 selectivity for *cis*-**2i**.

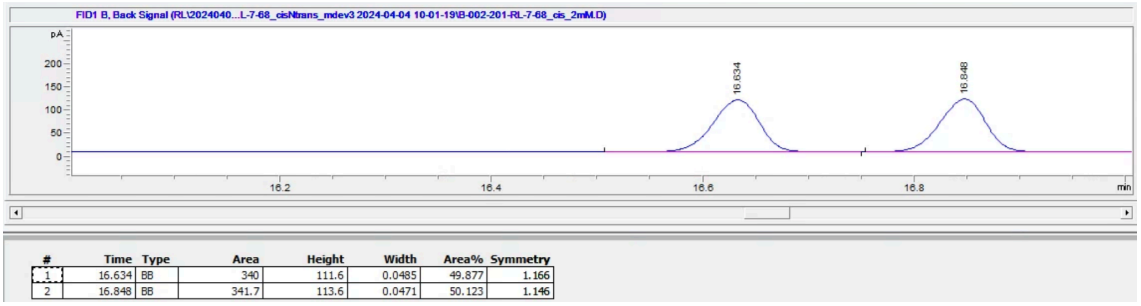
#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.464	BB	519.3	900.9	9.1252E-3	21.584	0.974
2	4.798	VB R	551.7	800.3	0.0106	22.931	1.569
3	4.998	BB	1334.9	1872.1	0.0107	55.485	1.27

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.47	BB	516.6	897.3	9.117E-3	13.677	0.968
2	4.822	VB R	3111.5	3776.4	0.0128	82.370	1.629
3	5.009	BB	149.3	227.5	0.0109	3.952	0.933

Figure C-66. (A) Representative GC-FID trace for reactions of ParLQ with **1i** and EDA. **(B)** GC-FID trace for the top-performing variant, HKFNY, in reactions with **1i** and EDA.

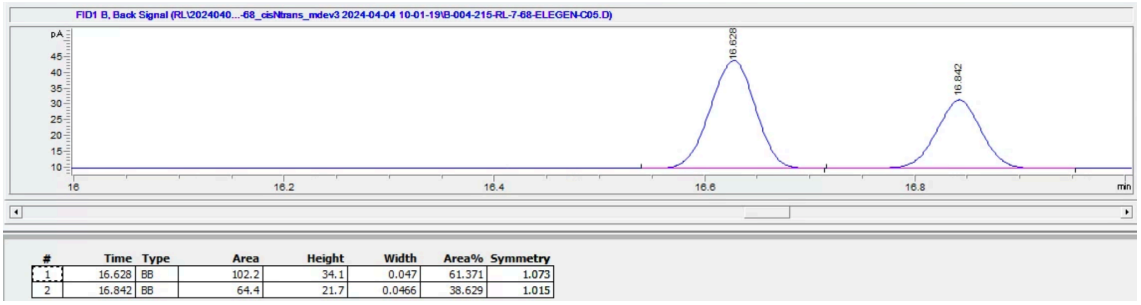
C.8. Chiral Traces for Determination of Enantiopurity

For reactions of **1a** with the libraries of ALDE-predicted variants, all samples were additionally screened for the enantiomeric ratio (er) of the *cis*- products. After analysis by achiral GC-FID samples were directly analyzed by chiral GC-FID using a chiral Agilent J&W CycloSil-B column.



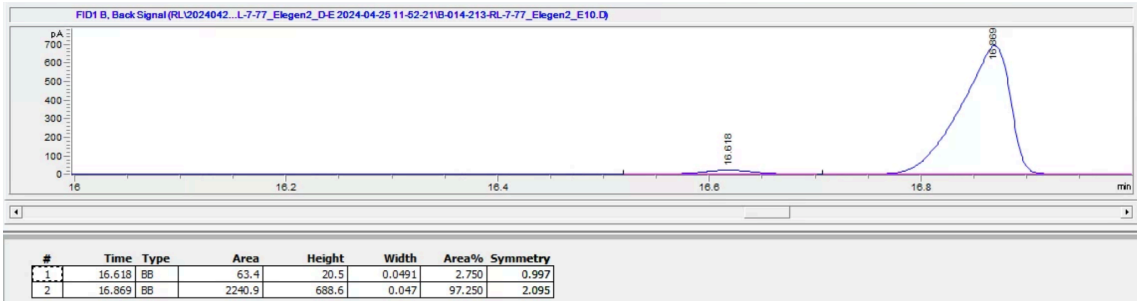
#	Time	Type	Area	Height	Width	Area%	Symmetry
1	16.634	BB	340	111.6	0.0485	49.877	1.166
2	16.848	BB	341.7	113.6	0.0471	50.123	1.146

Figure C-67. Chiral GC trace of authentic standard for *cis*-**2a**



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	16.628	BB	102.2	34.1	0.047	61.371	1.073
2	16.842	BB	64.4	21.7	0.0466	38.629	1.015

Figure C-68. Chiral GC trace of products from reaction of **1a** with ParLQ



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	16.618	BB	63.4	20.5	0.0491	2.750	0.997
2	16.869	BB	2240.9	688.6	0.047	97.250	2.095

Figure C-69. Chiral GC trace of products from reaction of **1a** with MPFDY

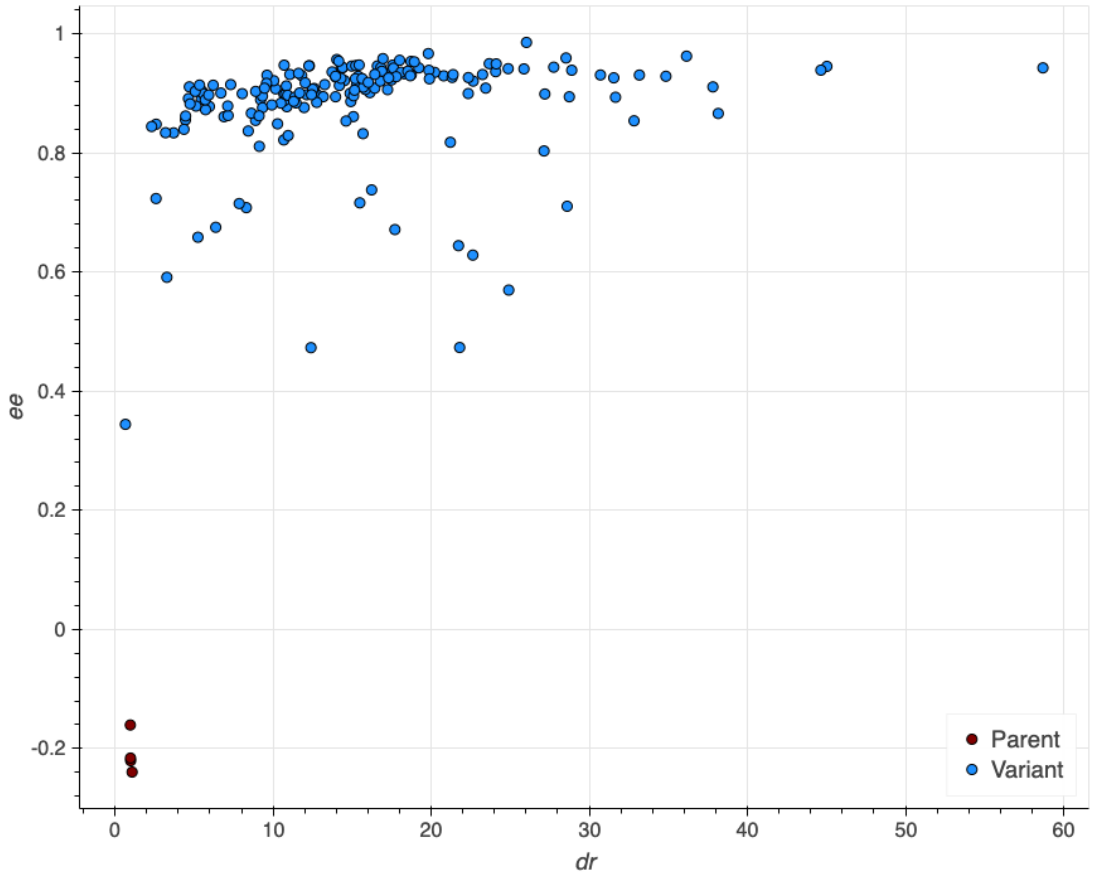


Figure C-70. Enantioselectivity data for all ALDE predicted variants for the production of **2a**. ee values are plotted against dr values for each sample.

For all other substrates the enantioselectivity of the catalyst was only measured for the top-performing variant. All samples were measured using the same method as for the chiral separation of *cis*-**2a** by GC.

C.9. Additional ML Model Analyses

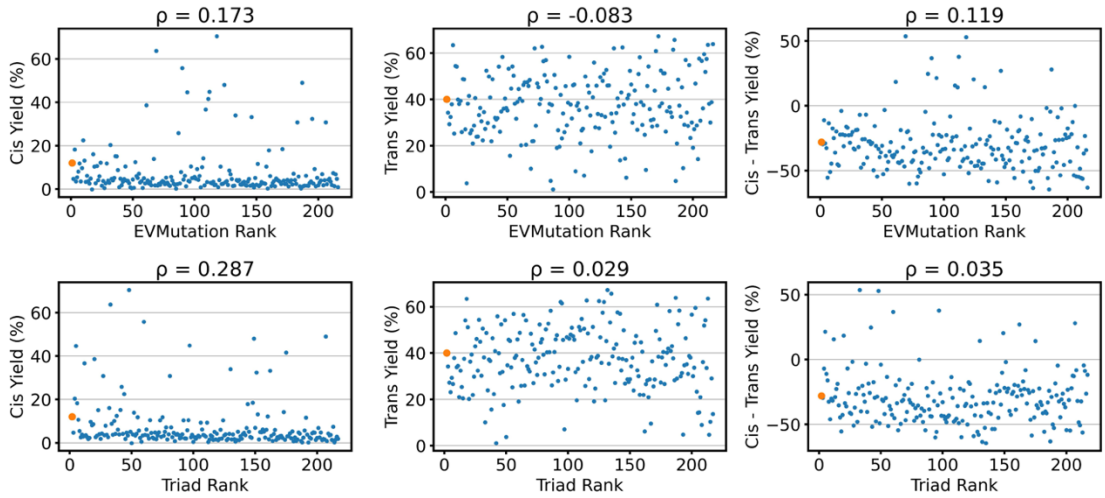


Figure C-71. Correlation between zero-shot predictors (EVMutation Rank and Triad Rank) and different fitness metrics (Cis, Trans, and Cis - Trans Yield) for the initial random library of variants used in the *ParPgb* wet-lab campaign. EVMutation rank refers to the evolutionary likelihood of a variant (1 is the most likely), and Triad rank refers to the computationally predicted stability of a variant as a ΔΔG value (1 is the most stable). Orange dot refers to the parent sequence, WYLQF. Each title shows the spearman correlation between the zero-shot predictor and the fitness metric. Cis yield is weakly correlated to the zero-shot predictors, but the overall objective is not.

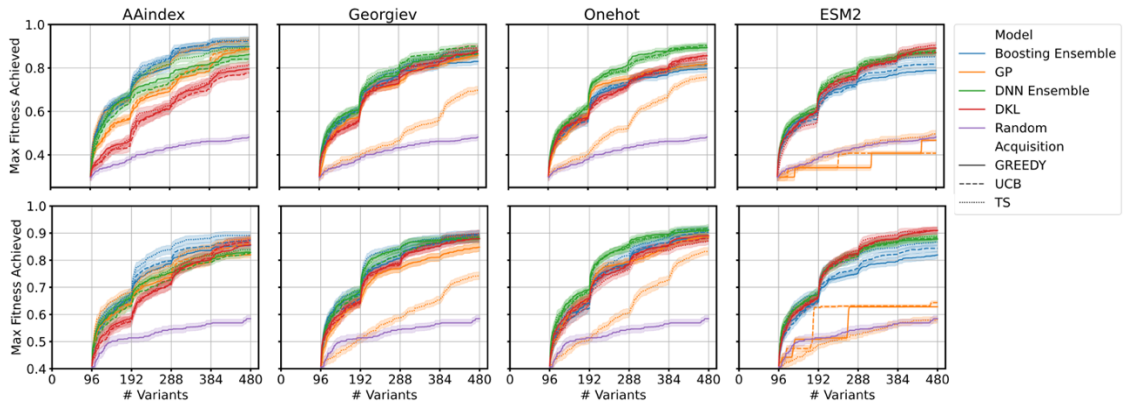


Figure C-72. Optimization trajectories for ALDE campaigns for 4 encodings, 4 models, and 3 acquisition functions. Simulation involved batch BO with an initial batch of 96 samples, followed by 4 batches of 96 samples each. Top row is GB1 and bottom row is TrpB. Error bars indicate standard deviation across 70 random initializations.

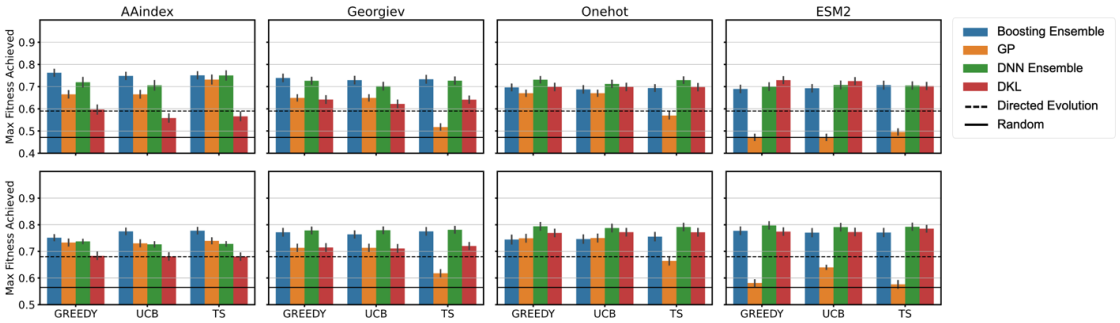


Figure C-73. Performance of MLDE baseline for 4 encodings, 4 models, and 3 acquisition functions, compared to the average DE simulation and to random sampling. Performance is quantified as the normalized maximum fitness achieved by MLDE, where the training set is 384 random samples and the test set in 96 samples proposed by the model using a greedy acquisition function. Top row is GB1 and bottom row is TrpB. Error bars indicate standard deviation across 70 random initializations.

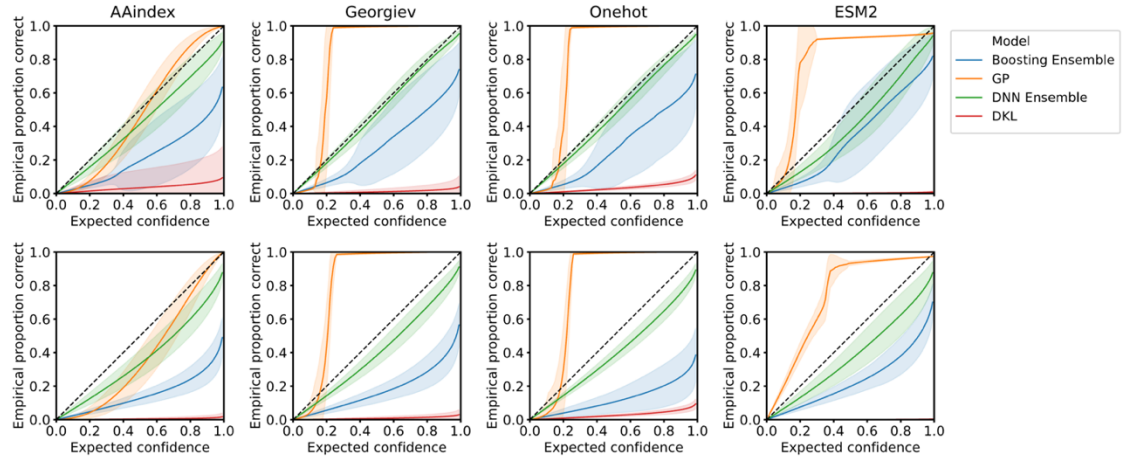
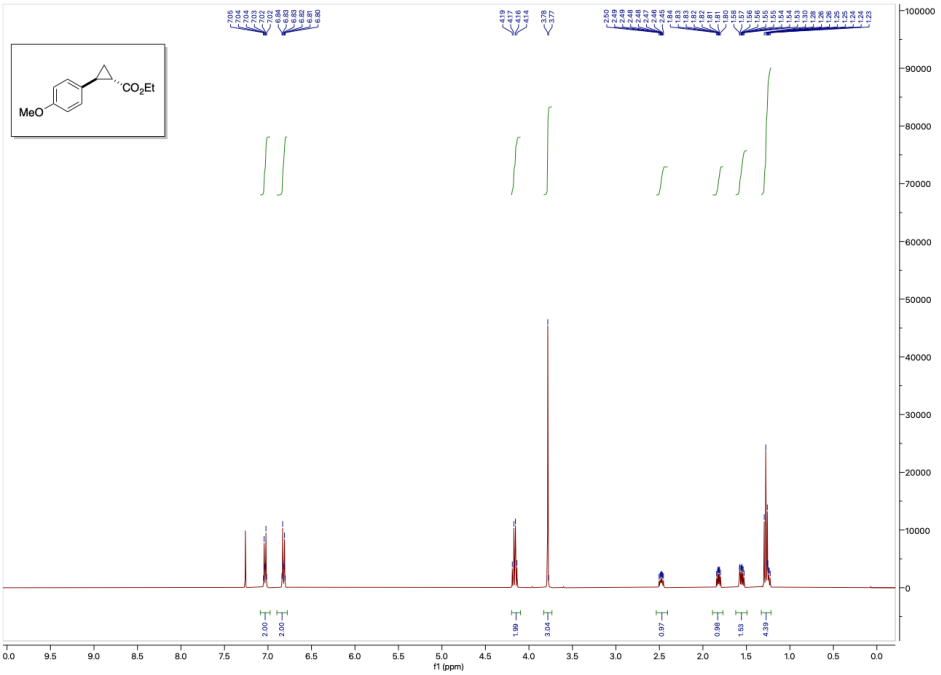
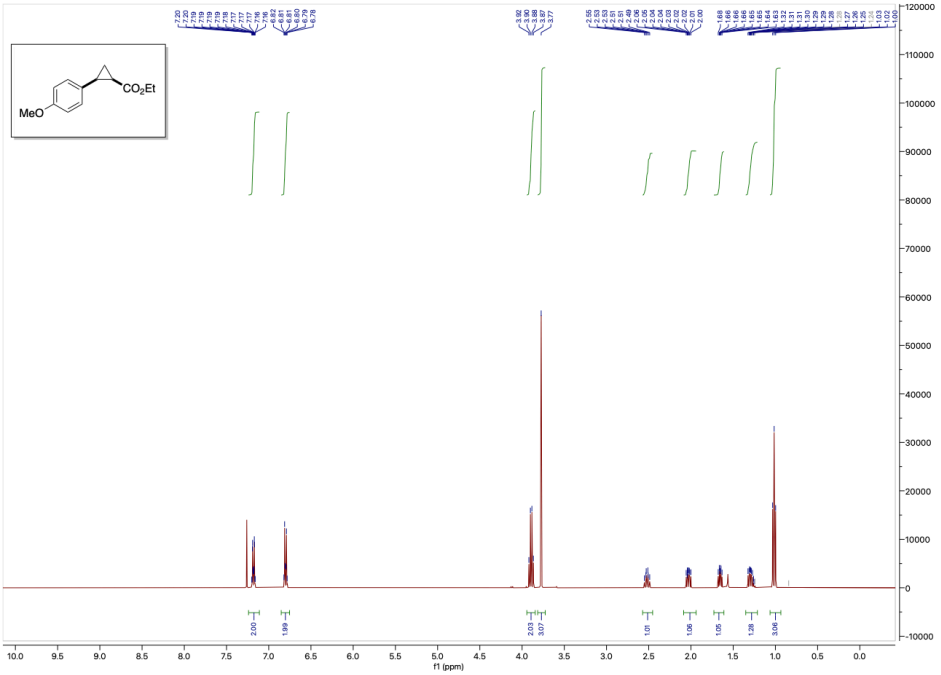


Figure C-74. Calibration curves for 4 encodings and 4 models. The x axis is the expected confidence from the posterior, given a certain confidence interval and the y value is the actual proportion of true labels that fall within the confidence interval. Calibration curve is evaluated across all sequences in the design space with labels, on models trained on the final batch (384 train samples) from ALDE campaigns using UCB. Top row is GB1 and bottom row is TrpB.

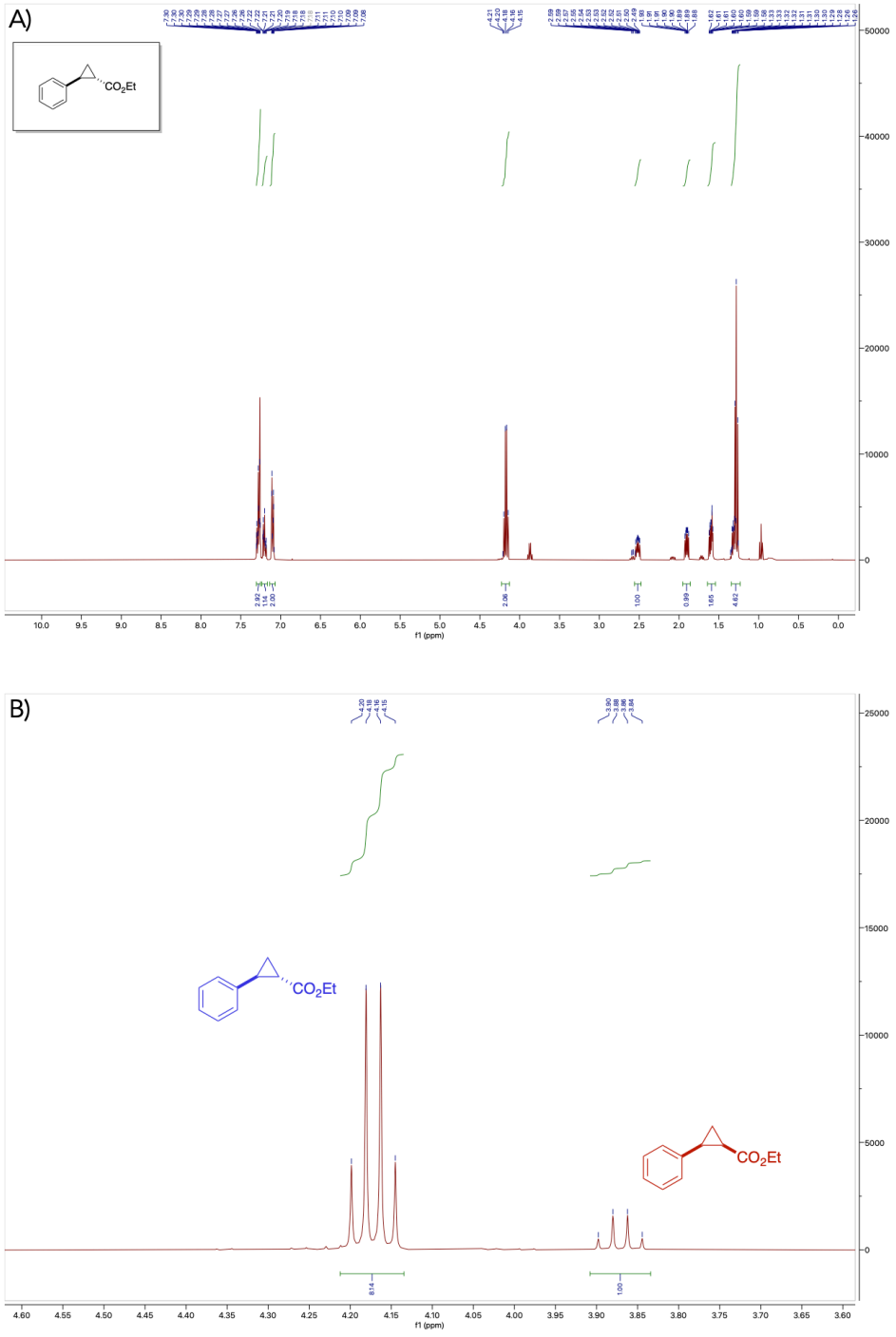
C.10. ^{1}H NMR Spectra of Authentic Standards



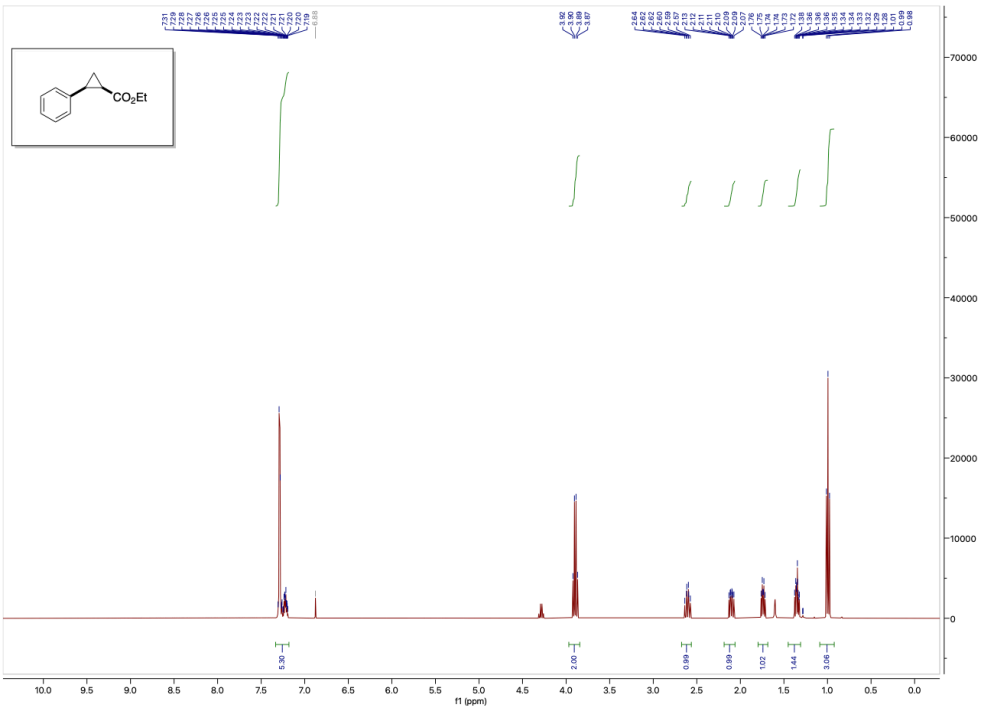
^{1}H NMR spectrum of racemic *trans*-**2a**.



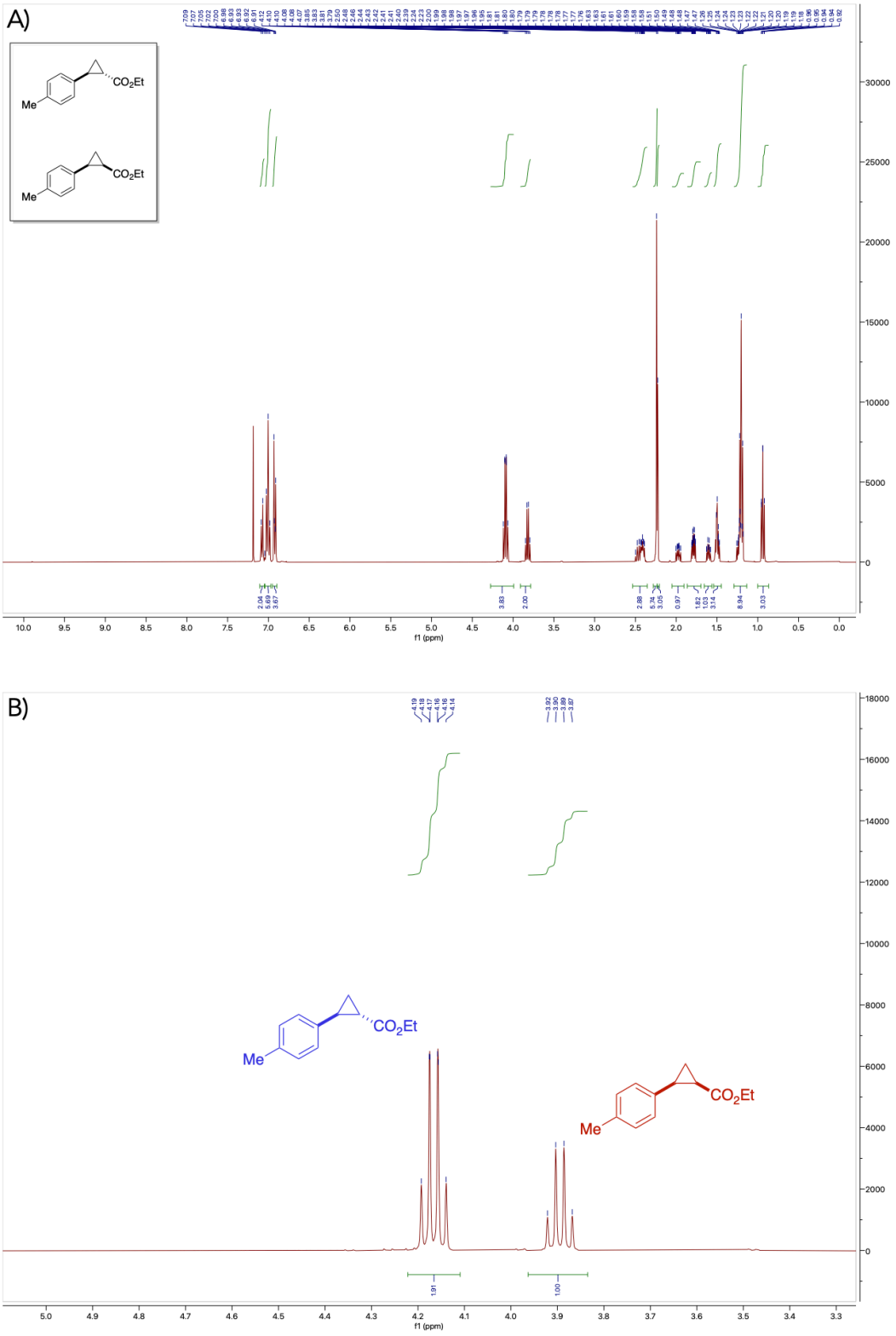
^{1}H NMR spectrum of racemic *cis*-**2a**.



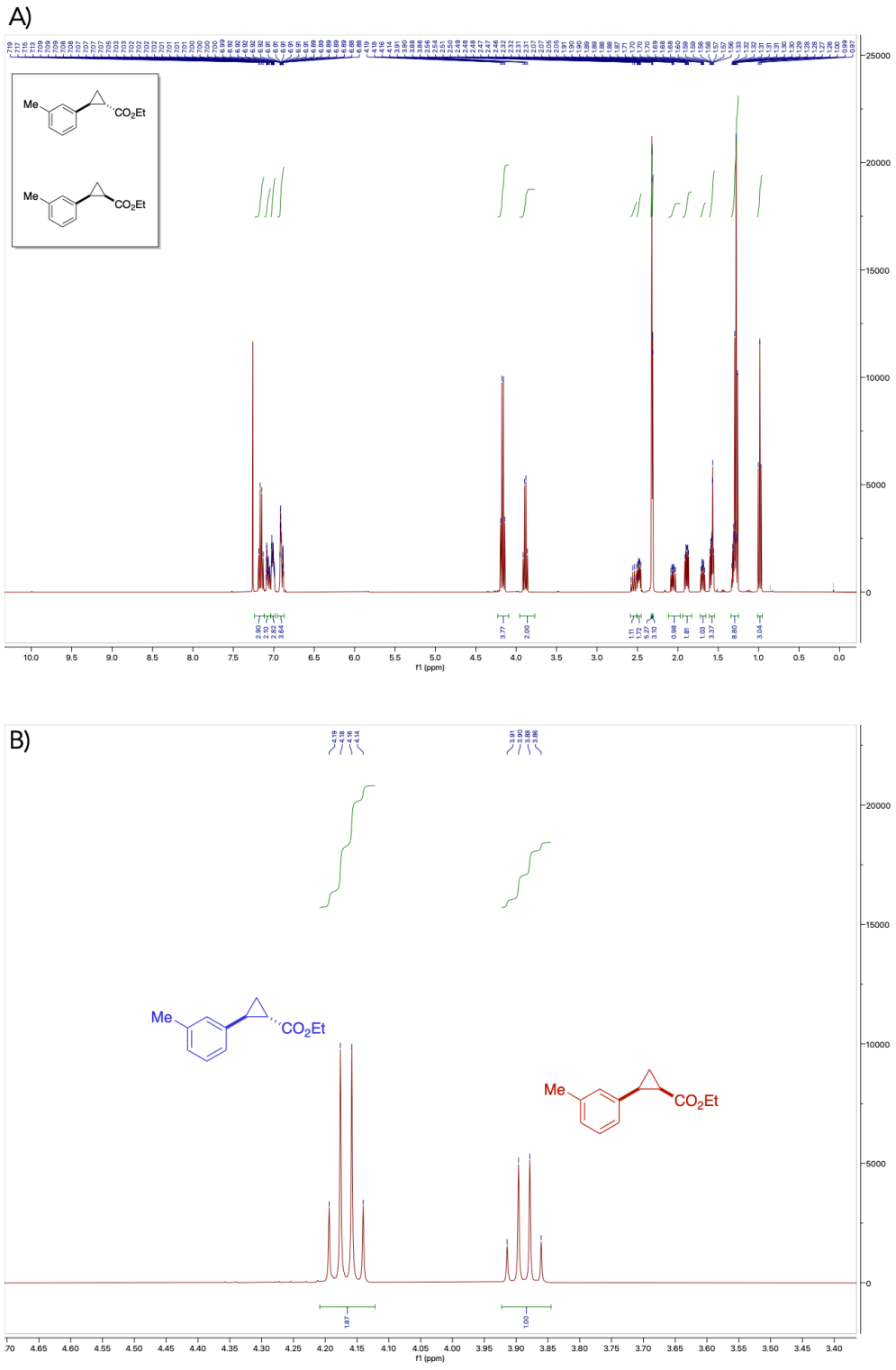
(A) ^{1}H NMR spectrum of a 8:1 racemic *trans*-**2b**/*cis*-**2b**. **(B)** Integrations of product ester methylene protons for *trans*-**2b** and *cis*-**2b** to determine the isolated molar ratio.



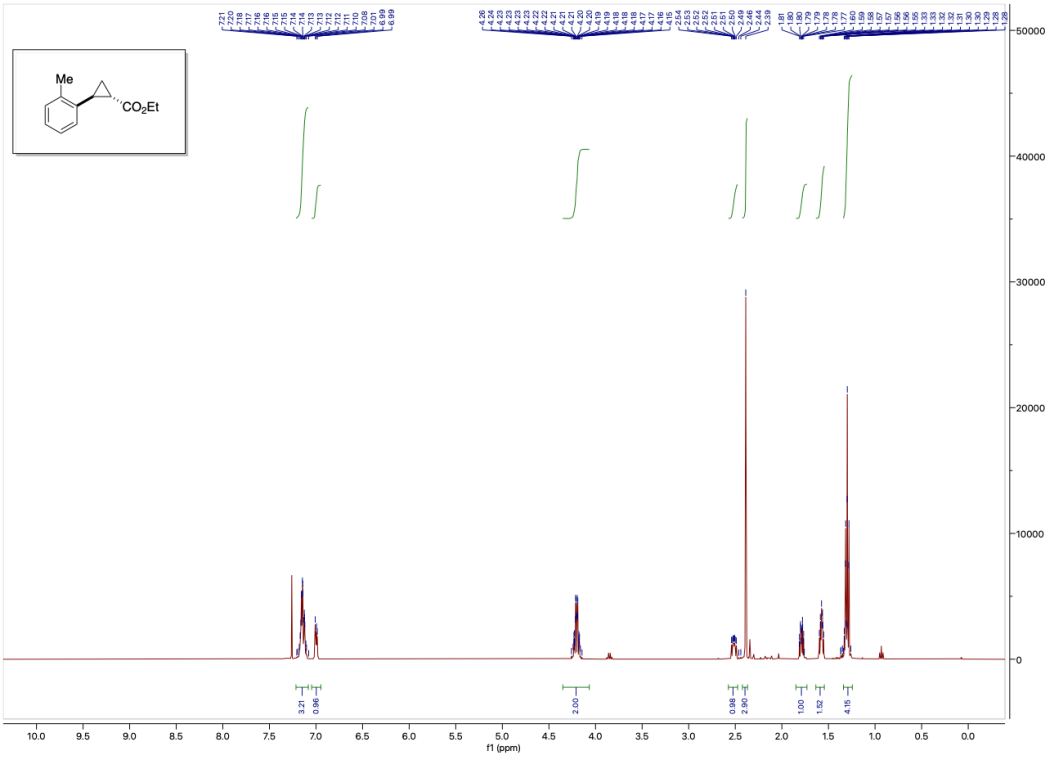
^{1}H NMR spectrum of racemic *cis*-**2b.** A trace impurity of diethyl fumarate can be observed.



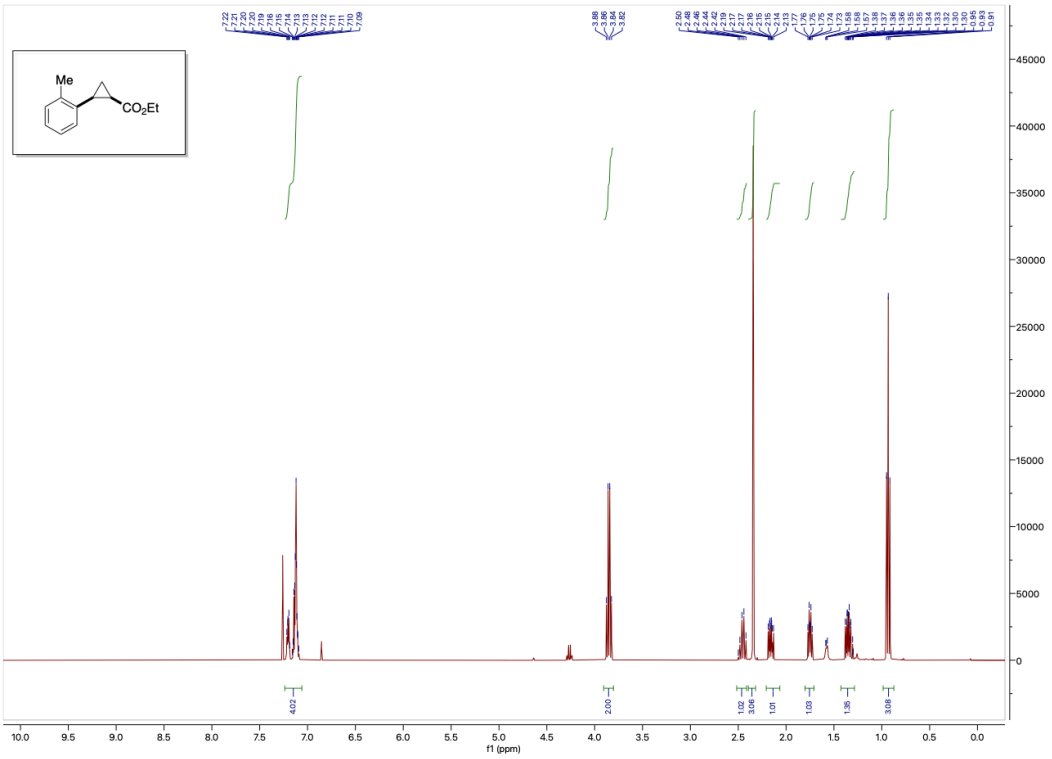
(A) ^{1}H NMR spectrum of a 1.91:1 racemic *trans*-**2c**/*cis*-**2c**. **(B)** Integrations of product ester methylene protons for *trans*-**2c** and *cis*-**2c** to determine the isolated molar ratio.



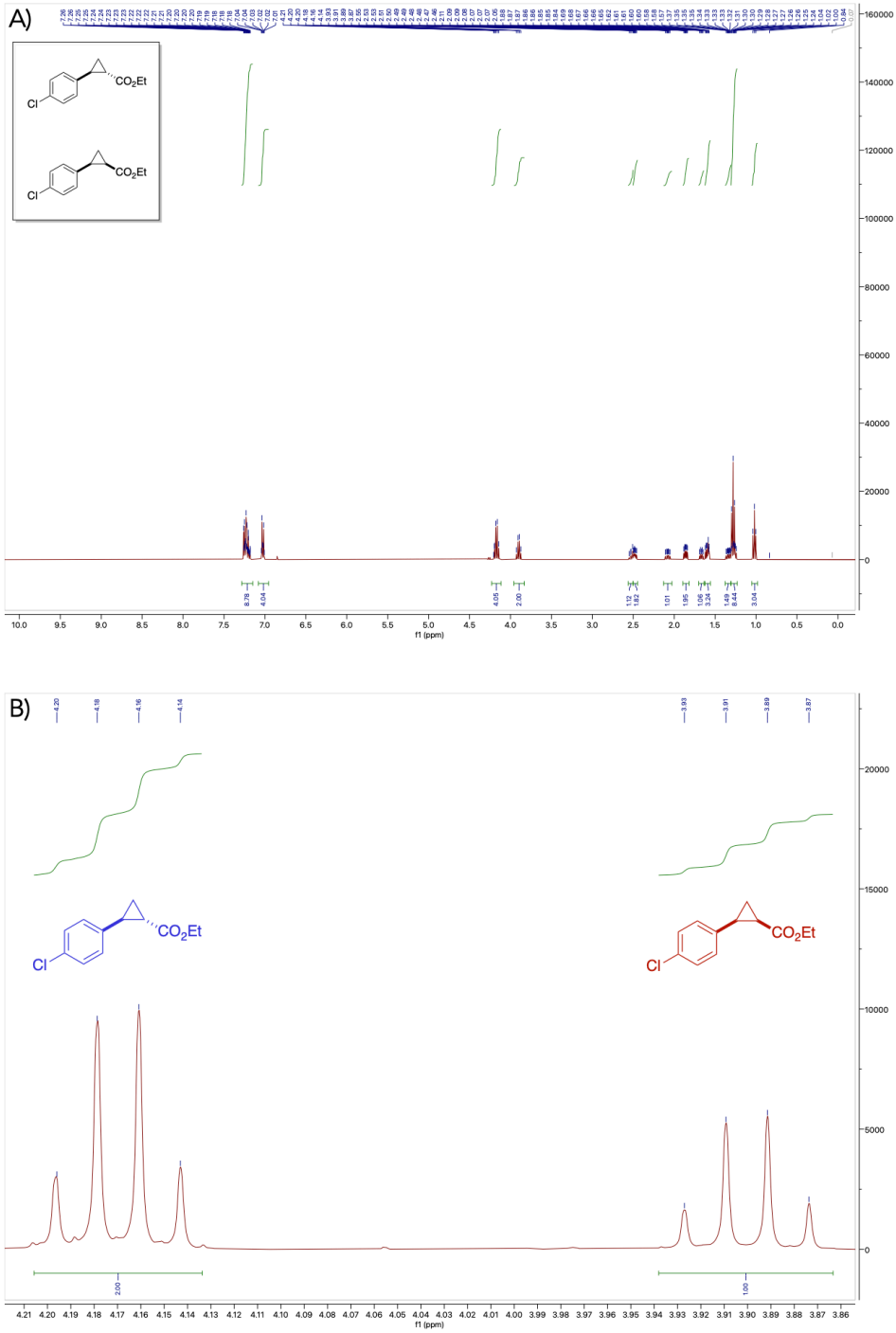
(A) ^{1}H NMR spectrum of a 1.87:1 racemic *trans*-**2d**/*cis*-**2d**. **(B)** Integrations of product ester methylene protons for *trans*-**2d** and *cis*-**2d** to determine the isolated molar ratio.



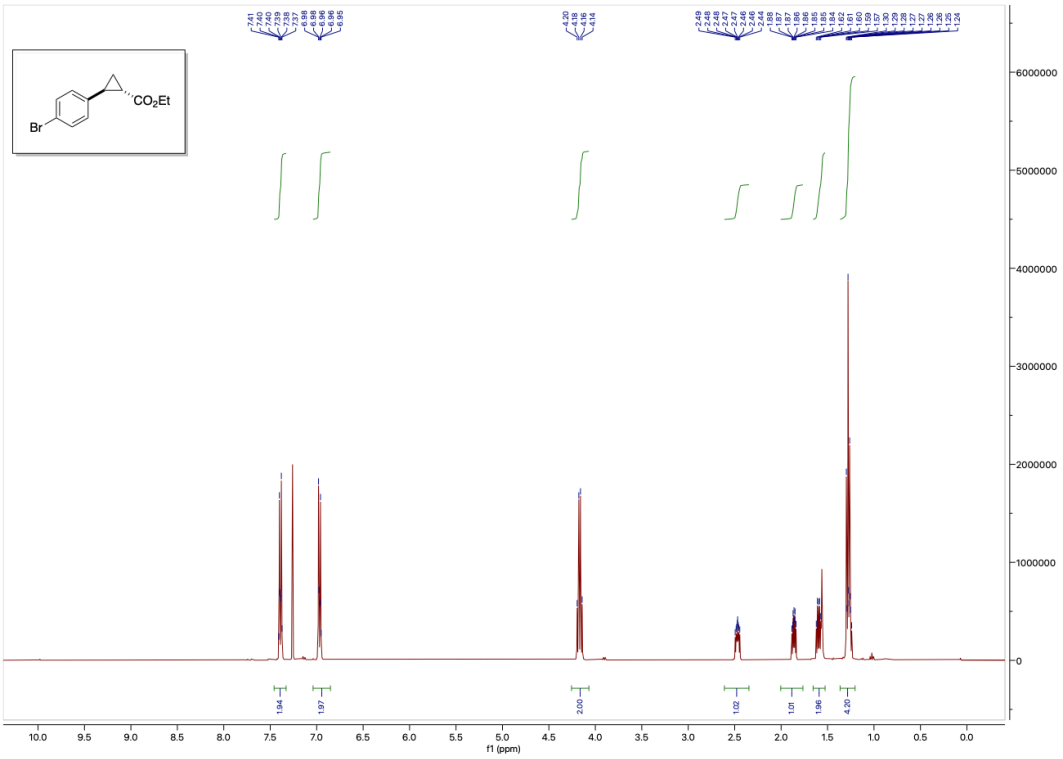
^{1}H NMR spectrum of racemic *trans*-**2e**.



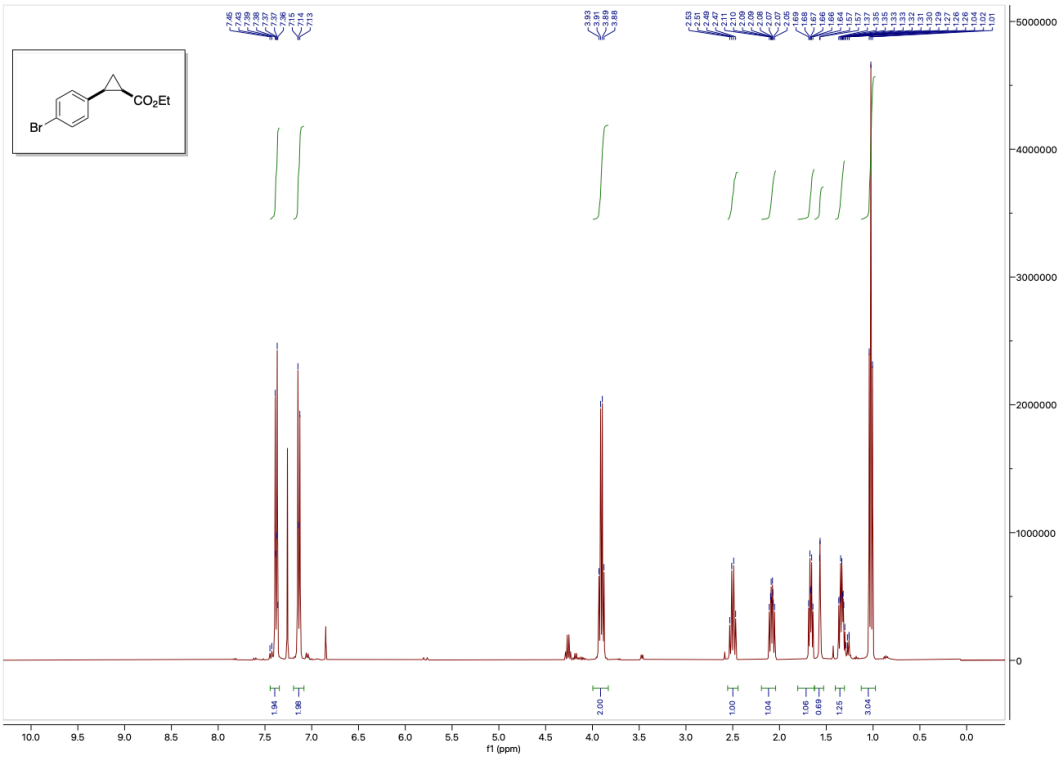
^{1}H NMR spectrum of racemic *cis*-**2e**. A trace impurity of diethyl fumarate can be observed.



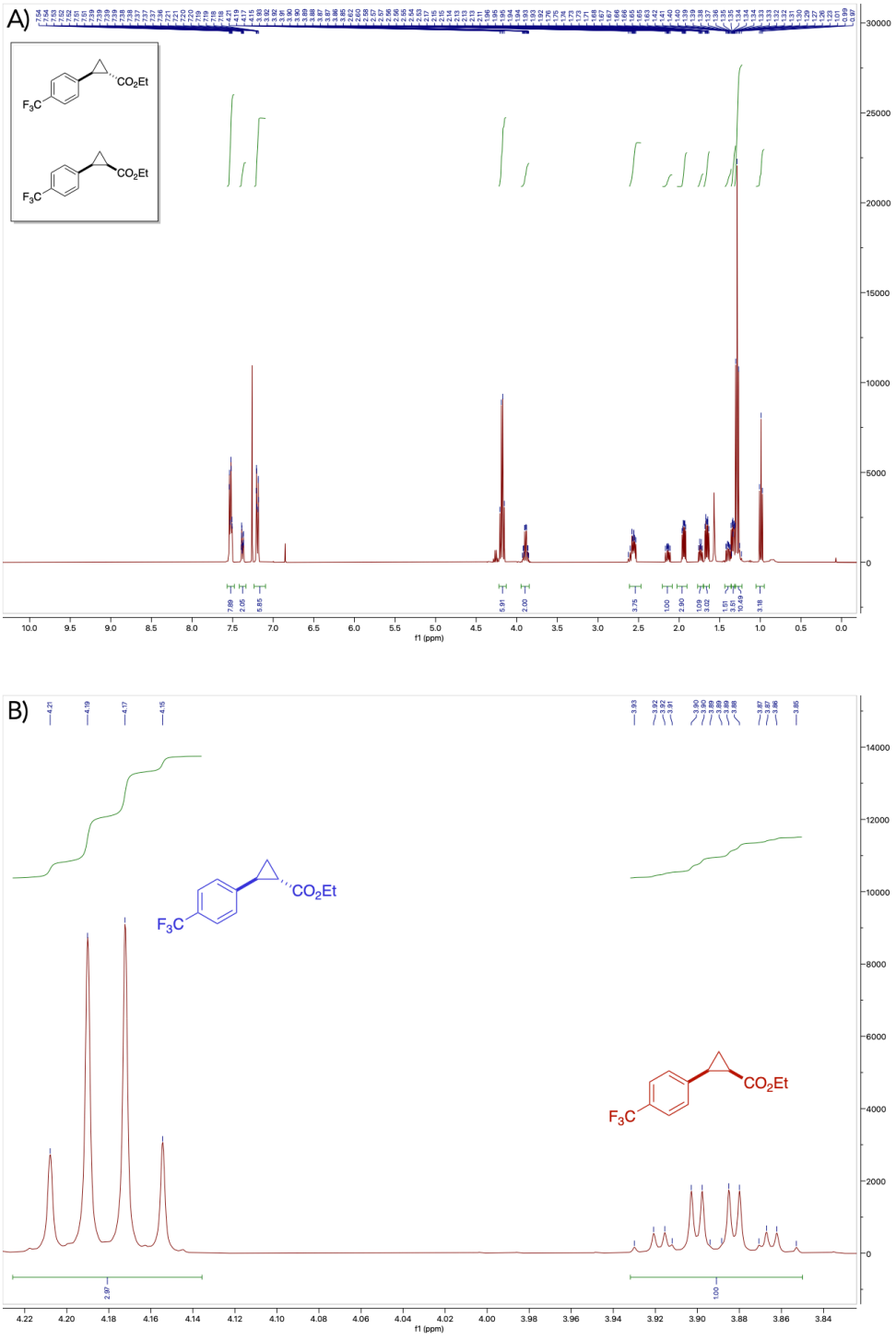
(A) ^{1}H NMR spectrum of a 2:1 racemic *trans*-**2f**/*cis*-**2f**. A trace impurity of diethyl fumarate can be observed. **(B)** Integrations of product ester methylene protons for *trans*-**2f** and *cis*-**2f** to determine the isolated molar ratio.



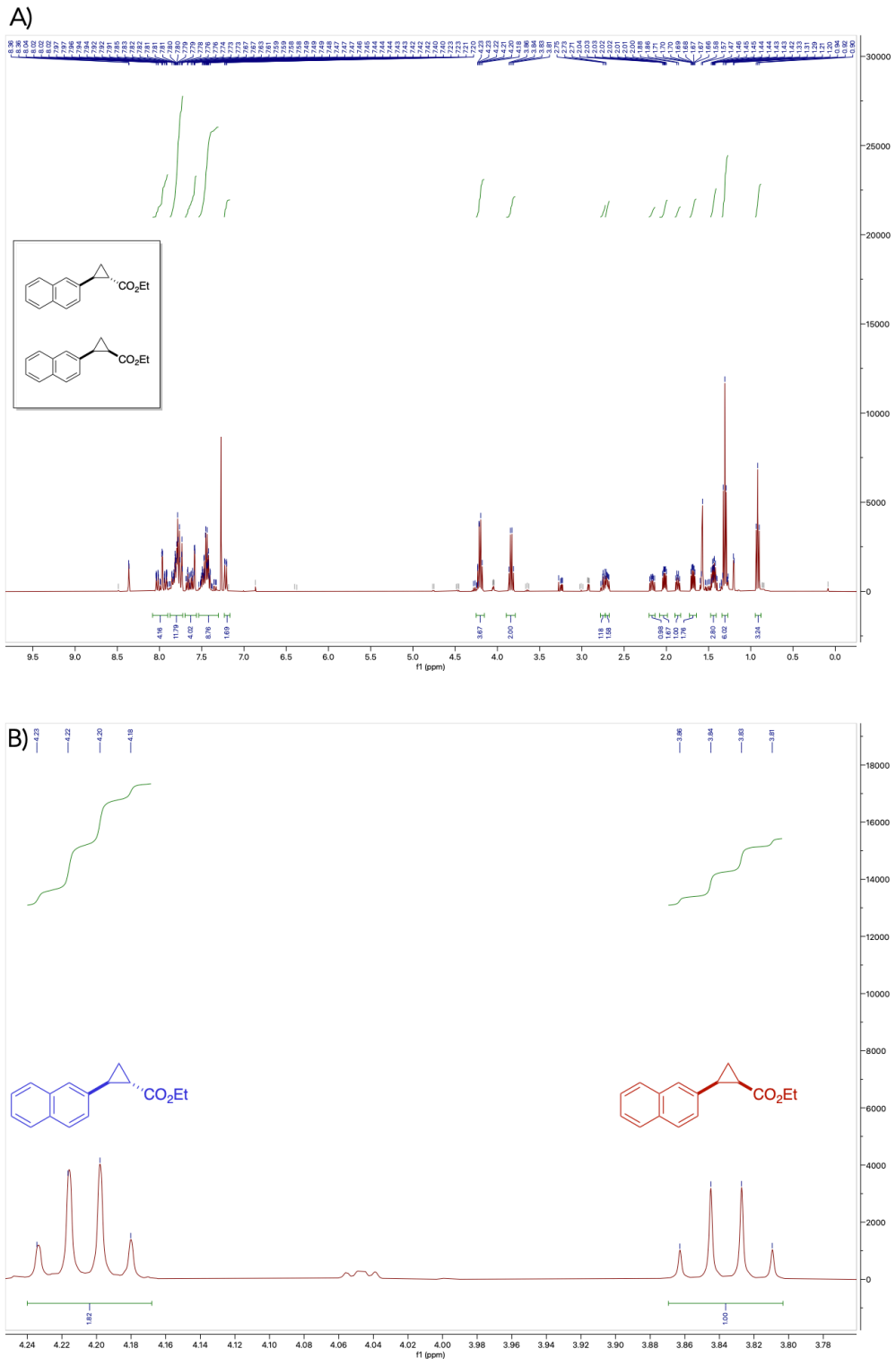
^{1}H NMR spectrum of racemic *trans*-**2g**.



^{1}H NMR spectrum of racemic *cis*-**2g**. A trace impurity of diethyl fumarate can be observed.



(A) ^{1}H NMR spectrum of a 2.97:1 racemic *trans*-**2h**/*cis*-**2h**. A trace impurity of diethyl fumarate can be observed. **(B)** Integrations of product ester methylene protons for *trans*-**2h** and *cis*-**2h** to determine the isolated molar ratio.



(A) ^{1}H NMR spectrum of a 1.82:1 racemic *trans*-**2i**/*cis*-**2i**. Several trace impurities are observed. **(B)** Integrations of product ester methylene protons for *trans*-**2i** and *cis*-**2i** to determine the isolated molar ratio.

C.11. References

1. Abramson, J. *et al*. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. **630**, 493-500 (2024).
2. Studier, F.W. *et al*. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol*. **189**, 113-130 (1986).
3. Nov, Y. When Second Best Is Good Enough: Another Probabilistic Look at Saturation Mutagenesis. *Appl. Environ. Microbiol.* **78**, 258-262 (2012).
4. Chester, N. *et al*. Dimethyl sulfoxide-mediated primer Tm reduction: a method for analyzing the role of renaturation temperature in the polymerase chain reaction. *Analytical Biochemistry*. **209**, 284-290 (1993).
5. Gibson, D.G. *et al*. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* **6**, 343-345 (2009).
6. Wittmann, B.J. *et al.* evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. *ACS Synth. Biol*. **11**, 1313-1324 (2022).
7. Long, Y. *et al.* LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *Under Review.*
8. Doyle, M.P. *et al*. Efficient Alternative Catalysts and Methods for the Synthesis of Cyclopropanes from Olefins and Diazo Compounds. *Synthesis*. **10**, 787-789 (1981).
9. Li, P. *et al*. Iodine-catalyzed diazo activation to access radical reactivity. *Nat. Commun.* **9**, Article number: 1972 (2018).
10. Morandi, B. *et al*. Iron-Catalyzed Cyclopropanation with Glycine Ethyl Ester Hydrochloride in Water. *Org. Lett.* **14**, 2162-2163 (2012).
11. Liang, Y. *et al*. Using Soluble Polymers to Enforce Catalyst-Phase-Selective Solubility and as Antileaching Agents to Facilitate Homogeneous Catalysis. *Angew. Chem. Int. Ed.* **53**, 8084-8087 (2014).
12. Huang, Y. *et al*. Diastereoselective and Enantioselective Cyclopropanation of Alkenes Catalyzed by Cobalt Porphyrins. *J. Org. Chem.* **68**, 8179-8184 (2003).
13. Rosenberg, M.L. *et al*. Highly cis-selective cyclopropanations with ethyl diazoacetate using a novel Rh(I) catalyst with a chelating N-heterocyclic iminicarbene ligand. *Org. Lett.* **11**, 547-550 (2009).
14. Hao, J. *et al*. Enantioselective Olefin Cyclopropanation with G-Quadruplex DNA-Based Biocatalysts. *ACS Catal.* **10**, 6561-6567 (2020).
15. Chen, Y. *et al*. Vitamin B_{12} Derivatives as Natural Asymmetric Catalysts: Enantioselective Cyclopropanation of Alkenes. *J. Org. Chem*. **69,** 2431-2435 (2004).

Chapter V

MACHINE LEARNING-ASSISTED EVOLUTION OF BROADLY FUNCTIONAL ENZYME LIBRARIES

Lal, R.G.; Yang, J.; Zhang, Z.; Arnold, F.H.* Machine Learning-Assisted Evolution of Broadly Functional Enzyme Libraries. *ACS Central Science*. In review.

R.G.L participated in the conceptualization of the project, as well as execution of the research, including enzymatic reactions, substrate scope studies, and synthetic applications. R.G.L participated in writing and reviewing the manuscript.

ABSTRACT

Biocatalysis offers sustainable solutions to pressing challenges in chemical synthesis by exploiting the remarkable efficiency and selectivity of enzymes. Importantly, enzymes are able to accommodate non-native substrates and mediate transformations outside of their natural repertoire. Enzymes can be repurposed for diverse applications by harnessing these "promiscuous" activities and optimizing them using directed evolution (DE). The success of a DE campaign, however, depends on the availability of a protein starting point that displays detectable levels of the desired function. The identification of such starting points can be a challenge, as non-natural protein function is difficult to predict from sequence alone. Researchers thus typically rely on intuition and prior engineering experience to compile libraries of proteins that they then screen for novel activities. Here, we directly generated libraries of promiscuous heme enzymes which can access an expanded scope of chemical reactions through active-site diversification of a single 'parent' sequence. We applied active learning-assisted directed evolution (ALDE) models, trained on sequence-function data for variants of a protoglobin protein tested against up to 26 different carbene and nitrene transfer reactions, to rank multi-mutation variants predicted to exhibit superior performance across multiple reactions. We observed improvements in activity and selectivity for every reaction performed by the parent enzyme in at least one member of the predicted libraries. Moreover, variants from these libraries could catalyze five out of 10 reactions not catalyzed by the parent protoglobin. These results indicate that ALDE can efficiently guide the construction of high-value libraries with expanded catalytic scope, helping to alleviate a central bottleneck in biocatalyst discovery.

5.1. Introduction

Biocatalysis harnesses enzymes, Nature's catalysts, to perform chemical transformations with remarkable efficiency and selectivity, offering sustainable solutions to challenges in medicine, agriculture, and industry.[1] A key feature underpinning this utility is the capacity of enzymes to accommodate non-native substrates (substrate promiscuity) or to catalyze reactions via non-native mechanisms (catalytic promiscuity).[2,3] Researchers have leveraged enzyme promiscuity to develop valuable biocatalytic transformations. However, promiscuous activities are rarely optimal and thus require improvement using directed evolution (DE), in which sequential rounds of diversification and screening are used to enhance the desired activity. Such approaches have delivered biocatalysts with broadened substrate scopes[4,5] and even entirely new-to-nature reaction modes[6-10] with high yields and selectivities.

Directed evolution of a novel enzymatic activity starts with a protein sequence that displays at least trace levels of the desired function such that a round of mutation and screening can reliably generate improvements. While significant advances have been made in optimization strategies and laboratory practices to accelerate and improve DE outcomes,[11] the discovery of starting points displaying a target function remains a central bottleneck, as it is challenging to predict promiscuous functions from sequence alone.[12,13] This challenge is especially pronounced for new-to-nature biocatalytic transformations, where protein fitness data, which could be used to infer promiscuous functions, are scarce and rarely describe activity for more than one substrate.[14] Given these challenges, the identification of new activities commonly requires extensive experimental testing of diverse protein libraries. Such protein sets are typically compiled based on biochemical intuition and prior protein engineering experience, with the aim of maximizing the likelihood of identifying a protein capable of catalyzing the desired reaction (**Figure 5-1A**).[15]

Recently, there have been efforts to design and engineer libraries of enzymes which display broad promiscuity, with the goal of improving the probability of discovering novel enzymatic functions upon screening.[14,16-21] One promising strategy has been to generate libraries of enzyme variants which contain diverse active-site mutations but still retain their

functionality. The driving hypothesis is that diverse active sites within an enzyme scaffold already capable of enacting a desired reaction mode could display a sort of *ensemble promiscuity* by offering variants which span a greater breadth of the protein fitness landscapes.[22] However, the efficient development of such libraries is nontrivial as simultaneous incorporation of multiple mutations can greatly destabilize a protein.[23,24] Furthermore, epistasis, the phenomenon whereby the effects of a mutation are dependent on the context in which it is introduced, makes it challenging to predict how several mutations will behave in conjunction.[25] The computational package FuncLib developed by Fleishman and coworkers has been one effective answer to these challenges.[26] Broadly, this tool predicts functional enzyme variants containing mutations at user-defined active-site residues by (1) assessing the stability of all possible variants using physics-based models and (2) designing combinations of interacting residues that will be compatible with function based on these same models and evolutionary information inferred from existing sequence diversity. Although effective,[27,28] this method performs zero-shot design without the ability to learn from real-world assay labels or specific substrate/promiscuity goals.

Emerging machine learning (ML) methods have the potential to enable the generation or optimization of more promiscuous enzymes.[29,30] For example, machine learning-assisted directed evolution (MLDE) methods have arisen as a technique for predicting combinations of beneficial mutations, including sites which leverage positive epistatic interactions.[31-36] Recently, we described active learning-assisted directed evolution (ALDE), a framework in which an ML model is updated in an iterative manner after collecting new data, and applied it to optimize yield for non-native enzymatic carbene transfer activity in protoglobins, a class of microbial globin.[37,38] While previous MLDE efforts have been aimed toward the navigation of sequence-function landscapes for a single catalytic task, we reasoned that sequence optimization across several reaction types would enable the discovery of enzyme variants which exist at overlaps of these landscapes, which we propose are ‘zones of promiscuity’ (**Figure 5-1B**). Our hypothesis in this work was that such variants would have a higher probability of displaying activity for related reactions and substrates in future initial activity screens.

Here, we extend ALDE to multi-objective optimization and propose ‘multi-reaction Active Learning-assisted Directed Evolution’ (mr-ALDE), in which a machine learning model is trained on functional data spanning multiple reactions assayed using a library of variants harboring active-site mutations. This model is then prompted to predict multi-mutation variants that will be generally more functional toward all reactions studied. We reasoned that this framework could furnish enzyme collections well suited for initial screening campaigns seeking activities related to the training reactions (**Figure 5-1B**). To test this strategy, we used mr-ALDE to develop protoglobin variants broadly functional toward carbene and nitrene transfer chemistries. Using an ML-model trained on function data for over 500 multi-mutation variants originating from a single ‘parent’ protoglobin, we predicted and generated 127 protoglobins with mutations in the active site. These variants, when considered as an ensemble, were capable of performing nearly every tested reaction in higher yield than the starting sequence and could even access reactions which the initial sequence could not deliver with detectable yield.

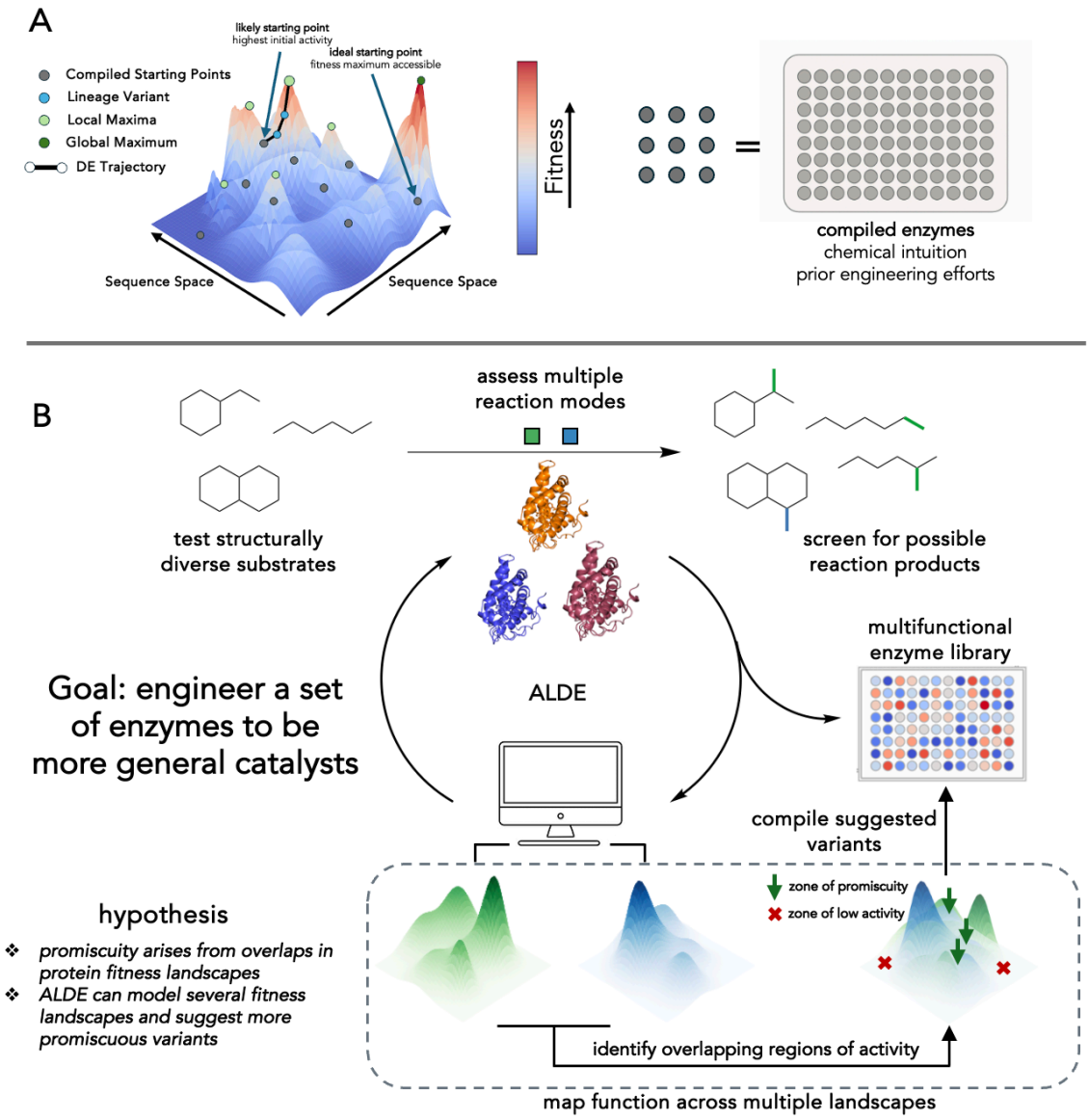


Figure 5-1. Promiscuity in directed evolution. (A) Directed evolution is a greedy uphill walk in a protein sequence-function landscape. Starting points for DE are selected from compiled libraries of enzymes. The starting point influences the outcomes of DE. **(B)** Proposed framework for using ALDE to optimize a family of enzymes for multiple activities. An ALDE model is trained using data from active-site mutants of a parent enzyme assayed for multiple reactions. The model then suggests variants which it predicts have improved activity across reactions seen in training. The predicted enzymes are compiled and assessed for their ability to catalyze reactions both in and out of the training set.

5.2. Study Design

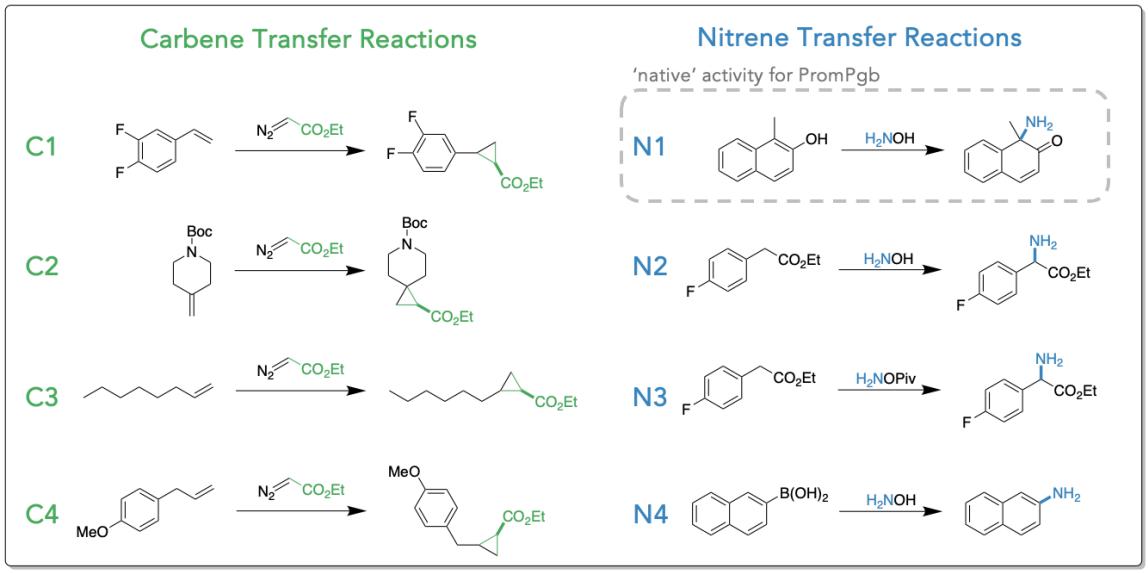
The goal of this study was to develop an ALDE-based workflow for constructing enzyme libraries exhibiting broad activity across new-to-nature enzymatic carbene and nitrene transfer reactions.[39] We proposed an approach which uses sequence–function data from multi-mutation variants to train predictive models that guide the selection of broadly functional enzymes. Our implementation of this strategy began with the selection of a parent enzyme and proceeded through three stages: 1) Training Data Collection, 2) Round 1 Prediction Assessment, and 3) Round 2 Prediction Assessment, each separated by *in silico* model training and proposal of new variants to test. We set out to apply ALDE-based enzyme diversification to simultaneously access variants with improved activity toward an expanded scope of both carbene and nitrene transfer reaction modes. This engineering goal would enable us to investigate mr-ALDE's ability to deliver enzymes with both expanded substrate range and mechanistic scope. To initiate the study, we first determined a 'parent' enzyme which could (1) access a broad scope of chemical reactivities and (2) stand up to a mutational load of up to eight amino acid substitutions. Protoglobins originating from Archaea have been engineered for a variety of carbene and nitrene transfer activities in previous studies[40-45] and are highly thermostable.[24,38,46] Furthermore, we have shown that active-site residues in protoglobins exhibit epistasis toward non-native activities and that ALDE can effectively predict synergistic mutation combinations for specific catalytic tasks.[37] Because protoglobins have no known native catalytic function, we believed that zero-shot mutation effect prediction methods for designing variants of these proteins would not be effective for accessing catalytically functional enzymes, particularly for activities such as carbene and nitrene transfer, for which there is no fitness information encoded in protein sequences produced by natural evolution.

To identify a parent protoglobin sequence, we screened a compilation of 84 wild-type and engineered protoglobin analogs, interrogating their ability to catalyze a variety of new-to-nature reactions which had previously been demonstrated with hemoproteins.[40,42-45,47,48] This set of Initial Training Reactions (ITRs) comprised four carbene-mediated cyclopropanation reactions and four nitrene-based reactions spanning C–H bond functionalization, aromatic desymmetrization, and boronic acid amination mechanisms

(**Scheme 5-1**). These reactions were selected to challenge the active site along two axes: the cyclopropanation set tests whether a common carbene-transfer mechanism can be applied to alkene substrates of varying steric and electronic character, while the interrogation of both carbene and nitrene transfer reactions examines the enzyme's catalytic promiscuity toward distinct mechanistic manifolds.

Through screening the ITRs with the precompiled panel of protoglobins, we identified a variant of *Aeropyrum pernix* protoglobin (*Ape*Pgb) which was evolved for dearomatization chemistry (**N1 in Scheme 5-1**).[45] This variant, designated PromPgb (Promiscuous Protoglobin), demonstrated measurable activity for all ITRs except **N3** (C–H insertion with a hydroxylamine-derived nitrenoid), with product yields varying across the different transformations. Given PromPgb's broad range of activities, we selected it as the parent enzyme for application of mr-ALDE.

Scheme 5-1. Initial training reactions (ITRs) used to identify a promiscuous protoglobin and first train the ALDE model.



Before ALDE is initiated, the design space—defined by a selected set of k residues—is specified, corresponding to 20^k potential sequence variants. For this study, we identified eight positions, all of which are facing toward the distal side of the heme cofactor and are

known to impact non-native function based on previous protoglobin engineering efforts: L59, W62, F63, F73, L86, S90, F93, and S149 (denoted LWFFLSFS, **Figure 5-2A**).[43,44] The combinatorial space of all variants containing mutations at these positions is > 2.5 x 10^{10}. Although a library of variants containing mutations at all design sites would better capture the sequence–function landscapes we sought to model, random sampling of this space was expected to produce mostly non-functional enzymes. Thus, for the training library we tested double- and triple-site variants containing mutations at subsets of these positions (**Figure 5-2B**) in order to reduce the overall mutational load of the initial training set. Because the majority of functional epistatic outcomes in proteins can be attributed to pairwise interactions,[49,50] we reasoned that randomly selected protoglobin variants with 2–3 mutations would be less likely to be inactive toward the ITRs while still reporting on key relationships between active-site residues. Twelve double-site libraries and six triple-site libraries were designed based on structural proximity (assessed through homology modeling) and ease of construction (**Table D-8** of **Appendix D**). These libraries covered 22 of the 28 possible combinations of the selected eight residues. Libraries were constructed using typical multi-site saturation mutagenesis techniques (**Cloning Protocols and Results** of **Appendix D**).[51] Twenty-two unique variants from each double-site library and 44 from each triple-site library, characterized by nanopore-based multiplexed sequencing (LevSeq),[52] were arrayed, affording 528 sequences (1% of the possible search space) for initial reaction screening (**Table D-9** of **Appendix D**).

This collection of 528 PromPgb variants was evaluated against the eight ITRs, yielding a comprehensive dataset of >4,000 sequence-function pairs (Training Data Collection, **Figure 5-3B**). The parent-normalized fitness data for the ITRs were grouped into two objectives, carbene and nitrene reactions, for ALDE model training. Essentially, for each training library variant, the average change in activity from PromPgb for the carbene transfer ITRs (**C1**-**C4**) and the nitrene transfer ITRs (**N1**-**N4**) were separately computed as fitness objectives (**Machine Learning Details** of **Appendix D**). We then trained an ensemble of multi-task supervised ML models to predict both fitness objectives from encodings of sequence. Afterward, the expected hypervolume improvement acquisition function was applied to the trained model to rank candidate sequences in the design space from most to least likely to improve both fitness objectives, balancing *exploration* of new

areas of sequence space with *exploitation* of variants that are predicted to have high fitness. We focused on the mean improvement to carbene and nitrene activity as the two objectives to optimize for two reasons: (1) to increase regularization during training of the multi-task ensemble of surrogate models and (2) to reduce the computational cost of calculating the expected hypervolume improvement acquisition function.[53] The proposed variants (63 distinct sequences; Round 1 Predictions) were tested in the wet lab with 14 new reactions alongside the ITRs. The function data for these 22 reactions (13 carbene transfer reactions and nine nitrene transfer reactions) were then used to update the learning model for a second round to obtain further improved variants (**Figure 5-3A**). This final round of variants (64 distinct sequences; Round 2 Predictions) was then interrogated against a panel of the previous 22 reactions and four additional reactions (**Scheme 5-2A**). Across these two rounds of model-guided experimentation, we evaluated the catalytic performance of predicted protoglobin variants and used these data to assess mr-ALDE's ability to identify libraries of functional enzymes which can access a broader range of activities than the parent.

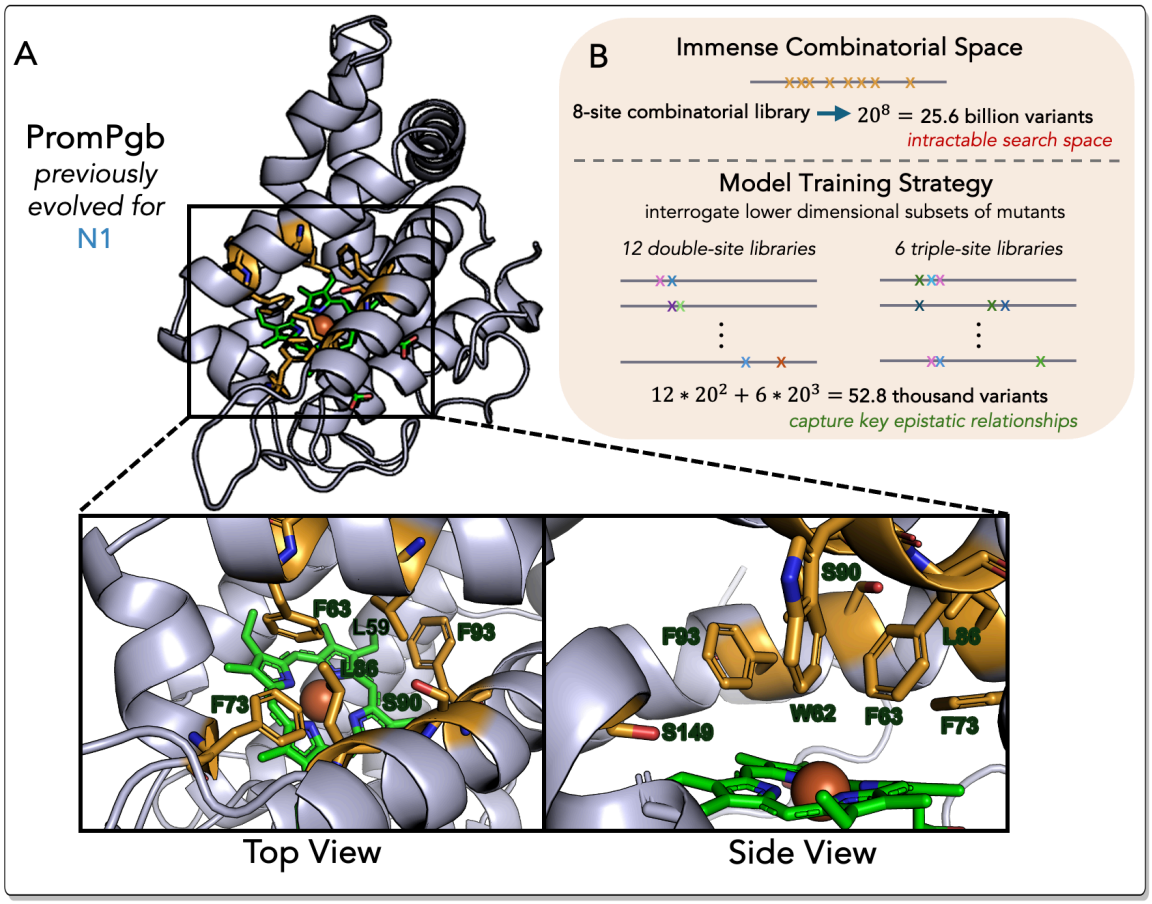


Figure 5-2. (A) The active site of PromPgb has several interacting residues above the distal face of the heme cofactor. Eight residues were selected for the design space in this study: L59, W62, F63, F73, L86, S90, F93, and S149. **(B)** To reduce the deleterious effects of increasing numbers of mutations, the initial round of sequences was built as a combined library of selected double- and triple-site combinatorial libraries.

5.3. Results

5.3.1. Model Training is Informed by an Extensive Functional Landscape

Initial model training was based on experimental data obtained by screening a library of 264 double-site and 264 triple-site variants of PromPgb against the eight ITRs. Among the 8 x 19 = 152 possible single amino-acid substitutions at the eight design-space positions, 151 were present in the constructed library at least once (**Figure D-21** of **Appendix D**). Screening revealed that mutations within the training library substantially altered catalytic

performance across all eight ITRs (**Figure 5-3B**). Additionally, we found that the identities of residues in the active site could drastically change the stereochemical outcomes of the cyclopropanation reactions **C1**, **C3**, and **C4** (**Figures D-24**, **D-28**, and **D-31** of **Appendix D**). Excitingly, nearly 10% of variants in the training library were capable of catalyzing C–H amination using hydroxylamine as a nitrene precursor (reaction **N2**), the only ITR not catalyzed by PromPgb. While members of the training library already displayed broadened specificity, applying ALDE should allow us to distill this information into a smaller, more diverse set of enzyme designs requiring fewer resources to screen. Overall, we were encouraged to see that combinations of mutations at these positions could positively impact activities toward all eight of the ITRs.

Before proceeding with model training, we examined how changes in activity correlated across reactions for each tested variant. A Pearson correlation analysis (**Figure D-68** of **Appendix D**) revealed that the seven ITRs accessible to PromPgb were generally, albeit weakly, positively correlated (average $r = 0.28$). This suggests that while mutations influencing enzyme performance often affect multiple reactions similarly, the underlying structural and mechanistic determinants remain distinct. These correlations likely also capture global effects of mutations on enzyme expression and stability, suggesting that part of the shared activity behavior arises from changes in overall protein abundance or folding. Such information remains valuable for model training, as these are desirable traits that we aim to retain and propagate in model-suggested variants. Reaction **N2**—the only ITR that PromPgb could not catalyze with detectable yield—displayed the weakest correlations overall, including below-average correlations with **N1** and **N3**, which share its nitrene precursor and organic co-substrate, respectively. These trends highlight the mechanistic complexity of introducing new chemistries into enzymes and emphasize the importance of sampling diverse active-site configurations to uncover novel functions.

5.3.2. ALDE Generates Diverse Libraries of Functional Protoglobins

We performed two rounds of mr-ALDE, which included model training, experimental validation, and model updating, to explore the defined design space (**Figure 5-3A**). In each round, an oligo pool-based cloning strategy was applied to achieve exact model-predicted

multi-mutants of PromPgb. Details regarding DNA sequence design are described in the supplementary materials (**Cloning Protocols and Results** of **Appendix D**). In the first round, 63 of the 96 top-ranked variants were successfully cloned and characterized by LevSeq sequencing.[52] These sequences had a mean Hamming distance of 6.49 amino acids from PromPgb and a mean pairwise distance of 4.19 amino acids. In the second round, informed by updated activity data from Round 1, 64 of the 96 top-ranked variants were obtained and characterized, with a mean Hamming distance of 4.89 amino acids from PromPgb and a mean pairwise distance of 3.30 amino acids. While sequence diversity varied between the two rounds of prediction, the mean number of mutations per variant in both libraries remained higher than the maximum of three mutations present in the training set.

To understand the performance of the modeling strategy, we first evaluated how well ALDE-predicted protoglobin variants retained their function by comparing their activities on ITRs with those of the initial training variants. Here we only consider the seven ITRs for which PromPgb demonstrates detectable yield (reactions **C1–C4**, **N1**, **N3**, and **N4**). Examination of these seven reactions showed that, in general, predicted variants demonstrated improved average activity toward carbene and nitrene chemistries when compared to the parent PromPgb and members of the training library (**Figure 5-3C**). Interestingly, predictions made in Round 1 appear to be more competent toward nitrene transfer reactions while those in Round 2 see greater improvements in carbene transfer activity. Across both rounds, ALDE-predicted variants more frequently retained basal activity (≥5% of parent yield) toward the parent-accessible ITRs than did the random multi-mutants in the training library, with median accessibility of six and five reactions in the first and second rounds, respectively, versus four in the training library (**Figure 5-3D**). Even more excitingly, members of the designed libraries were far more likely to demonstrate activities higher than parent for a greater number of reactions than training variants (**Figure 5-3E**). While only 33% of training library members were capable of performing any ITR with greater yield than PromPgb, nearly 70% of protoglobin variants in both predicted libraries were capable of performing at least one ITR with greater yield than PromPgb. This enrichment of protoglobin variants which have improved retention of function and even heightened activities demonstrates that MLDE is effective at

synthesizing information from a multi-reaction dataset to predict diverse combinations of active-site mutations.

PromPgb was previously engineered for reaction **N1,** and this activity can be considered as the "native" function of the protein. We found no predicted sequences which could catalyze this dearomatization transformation as well as PromPgb. The best mutant for reaction **N1** (VAVFIAFN; **Table 5-1**, Entry 6) could only catalyze this reaction with 11% yield, corresponding to 20% of the activity of PromPgb, indicating a tradeoff between the improvement of a promiscuous activity and the "native" function.[54,55] In the training data for reaction **N1**, we only observe a single multi-mutant capable of catalyzing this reaction with higher yield than PromPgb. As a result, the model has not been trained with sufficient data to predict mutations which will improve this activity. Nevertheless, the identification of variants retaining activity toward **N1** while gaining activity in other reactions underscores the potential to balance tradeoffs through strategic library design to achieve a compilation of enzymes with generally heightened catalytic competency.

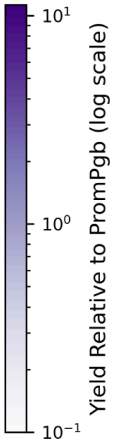
Consideration of the top-performing ALDE design for each of the eight ITRs reveals varying degrees of specialization. The variants which demonstrated the highest activities for carbene transfer ITRs (reactions **C1**–**C4**) tended to maintain their activities for other carbene transfer reactions, while taking major losses in activity for all nitrene transfer ITRs (**N1**–**N4**) besides **N4 (Table 5-1**, Entries 2–5). In contrast, multi-mutants which displayed the highest yields toward nitrene transfer chemistries tended to maintain carbene transfer activities to some degree **(Table 5-1**, Rows 6-9). Notably, the two designs which catalyzed the greatest number of ITRs with higher yields than PromPgb (VAFFMAFQ and TNVFITFQ; **Table 5-1**, Entries 10 and 11) were not the highest performers for any single ITR. This suggests that within this design space there exist combinations of residues which make PromPgb a more competent general catalyst toward multiple substrate and reaction types.

Together, these findings underscore the power of ALDE-based library design to uncover both specialized and broadly functional catalysts within a single protein scaffold. By sampling mutational combinations that enhance reactivity across mechanistically distinct reactions, this approach identifies variants that not only expand the accessible catalytic

repertoire but also reveal generalist sequences from which new activities can be evolved. Achieving comparable functional breadth through conventional directed evolution would demand several separate campaigns, each targeting a distinct reaction, whereas this single design cycle furnished 127 catalytically active and mechanistically diverse variants.

Table 5-1. Heatmap of fold-improvements for ITRs for select variants from predicted protoglobin libraries. Cell coloring is shown as the log-fold change in yield relative to PromPgb. Absolute fold-changes in reaction yield are given in each cell. Rows 2–9 represent the top-yielding variant for reactions **C1**–**C4** and **N1**–**N4,** respectively (top reaction boxed). Rows 10 and 11 show variants show the variants for which the greatest number of activities were improved relative to PromPgb by the greatest degree from each round of predictions. For variants displaying activity for reaction **N2**, a check mark indicates the formation of the C–H amination product. Shading in cells for reaction **N2** is calculated based on a variant's yield for this reaction normalized to the yield of PromPgb for reaction **N3**, which shares the same product.

Entry	Mutant	Note	C1	C2	C3	C4	N1	N2	N3	N4
1	LWFFLSFS	Parent	1.0	1.0	1.0	1.0	1.0	×	1.0	1.0
2	IAFFMVFN	Top C1 Round 2	1.7	1.2	0.3	0.8	0.1	×	×	0.2
3	LNFFMVFQ	Top C2 Round 2	1.2	1.4	0.7	1.4	0.1	×	0.6	3.5
4	VYFFMAFQ	Top C3 Round 2	1.0	1.2	2.7	1.2	0.1	×	0.1	3.8
5	VAFFIAFN	Top C4 Round 2	1.2	0.8	1.4	4.7	0.1	×	0.1	0.5
6	VAVFIAFN	Top N1 Round 2	0.8	0.7	2.2	3.7	0.3	✓	0.4	3.0
7	LYVFIVFQ	Top N2 Round 1	1.3	0.6	1.5	0.9	0.2	✓	0.9	5.1
8	TAVFMCVQ	Top N3 Round 1	0.8	0.2	0.2	0.3	0.1	✓	8.5	5.3
9	LGFFMAVN	Top N4 Round 2	0.8	0.4	0.9	4.4	0.1	×	0.1	11.3
10	VAFFMAFQ	Most Improved Reactions	1.2	1.3	1.4	3.1	0.1	×	0.1	1.7
11	TNVFITFQ	Most Improved Reactions	1.5	0.6	1.6	1.1	0.1	✓	4.0	2.5



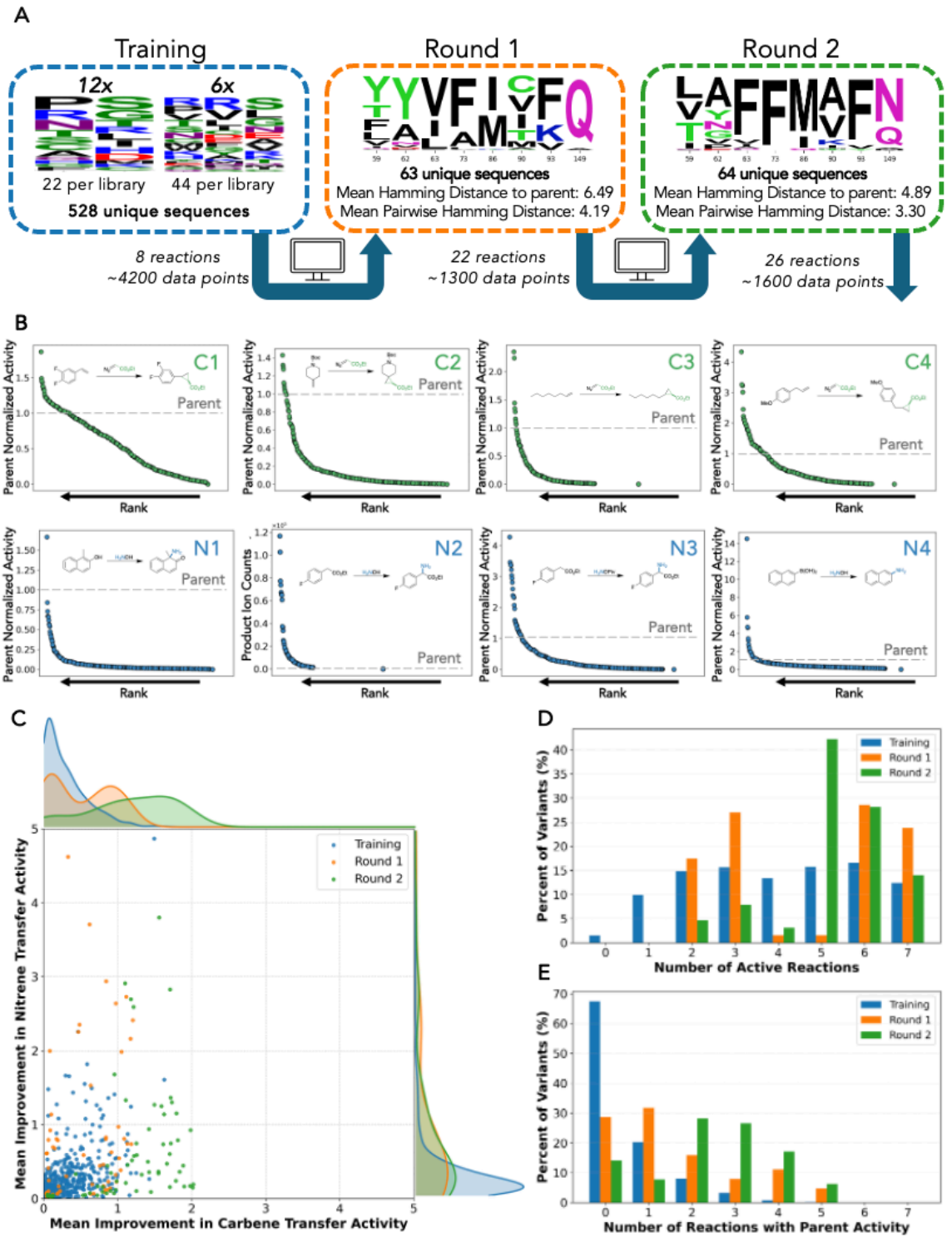


Figure 5-3. (A) After training data were collected, a panel of PromPgb variants harboring mutations at the eight positions were proposed through the ALDE algorithm. These multi-mutants (Round 1) were screened on 22 carbene and nitrene transfer reactions for further model training. After a second round (Round 2) of mutants was predicted, a total of 26 reactions was assessed. **(B)** Initial round of activity data collected for eight chemical

reactions with random 2–3 site mutants of PromPgb. Reactions C1–C4 were assessed by gas chromatography with detection with a flame ionization detector (GC-FID). Reactions N1–N4 were assessed by liquid chromatography with mass-spectrometric detection (LC-MS). For reactions C1, C3, and C4 activities are taken as the sum of the yields of the two possible cyclopropane diastereomers. **(C)** The predicted variants display yields that generally represent higher values of the objective functions used to train the model. The evaluated objective functions shown here are computed only with data for reactions seen only by the original training set. Improvement is normalized across all reactions, with 1 referring to parent activity. **(D)** Counts of the total number of ITRs which each training-library or model-predicted design was capable of facilitating with detectable yield. Reaction **N2** is not considered in these counts as PromPgb does not perform this reaction. **(E)** Counts of the number of ITRs which each training-library or model-predicted design catalyzed with greater yield than the parent sequence, PromPgb.

5.3.3. Model-Predicted Variants Demonstrate Broadened Substrate Range as an Ensemble

Having established that the mr-ALDE framework generates protoglobin variants with enhanced activity toward catalytic functions accessible to PromPgb, we next examined how this functional diversity translates to substrate range. To this end, we evaluated the ensemble of model-predicted variants across an expanded substrate panel representing electronically and sterically diverse carbene and nitrene acceptors (**Scheme 5-2A**). This set of reactions comprised 18 distinct transformations not seen in initial model training. Broadly, the reactions interrogated can be classified into two primary mechanistic categories: additions to π-systems and insertions into X–H bonds. Four additional transformations (reactions **N1**, **N4**, **C10**, and **C11**) likely proceed through distinct mechanisms that fall outside of these two classes. The collected data for all transformations are provided in the Supporting Information and are available in full on GitHub.

For every transformation that PromPgb could catalyze with detectable product formation, we identified at least one model-predicted variant that exhibited enhanced activity relative to the parent. Notably, these improvements extended even to reactions that are mechanistically distinct from those represented among the ITRs, indicating that the sequence features enriched during mr-ALDE can generalize to new reactivity modes. For example, variant FYIFMMFQ catalyzes the intramolecular C(sp^3)–H amination of alkyl azide **1** to afford pyrrolidine **2** in 3% yield (**Reaction N6** in **Scheme 5-2B**), whereas PromPgb furnishes only trace product under identical conditions. Strikingly, this variant

emerged in the first round of ALDE, trained solely on ITR data, despite the fact that reaction **N6** requires both azide activation and an intramolecular cyclization—mechanistic steps not present in the ITR panel. We also observe cases that highlight the value of the iterative model-updating component of ALDE. For reactions **C4** and **C8**, second-round predictions yielded improvements in a larger fraction of variants compared to the first round (**Figures D-30** and **D-40** of **Appendix D**). In the case of reaction **C13** (intermolecular carbene C–H insertion of pyrrolidine **3**), the second round yielded variant VVFFMAFN, which delivers product **4** in 6% yield, representing a six-fold increase relative to PromPgb (**Reaction C13** in **Scheme 5-2B**). To achieve the observed levels of activity for reactions **N6** and **C13** through conventional directed evolution of PromPgb would likely require multiple rounds of mutagenesis and screening. Thus, although the absolute yields for these reactions remain modest, the top-performing variants identified here constitute improved starting points for subsequent directed evolution efforts.

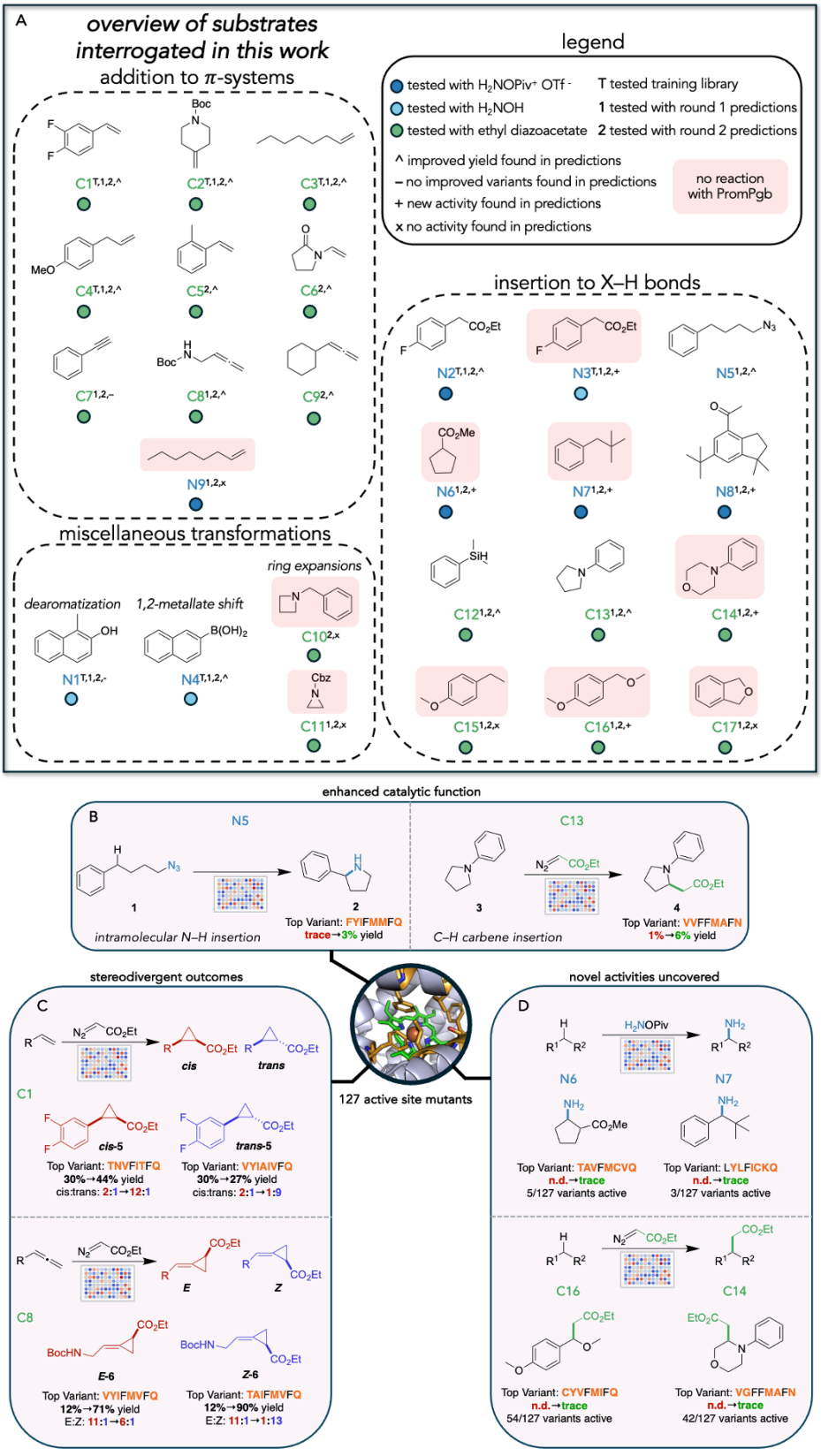
None of the data used in either round of ALDE model training encoded the stereochemical outcomes of any training reactions. All reactions with multiple diastereomeric products were recorded as the sum of both stereoisomers. Nevertheless, by training on substrates with varied steric and electronic profiles, ALDE guided the exploration of active-site environments that differ in configuration, ultimately producing variants that favor different stereochemical trajectories of the same reaction. For nearly all of the cyclopropanation reactions tested in this work, the complete ensemble of model-predicted variants is capable of delivering stereodivergent outcomes toward distinct diastereomeric products (**Scheme 5-2C**). For reaction **C1,** representing the formation of a cyclopropane core present in the antithrombotic therapeutic ticagrelor,[56] PromPgb displays little to no selectivity for formation of the possible *cis*-**5** and *trans*-**5** cyclopropane products. Across both predicted libraries we find variants that are capable of selectively forming both possible diastereomers without significant losses in yield. Importantly, the synthesis of ticagrelor requires *trans*-**5**. To our delight, we identified variant VYIAIVFQ, which could deliver this diastereomer with a 9:1 preference for the *trans*- product. This emergence of stereodivergent behavior extended to other substrate classes: for the formation of alkylidene cyclopropanes ***Z*-6** and ***E*-6**, several variants not only display significantly improved yields for the PromPgb-favored *E*-isomer, but we also found that variant

TAIFMVFQ inverts the selectivity toward formation of the *Z*-product (**Reaction C8** in **Scheme 5-2C**). Access to stereodivergent outcomes early in screening is particularly advantageous, as it provides multiple differentiated starting points from which independent DE campaigns can be initiated for desired stereoisomers, as would be the case for the synthesis of the ticagrelor cyclopropane precursor *trans*-**5**.

There were 10 transformations in the full set of tested reactions which PromPgb could not catalyze with detectable yield (**Scheme 5-2A**). We were able to identify variants that catalyzed five of these previously inaccessible transformations. For reactions **N6** and **N7**, which involve C–H nitrene insertion using NH_2OPiv at positions less activated or more hindered than for reaction **N3**, we identified a few variants in the first round of predictions that are capable of catalyzing these reactions in trace yield. The three designs capable of catalyzing reaction **N6** all follow the motif LYLFXXKQ, notably containing Phe to Lys substitutions at position 93, a site which was particularly recalcitrant to mutation for ITRs. This result highlights the importance of the balance of exploiting training data versus the exploration of sequence space in the ALDE algorithm. Additionally, we find that over a third of model-predicted mutants are capable of catalyzing the α-heteroatom C–H carbene insertion reactions **C14** and **C16**. Potentially, one could access reactions **C14** and **C16** through DE using a strategy known as a 'substrate walk' to enhance yield for the existing activity **C13**, which is mechanistically similar to these transformations. Here, however, ALDE directly proposes a diverse set of variants that are broadly competent for carbene transfer, derived solely from information encoded in a predefined reaction panel. As a result, directed evolution can be initiated with ALDE designs on substrates of greater synthetic or mechanistic interest without the need to first optimize an easier reaction (or intermediate substrate).

Scheme 5-2. (A) Twenty-six carbene and nitrene transfer reactions representing electronically and sterically diverse acceptors were evaluated, including additions to π-systems and X–H insertion reactions. Symbols denote co-substrates for each reaction, rounds in which reaction was tested, and outcomes of library screening. **(B)** For nearly every tested reaction which PromPgb could perform, ALDE delivered improved variants. **(C)** Model-predicted variants unlock access to stereodivergent outcomes as an ensemble.

(D) ALDE designs were capable of performing four of the nine assayed reactions which PromPgb could not perform.



5.4. Discussion

In this work, we present mr-ALDE, an implementation of ALDE that can be used to generate enzyme libraries which are capable of accessing a broad scope of chemical transformations. Starting from a thermostable protoglobin scaffold and a training set that captured double- and triple-mutant epistasis across a limited "basis set" of carbene and nitrene transfer reactions, ALDE generated multi-mutation variants that (1) retained basal activity across more training reactions than did multi-mutants seen in initial training, despite a higher mutational load, (2) frequently exceeded parent yields for at least one tested reaction, and (3) in several cases outperformed the parent on mechanistically distinct transformations not represented in initial training. In short, this data-driven approach effectively delivers enzyme variants that maintain functionality toward reactions represented in training while introducing diverse active-site environments that provide distinct solutions to related catalytic challenges. Thus, we propose that mr-ALDE represents an exciting methodology for generating enzyme libraries which are well suited for initial screening campaigns, where the primary challenge is identifying any detectable activity for a target reaction.

While this method is inherently reliant on the construction of large sequence-function datasets, we believe that it presents many advantages in comparison to computational library generation methods which rely on zero-shot prediction or design. Firstly, this method is particularly applicable to the generation of enzyme libraries primed for the discovery of new-to-nature biocatalytic reactions, for which there is presumably little evolutionary information contained in existing protein sequence information. Methods such as FuncLib or the MODIFY software package developed by Ding and coworkers[14] use sequence patterns from proteins found in nature to identify combinations of mutations which maintain stability and other biochemical properties likely important for function. However, new-to-nature chemistries often operate through catalytic principles and transition-state geometries that have not been sampled during natural evolution and therefore may not conform to the sequence–structure relationships encoded in existing protein families. Directed evolution is also not well suited for this optimization task, due its multi-objective nature and the limited scope of single mutations when interrogating a

diverse protein search space. By contrast, a direct MLDE-based approach leverages experimentally measured sequence–function data to identify variants that are empirically optimized for the desired reactivity, enabling the discovery of sequence solutions that lie outside the constraints imposed by evolutionary precedent.

Our findings regarding changes in population behavior in the first and second rounds of ALDE-predicted variants illustrate a practical strength of mr-ALDE: once an initial model is trained on a broad dataset, later rounds can focus on small, model-informed panels. Data collected from future engineering campaigns on members of ALDE-predicted libraries can be used to further update the model, enabling the continuous design of functional, diverse enzyme sequences. Similarly, the data generated through this ALDE framework have the potential to inform future machine learning-assisted directed evolution campaigns initiated from the designs generated in this work. In the course of this work, we assessed the function of 653 protoglobin variants harboring multiple active-site mutations against a panel of up to 26 distinct carbene and nitrene transfer reactions, resulting in over 7,000 unique sequence-function paired datapoints, which we have made publicly available at github.com/fhalab. To our knowledge this is the largest annotated dataset of non-native enzyme function containing multiple substrate and reaction classes. This dataset may provide a useful reference for future MLDE training and validation studies, helping to inform broader efforts in data-driven enzyme engineering and design.

This initial investigation is the first application of MLDE methods to improve enzyme promiscuity using data from several reaction classes. We recognize that our model system was particularly well positioned for success using the present engineering approach. Our efforts were initiated with a thermostable protein which we knew to be capable of promiscuously catalyzing both carbene and nitrene transfer chemistries. Furthermore, our lab has extensive knowledge of engineering *Ape*Pgb variants, and we have previously applied ALDE to the active site of a related protoglobin homolog. As a result, we were able to confidently select active-site residues for the ALDE design space that influence non-native reactivity and exhibit epistatic effects.

Future studies will determine how well mr-ALDE can be extended to other enzyme and reaction classes and determine best practices for choosing training reaction sets, library

designs, and ML-model parameters. Firstly, it is not clear how one should optimally design the original set of training reactions in mr-ALDE. It is possible that including fewer initial training reactions, or reactions spanning a greater chemical scope could lead to improved model training for predicting promiscuous enzyme variants. One could also envision a design framework in which the model is provided with the chemical details for training reactions. Such a model could possibly be used to explicitly suggest variants with activity for a specific substrate by combining existing function data with information on chemical similarity to the desired transformation. Additionally, we are curious to see if mr-ALDE can be applied to broaden the capabilities of an enzyme to accommodate unnatural substrates for its native reaction mechanism, a common task in biocatalysis. While we expect that mr-ALDE can be applied broadly to other enzyme classes, future studies will require the investigation of datasets spanning diverse substrates within a *single* reaction class to generate libraries with expanded substrate promiscuity. Finally, in this study we selected the mean improvements in carbene and nitrene transfer activities as objective functions for multi-objective optimization. However, this encoding is apparently sensitive to variants seen in training which display atypically large improvements in activity for a particular reaction. This is evidenced by the strong representation of the amide-containing amino acids asparagine and glutamine at position 149 in model-predicted designs, likely due to the presence of these mutations in high-performing variants for reaction **N4**. We anticipate that subsequent studies must be conducted to develop ideal objective-function encodings which can avoid bias from outliers seen in model training data and to more effectively deliver diverse enzyme libraries.

Altogether, the discovery of 127 catalytically active variants encompassing both carbene and nitrene transfer reactions illustrates the remarkable efficiency of coupling machine learning with laboratory evolution principles. Where conventional directed evolution would have required several discrete engineering campaigns to achieve similar functional coverage, mr-ALDE accomplishes this in a single, integrated workflow. The ability of mr-ALDE to rapidly generate stable, mechanistically diverse, and functionally enriched enzyme libraries highlights its potential for accelerating the discovery of new biocatalytic activities across an expanding range of chemical transformations.[57]

5.5. Contributions

R.G.L.: conceptualization, methodology, investigation, analysis, writing – original draft, writing – editing

J.Y. : conceptualization, methodology, software, investigation, analysis, writing – original draft, writing – editing

Z.Z.: investigation, writing – original draft, writing – editing

F.H.A.: resources, writing – editing, supervision, funding

5.6. Acknowledgements

This work was supported by the U.S. Army Research Office cooperative agreement for the Institute for Collaborative Biotechnologies (W911NF-19-2-0026 to F.H.A.). This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. This work was also supported by the NSF Division of Chemical, Bioengineering, Environmental and Transport Systems (CBET 1937902). J.Y. and R.G.L. are partially supported by National Science Foundation Graduate Research Fellowships and J.Y. is supported by the Google PhD Fellowship. The authors thank Kathleen M. Sicinski, Jae L. Kennemur, Yueming Long, Edwin Alfonzo, and Deirdre M. Hanley for providing guidance on analytical method development for training reactions. The authors thank Raul Astudillo for providing guidance on the use and implementation of expected hypervolume improvement. The authors thank Francesca Zhoufan-Li for help with data visualization and analysis.

5.7. References

1. Wu, S.; Snajdrova, R.; Moore, J.C.; Baldenius, K.; Bornscheuer, U.T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem. Int. Ed.* **2020**, *60*, 88-119.
2. Kheronsky, O.; Tawfik, D.S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **2010**, *79*, 471-505.
3. Copley, S.D. Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.* **2017**, *47*, 167-175.
4. 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.* **2016**, *55*, 11577-15781.
5. Zetzsche, L.E.; Yazarians, J.A.; Chakrabarty, S.; Hinze, M.E.; Murray, L.A.M.; Lukowski, A.L.; Joyce, L.A.; Narayan, A.R.H. Biocatalytic oxidative cross-coupling reactions for biaryl bond formation. *Nature* **2022**, *603*, 79-85.
6. Raps, F.C.; Hyster, T.K. Emergent Mechanisms in Biocatalysis. *ACS Cent. Sci.* **2025**, *11*, 1029-1040.
7. Zhang, J.G.; Huls, A.J.; Palacios, P.M.; Guo, Y.; Huang, X. Biocatalytic Generation of Trifluoromethyl Radicals by Nonheme Iron Enzymes for Enantioselective Alkene Difunctionalization. *J. Am. Chem. Soc.* **2024**, *146*, 34878-34886.
8. Fu, W.; Murcek, K.; Chen, J.; Liu, A.; Zhao, Y.; Liu, P.; Yang, Y. Catalytic Enantioselective Smiles Rearrangement Enabled by the Directed Evolution of P450 Radical Aryl Migratases. *J. Am. Chem. Soc.* **2025**, *147*, 12197-12205.
9. Trimble, J.S.; Crawshaw, R.; Hardy, F.J.; Levy, C.W.; Brown, M.J.B.; Fuerst, D.E.; Heyes, D.J.; Obexer, R.; Green, A.P. A designed photoenzyme for enantioselective [2+2] cycloadditions. *Nature* **2022**, *611*, 709-714.
10. Raps, F.C.; Rivas-Souchet, A.; Jones, C.M.; Hyster, T.K. Emergence of a distinct mechanism of C–N bond formation in photoenzymes. *Nature* **2025**, *637*, 362-368.
11. Vidal, L.S.; Isalan, M.; Heap, J.T.; Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. *RSC Chem. Biol.* **2023**, *4*, 271-291.
12. Jeffery, C.J. Current successes and remaining challenges in protein function prediction. *Front. Bioinform.* **2023**, *3*, 1222182.
13. Marshall, J.R.; Mangas-Sanchez, J.; Turner, N.J. Expanding the synthetic scope of biocatalysis by enzyme discovery and protein engineering. *Tetrahedron* **2021**, *82*, 131926.
14. Ding, K.; Chin, M.; Zhao, Y.; Huang, W.; Mai, B.K.; Wang, H.; Liu, P.; Yang, Y.; Luo, Y. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nat. Commun.* **2024**, *15*, 6392.
15. Arnold, F.H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed.* **2017**, *57*, 4143-4148.
16. Paton, A.E.; Boiko, D.A.; Perkins, J.C.; Cemalovic, N.I.; Reschützegger, T.; Gomes, G.; Narayan, A.R.H. Connecting chemical and protein sequence space to predict biocatalytic reactions. *Nature* **2025**, *646*, 108-116.

17. Devamani, T.; Rauwerdink, A.M.; Lunzer, M.; Jones, B.J.; Mooney, J.L.; Tan, M.A.O.; Zhang, Z.-J.; Xu, J.-H.; Dean, A.M.; Kazlauskas, R.J. Catalytic Promiscuity of Ancestral Esterases and Hydroxynitrile Lyases. *J. Am. Chem. Soc.* **2016**, *138*, 1046-1056.
18. McDonald, A.D.; Higgins, P.M.; Buller, A.R. Substrate multiplexed protein engineering facilitates promiscuous biocatalytic synthesis. *Nat. Commun.* **2022**, *13*, 5242.
19. Campbell, M.E.; Ohler, A.R.; McGill, M.J.; Buller, A.R. Promiscuity Guided Evolution of Decarboxylative Aldolases for Synthesis of Tertiary γ-Hydroxy Amino Acids. *Angew. Chem. Int. Ed.* **2025**, *64*, e202422109.
20. Shen, Z.; Siriboe, M.G.; Ren, X.; Dayananda, T.; Jenkins, J.L.; Khare, S.D.; Fasan, R. Computational Design of Generalist Cyclopropanases with Stereodivergent Selectivity. *ChemRxiv* **2025** DOI: 10.26434/chemrxiv-2025-wzmzg
21. Chiang, C.-H.; Wang, Y.; Hussain, A.; Brooks III, C.L.; Narayan, A.R.H. Ancestral Sequence Reconstruction to Enable Biocatalytic Synthesis of Azaphilones. *J. Am. Chem. Soc.* **2024**, *146*, 30194-30203.
22. Shahmoradi, A.; Wilke, C.O. Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. *Proteins* **2016**, *84*, 841-854.
23. Bershtein, S.; Segal, M.; Bekerman, R.; Tokuriki, N.; Tawfik, D.S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **2006**, *444*, 929-932.
24. Bloom, J.D.; Labthavikul, S.T.; Otey, C.R.; Arnold, F.H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.* **2006**, *103*, 5869-5874.
25. Miton, C.M.; Buda, K.; Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 160.
26. Kheronsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D.L.; Prilusky, J.; Amengual-Rigo, P.; Guallar, V.; Tawfik, D.S.; Fleishman, S.J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Molecular Cell* **2018**, *72*, 178-186.
27. Gomez de Santos, P.; Mateljak, I.; Hoang, M.D.; Fleishman, S.J.; Hollmann, F.; Alcalde, M. Repertoire of Computationally Designed Peroxygenases for Enantiodivergent C–H Oxyfunctionalization Reactions. *J. Am. Chem. Soc.* **2023**, *145*, 3443-3453.
28. Listov, D.; Vos, E.; Hoffka, G.; Hoch, S.Y.; Berg, A.; Hamer-Rogotner, S.; Dym, O.; Kamerlin, S.C.L.; Fleishman, S.J. Complete computational design of high-efficiency Kemp elimination enzymes. *Nature* **2025**, *643*, 1421-1427.
29. Lambert, T.; Tavakoli, A.; Dharuman, G.; Yang, J.; Bhethanabotla, V.; Kaur, S.; Hill, M.; Ramanathan, A.; Anandkumar, A.; Arnold, F.H. Sequence-based generative AI design of versatile tryptophan synthases. *Nat. Commun.* **2026** DOI: 10.1038/s41467-026-68384-6
30. Rapp, J.T.; Bremer, B.J.; Romero, P.A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **2024**, *1*, 97-107.
31. Yang, K.K.; Wu, Z.; Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687-694.
32. Freschlin, C.R.; Fahlberg, S.A.; Romero, P.A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **2022**, *75*, 102713.

33. Yang, J.; Li, F.-Z.; Arnold, F.H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **2024**, *10*, 226-241.
34. Wittmann, B.J.; Yue, Y.; Arnold, F.H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **2021**, *12*, 1026-1045.
35. Li, F.-Z.; Yang, J.; Johnston, K.E.; Gürsoy, E.; Yue, Y.; Arnold, F.H. Evaluation of machine learning-assisted directed evolution across diverse combinatorial landscapes. *Cell Syst.* **2025**, *16*, 101387.
36. Vornholt, T.; Mutný, M.; Schmidt, G.W.; Schellhaas, C.; Tachibana, R.; Panke, S.; Ward, T.R.; Krause, A.; Jeschek, M. Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. *ACS Cent. Sci.* **2024**, *10*, 1357-1370.
37. Yang, J.; Lal, R.G.; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.; Arnold, F.H. Active learning-assisted directed evolution. *Nat. Commun.* **2025**, *16*, 714.
38. Pesce, A.; Bolognesi, M.; Nardini, M. Protoglobin: structure and ligand-binding properties. *Microbial Globins – Status and Opportunities*; Poole, R.K., Ed.; Elsevier, 2013; pp 79-96. DOI: 10.1016/B978-0-12-407693-8.00003-0
39. Yang, Y.; Arnold, F.H. Navigating the Unnatural Reaction Space: Directed Evolution of Heme Proteins for Selective Carbene and Nitrene Transfer. *Acc. Chem. Res.* **2021**, *54*, 1209-1225.
40. Knight, A.M.; Kan, S.B.J.; Lewis, R.D.; Brandenberg, O.F.; Chen, K.; Arnold, F.H. Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes. *ACS Cent. Sci.* **2018**, *4*, 372-377.
41. Porter, N.J.; Danelius, E.; Gonen, T.; Arnold, F.H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **2022**, *144*, 8892-8896.
42. Hanley, D.; Li, Z.-Q.; Gao, S.; Virgil, S.C.; Arnold, F.H.; Alfonzo, E.A. Stereospecific Enzymatic Conversion of Boronic Acids to Amines. *J. Am. Chem. Soc.* **2024**, *146*, 19160-19167.
43. Alfonzo, E.; Hanley, D.; Li, Z.-Q.; Sicinski, K.M.; Gao, S.; Arnold, F.H. Biocatalytic Synthesis of α-Amino Esters via Nitrene C–H Insertion. *J. Am. Chem. Soc.* **2024**, *146*, 27267-27273.
44. Kennemur, J.L.; Long, Y.; Ko, C.J.; Das, A.; Arnold, F.H. Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]alkanes. *J. Am. Chem. Soc*. **2025**, *147*, 27165-27171.
45. Sicinski, K.M.; Ritts, C.B.; Lin, E.; Long, Y.; Arnold, F.H. Transformation of Aromatic Feedstocks into Value-Added Products via Biocatalytic Primary Amination. *Unpublished*.
46. Freitas, T.A.K.; Hou, S.; Dioum, E.M.; Saito, J.A.; Newhouse, J.; Gonzalez, G.; Gilles-Gonzalez, M.-A.; Alam, M. Ancestral hemoglobins in Archaea. *Proc. Natl. Acad. Sci.* **2004**, *101*, 6675-6680.
47. Zhang, R.K.; Chen, K.; Huang, X.; Wohlschlager, L.; Renata, H.; Arnold, F.H. Enzymatic assembly of carbon-carbon bonds via iron-catalysed sp^3 C–H functionalization. *Nature* **2018**, *565*, 67-72.
48. Hernandez, K.E.; Renata, H.; Lewis, R.D.; Kan, S.B.J.; Zhang, C.; Forte, J.; Rozzell, D.; McIntosh, J.A.; Arnold, F.H. Highly Stereoselective Biocatalytic

Synthesis of Key Cyclopropane Intermediate to Ticagrelor. *ACS Catal.* **2016**, *6*, 7810-7813.

49. Olson, C.A.; Wu, N.C.; Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Curr. Biol.* **2014**, *24*, 2643-2651.
50. Park, Y.; Metzger, B.P.H.; Thornton, J.W. The simplicity of protein sequence-function relationships. *Nat. Commun.* **2024**, *15*, 7953.
51. Acevedo-Rocha, C.G.; Hoebenreich, S.; Reetz, M.T. Iterative Saturation Mutagenesis: A Powerful Approach to Engineer Proteins by Systematically Simulating Darwinian Evolution. *Directed Evolution Library Creation*; Gillam, E.M.J.; Copp, J.N.; Ackerley, D., Eds.; Humana Press, 2014; pp 103-128. DOI: https://doi.org/10.1007/978-1-4939-1053-3_7
52. Long, Y.; Mora, A.; Li, F.-Z.; Gürsoy, E.; Johnston, K.E.; Arnold, F.H. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *ACS Synth. Biol.* **2025**, *14*, 230-238.
53. Daulton, S.; Eriksson, D.; Balandat, M.; Bakshy, E. Multi-Objective Bayesian Optimization over High-Dimensional Search Spaces. *arXiv* **2021** DOI: 10.48550/arXiv.2109.10964
54. Kondrashov, F.A. In search of the limits of evolution. *Nat. Genet.* **2005**, *37*, 9-10.
55. Kaltenbach, M.; Emond, S.; Hollfelder, F.; Tokuriki, N. Functional Trade-Offs in Promiscuous Enzymes Cannot Be Explained by Intrinsic Mutational Robustness of the Native Activity. *PLOS Genetics* **2016**, *12*, e1006305.
56. Zhang, H.; Liu, J.; Zhang, L.; Kong, L.; Yao, H.; Sun, H. Synthesis and biological evaluation of ticagrelor derivatives as novel antiplatelet agents. *Bioorg. Med. Chem. Let.* **2012**, *22*, 3598-3602.
57. Yang, J.; Li, F.-Z.; Long, Y.; Arnold, F.H. Illuminating the universe of enzyme catalysis in the era of artificial intelligence. *Cell Syst.* **2025** DOI: 10.1016/j.cels.2025.101372

Appendix D

SUPPLEMENTARY INFORMATION FOR CHAPTER V

D.1. General Information

D.1.1. Safety Statement

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure. Other than that, no unexpected or unusually high safety concerns were raised with these methods. Safety notes for individual synthetic procedures will be documented alongside the procedure.

D.1.2. Chemicals

All chemical transformations were performed in a well-ventilated fume hood to avoid inhalation and exposure to chemicals. Reagents and solvents were obtained commercially (Sigma-Aldrich, Alfa Aesar, VWR, Fisher Scientific, Matrix Scientific, Oakwood Chemical, TCI America, and other suppliers) and used without prior purification unless otherwise stated.

D.1.3. Instrumentation

Organic solutions were concentrated under reduced pressure on an IKA RV 10 rotary evaporator. Thin-layer chromatography (TLC) was performed on commercial Millipore Silica Gel 60 plates containing the F254 fluorescent indicator. Visualization of the developed chromatographs was performed by irradiation with UV light or by treatment with an appropriate TLC staining solution (e.g., Ceric Ammonium Molybdate, $KMnO_4$, or Bromocresol Green) followed by heating if necessary. Chromatographic purification was accomplished by flash chromatography using a Biotage Isolera One instrument.

All NMR spectra were obtained at the Caltech Liquid NMR Facility. NMR spectra were collected on a Bruker Prodigy 400 MHz instrument equipped with a cryoprobe operating at 400 MHz and 101 MHz for ^{1}H and ^{13}C, respectively. ^{1}H spectra are referred to residual $CDCl_3$ solvent signals referenced at δ 7.26 ppm. Data for ^{1}H NMR are reported as follows: chemical shift (δ ppm), integration, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentad, sext = sextet, hept = heptet, m = multiplet, br s = broad singlet), and coupling constant (Hz).

Gas chromatography (GC) was performed on an Agilent Technologies 7820A GC system equipped with a split-mode capillary injection system. For achiral analyses, an Agilent J&W HP-5 Column was used as the stationary phase. Reaction systems displaying sufficient yield were subsequently analyzed with a flame-ionization detector. When target analytes were only produced in trace yield, samples were analyzed by an Agilent Technologies 5977B mass spectrometer.

Liquid chromatography (LC) was performed on an Agilent Technologies 1260 Infinity HPLC-MS system (Agilent 6120 quadrupole mass spectrometer) equipped with an Agilent C18 column (Poroshell 120 ESC18, 4.6 x 50 mm, 2.7-μm packing) with a Poroshell 120 guard column (2.7-μm packing, 2.1 x 5 mm). Water and acetonitrile modified with 0.1% acetic acid were used as eluents.

D.2. Cloning and Sequence Design

D.2.1. Relevant Protoglobin Sequences

Table D-1. The amino acid sequences of wild-type *Ape*Pgb and the previously engineered variant, PromPgb. Residues in each sequence highlighted in yellow are the sites which were investigated in this work for wet-lab experimentation: L59, W62, F63, F73, L86, S90, F93, and S149 in PromPgb.

Protein Variant	Amino Acid Sequence
wild-type *Aeropyrum pernix* Protoglobin (*Ape*Pgb)	MTPSDIPGYDYGRVEKSPITDLEFDLLKKTVM LGEKDVMYLKKACDVLKDQVDEILDLWYG WVASNEHLIYYFSNPDTGEPIKEYLERVRARF GAWILDTTCRDYNREWLDYQYEVGLRHHRS KKGVTDGVRTVPHIPLRYLIAFIYPITATIKPFL AKKGGSPEDIEGMYNAWFKSVVLQVAIWSHP YTKENDW
Aeropyrum pernix Protoglobin D10G C45Y L56I W59L Y60G V63F R90S F145G I149S F156L Y189H (PromPgb)	MTPSDIPGYGYGRVEKSPITDLEFDLLKKTVM LGEKDVMYLKKAYDVLKDQVDEIIDLLGGW FASNEHLIYYFSNPDTGEPIKEYLERVSARFGA WILDTTCRDYNREWLDYQYEVGLRHHRSKK GVTDGVRTVPHIPLRYLIAGIYPSTATIKPLLA KKGGSPEDIEGMYNAWFKSVVLQVAIWSHPH TKENDW

Table D-2. DNA Sequence of PromPgb, including C-terminal 6xHis-tag and stop codon. The region highlighted in green represents the region of the gene in which exact, predicted sets of mutations were incorporated in an oligo pool.

ATGACTCCCTCGGACATCCCGGGATATGGTTATGGGCGTGTCGAGAAGTCACCCAT
CACGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGAC
GTAATGTACTTGAAAAAGGCGTATGACGTTCTGAAAGATCAAGTTGATGAGATCA
TTGACTTGCTGGGTGGTTGGTTTGCATCAAATGAGCATTTGATTTATTACTTCTCCA
ATCCGGATACAGGAGAGCCTATTAAGGAATACCTGGAACGTGTAAGCGCTCGCTT
TGGAGCCTGGATTCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACT
ACCAGTACGAAGTTGGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGG
AGTACGCACCGTGCCCCATATCCCACTTCGTTATCTTATCGCAGGTATCTATCCTAG
TACCGCCACTATCAAGCCACTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCG
AAGGGATGTACAACGCTTGGTTCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCA
CACCCTCATACTAAGGAGAATGACTGGCTCGAGCACCACCACCACCACCACTGAG

Table D-3. Codons utilized when ALDE-predicted mutations were incorporated into the PromPgb DNA sequence. Codons which are generally recognized as the most prevalently found in the *E. coli* genome were selected.

Amino Acid	Codon
A	GCG
C	TGC
D	GAT
E	GAA
F	TTT
G	GGC
H	CAT
I	ATT
K	AAA
L	CTG
M	ATG
N	AAC
P	CCG
Q	CAG
R	CGT
S	AGC
T	ACC
V	GTG
W	TGG
Y	TAT

D.2.2. Nomenclature for Variant Naming

Single-site mutants are named using the standard nomenclature: (original AA)(site)(new AA). The mutant L59F refers to a variant of PromPgb mutated from leucine to phenylalanine at position 59. Eight-site multi-mutants are named as a string of the eight amino acids to which the positions of interest have been mutated. The variant IYVAIIKQ

refers to a variant of PromPgb bearing the mutations F59I, W62Y, F63V, F73A, L86I, S90I, F93K, and S149Q. In this nomenclature, PromPgb is named LWFFLSFS.

D.2.3. Primer Design for Site-Saturation Mutagenesis (SSM)

D.2.3.1. General Cloning Primers

Table D-4. Primers 007 and primers 008 are used to generate amplicons of linearized pET-22b(+)[1] backbone. Primers 005 and 006 were used with primers internal to the protoglobin gene (**Table D-5**) to generate mutant protoglobin genes. All primers were ordered from IDT (Coralville, IA).

Primer Name	**Direction**	**Sequence**	**Description**
005	Forward	5'-gaaataattttgtttaactttaagaaggagatatacatatg-3'	Upstream of N-term, anneals with 007
006	Reverse	5'-gccggatctcagtggtggtggtggtggtgctcgag-3'	Downstream of C-term, anneals with 008
007	Reverse	5'-catatgtatatctccttcttaaagttaaacaaaattatttc-3'	Upstream of N-term, anneals with 005
008	Forward	5'-ctcgagcaccaccaccaccaccactgagatccggc-3'	Downstream of C-term, anneals with 006

D.2.3.2. Cloning Primers for Site-Saturation Mutagenesis

Table D-5. Primers containing degenerate codons at sites of interest for mutagenesis. Primers were used in conjunction with either primer 005 or 006 (**Table D-4**) to generate mutant protoglobin genes. All primers were ordered from IDT (Coralville, IA). Abbreviations: SSM – site saturation mutagenesis, dSSM – double site saturation mutagenesis, tSSM – triple site saturation mutagenesis.

Primer Name	**Sequence**	**Description**
59_fwd	5'-attgacttgNNKggtggttggtttgcatcaaatgagc-3'	SSM for site 59
59_rev	5'-atttgatgcaaaccaaccaccMNNcaagtcaatgatctcatc-3'	SSM for site 59
73_fwd	5'-gcatttgatttattacNNKtccaatccggatacaggagag-3'	SSM for site 73
73_rev	5'-tgtatccggattggaMNNgtaataaatcaaatgctcatttg-3'	SSM for site 73
86_fwd	5'-ttaaggaatacNNKgaacgtgtaagcgctcgctttg-3'	SSM for site 86
86_rev	5'-aagcgagcgcttacacgttcMNNgtattccttaatag-3'	SSM for site 86
90_fwd	5'-gcctattaaggaatacctggaacgtgtaNNKgctcgctttggagcctgg-3'	SSM for site 90
90_rev	5'-ccaggctccaaagcgagcMNNtacacgttccaggtattccttaataggc-3'	SM for site 90
93_fwd	5'-ctggaacgtgtaagcgctcgcNNKggagcctggattctggacac-3'	SSM for site 93
93_rev	5'-cagaatccaggctccMNNgcgagcgcttacacgttccaggtattc-3'	SSM for site 93
149_fwd	5'-tcgcaggtatctatcctNNKaccgccactatcaagccac-3'	SSM for site 149
149_rev	5'-gatagtggcggtMNNaggatagatacctgcgataagataacgaag-3'	SSM for site 149
59+62_fwd	5'-atgagatcattgacttgNNKggtggtNNKtttgcatcaaatgagc-3'	dSSM for sites 59 and 62

59+62_rev	5’-tttgatgcaaaMNNaccaccMNNcaagtcaatgatctcatcaac-3’	dSSM for sites 59 and 62
59+63_fwd	5’-atgagatcattgacttgNNKggtggttggNNKgcatcaaatgagc-3’	dSSM for sites 59 and 63
59+63_rev	5’-tttgatgcMNNccaaccaccMNNcaagtcaatgatctcatcaac-3’	dSSM for sites 59 and 63
62+63_fwd	5’-gacttgctgggtggtNNKNNKgcatcaaatgagcatttgatttattacttc-3’	dSSM for sites 62 and 63
62+63_rev	5’-aatcaaatgctcatttgatgcMNNMNNaccacccagcaagtcaatgatc-3’	dSSM for sites 62 and 63
63+73_fwd	5’-gtggttggNNKgcatcaaatgagcatttgatttattacNNKtccaatccggatacagga g-3’	dSSM for sites 63 and 73
63+73_rev	5’-ggattggaMNNgtaataaatcaaatgctcatttgatgcMNNccaaccacccagcaagt c-3’	dSSM for sites 63 and 73
86+90_fwd	5’-ctattaaggaatacNNKgaacgtgtaNNKgctcgctttggagcctg-3’	dSSM for sites 86 and 90
86+90_rev	5’-caggctccaaagcgagcMNNtacacgttcMNNgtattccttaatag-3’	dSSM for sites 86 and 90
86+93_fwd	5’-tattaaggaatacNNKgaacgtgtaagcgctcgcNNKggagcctggattctggac-3’	dSSM for sites 86 and 93
86+93_rev	5’-aatccaggctccMNNgcgagcgcttacacgttcMNNgtattccttaataggctc-3’	dSSM for sites 86 and 93
90+93_fwd	5’-ggaacgtgtaNNKgctcgcNNKggagcctggattctggacactacctgc-3’	dSSM for sites 90 and 93

90+93_rev	5’-ccagaatccaggctccMNNgcgagcMNNtacacgttccaggtattccttaatag-3’	dSSM for sites 90 and 93
59+62+63_fwd	5’-caagttgatgagatcattgacttgNNKggtggtNNKNNKgcatcaaatgagcatttg-3’	tSSM for sites 59, 62, and 63
59+62+63_rev	5’-gctcatttgatgcMNNMNNaccaccMNNcaagtcaatgatctcatcaacttgatc-3’	tSSM for sites 59, 62, and 63

D.2.3.3. Cloning Primers for Oligo Libraries

Table D-6. Primers used for incorporating adaptors to the ends of oligo fragments for incorporation into the pET-22b(+) vector.

Primer Name	**Sequence**	**Description**
oligo_fwd	5’-gttctgaaagatcaagttgatgagatcattgacttg-3’	Anneals to multi-mutant oligo fragments
oligo_rev	5’-ccaaaagtggcttgatagtggcggt-3’	Anneals to multi-mutant oligo fragments
pET_BB_fwd	5’-accgccactatcaagccacttttgg-3’	Anneals to base pairs internal to PromPgb

pET_BB_rev	5'-caagtcaatgatctcatcaacttgatctttcagaac-3'	Anneals to base pairs internal to PromPgb

D.3. Cloning Protocols and Results

D.3.1. Protocols for the Cloning of Random PromPgb Variants

D.3.1.1. Cloning for Single Site-Saturation Mutagenesis (SSM)

Chemically competent *Escherichia coli* (*E. coli*) cells (T7 Express Competent *E. coli*) were purchased from New England Biolabs (NEB, Ipswich, MA). Additionally, Phusion polymerase and *Dpn*I were purchased from NEB. SSM experiments were performed using primers bearing degenerate codons (NNK) using a modified QuikChange™ protocol (**Tables D-4** and **D-5**).[2]

The PCR conditions were as follows (final concentrations): Phusion HF Buffer 1x, 0.2 mM dNTPs each, 0.5 μM of forward primers, 0.5 μM reverse primer, and 0.02 U/μL of Phusion polymerase. The standard Phusion PCR protocol was used.[3] Upon completion of PCRs, the remaining template was digested with *Dpn*I. Gel purification was performed with a Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA). The purified PCR product was then assembled using the Gibson assembly protocol.[4]

Table D-7. Primer combinations for the generation of PCR amplicons for the construction of expression plasmids containing mutagenized PromPgb variants. '*site*' refers to the specific site or set of sites being saturated.

Fragment Name	**Forward Primer**	**Reverse Primer**
SSM_Frag1	005	*site*_rev
SSM_Frag2	*site*_fwd	006
pET_Backbone	008	007

D.3.1.2. Construction of Random Multi-Mutant Libraries

The primers described in **Table D-5** were designed to provide access to all 18 of the random multi-mutation variant libraries (**Table D-8**). These libraries represent the set of variants used to initially train machine learning models. Libraries requiring multiple mutagenesis steps to construct the final multi-mutants were generated as follows: Site saturation or multi-site saturation mutagenesis was performed on the first set of sites within the defined library according to the above PCR and assembly protocol (*Cloning for Site-Saturation Mutagenesis (SSM)*). The assembly products obtained were used to transform T7 Express Competent *E. coli* (High Efficiency) cells following the protocol recommended by the manufacturer. Upon heat-shock, freshly transformed *E. coli* cells were recovered in 0.4 mL Luria-Bertani medium (LB) (Research Products Int.) at 37 ºC with shaking at 220 rpm for 30 minutes. For libraries requiring only a single mutagenesis step, this transformation mixture was directly applied to the protocol described below for plating on LB-Amp agar plates. For libraries requiring two mutagenesis steps, 50 µL was transferred into 6 mL of LB with 100 µg/mL ampicillin (LB-Amp) in a 15-mL culture tube. This culture was allowed to shake overnight at 37 °C and 220 rpm. The following morning, this library overnight culture was miniprepped using a QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany). The miniprepped plasmid DNA pool was used as the new template for mutagenesis with the primers for the remaining sites in the random library. T7 Express Competent *E. coli* were transformed with the Gibson products for the new multi-site library using the recommended protocol. Upon heat-shock transformation, the freshly transformed

E. coli cells were recovered in 0.4 mL LB medium at 37 ºC with shaking at 220 rpm for 30 minutes, yielding a transformation mixture with *E. coli* harboring the final multi-site mutagenesis library.

Transformation mixtures with *E. coli* harboring one of the final desired libraries were plated on LB-agar plates with 100 μg/mL ampicillin (LB-Amp agar plates). The plates were incubated overnight at 37 °C until colony formation was observed. For each of the 18 libraries, single colonies from LB-Amp agar plates were picked with sterilized toothpicks to individually inoculate the wells of a 2-mL 96-well deep-well plate charged with 400 μL of LB-Amp. The plates were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. 96-Well deep-well plates were shaken in an INFORS HT Multitron Shaker in all instances. The following morning, 50 μL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 μL of 50% glycerol solution. These glycerol stocks were stored at -80 °C for future inoculation. Additionally, the sequences of protoglobin genes contained in every well were sequenced using LevSeq sequencing (**Figures D-1–D-20**).[5]

Table D-8. Combinatorial libraries tested used to collect training data and the assembly methods used in their generation. Residue distances were calculated from Cα positions in a homology model of PromPgb co-folded with heme using AlphaFold3.[6]

Library	**Sites**	**Residue Distance (Å)**	**Construction Steps**
1	59, 62	5.0	single dSSM step with 59+62 primers
2	62, 63	3.8	single dSSM step with 62+63 primers
3	63, 73	9.0	single dSSM step with 63+73 primers
4	73, 86	6.3	sSSM with 73 primers, then sSSM with 86 primers
5	86, 90	5.9	single dSSM step with 86+90 primers
6	90, 93	4.8	single dSSM step with 90+93 primers

7	93,149	11.3	sSSM with 93 primers, then sSSM with 149 primers
8	59, 149	9.8	sSSM with 59 primers, then sSSM with 149 primers
9	59, 73	12.8	single dSSM step with 59+73 primers
10	73, 90	11.2	sSSM with 73 primers, then sSSM with 90 primers
11	86, 93	10.1	single dSSM step with 86+93 primers
12	59, 93	9.0	sSSM with 59 primers, then sSSM with 93 primers
13	62, 63, 86	3.8 (62-63), 11.8 (62-86), 9.2 (63-86)	dSSM with 62+63 primers, then sSSM with 86 primers
14	59, 62, 63	5.0 (59-62), 6.3 (59-63), 3.8 (62-63)	single tSSM step with 59+62+63 primers
15	59, 63, 90	6.3 (59-63), 7.6 (59-90), 10.7 (63-90)	dSSM with 59+63 primers, then sSSM with 90 primers
16	59, 90, 93	7.6 (59-90), 9.0 (59-93), 4.8 (90-93)	dSSM with 90+93 primers, then sSSM with 59 primers
17	62, 63,149	3.8 (62-63), 11.2 (62-149), 12.6 (63-149)	dSSM with 62+63 primers, then sSSM with 149 primers
18	86, 90, 149	5.9 (86-90), 17.1 (86-149), 13.9 (90-149)	dSSM with 86+90 primers, then sSSM with 149 primers

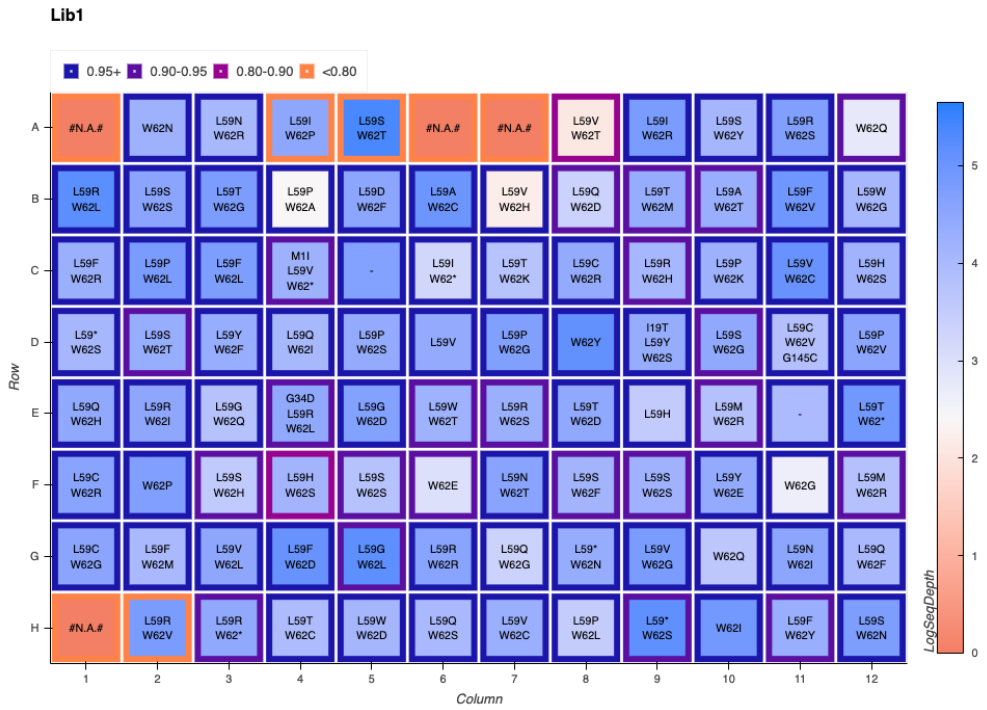


Figure D-1. LevSeq sequencing plate map for random variants picked from library 1.

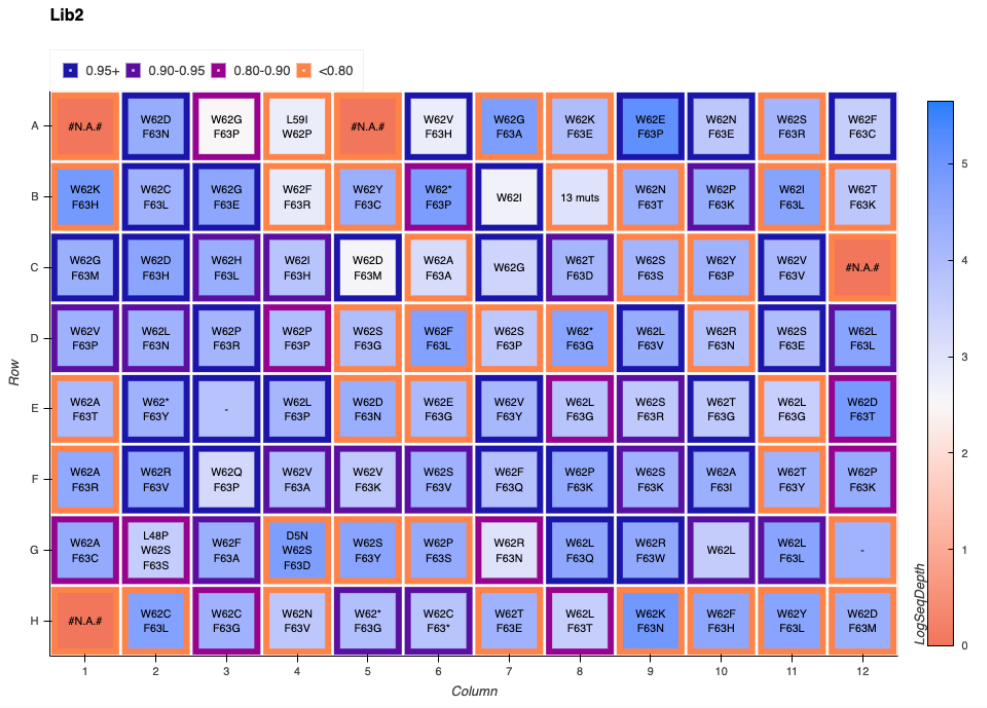


Figure D-2. LevSeq sequencing plate map for random variants picked from library 2.

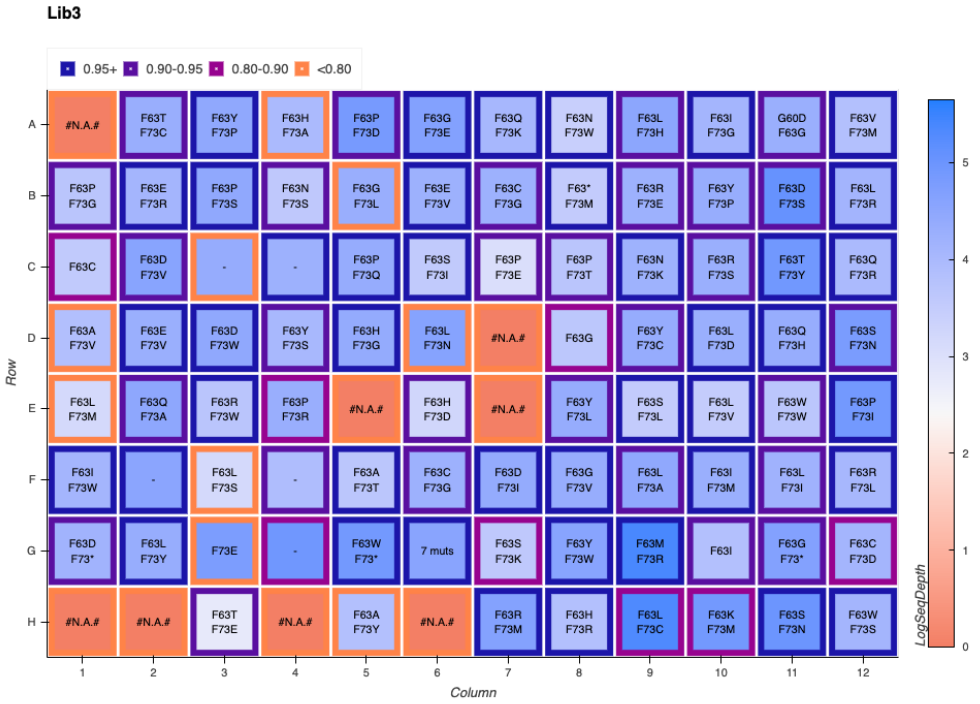


Figure D-3. LevSeq sequencing plate map for random variants picked from library 3.

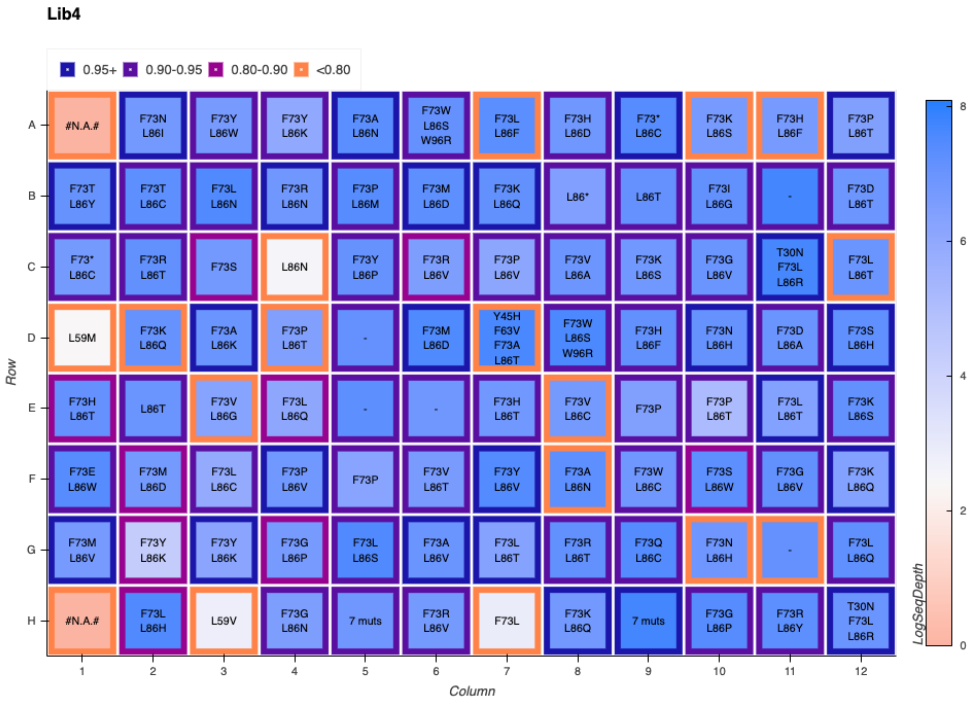


Figure D-4. LevSeq sequencing plate map for random variants picked from library 4.

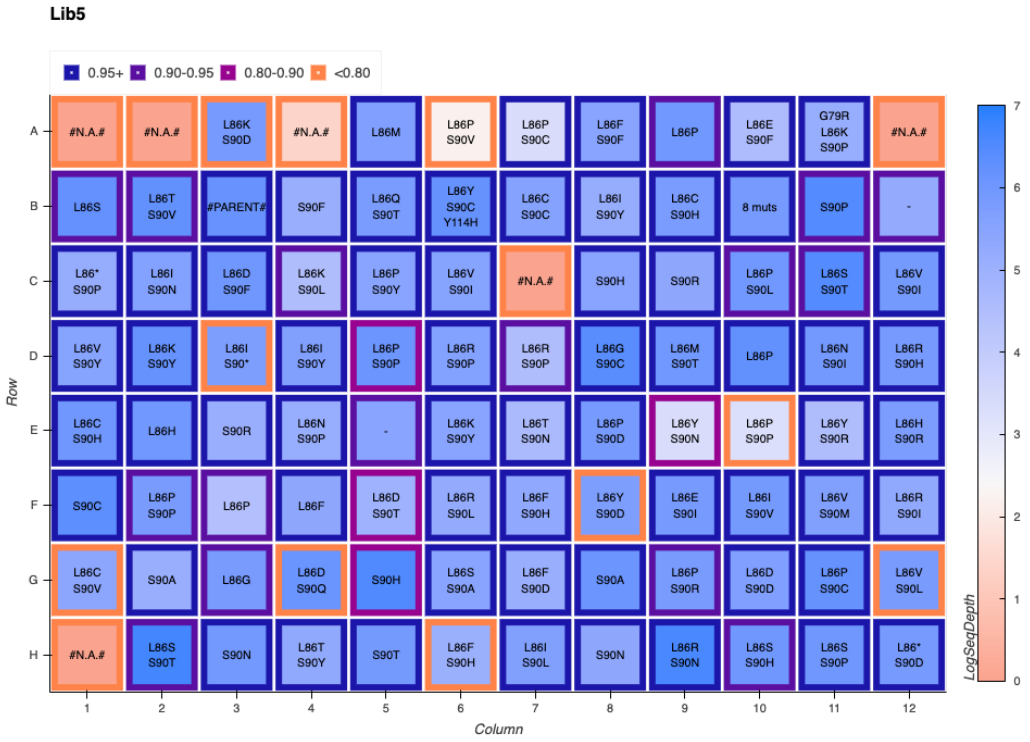


Figure D-5. LevSeq sequencing plate map for random variants picked from library 5.

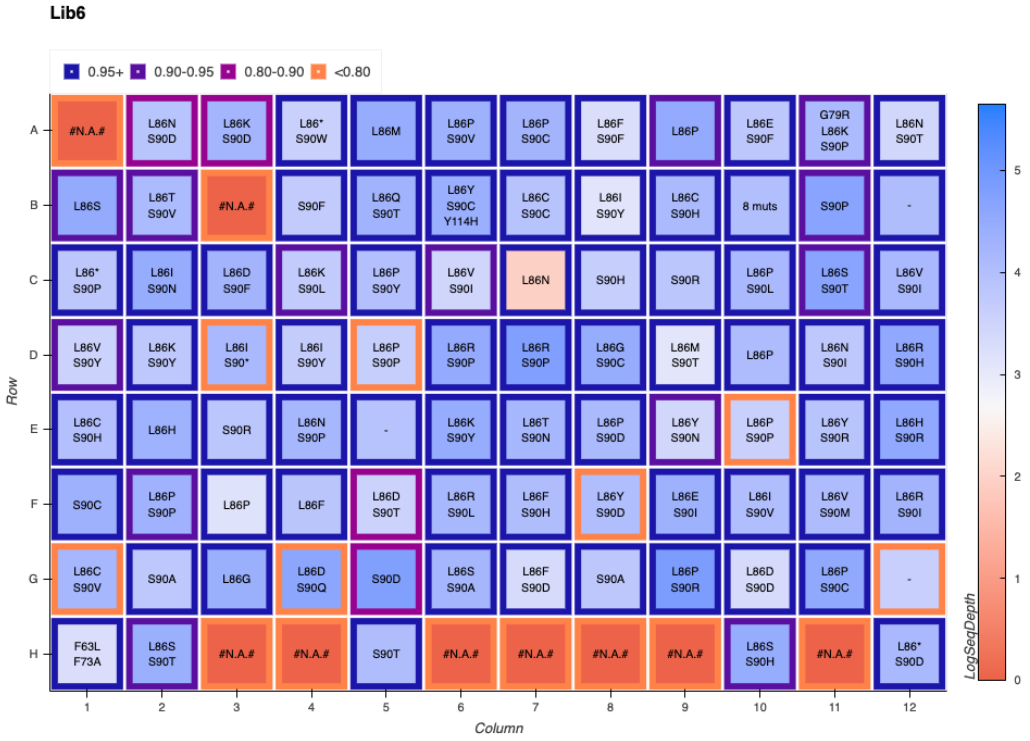


Figure D-6. LevSeq sequencing plate map for random variants picked from library 6.

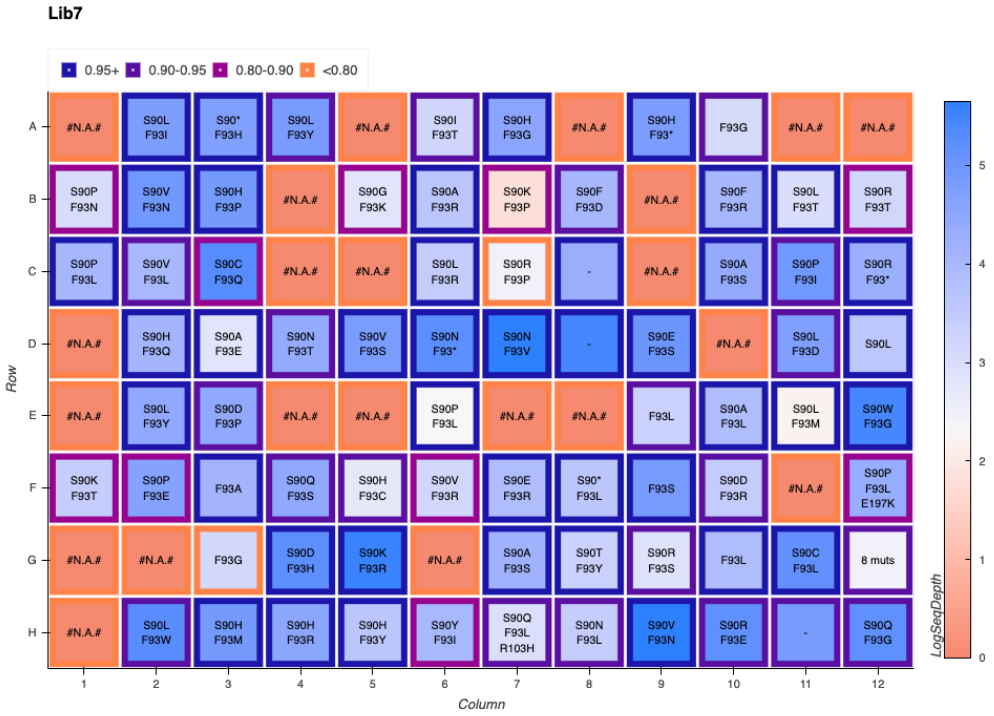


Figure D-7. LevSeq sequencing plate map for random variants picked from library 7.

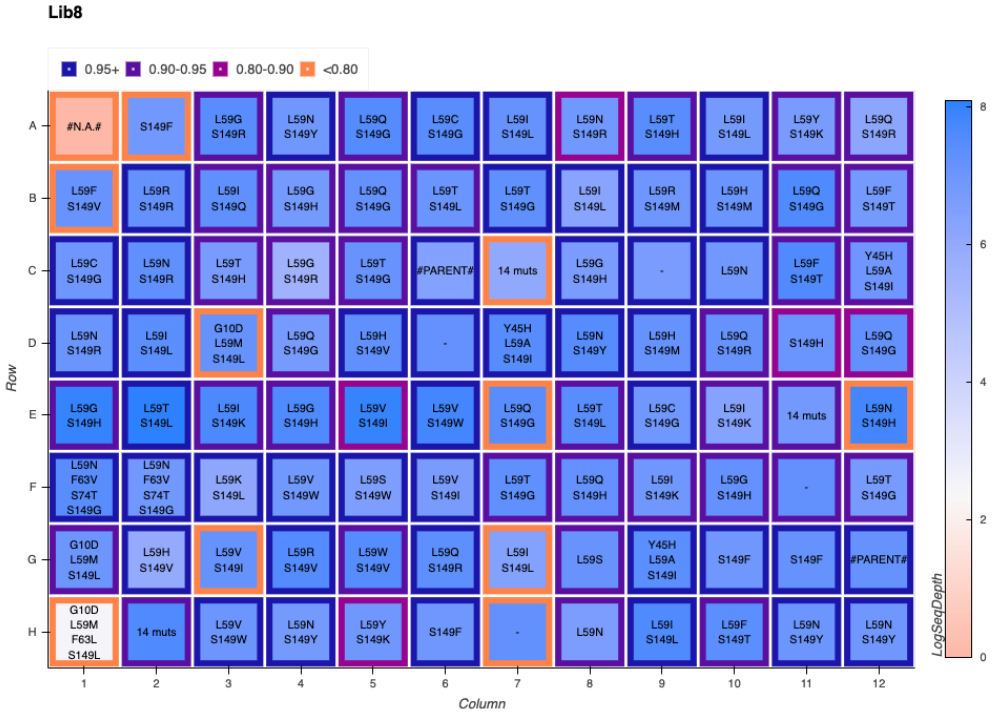


Figure D-8. LevSeq sequencing plate map for random variants picked from library 8.

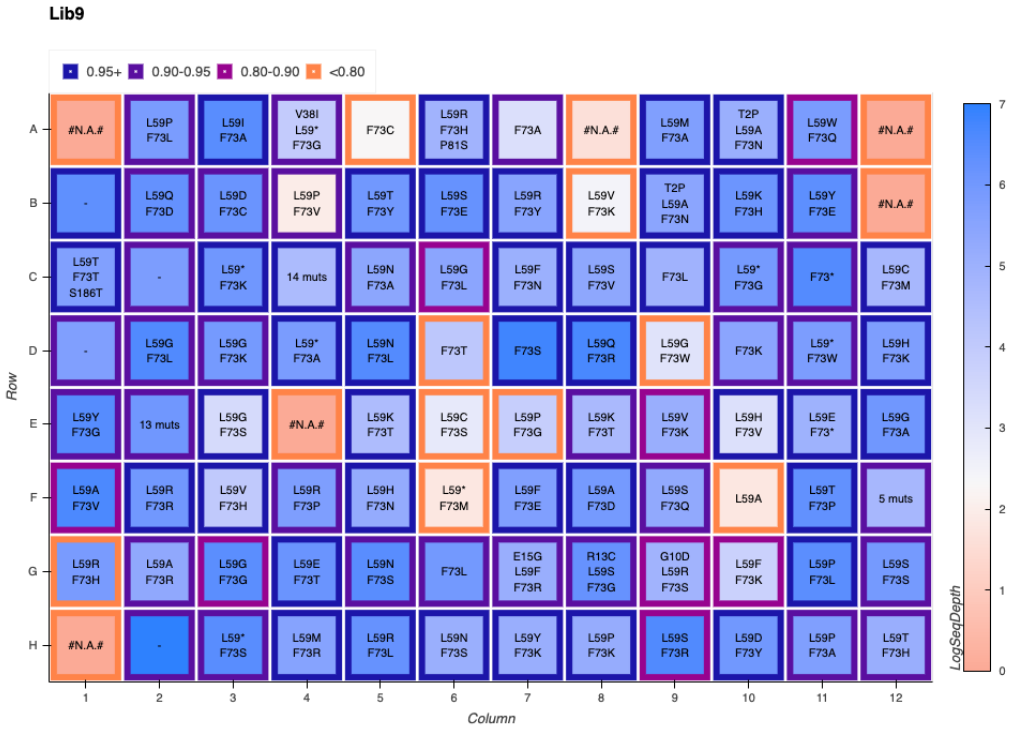


Figure D-9. LevSeq sequencing plate map for random variants picked from library 9.

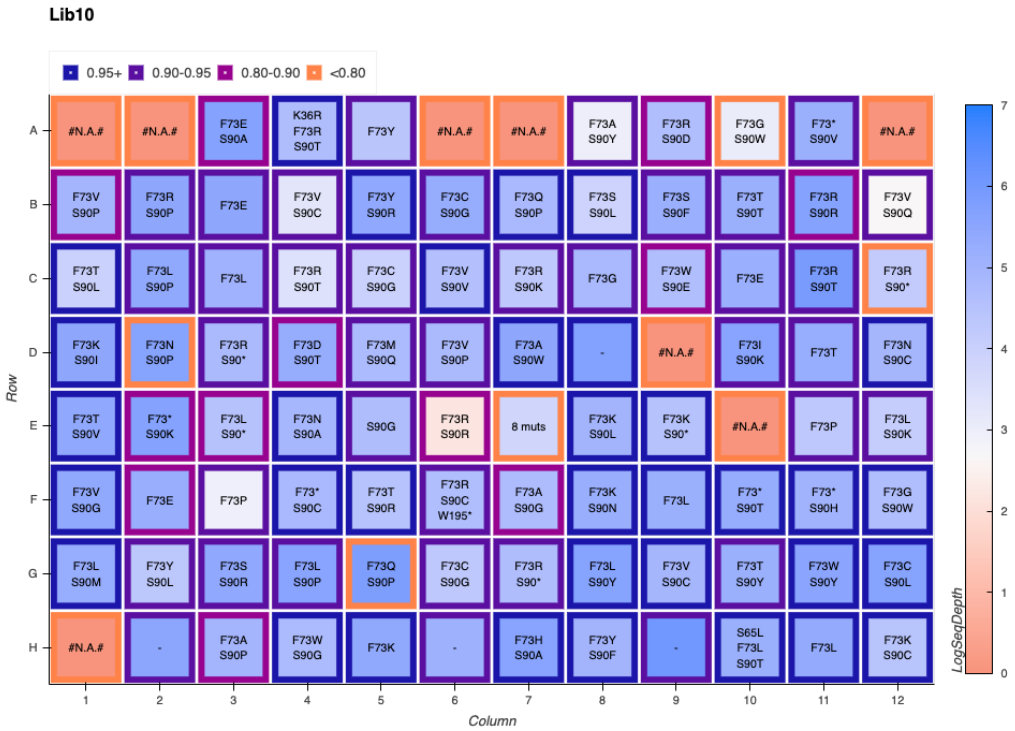


Figure D-10. LevSeq sequencing plate map for random variants picked from library 10.

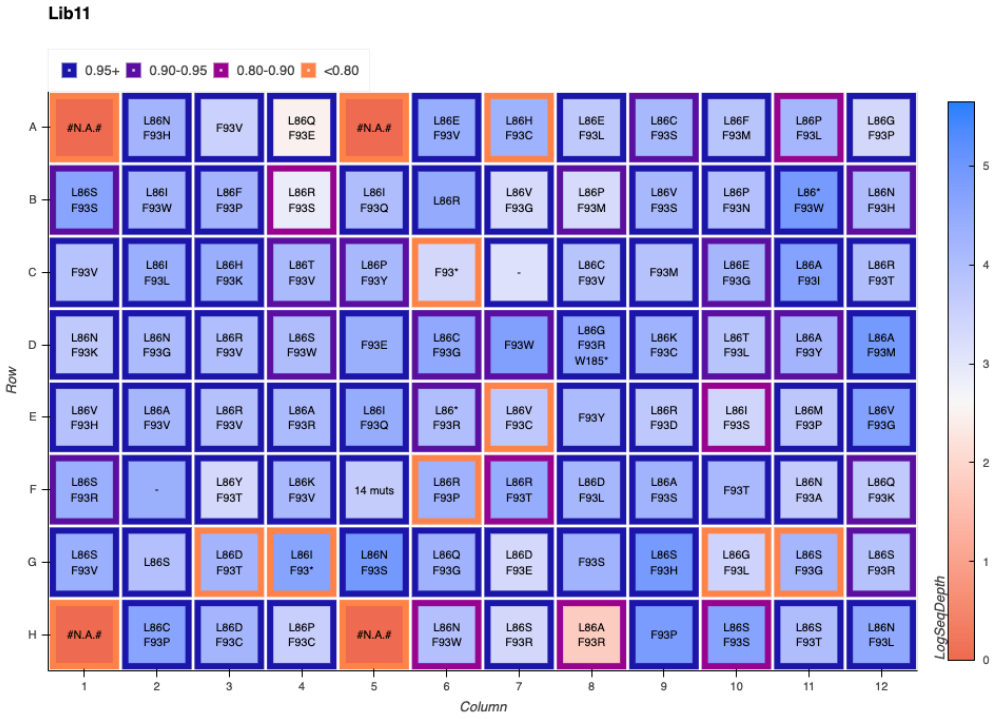


Figure D-11. LevSeq sequencing plate map for random variants picked from library 11.

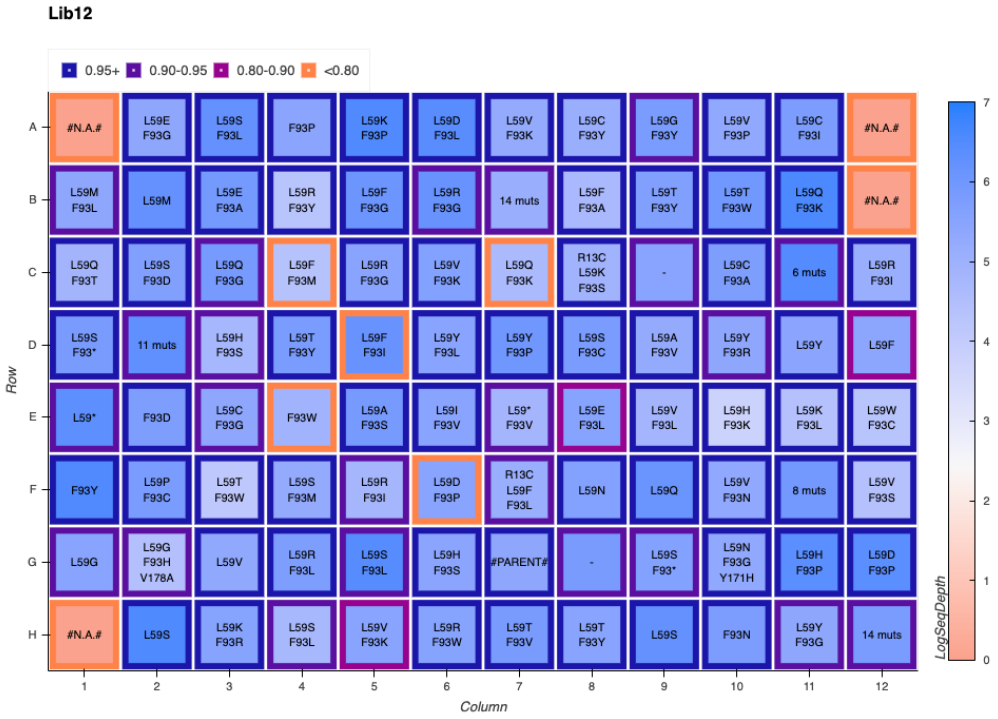


Figure D-12. LevSeq sequencing plate map for random variants picked from library 12.

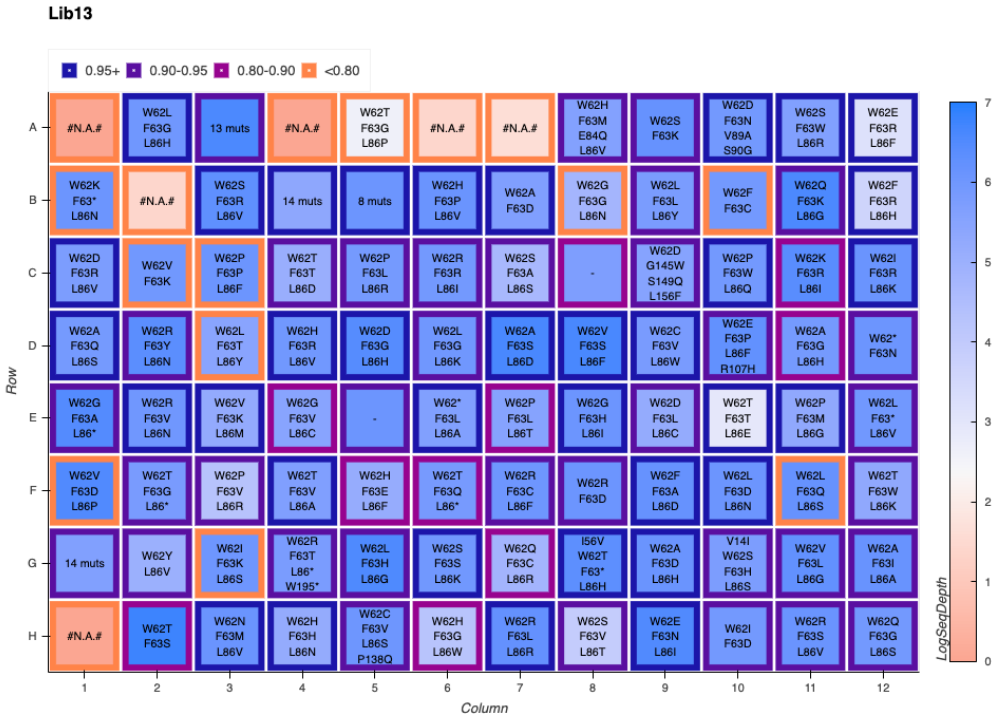


Figure D-13. LevSeq sequencing plate map for random variants picked from library 13.

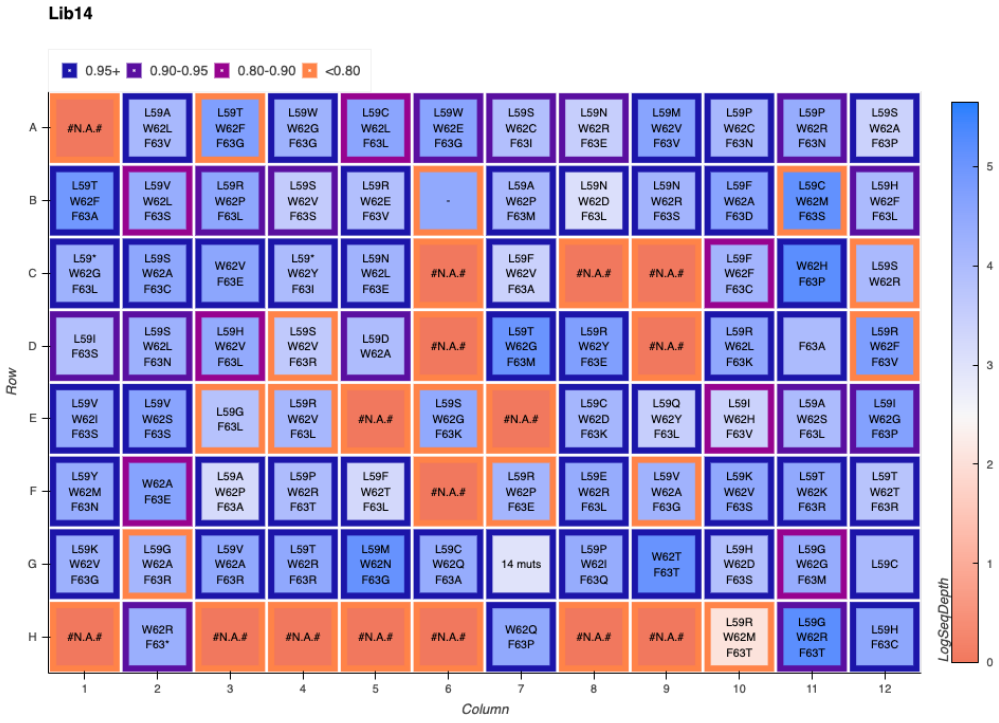


Figure D-14. LevSeq sequencing plate map for random variants picked from library 14.

Figure D-15. LevSeq sequencing plate map for additional random variants picked from libraries 13 and 14.

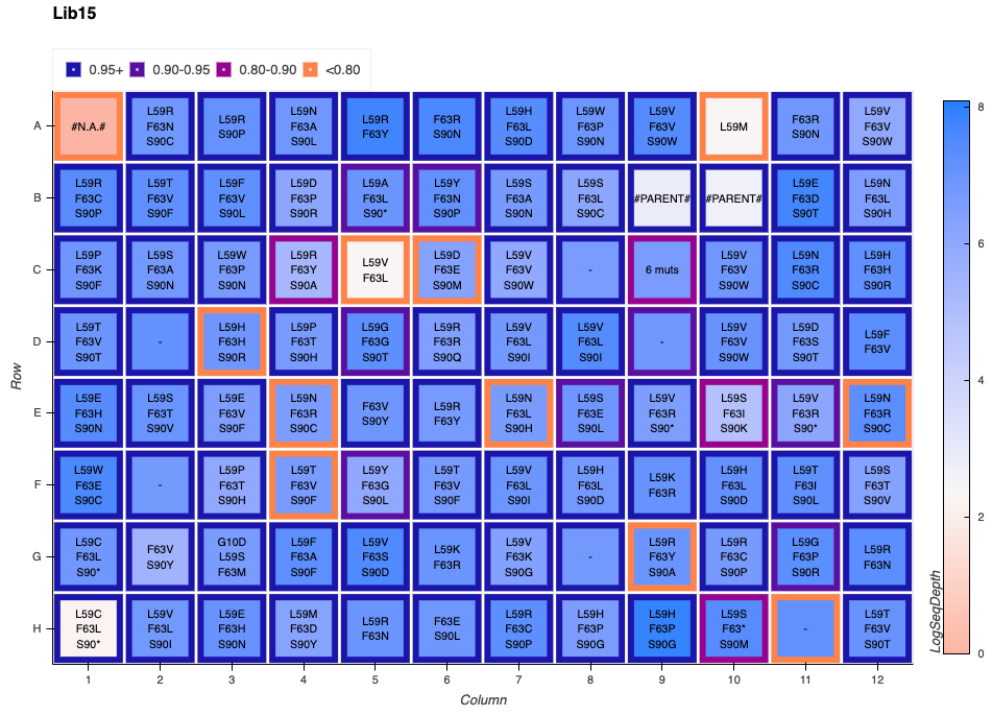


Figure D-16. LevSeq sequencing plate map for random variants picked from library 15.

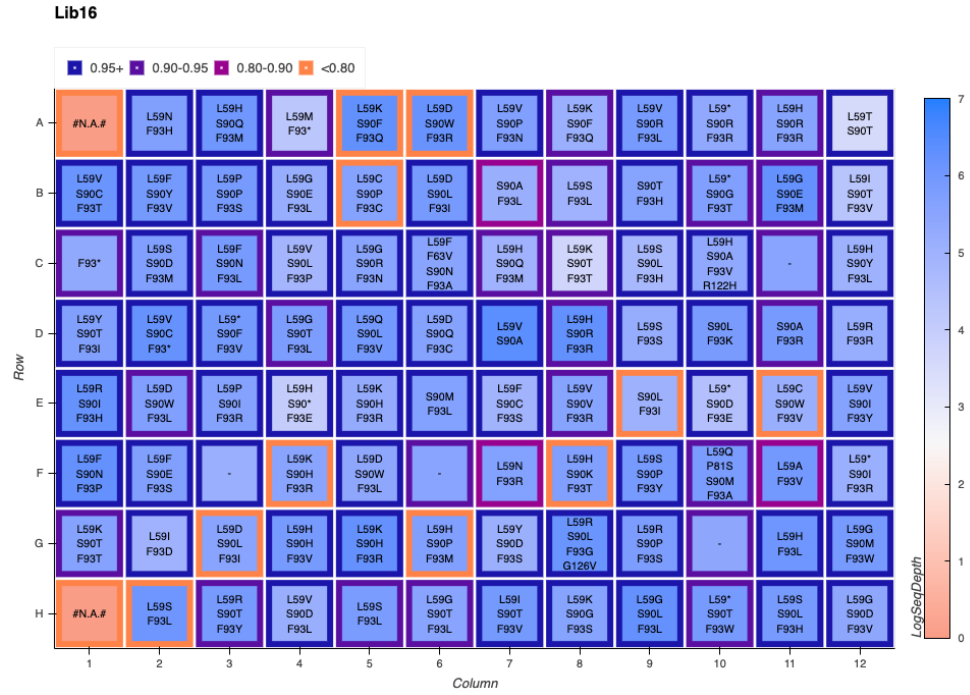


Figure D-17. LevSeq sequencing plate map for random variants picked from library 16.

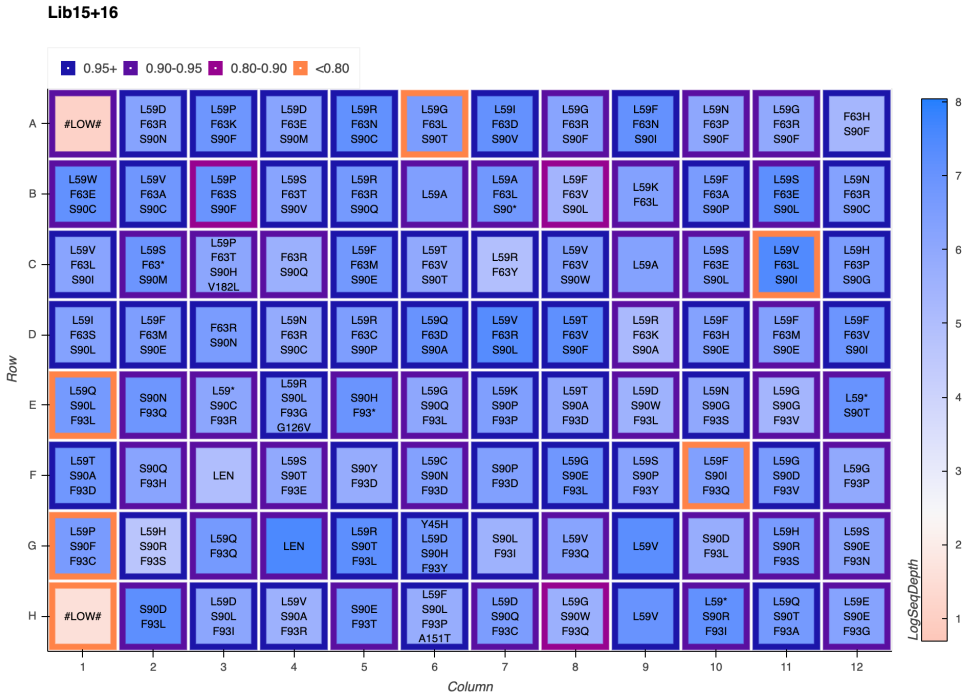


Figure D-18. LevSeq sequencing plate map for additional random variants picked from libraries 15 and 16.

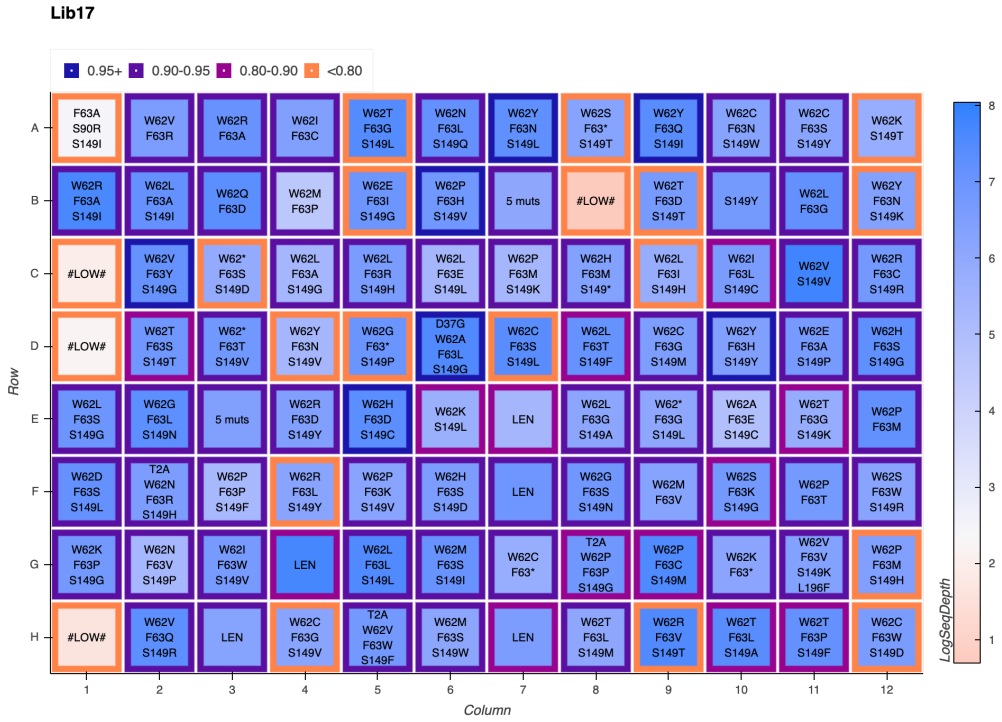


Figure D-19. LevSeq sequencing plate map for random variants picked from library 17.

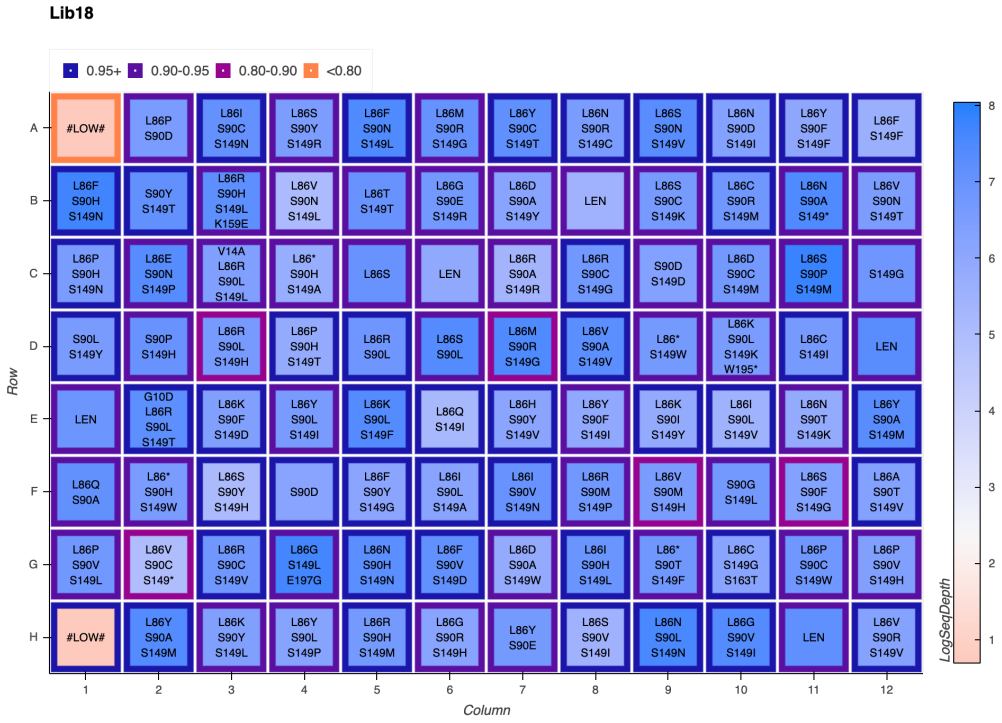


Figure D-20. LevSeq sequencing plate map for random variants picked from library 18.

D.3.1.3. Rearray of Random Mutants

Multi-mutation variants were rearrayed from the previously described, randomly picked, plates to reduce screening burden for model training data collection. A Python script was used to randomly select 22 variants from each of the 12 double-mutant libraries and 44 variants from each of the six triple-site libraries, yielding 528 unique multi-mutants. These variants were rearrayed over six 96-well deep-well plates in the following manner: The wells of a 2-mL 96-well deep-well plate were filled with 400 µL LB-Amp. Previously generated 96-well plates were removed from -80 °C storage and placed on dry ice. Pipet tips were used to scratch the frozen glycerol stock surface and used to inoculate 88 wells of the aforementioned deep-well plate. Additionally, to each of these deep-well plates, six wells were inoculated with *E. coli* harboring the parent gene, PromPgb, one well was inoculated with *E. coli* harboring a gene encoding a tryptophan synthase variant (TrpB, UniProt: P0A879), and one well was left sterile. These overnight cultures were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 µL of overnight culture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 µL of 50% glycerol solution. These glycerol stocks were stored at -80°C for future inoculation.

Table D-9. Variants selected as members of the initial training library for model training. A Python script was used to randomly select sequenced variants generated through multi-NNK cloning.

Library	**Variant**
1	['L59L', 'W62N']
1	['L59T', 'W62G']
1	['L59A', 'W62T']
1	['L59C', 'W62R']
1	['L59R', 'W62L']
1	['L59G', 'W62D']
1	['L59P', 'W62S']
1	['L59C', 'W62G']
1	['L59R', 'W62S']
1	['L59P', 'W62K']

1	['L59T', 'W62D']
1	['L59S', 'W62H']
1	['L59N', 'W62R']
1	['L59N', 'W62I']
1	['L59H', 'W62S']
1	['L59P', 'W62G']
1	['L59Q', 'W62H']
1	['L59R', 'W62I']
1	['L59A', 'W62C']
1	['L59F', 'W62M']
1	['L59N', 'W62T']
1	['L59P', 'W62L']
2	['W62P', 'F63R']
2	['W62V', 'F63K']
2	['W62T', 'F63D']
2	['W62L', 'F63L']
2	['W62A', 'F63I']
2	['W62G', 'F63E']
2	['W62F', 'F63C']
2	['W62L', 'F63V']
2	['W62F', 'F63Q']
2	['W62N', 'F63E']
2	['W62I', 'F63H']
2	['W62S', 'F63E']
2	['W62S', 'F63K']
2	['W62R', 'F63W']
2	['W62E', 'F63P']
2	['W62V', 'F63V']
2	['W62V', 'F63A']
2	['W62D', 'F63N']
2	['W62S', 'F63R']
3	['F63P', 'F73I']
3	['F63R', 'F73M']

3	['F63W', 'F73W']
3	['F63H', 'F73R']
3	['F63Q', 'F73H']
3	['F63Y', 'F73C']
3	['F63N', 'F73S']
3	['F63A', 'F73T']
3	['F63D', 'F73W']
3	['F63P', 'F73T']
3	['F63D', 'F73V']
3	['F63P', 'F73S']
3	['F63T', 'F73Y']
3	['F63D', 'F73S']
3	['F63M', 'F73R']
3	['F63E', 'F73V']
3	['F63T', 'F73C']
3	['F63Y', 'F73P']
3	['F63G', 'F73E']
3	['F63Q', 'F73K']
3	['F63S', 'F73I']
3	['F63L', 'F73Y']
4	['F73G', 'L86P']
4	['F73G', 'L86N']
4	['F73R', 'L86Y']
4	['F73A', 'L86V']
4	['F73V', 'L86A']
4	['F73V', 'L86T']
4	['F73L', 'L86T']
4	['F73P', 'L86T']
4	['F73R', 'L86V']
4	['F73H', 'L86T']
4	['F73H', 'L86D']
4	['F73K', 'L86S']
4	['F73D', 'L86T']

4	['F73L', 'L86S']
4	['F73N', 'L86I']
4	['F73Y', 'L86K']
4	['F73R', 'L86N']
4	['F73K', 'L86Q']
4	['F73P', 'L86M']
4	['F73D', 'L86A']
4	['F73T', 'L86Y']
4	['F73P', 'L86V']
5	['L86H', 'S90R']
5	['L86I', 'S90L']
5	['L86V', 'S90Y']
5	['L86P', 'S90Y']
5	['L86I', 'S90Y']
5	['L86F', 'S90F']
5	['L86S', 'S90P']
5	['L86K', 'S90L']
5	['L86Q', 'S90T']
5	['L86T', 'S90Y']
5	['L86F', 'S90H']
5	['L86S', 'S90A']
5	['L86C', 'S90H']
5	['L86D', 'S90D']
5	['L86D', 'S90F']
5	['L86C', 'S90C']
5	['L86N', 'S90I']
5	['L86M', 'S90T']
5	['L86F', 'S90D']
5	['L86P', 'S90P']
5	['L86T', 'S90N']
5	['L86V', 'S90I']
6	['S90L', 'F93D']
6	['S90H', 'F93G']

6	['S90N', 'F93T']
6	['S90H', 'F93R']
6	['S90A', 'F93R']
6	['S90E', 'F93R']
6	['S90H', 'F93P']
6	['S90Q', 'F93S']
6	['S90L', 'F93I']
6	['S90Q', 'F93G']
6	['S90R', 'F93E']
6	['S90C', 'F93L']
6	['S90H', 'F93Y']
6	['S90N', 'F93L']
6	['S90V', 'F93L']
6	['S90L', 'F93W']
6	['S90D', 'F93H']
6	['S90N', 'F93V']
6	['S90F', 'F93R']
6	['S90H', 'F93M']
6	['S90W', 'F93G']
6	['S90P', 'F93I']
7	['F93K', 'S149K']
7	['F93E', 'S149H']
7	['F93T', 'S149T']
7	['F93Q', 'S149V']
7	['F93P', 'S149T']
7	['F93R', 'S149P']
7	['F93V', 'S149Y']
7	['F93C', 'S149H']
7	['F93T', 'S149G']
7	['F93S', 'S149F']
7	['F93Y', 'S149R']
7	['F93S', 'S149N']
7	['F93W', 'S149I']

7	['F93L', 'S149A']
7	['F93R', 'S149T']
7	['F93Y', 'S149L']
7	['F93D', 'S149G']
7	['F93S', 'S149L']
7	['F93H', 'S149I']
7	['F93T', 'S149Q']
7	['F93M', 'S149V']
7	['F93P', 'S149I']
8	['L59T', 'S149G']
8	['L59Q', 'S149G']
8	['L59V', 'S149W']
8	['L59F', 'S149T']
8	['L59H', 'S149M']
8	['L59K', 'S149L']
8	['L59Y', 'S149K']
8	['L59S', 'S149W']
8	['L59N', 'S149R']
8	['L59T', 'S149L']
8	['L59R', 'S149M']
8	['L59G', 'S149R']
8	['L59Q', 'S149R']
8	['L59T', 'S149H']
8	['L59I', 'S149Q']
8	['L59Q', 'S149H']
8	['L59R', 'S149V']
8	['L59R', 'S149R']
8	['L59W', 'S149V']
8	['L59C', 'S149G']
8	['L59H', 'S149V']
8	['L59V', 'S149I']
9	['L59M', 'F73R']
9	['L59D', 'F73Y']

9	['L59V', 'F73H']
9	['L59R', 'F73L']
9	['L59G', 'F73K']
9	['L59S', 'F73E']
9	['L59F', 'F73N']
9	['L59S', 'F73V']
9	['L59R', 'F73R']
9	['L59S', 'F73S']
9	['L59P', 'F73A']
9	['L59C', 'F73M']
9	['L59I', 'F73A']
9	['L59H', 'F73N']
9	['L59A', 'F73R']
9	['L59Y', 'F73E']
9	['L59S', 'F73Q']
9	['L59G', 'F73L']
9	['L59T', 'F73H']
9	['L59P', 'F73L']
9	['L59T', 'F73P']
9	['L59Y', 'F73K']
10	['F73L', 'S90M']
10	['F73S', 'S90R']
10	['F73S', 'S90F']
10	['F73T', 'S90T']
10	['F73S', 'S90L']
10	['F73W', 'S90G']
10	['F73L', 'S90Y']
10	['F73V', 'S90G']
10	['F73T', 'S90R']
10	['F73K', 'S90I']
10	['F73A', 'S90W']
10	['F73Y', 'S90R']
10	['F73M', 'S90Q']

10	['F73N', 'S90C']
10	['F73T', 'S90V']
10	['F73T', 'S90Y']
10	['F73L', 'S90K']
10	['F73V', 'S90P']
10	['F73R', 'S90T']
10	['F73L', 'S90P']
10	['F73Y', 'S90L']
10	['F73N', 'S90A']
11	['L86S', 'F93V']
11	['L86K', 'F93V']
11	['L86K', 'F93C']
11	['L86R', 'F93T']
11	['L86S', 'F93T']
11	['L86P', 'F93Y']
11	['L86S', 'F93W']
11	['L86I', 'F93W']
11	['L86S', 'F93H']
11	['L86A', 'F93R']
11	['L86D', 'F93C']
11	['L86A', 'F93Y']
11	['L86V', 'F93G']
11	['L86A', 'F93I']
11	['L86N', 'F93S']
11	['L86R', 'F93D']
11	['L86A', 'F93M']
11	['L86N', 'F93A']
11	['L86E', 'F93G']
11	['L86N', 'F93L']
11	['L86N', 'F93G']
11	['L86C', 'F93V']
12	['L59R', 'F93L']
12	['L59S', 'F93M']

12	['L59T', 'F93Y']
12	['L59Q', 'F93G']
12	['L59Y', 'F93G']
12	['L59W', 'F93C']
12	['L59M', 'F93L']
12	['L59S', 'F93D']
12	['L59E', 'F93A']
12	['L59C', 'F93Y']
12	['L59V', 'F93L']
12	['L59Q', 'F93K']
12	['L59V', 'F93P']
12	['L59E', 'F93G']
12	['L59C', 'F93I']
12	['L59I', 'F93V']
12	['L59R', 'F93Y']
12	['L59V', 'F93K']
12	['L59S', 'F93C']
12	['L59H', 'F93P']
12	['L59T', 'F93W']
12	['L59A', 'F93S']
13	['W62D', 'F63G', 'L86H']
13	['W62R', 'F63Y', 'L86N']
13	['W62H', 'F63R', 'L86V']
13	['W62L', 'F63G', 'L86H']
13	['W62V', 'F63S', 'L86F']
13	['W62Q', 'F63K', 'L86G']

13	['W62V', 'F63K', 'L86M']
13	['W62A', 'F63I', 'L86A']
13	['W62S', 'F63S', 'L86K']
13	['W62S', 'F63V', 'L86T']
13	['W62L', 'F63D', 'L86N']
13	['W62P', 'F63M', 'L86G']
13	['W62H', 'F63P', 'L86V']
13	['W62G', 'F63H', 'L86I']
13	['W62A', 'F63S', 'L86D']
13	['W62P', 'F63L', 'L86R']
13	['W62R', 'F63V', 'L86N']
13	['W62S', 'F63A', 'L86S']
13	['W62P', 'F63V', 'L86R']
13	['W62E', 'F63N', 'L86I']
13	['W62P', 'F63W', 'L86Q']
13	['W62F', 'F63A', 'L86D']
13	['W62R', 'F63L', 'L86R']

13	['W62L', 'F63G', 'L86K']
13	['W62N', 'F63M', 'L86V']
13	['W62R', 'F63C', 'L86F']
13	['W62Q', 'F63G', 'L86S']
13	['W62L', 'F63H', 'L86G']
13	['W62T', 'F63T', 'L86D']
13	['W62R', 'F63S', 'L86V']
13	['W62A', 'F63D', 'L86H']
13	['W62S', 'F63R', 'L86V']
13	['W62F', 'F63R', 'L86H']
13	['W62I', 'F63R', 'L86K']
13	['W62V', 'F63L', 'L86G']
13	['W62C', 'F63V', 'L86W']
13	['W62F', 'F63Y', 'L86G']
13	['W62K', 'F63R', 'L86E']
13	['W62Y', 'F63E', 'L86V']
13	['W62K', 'F63S', 'L86S']

13	['W62A', 'F63T','L86G']
13	['W62A', 'F63P', 'L86S']
13	['W62S', 'F63N', 'L86V']
13	['W62S', 'F63S', 'L86Y']
14	['L59S', 'W62C', 'F63I']
14	['L59V', 'W62I', 'F63S']
14	['L59S', 'W62V', 'F63S']
14	['L59K', 'W62V', 'F63S']
14	['L59P', 'W62C', 'F63N']
14	['L59C', 'W62Q', 'F63A']
14	['L59N', 'W62R', 'F63E']
14	['L59T', 'W62F', 'F63A']
14	['L59H', 'W62D', 'F63S']
14	['L59C', 'W62D', 'F63K']
14	['L59N', 'W62R', 'F63S']
14	['L59G', 'W62R', 'F63T']
14	['L59R', 'W62E', 'F63V']

14	['L59W', 'W62G', 'F63G']
14	['L59H', 'W62F', 'F63L']
14	['L59Y', 'W62M', 'F63N']
14	['L59Q', 'W62Y', 'F63L']
14	['L59T', 'W62K', 'F63R']
14	['L59P', 'W62R', 'F63T']
14	['L59M', 'W62N', 'F63G']
14	['L59T', 'W62G', 'F63M']
14	['L59V', 'W62A', 'F63R']
14	['L59P', 'W62I', 'F63Q']
14	['L59R', 'W62L', 'F63K']
14	['L59R', 'W62Y', 'F63E']
14	['L59N', 'W62L', 'F63E']
14	['L59W', 'W62E', 'F63G']
14	['L59E', 'W62R', 'F63L']
14	['L59M', 'W62V', 'F63V']
14	['L59A', 'W62L', 'F63V']

14	['L59F', 'W62V', 'F63A']
14	['L59V', 'W62S', 'F63S']
14	['L59S', 'W62A', 'F63C']
14	['L59K', 'W62V', 'F63G']
14	['L59R', 'W62P', 'F63L']
14	['L59P', 'W62R', 'F63N']
14	['L59I', 'W62R', 'F63N']
14	['L59Q', 'W62K', 'F63V']
14	['L59A', 'W62S', 'F63G']
14	['L59P', 'W62H', 'F63S']
14	['L59F', 'W62P', 'F63A']
14	['L59H', 'W62M', 'F63G']
14	['L59C', 'W62Q', 'F63T']
14	['L59C', 'W62G', 'F63G']
15	['L59R', 'F63N', 'S90C']
15	['L59P', 'F63K', 'S90F']
15	['L59H', 'F63P', 'S90G']
15	['L59T', 'F63I', 'S90L']
15	['L59Y', 'F63G', 'S90L']
15	['L59S', 'F63L', 'S90C']

15	['L59Y', 'F63N', 'S90P']
15	['L59T', 'F63V', 'S90T']
15	['L59S', 'F63T', 'S90V']
15	['L59N', 'F63R', 'S90C']
15	['L59N', 'F63L', 'S90H']
15	['L59F', 'F63A', 'S90F']
15	['L59F', 'F63V', 'S90L']
15	['L59E', 'F63H', 'S90N']
15	['L59D', 'F63S', 'S90T']
15	['L59N', 'F63A', 'S90L']
15	['L59P', 'F63T', 'S90H']
15	['L59G', 'F63G', 'S90T']
15	['L59V', 'F63V', 'S90W']
15	['L59S', 'F63A', 'S90N']
15	['L59W', 'F63P', 'S90N']
15	['L59G', 'F63P', 'S90R']
15	['L59V', 'F63K', 'S90G']
15	['L59D', 'F63P', 'S90R']
15	['L59E', 'F63D', 'S90T']
15	['L59R', 'F63R', 'S90Q']
15	['L59V', 'F63L', 'S90I']
15	['L59V', 'F63S', 'S90D']
15	['L59T', 'F63V', 'S90F']
15	['L59M', 'F63D', 'S90Y']
15	['L59R', 'F63C', 'S90P']
15	['L59S', 'F63E', 'S90L']
15	['L59H', 'F63L', 'S90D']
15	['L59H', 'F63H', 'S90R']
15	['L59E', 'F63V', 'S90F']

15	['L59W', 'F63E', 'S90C']
15	['L59V', 'F63A', 'S90C']
15	['L59F', 'F63M', 'S90E']
15	['L59F', 'F63A', 'S90P']
15	['L59I', 'F63S', 'S90L']
15	['L59Q', 'F63D', 'S90A']
15	['L59G', 'F63R', 'S90F']
15	['L59I', 'F63D', 'S90V']
15	['L59D', 'F63E', 'S90M']
16	['L59V', 'S90L', 'F93P']
16	['L59H', 'S90Y', 'F93L']
16	['L59K', 'S90G', 'F93S']
16	['L59P', 'S90P', 'F93S']
16	['L59Y', 'S90T', 'F93I']
16	['L59K', 'S90H', 'F93R']
16	['L59R', 'S90T', 'F93Y']
16	['L59V', 'S90P', 'F93N']
16	['L59I', 'S90T', 'F93V']
16	['L59G', 'S90E', 'F93L']
16	['L59Q', 'S90L', 'F93V']
16	['L59F', 'S90N', 'F93L']
16	['L59R', 'S90P', 'F93S']
16	['L59R', 'S90I', 'F93H']
16	['L59D', 'S90L', 'F93I']
16	['L59F', 'S90N', 'F93P']
16	['L59G', 'S90R', 'F93N']
16	['L59S', 'S90D', 'F93M']
16	['L59V', 'S90I', 'F93Y']
16	['L59D', 'S90Q', 'F93C']

16	['L59H', 'S90R', 'F93R']
16	['L59F', 'S90Y', 'F93V']
16	['L59S', 'S90P', 'F93Y']
16	['L59G', 'S90M', 'F93W']
16	['L59G', 'S90E', 'F93M']
16	['L59P', 'S90I', 'F93R']
16	['L59D', 'S90W', 'F93L']
16	['L59G', 'S90T', 'F93L']
16	['L59K', 'S90F', 'F93Q']
16	['L59H', 'S90Q', 'F93M']
16	['L59H', 'S90H', 'F93V']
16	['L59F', 'S90E', 'F93S']
16	['L59G', 'S90L', 'F93L']
16	['L59V', 'S90D', 'F93L']
16	['L59S', 'S90L', 'F93H']
16	['L59V', 'S90R', 'F93L']
16	['L59K', 'S90P', 'F93P']
16	['L59G', 'S90Q', 'F93L']
16	['L59Q', 'S90T', 'F93A']
16	['L59H', 'S90R', 'F93S']
16	['L59G', 'S90G', 'F93V']
16	['L59T', 'S90A', 'F93D']
16	['L59C', 'S90N', 'F93D']
16	['L59N', 'S90G', 'F93S']
17	['W62C', 'F63G', 'S149M']

17	['W62N', 'F63L', 'S149Q']
17	['W62L', 'F63E', 'S149L']
17	['W62I', 'F63L', 'S149C']
17	['W62S', 'F63K', 'S149G']
17	['W62T', 'F63G', 'S149K']
17	['W62P', 'F63C', 'S149M']
17	['W62M', 'F63S', 'S149W']
17	['W62L', 'F63S', 'S149G']
17	['W62L', 'F63R', 'S149H']
17	['W62P', 'F63K', 'S149V']
17	['W62H', 'F63D', 'S149C']
17	['W62P', 'F63H', 'S149V']
17	['W62G', 'F63L', 'S149N']
17	['W62L', 'F63T', 'S149F']
17	['W62R', 'F63D', 'S149Y']
17	['W62N', 'F63V', 'S149P']

17	['W62H', 'F63S', 'S149D']
17	['W62G', 'F63S', 'S149N']
17	['W62P', 'F63P', 'S149F']
17	['W62Y', 'F63H', 'S149Y']
17	['W62E', 'F63A', 'S149P']
17	['W62C', 'F63N', 'S149W']
17	['W62T', 'F63P', 'S149F']
17	['W62L', 'F63L', 'S149L']
17	['W62L', 'F63A', 'S149I']
17	['W62T', 'F63L', 'S149M']
17	['W62L', 'F63A', 'S149G']
17	['W62H', 'F63S', 'S149G']
17	['W62R', 'F63C', 'S149R']
17	['W62I', 'F63W', 'S149V']
17	['W62V', 'F63Q', 'S149R']
17	['W62C', 'F63S', 'S149Y']

17	['W62T', 'F63S', 'S149T']
17	['W62M', 'F63S', 'S149I']
17	['W62K', 'F63P', 'S149G']
17	['W62R', 'F63A', 'S149I']
17	['W62V', 'F63Y', 'S149G']
17	['W62D', 'F63S', 'S149L']
17	['W62Y', 'F63N', 'S149L']
17	['W62A', 'F63E', 'S149C']
17	['W62S', 'F63W', 'S149R']
17	['W62L', 'F63G', 'S149A']
17	['W62P', 'F63M', 'S149K']
18	['L86V', 'S90N', 'S149T']
18	['L86E', 'S90N', 'S149P']
18	['L86S', 'S90Y', 'S149H']
18	['L86D', 'S90A', 'S149Y']
18	['L86S', 'S90C', 'S149K']

18	['L86V', 'S90M', 'S149H']
18	['L86K', 'S90I', 'S149Y']
18	['L86I', 'S90V', 'S149N']
18	['L86V', 'S90R', 'S149V']
18	['L86F', 'S90H', 'S149N']
18	['L86P', 'S90H', 'S149N']
18	['L86G', 'S90R', 'S149H']
18	['L86R', 'S90A', 'S149R']
18	['L86G', 'S90V', 'S149I']
18	['L86K', 'S90F', 'S149D']
18	['L86I', 'S90L', 'S149V']
18	['L86R', 'S90H', 'S149M']
18	['L86V', 'S90A', 'S149V']
18	['L86P', 'S90V', 'S149H']
18	['L86I', 'S90C', 'S149N']
18	['L86S', 'S90P', 'S149M']

18	['L86Y', 'S90A', 'S149M']
18	['L86M', 'S90R', 'S149G']
18	['L86Y', 'S90L', 'S149I']
18	['L86F', 'S90V', 'S149D']
18	['L86Y', 'S90F', 'S149F']
18	['L86C', 'S90R', 'S149M']
18	['L86R', 'S90C', 'S149G']
18	['L86N', 'S90T', 'S149K']
18	['L86R', 'S90M', 'S149P']
18	['L86Y', 'S90C', 'S149T']
18	['L86A', 'S90T', 'S149V']
18	['L86I', 'S90H', 'S149L']
18	['L86S', 'S90N', 'S149V']
18	['L86I', 'S90L', 'S149A']
18	['L86G', 'S90E', 'S149R']
18	['L86P', 'S90H', 'S149T']

18	['L86R', 'S90C', 'S149V']
18	['L86V', 'S90N', 'S149L']
18	['L86S', 'S90V', 'S149I']
18	['L86D', 'S90A', 'S149W']
18	['L86Y', 'S90L', 'S149P']
18	['L86S', 'S90Y', 'S149R']
18	['L86P', 'S90C', 'S149W']

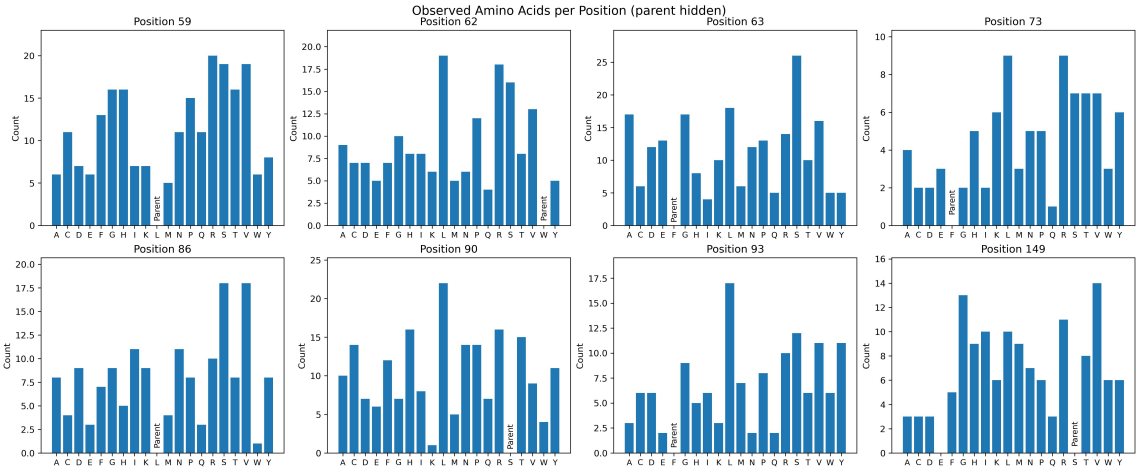


Figure D-21. Counts of amino acid identities observed at each position in variants selected for the initial training library. The only amino-acid substitution not observed in training was S149E.

D.3.2. Protocols for the Cloning of MLDE Predicted Sequences

D.3.2.1. Assembly of Plasmids Bearing Predicted Mutant Genes

Ninety-six multi-mutation variants were generated at each round of predictions. For each round, the DNA sequences for these mutants were achieved through mutagenesis of the target codons in the parent DNA sequence, and an oligo sequence (300 bp) was generated for each variant, containing mutations at up to all eight positions of interest (**Table D-2**). All 96 oligo sequences were synthesized and delivered by Twist Bioscience (South San Francisco, CA) as a single pooled sample for each round. Oligo pools were received as dry residues which were reconstituted in 10 mM Tris–HCl (pH=8.0) to a final oligo concentration of 10 ng/μL. Oligo pools were then amplified by PCR using the primers described in **Table D-6**. This fragment library was assembled with a pET-22b(+) backbone with overhangs designed for Gibson ligation to generate fully encoded protoglobin sequences bearing exact sets of desired mutations. The assembly products obtained were used to transform T7 Express Competent *E. coli* (High Efficency) cells. Upon heat-shock, freshly transformed *E. coli* cells were recovered in 0.4 mL Luria-Bertani medium (LB) (Research Products Int.) at 37 ºC with shaking at 220 rpm for 30 minutes. This transformation mixture was directly plated on LB-Amp agar plates. The plates were incubated overnight at 37 °C until colony formation was observed. For each round of predictions between 350–550 colonies from LB-Amp agar plates were picked with sterilized toothpicks to individually inoculate the wells of 2-mL 96-well deep-well plates charged with 400 μL of LB-Amp. The plates were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. The following morning, 50 μL of preculture from each well were added to the wells of a 96-well flat-bottom tissue culture plate (ThermoFisher) preloaded with 50 μL of 50% glycerol solution. These glycerol stocks were stored at -80 °C for future inoculation. Additionally, the sequences of protoglobin genes contained in every well were sequenced using LevSeq sequencing.[5] In total, 63/96 predicted variants were assembled for the first round of oligo pool assembly, and 64/96 were found for achieved in the second round (**Table D-10**). The target variants from each round were rearrayed into the wells of a 96-well plate, along with 4-5 parent wells (PromPgb), one TrpB well, and one sterile well. These plates were stored as glycerol stocks at -80 °C for future inoculation.

Table D-10. PromPgb mutants which were predicted using MLDE to be able to access a broader range of new-to-nature reactions and were successfully cloned using oligo pools.

Round	**Mutant**
Predictions Round 1	CYVFMIFQ
Predictions Round 1	YAIFIIFQ
Predictions Round 1	YALFIVFQ
Predictions Round 1	YYIFMIFQ
Predictions Round 1	YYLFIIFQ
Predictions Round 1	TAVLMGKQ
Predictions Round 1	YYVFMCKQ
Predictions Round 1	QYVFMCFQ
Predictions Round 1	LYLFICKQ
Predictions Round 1	VALFMCVQ
Predictions Round 1	FNIFMAFQ
Predictions Round 1	FYFFMCFQ
Predictions Round 1	FYIFMMFQ
Predictions Round 1	LYIFIMKQ
Predictions Round 1	YAIFMIFQ
Predictions Round 1	FNIFMAKQ
Predictions Round 1	FVIFMVFQ
Predictions Round 1	YAVAIIKQ
Predictions Round 1	TDVFITFQ
Predictions Round 1	FYIFKMKQ
Predictions Round 1	YAVAIVKQ
Predictions Round 1	YAVFIVFQ
Predictions Round 1	VYIAIVFQ
Predictions Round 1	LYLFMMKQ
Predictions Round 1	YYIAIVFQ
Predictions Round 1	FYVFMMFQ
Predictions Round 1	LYIFIVKQ
Predictions Round 1	YYIAMIFQ
Predictions Round 1	YAVVMTFQ

Predictions Round 1	YAIFIVFQ
Predictions Round 1	YYVAICFQ
Predictions Round 1	VYIFMVFQ
Predictions Round 1	YYIFIMFQ
Predictions Round 1	YAVFMTFQ
Predictions Round 1	YYLFICFQ
Predictions Round 1	YYVFMVKQ
Predictions Round 1	FYVFITFQ
Predictions Round 1	IYVAIIKQ
Predictions Round 1	LYLFIIKQ
Predictions Round 1	LYLFIVKQ
Predictions Round 1	YYVFMIFQ
Predictions Round 1	LYVFIVFQ
Predictions Round 1	VALFMAFQ
Predictions Round 1	TAVLMIKQ
Predictions Round 1	TNVFITFQ
Predictions Round 1	TYVFICVQ
Predictions Round 1	TYVFITFQ
Predictions Round 1	YYVAICKQ
Predictions Round 1	YYVFIVFQ
Predictions Round 1	YYVAMIFQ
Predictions Round 1	YYVAIVKQ
Predictions Round 1	YYVFICFQ
Predictions Round 1	VALAMCVQ
Predictions Round 1	YAIFIIFA
Predictions Round 1	TYVFMIFQ
Predictions Round 1	TYVFMCVQ
Predictions Round 1	TYVFMCFQ
Predictions Round 1	TYVFICFQ
Predictions Round 1	TDVFTCFQ
Predictions Round 1	TAVFMCVQ
Predictions Round 1	LYVAIVFQ
Predictions Round 1	YALFMCVQ

Predictions Round 1	YYVLMIFQ
Predictions Round 2	IAFFMAFQ
Predictions Round 2	IAFFMVFN
Predictions Round 2	VPFFMIFN
Predictions Round 2	LAFFIAFN
Predictions Round 2	VGFFMVFQ
Predictions Round 2	LAFFMAFQ
Predictions Round 2	LYFFMAFN
Predictions Round 2	VAFFIAFN
Predictions Round 2	TAFFMAFN
Predictions Round 2	TNVFMIFN
Predictions Round 2	TAFFIAFN
Predictions Round 2	VEFFMAFN
Predictions Round 2	VGFFMAFN
Predictions Round 2	LAFFIVFN
Predictions Round 2	VPFFMAFQ
Predictions Round 2	LGFFMAVN
Predictions Round 2	LYFFMKFN
Predictions Round 2	VAVFIAFN
Predictions Round 2	LAFFMVFQ
Predictions Round 2	VAFFMAFQ
Predictions Round 2	TGFFMVFQ
Predictions Round 2	TNFFMVFQ
Predictions Round 2	LYFFMAFQ
Predictions Round 2	TGFFMVFN
Predictions Round 2	LYFFMVFN
Predictions Round 2	TNFFMKFN
Predictions Round 2	TNFFMVFN
Predictions Round 2	TGFFMVYQ
Predictions Round 2	VVFFIAFN
Predictions Round 2	TAFFMAVN
Predictions Round 2	TAFFMVVN
Predictions Round 2	TAFFMIFN

Predictions Round 2	LAFFQVFN
Predictions Round 2	LAFFQAFN
Predictions Round 2	LAFFMAYN
Predictions Round 2	LNFFMVFQ
Predictions Round 2	LYFFMVFQ
Predictions Round 2	LAFFIAFQ
Predictions Round 2	VYFFMAFQ
Predictions Round 2	VYFFMVFN
Predictions Round 2	TAIFMVFQ
Predictions Round 2	VVFFMAFN
Predictions Round 2	VVFFMAFQ
Predictions Round 2	VAFFMVFN
Predictions Round 2	TVFFMVFN
Predictions Round 2	IAQFMAFN
Predictions Round 2	IAFFMAFN
Predictions Round 2	TAFFMVFN
Predictions Round 2	TAFFMAFQ
Predictions Round 2	TNFFMMFN
Predictions Round 2	LNFFIVFN
Predictions Round 2	LGFFMAFQ
Predictions Round 2	LAVFIAFN
Predictions Round 2	VAFFMAFN
Predictions Round 2	VAFFMVFQ
Predictions Round 2	LAFFMVFN
Predictions Round 2	LAFFMAFN
Predictions Round 2	VAFFIAFQ
Predictions Round 2	TNAFMIFN
Predictions Round 2	VYFFMAFN
Predictions Round 2	TAFFMVFQ
Predictions Round 2	TAQFMAFN
Predictions Round 2	LAFFMKFN
Predictions Round 2	VYFFMVFQ

D.4. Protoglobin Library Screening Protocols

D.4.1. Protocols for the Screening of Protoglobin Variants in 96-Well Plate Format

D.4.1.1. 96-Well Plate Library Expression

The wells of a 2-mL 96-well deep-well plate were filled with 400 µL LB-Amp. Previously generated 96-well plate glycerol stocks were removed from -80 °C storage and placed on dry ice. Multichannel pipet tips were used to scratch the frozen glycerol stock surface and used to inoculate the aforementioned deep-well plate. These overnight cultures were incubated at 37 ºC and shaken at 220 rpm for 16–18 hours. For expression cultures, the following morning 50 µL of the precultures were used to inoculate 900 µL TB-amp per well in 96-well deep-well plates. The expression cultures were initially incubated at 37 ºC and 220 rpm for 2.5 hours, at which point they were allowed to sit at room temperature for 30 minutes. Expression of proteins was induced with IPTG, and cellular heme production was increased with ALA. An induction mixture containing IPTG and ALA in TB-amp (50 µL) was added to each well such that the final concentrations of IPTG and ALA were 0.5 mM and 1.0 mM respectively. The total culture volumes were 1 mL per well. The plates were then incubated at 22 ºC and 220 rpm overnight.

D.4.1.2. 96-Well Plate Reaction Preparation

Expression cultures containing *E. coli* expressing hemoproteins of interest were centrifuged at 4,000 × g for 10 minutes at 4 ºC. The supernatant was discarded, and nitrogen-free M9 minimal media (M9-N, 380 µL for carbene transfer reactions, 365 µL for nitrene transfer reactions) were added to each well. For nitrene transfer reactions, 5 µL of a solution of D-glucose dissolved in M9-N were added such that the final concentration of D-glucose in reactions was 20 mM. The pellets were resuspended in the medium via shaking at room temperature for 30 minutes. The plates were then pumped into a vinyl Coy anaerobic chamber (0–30 ppm O_2). For carbene transfer reactions, to each well were added 20 µL of a MeCN solution containing the reaction substrate and ethyl diazoacetate (EDA).

The final reaction volume was 400 μL. For nitrene transfer reactions, to each well were added 20 μL of an EtOH solution containing the reaction substrate followed by 10 μL of an aqueous solution containing the desired nitrene precursor. The final reaction volume was 400 μL. The plates were then sealed carefully with a foil cover and shaken at room temperature for 16 hours in the Coy chamber.

D.4.1.3. 96-Well Plate Library Reaction Preparation for Analysis – Carbene Transfer

Once the reactions were complete, plates were worked up for processing by adding 600 μL of a 1:1 solution of ethyl acetate:cyclohexane containing 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). A silicone sealing mat (AWSM1003S, ArcticWhite) was used to cover each of the plates, and the two layers were thoroughly mixed by rapid inversion of the plates. The plates were then centrifuged (5,000 × g for 5 minutes at room temperature) to separate the phases.

For reactions **C1**–**C4**, which were always run in parallel with one another for library screening, 50 μL from each reaction plate were mixed in a GC vial insert in a GC vial, and the pooled samples were analyzed by GC-FID. For all other carbene transfer reactions, a 200-μL aliquot of the organic layer was transferred to a GC vial insert in a GC vial, and the samples were assayed by GC-FID or GC-MS.

D.4.1.4. 96-Well Plate Library Reaction Preparation for Analysis – Nitrene Transfer

Once the reactions were complete, plates were worked up for processing by adding 400 μL of MeCN. The plates were then sealed with foil and stored at -20 °C for 1 hour. The plates were then centrifuged (5,000 × g for 5 minutes at room temperature) to clarify the processed solution. Afterwards, a 200-μL aliquot of the combined layers was transferred to microtiter plate which was subsequently sealed by heat-sealing foil. These solutions were assayed by LC-MS.

D.5. Library Screening Details and Data

The following section contains descriptions on how data was generated for each reaction system analyzed in this work. **Table D-11** describes the manner in which data were collected for each reaction. Note that in some cases the data used to train our learning model in either training round in this work differs from the final data collected for that reaction. Both the datasets used for model training in this work and the final collected activity data are made available at https://github.com/fhalab/mrALDE. For future applications of these data to MLDE model training and validation studies we recommend using the final collected activity data. Detailed descriptions of the data collection procedures for each individual reaction system are provided below.

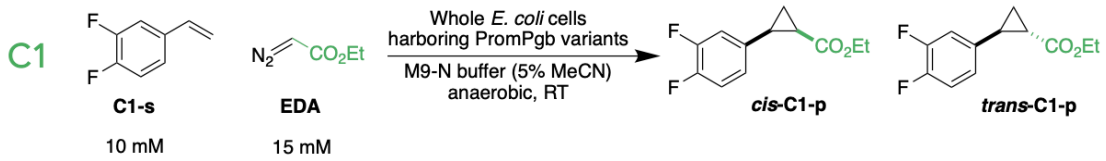
Table D-11. Overview of data collection methods for reactions investigated in this work. In some cases, data collected from the first round of predictions for model updating differed from a final round of data collection. A value of N/A in the "Data Collection Method – Final" column indicates that the same data were used for model training as are made available online.

Reaction Name	Data Collection Method – Training	Data Collection Method – Final	Notes
C1	GC-FID, pooled with reactions C2, C3, and C4	N/A	
C2	GC-FID, pooled with reactions C1, C3, and C4	N/A	
C3	GC-FID, pooled with reactions C1, C2, and C4	N/A	
C4	GC-FID, pooled with reactions C1, C2, and C3	N/A	
C5	GC-FID	N/A	
C6	GC-FID	N/A	
C7	GC-MS	N/A	
C8	GC-FID	N/A	
C9	GC-MS	N/A	
C10	LC-MS	N/A	
C11	GC-MS	N/A	
C12	GC-FID	N/A	Data for this reaction was not used to update the model after Round 1 Predictions

C13	GC-MS, pooled with reactions C14, C15, C16, and C17	GC-FID, unpooled	
C14	GC-MS, pooled with reactions C13, C15, C16, and C17	GC-MS, unpooled	
C15	GC-MS, pooled with reactions C13, C14, C16, and C17	GC-MS, unpooled	
C16	GC-MS, pooled with reactions C13, C14, C15, and C17	GC-MS, unpooled	
C17	GC-MS, pooled with reactions C13, C14, C15, and C16	GC-MS, unpooled	
N1	LC-MS	N/A	
N2	LC-MS	N/A	
N3	LC-MS	N/A	
N4	LC-MS	N/A	
N5	LC-MS	N/A	
N6	LC-MS	N/A	
N7	LC-MS	N/A	
N8	LC-MS		The exact product structure was not determined; activity was quantified based on an LC–MS peak corresponding to the expected [M+H]+ ion.
N9	LC-MS		During Round 1 data collection, several variants were initially assigned activity for reaction N9, and these measurements were incorporated into the model update. Subsequent reanalysis of these samples did not confirm product formation under the reported conditions.

Reaction C1:

Scheme D-1. Reaction conditions for reaction **C1** and expected products.



For library screening, reaction **C1** was analyzed by GC-FID using an equal part mixture of reaction extracts from reactions **C1**, **C2**, **C3**, and **C4**. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C1** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

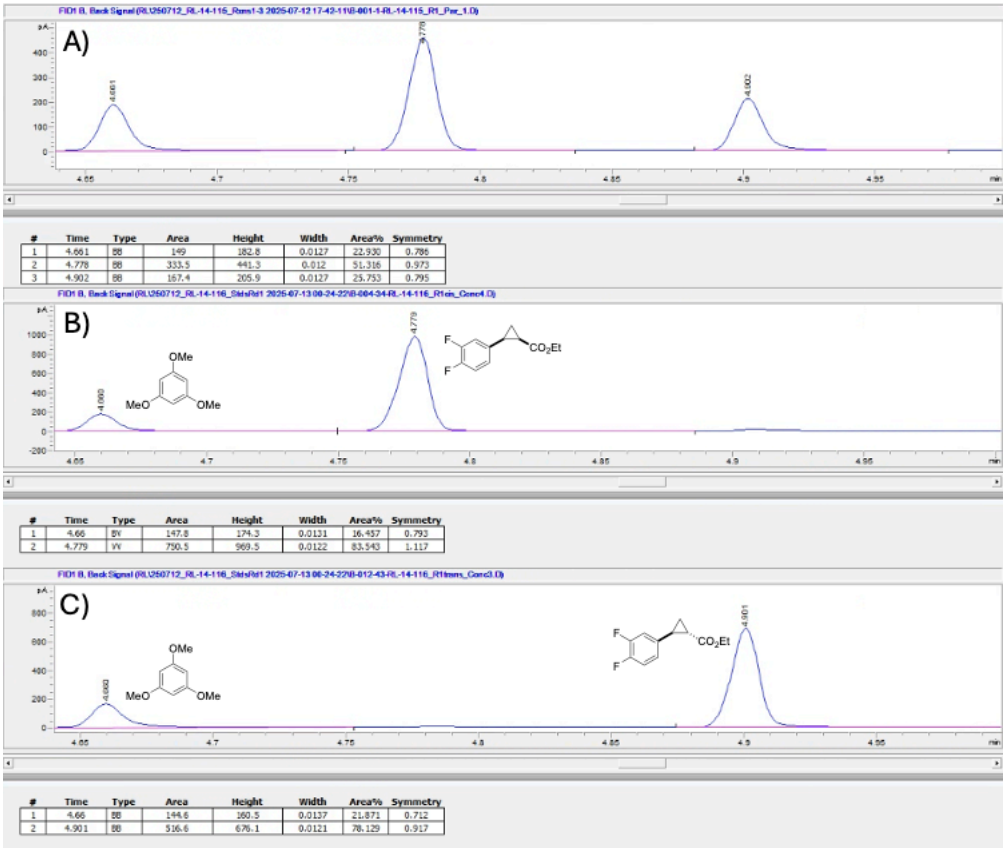


Figure D-22. (A) Representative GC-FID trace for reaction **C1** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of ***cis*-C1-p** with 1,3,5-trimethoxybenzene as an authentic standard. **(C)** GC-FID trace for a sample of the authentic standard of ***trans*-C1-p** with 1,3,5-trimethoxybenzene as an authentic standard.

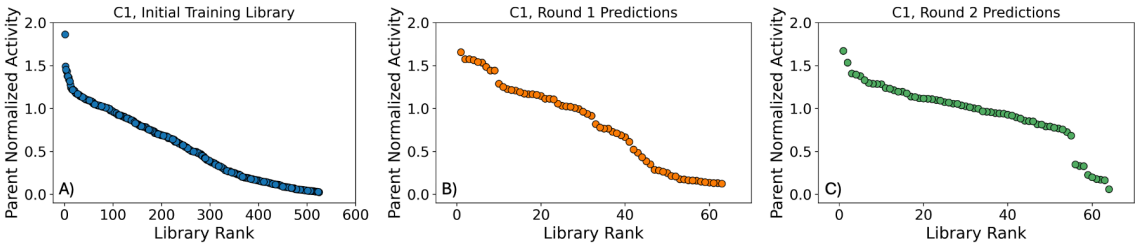


Figure D-23. Retention of function plots for reaction **C1** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

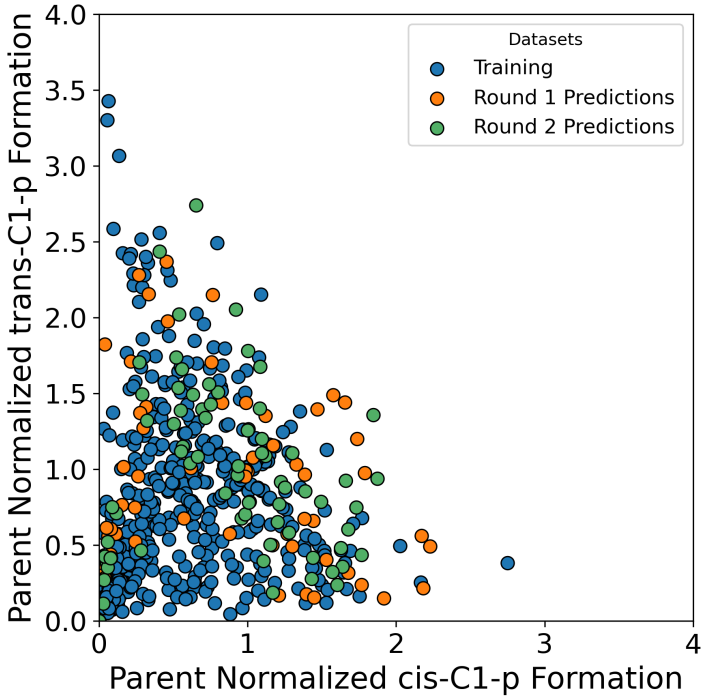
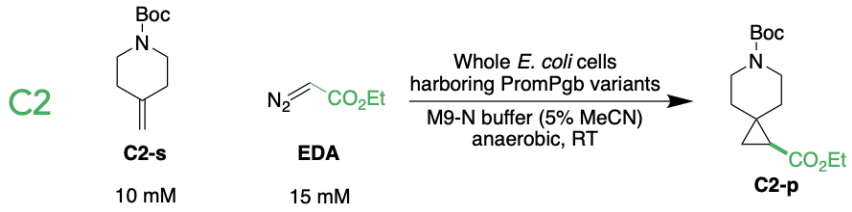


Figure D-24. Changes in the formation of ***cis*-C1-p** and ***trans*-C1-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C2:

Scheme S2. Reaction conditions for reaction **C2** and expected products.



For library screening, reaction **C2** was analyzed by GC-FID in an equal part mixture of reaction extracts from reactions **C1**, **C2**, **C3**, and **C4**. The authentic standard of **C2-p** was kindly provided by Dr. Jae L. Kennemur, who previously synthesized and characterized this compound in our laboratory. Full characterization data are reported in reference 7. Reaction **C2** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

Figure D-25. (A) Representative GC-FID trace for reaction **C2** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of **C2-p** with 1,3,5-trimethoxybenzene as an authentic standard.

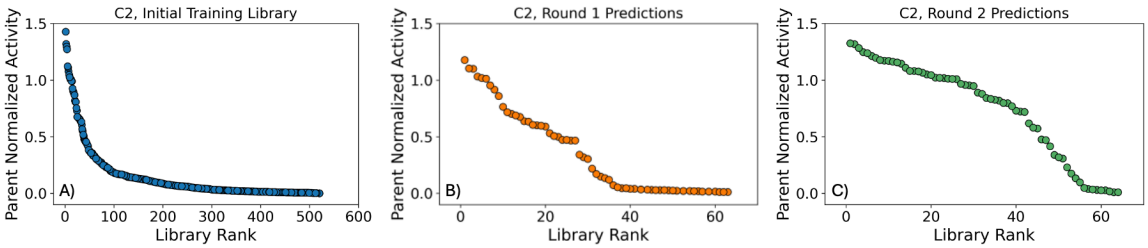
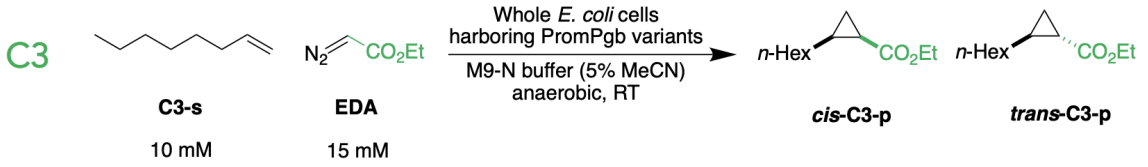


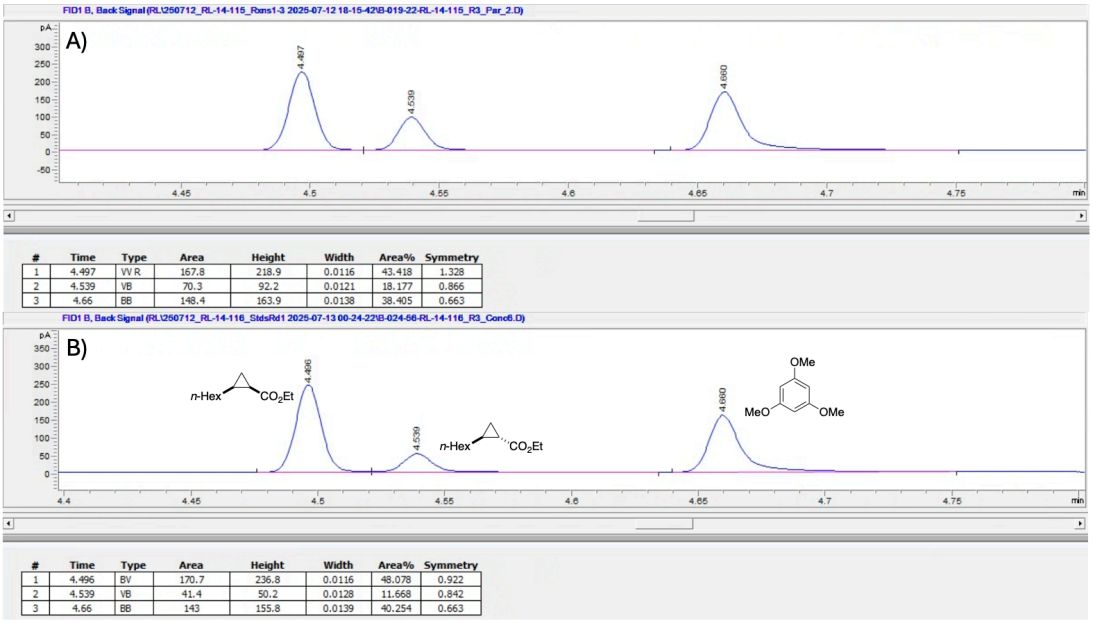
Figure D-26. Retention of function plots for reaction **C2** activity for variants in the **(A)** initial training library, **(B)** round 1 of predictions and **(C)** round 2 of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

Reaction C3:

Scheme D-3. Reaction conditions for reaction **C3** and expected products.



For library screening, reaction **C3** was analyzed by GC-FID in an equal part mixture of reaction extracts from reactions **C1**, **C2**, **C3**, and **C4**. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C3** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.497	VV R	167.8	218.9	0.0116	43.418	1.328
2	4.539	VB	70.3	92.2	0.0121	18.177	0.866
3	4.66	BB	148.4	163.9	0.0138	38.405	0.663

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.496	BV	170.7	236.8	0.0116	48.078	0.922
2	4.539	VB	41.4	50.2	0.0128	11.668	0.842
3	4.66	BB	143	155.8	0.0139	40.254	0.663

Figure D-27. (A) Representative GC-FID trace for reaction **C3** with PromPgb. **(B)** GC-FID trace for a sample of mixture of the authentic standards of ***cis*-C3-p** and ***trans*-C3-p** with 1,3,5-trimethoxybenzene as an authentic standard.

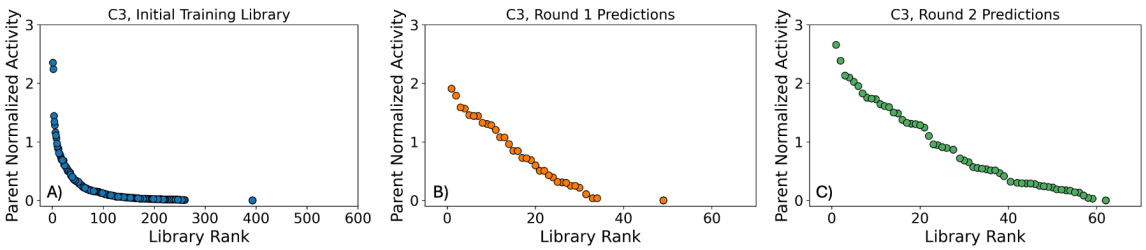


Figure D-28. Retention of function plots for reaction **C3** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

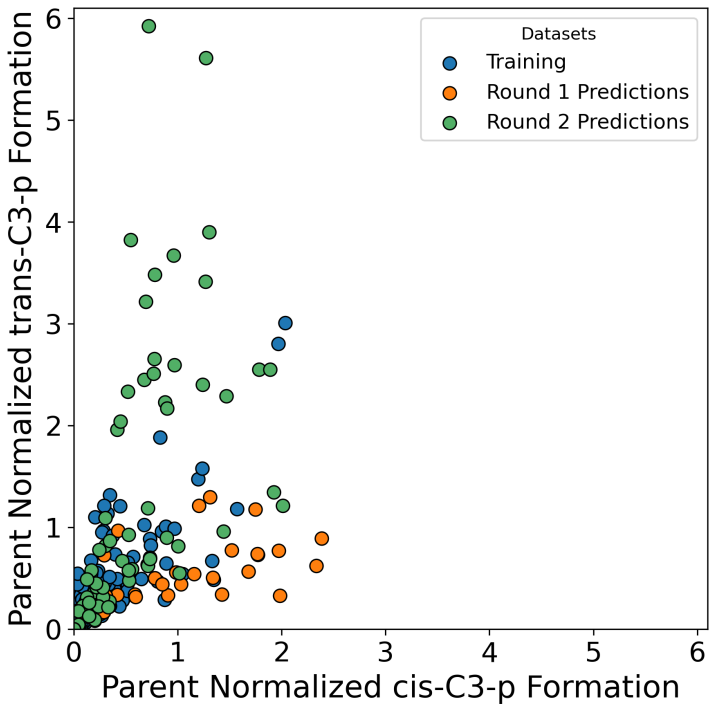
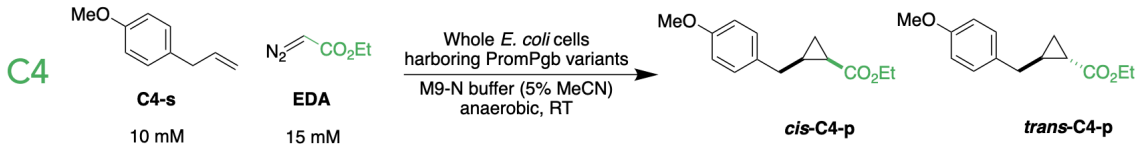


Figure D-29. Changes in the formation of ***cis*-C3-p** and ***trans*-C3-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C4:

Scheme D-4. Reaction conditions for reaction **C4** and expected products.



For library screening, reaction **C4** was analyzed by GC-FID in an equal part mixture of reaction extracts from reactions **C1**, **C2**, **C3**, and **C4**. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C4** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

Figure D-30. (A) Representative GC-FID trace for reaction **C4** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of ***cis*-C4-p** with 1,3,5-trimethoxybenzene as an authentic standard. **(C)** GC-FID trace for a sample of the authentic standard of ***trans*-C4-p** with 1,3,5-trimethoxybenzene as an authentic standard.

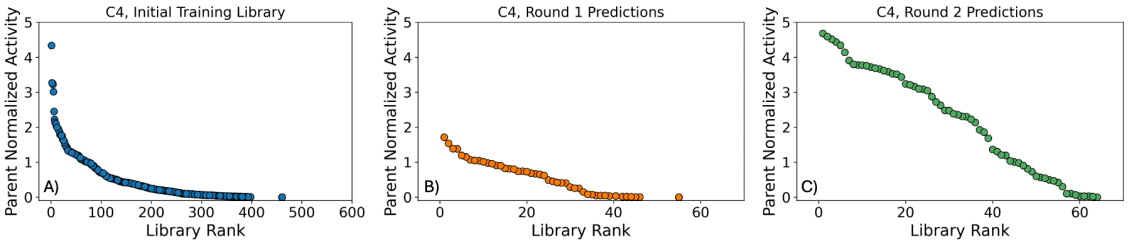


Figure D-31. Retention of function plots for reaction **C4** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

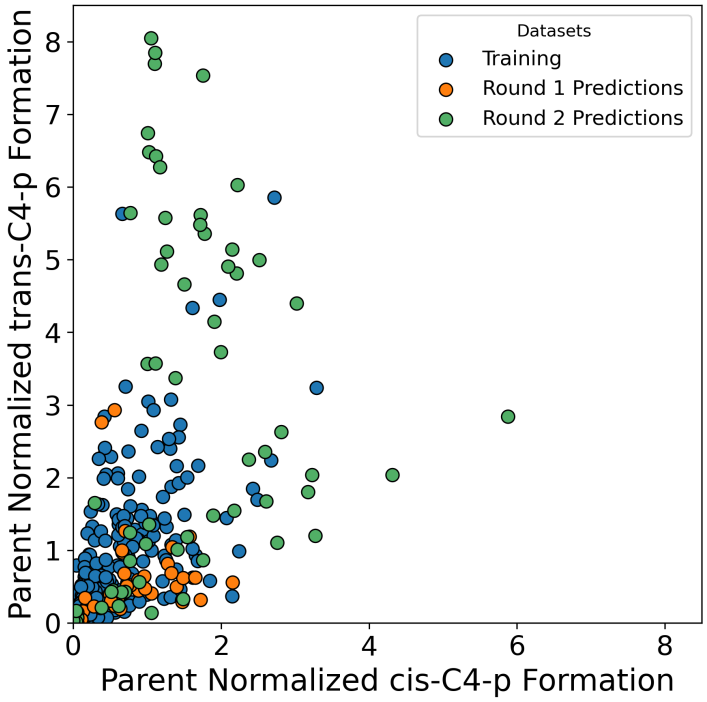
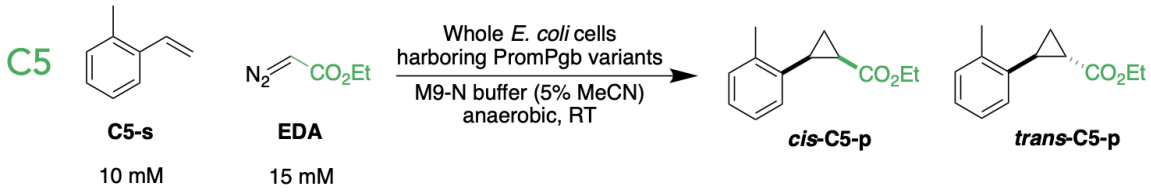


Figure D-32. Changes in the formation of ***cis*-C4-p** and ***trans*-C4-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C5:

Scheme D-5. Reaction conditions for reaction **C5** and expected products.



For library screening, reaction **C5** was analyzed by GC-FID. Authentic standards for this reaction were synthesized and characterized in our laboratory in a previous investigation. Full characterization data are reported in reference 8. Reaction **C5** was tested with round 2 of predictions.

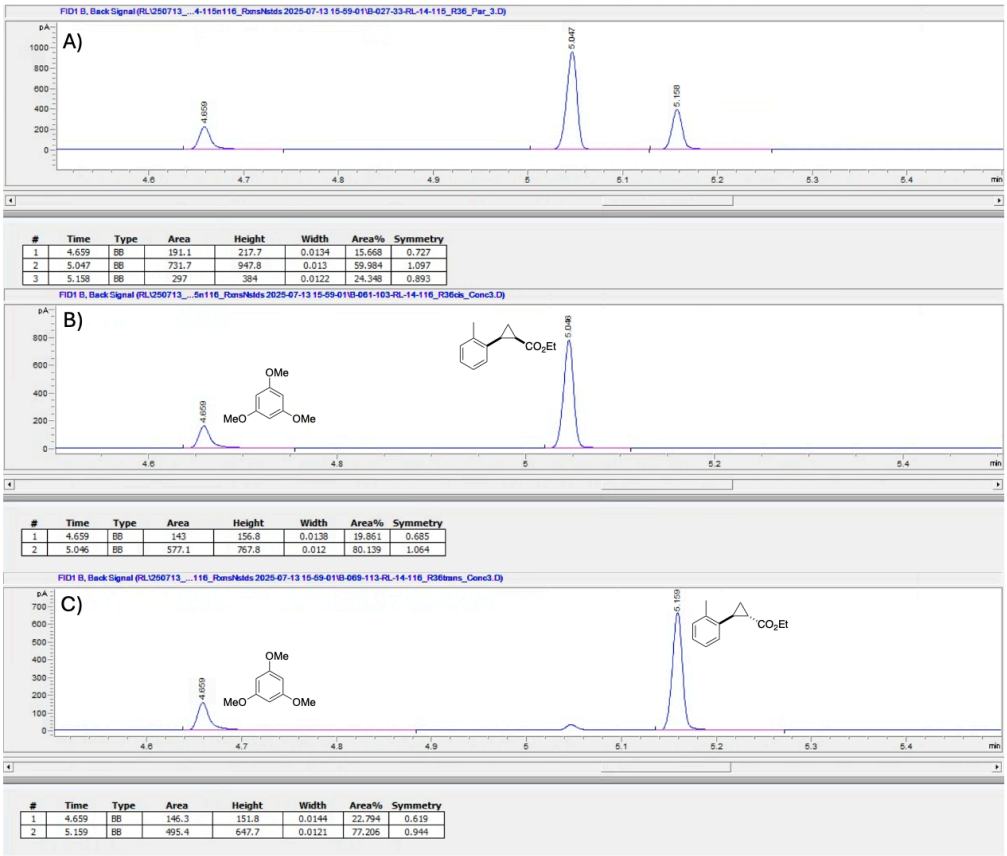


Figure D-33. (A) Representative GC-FID trace for reaction **C5** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of ***cis*-C5-p** with 1,3,5-trimethoxybenzene as an authentic standard. **(C)** GC-FID trace for a sample of the authentic standard of ***trans*-C5-p** with 1,3,5-trimethoxybenzene as an authentic standard.

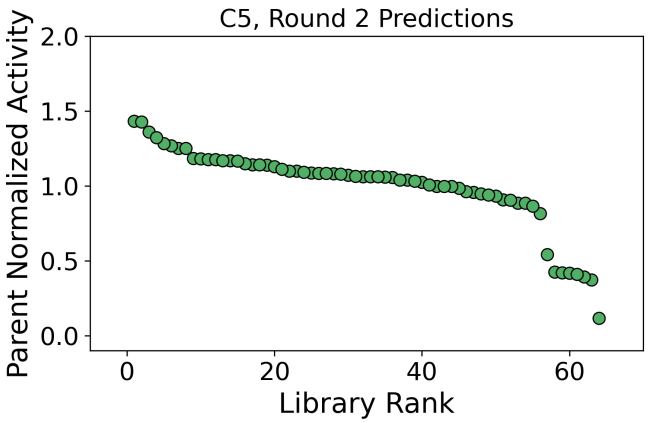


Figure D-34. Retention of function plot for reaction **C5** activity for variants in the second round of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

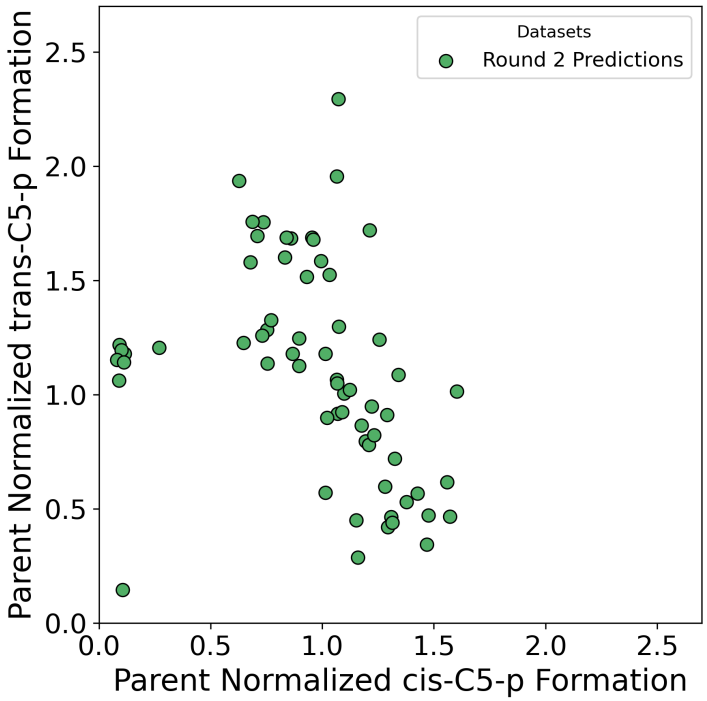
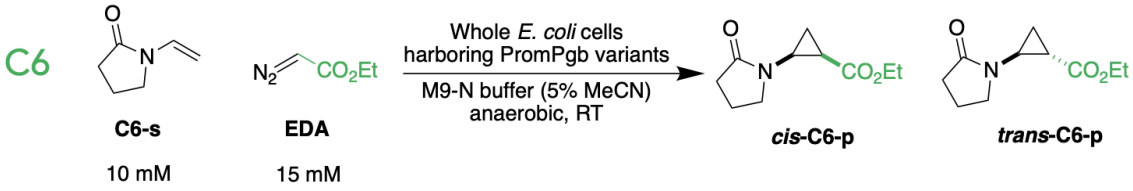


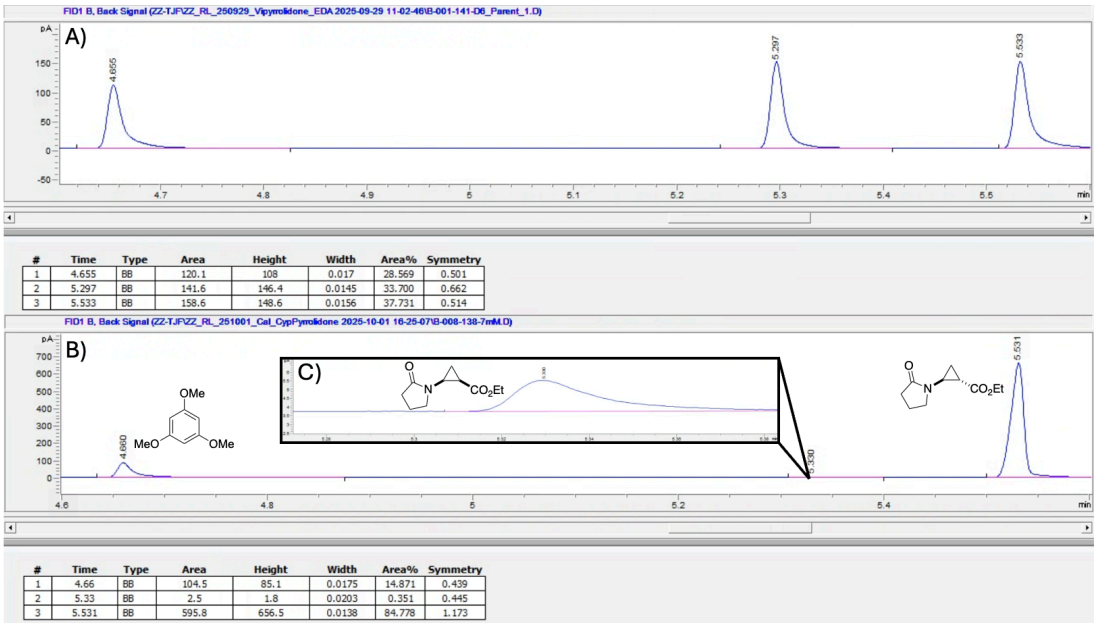
Figure D-35. Changes in the formation of ***cis*-C4-p** and ***trans*-C4-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C6:

Scheme D-6. Reaction conditions for reaction **C6** and expected products.



For library screening, reaction **C6** was analyzed by GC-FID. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C6** was tested with round 2 of predictions.



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.655	BB	120.1	108	0.017	28.569	0.501
2	5.297	BB	141.6	146.4	0.0145	33.700	0.662
3	5.533	BB	158.6	148.6	0.0156	37.731	0.514

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.66	BB	104.5	85.1	0.0175	14.871	0.439
2	5.33	BB	2.5	1.8	0.0203	0.351	0.445
3	5.531	BB	595.8	656.5	0.0138	84.778	1.173

Figure D-36. (A) Representative GC-FID trace for reaction **C5** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of ***cis*-C6-p** and ***trans*-C6-p** with 1,3,5-trimethoxybenzene as an authentic standard. **(C)** The authentic standard for ***cis*-C6-p** was isolated in trace quantities during synthesis.

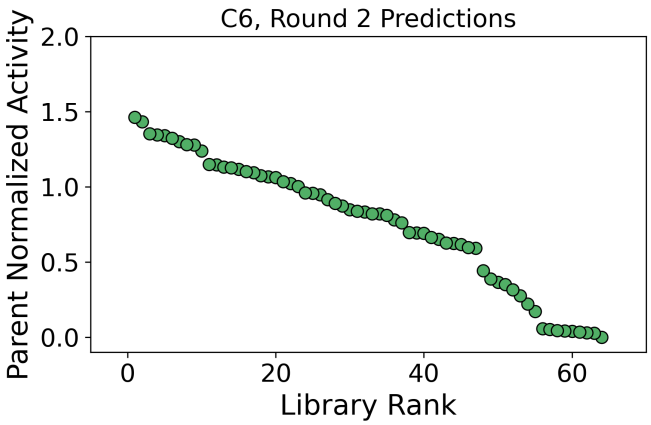


Figure D-37. Retention of function plot for reaction **C6** activity for variants in the second round of predictions. Parent normalized activities are computed from the total formation of cyclopropane product for each variant relative to PromPgb.

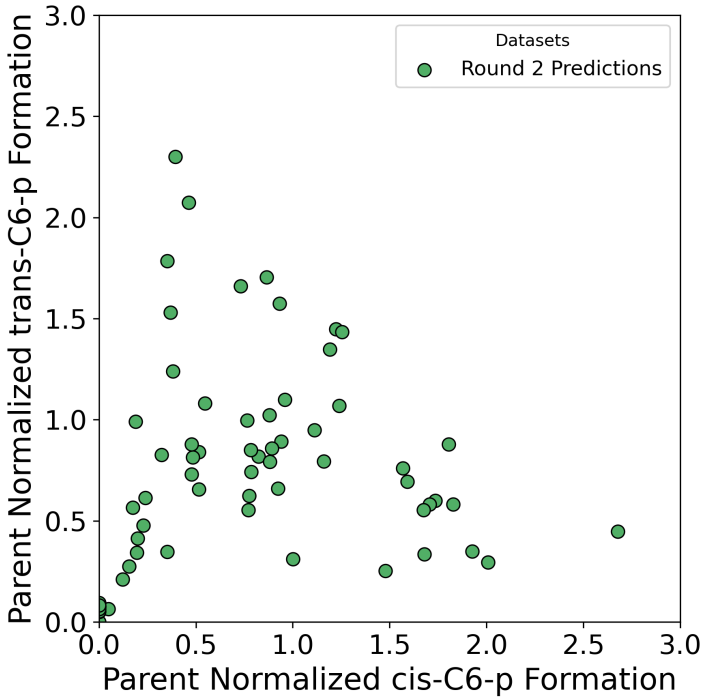
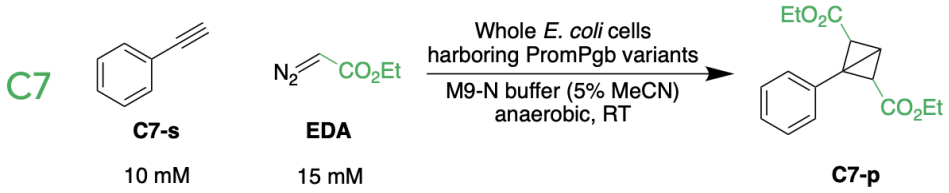


Figure D-38. Changes in the formation of ***cis*-C6-p** and ***trans*-C6-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C7:

Scheme D-7. Reaction conditions for reaction **C7** and expected products.



For library screening, reaction **C7** was analyzed by GC-MS. The authentic standard was obtained by scaling up the enzymatic reaction PromPgb, followed by purification and characterization. Reaction **C7** was tested with round 1 and round 2 of predictions. The bicyclobutane products obtained in reaction **C7** did not provide well-resolved peaks under the applied GC conditions; therefore, reaction yields were not quantified. Instead, for model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation.

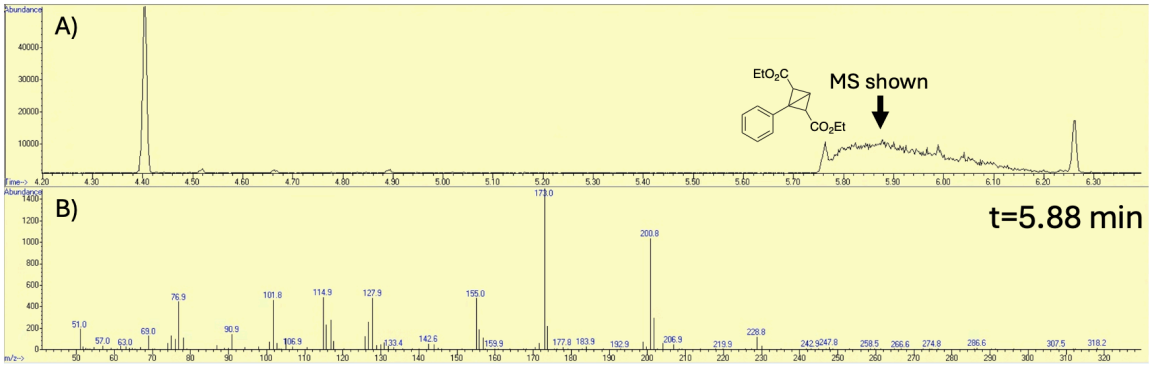
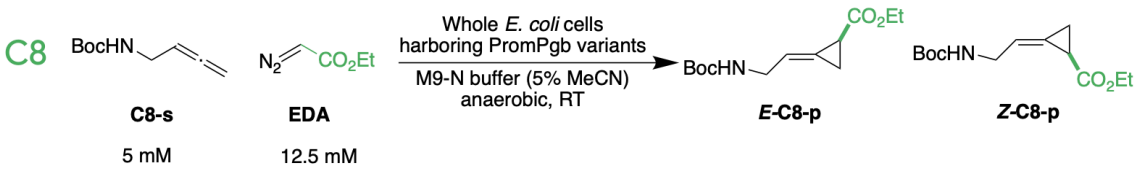


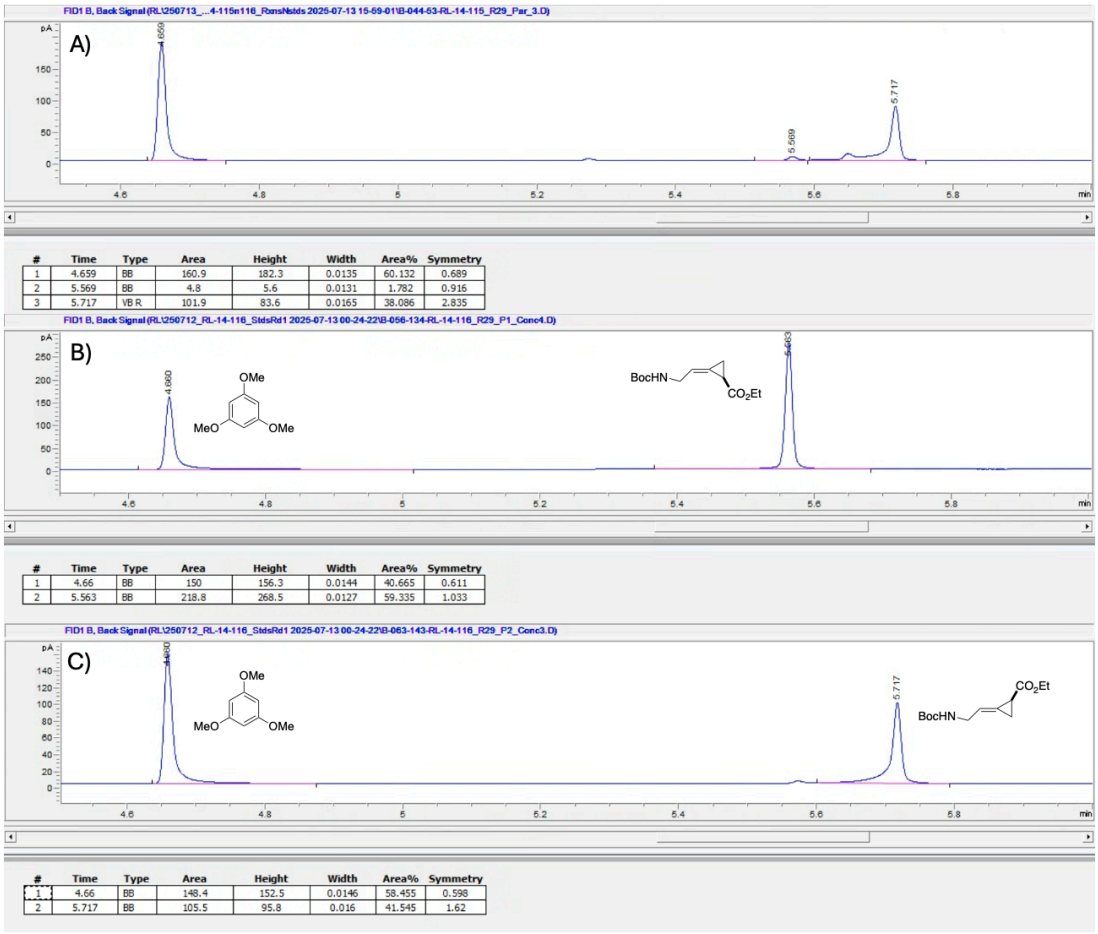
Figure D-39. (A) Representative GC-MS trace for reaction **C7** with PromPgb. This trace was measured from an aliquot taken from the scaled-up reaction of PromPgb under conditions for **C7** to isolate **C7-p**. The bicyclobutane product does not provide a well-resolved peak. **(B)** Selected mass spectrum from t = 5.88 min, showing representative mass peaks for the bicyclobutane **C7-p**.

Reaction C8:

Scheme D-8. Reaction conditions for reaction **C8** and expected products.



For library screening, reaction **C8** was analyzed by GC-FID. The authentic standards were obtained by scaling up the enzymatic reaction with variants VYIFMVFQ (***E*-C8-p**) and TAIFMVFQ (***Z*-C8-p**), followed by purification and characterization. Reaction **C8** was tested with round 1 and round 2 of predictions. The configuration of **C8-p** was assigned by analogy to reaction **C9**. NMR characterization of the product of reaction **C9** with PromPgb confirmed exclusive formation of ***E*-C9-p**. Given that PromPgb predominantly produces a single alkylidenecyclopropane isomer in both reactions **C8** and **C9** (**Figures D-39A** and **Figure D-42A**), the product for reaction **C8** with PromPgb was assigned as the *E*-configuration. The product isolated from variant TAIFMVFQ was thus assigned the *Z*-configuration (**Figure D-39B**).



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.659	BB	160.9	182.3	0.0135	60.132	0.689
2	5.569	BB	4.8	5.6	0.0131	1.782	0.916
3	5.717	VB R	101.9	83.6	0.0165	38.086	2.835

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.66	BB	150	156.3	0.0144	40.665	0.611
2	5.563	BB	218.8	268.5	0.0127	59.335	1.033

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.66	BB	148.4	152.5	0.0146	58.455	0.598
2	5.717	BB	105.5	95.8	0.016	41.545	1.62

Figure D-40. (A) Representative GC-FID trace for reaction **C8** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of ***Z*-C8-p** with 1,3,5-trimethoxybenzene as an authentic standard. The authentic standard was isolated from a scaled-up reaction with variant TAIFMVFQ. **(C)** GC-FID trace for a sample of the authentic standard of ***E*-C8-p** with 1,3,5-trimethoxybenzene as an authentic standard. The authentic standard was isolated from a scaled-up reaction with variant VYIFMVFQ.

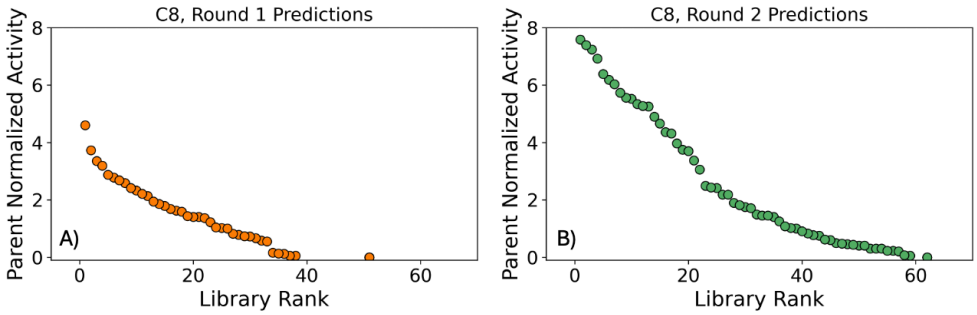


Figure D-41. Retention of function plots for reaction **C8** activity for variants in the **(A)** first round of predictions and **(B)** the second round of predictions. Parent normalized activities are computed from the total formation of alkylidene cyclopropane product for each variant relative to PromPgb.

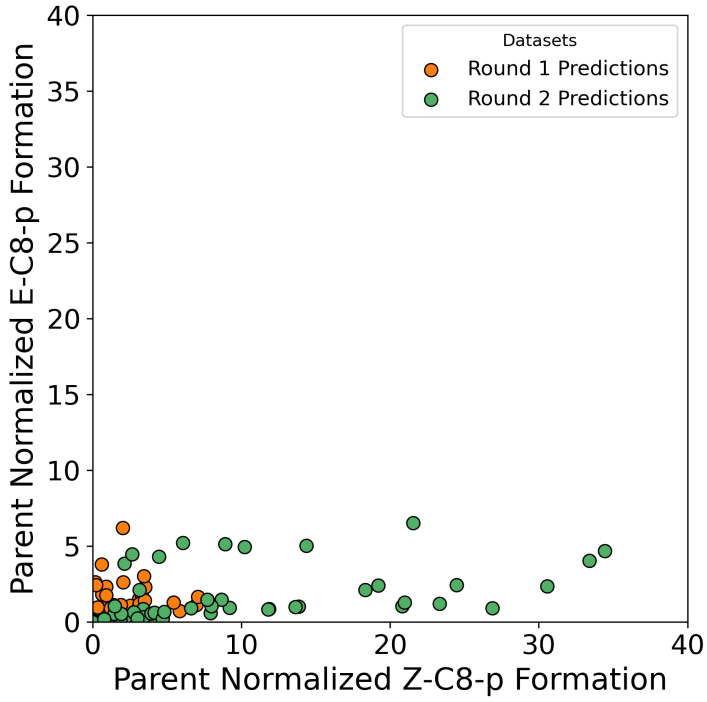
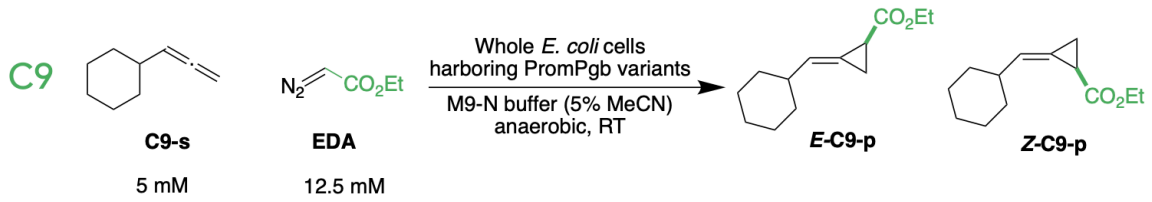


Figure D-42. Changes in the formation of ***Z*-C8-p** and ***E*-C8-p** for tested libraries. Parent normalized activities are computed from the formation of each cyclopropane diastereomer relative to PromPgb.

Reaction C9:

Scheme D-9. Reaction conditions for reaction **C9** and expected products.



For library screening, reaction **C9** was analyzed by GC-FID. The authentic standards were obtained by scaling up the enzymatic reaction with variant LAFFIAFN, followed by purification and characterization. Reaction **C9** was tested with round 2 of predictions. In the course of library screening, only ***E*-C9-p** formation could be found under enzymatic conditions, as confirmed by NMR.[9] For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation.

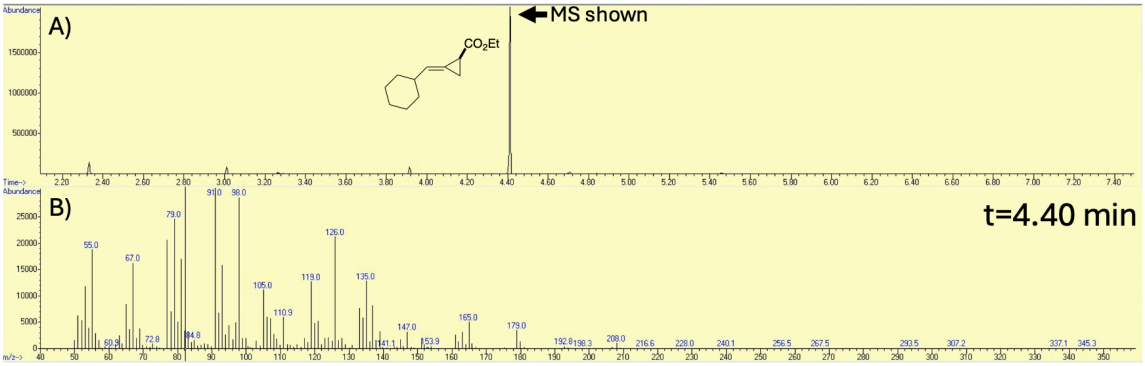
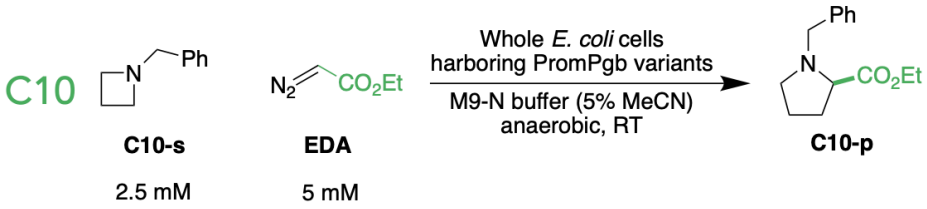


Figure D-43. (A) Representative GC-MS trace for reaction **C9** with variant LAFFIAFN. This trace was measured from an aliquot taken from the scaled-up reaction of LAFFIAFN under conditions for **C9** to isolate **C9-p** products. Only ***E*-C9-p** could be isolated. **(B)** Selected mass spectrum from t = 4.40 min, showing representative mass peaks for the alkylidene cyclopropane ***E*-C9-p**.

Reaction C10:

Scheme D-10. Reaction conditions for reaction **C10** and expected products.



For library screening, reaction **C10** was analyzed by LC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C10** was tested with round 1 and round 2 of predictions. PromPgb was not capable of performing this reaction, and no variants were found in either round of predictions which were capable of performing this transformation with detectable yield.

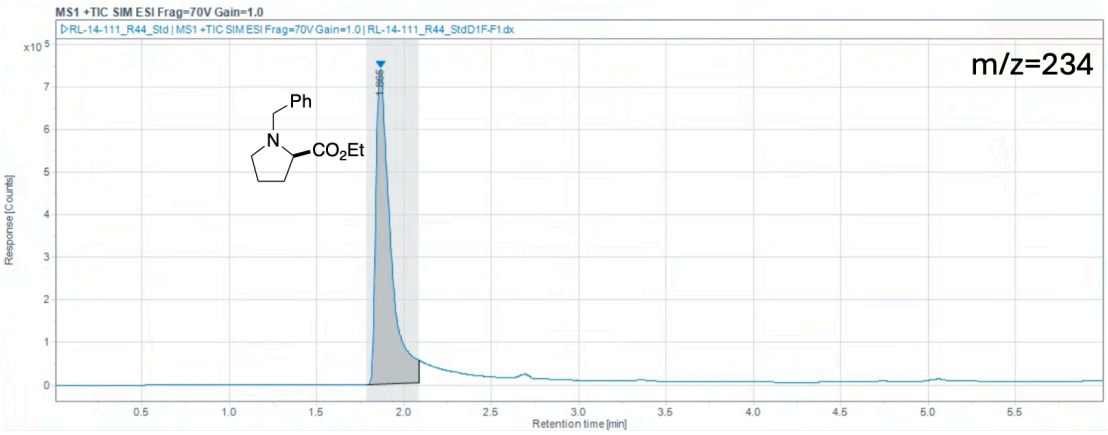
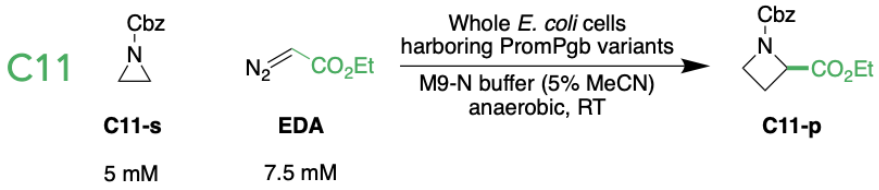


Figure D-44. LC-MS trace for a sample of the authentic standard of **C10-p**. The Selective Ion Monitoring channel for the [M+H]+ ion of **C10-p** (m/z = 234) is shown.

Reaction C11:

Scheme D-11. Reaction conditions for reaction **C11** and expected products.



For library screening, reaction **C11** was analyzed by GC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory in a previous investigation. Full characterization data are reported in reference 10. Reaction **C11** was tested with round 1 and round 2 of predictions. PromPgb was not capable of performing this reaction, and no variants were found in either round of predictions which were capable of performing this transformation with detectable yield.

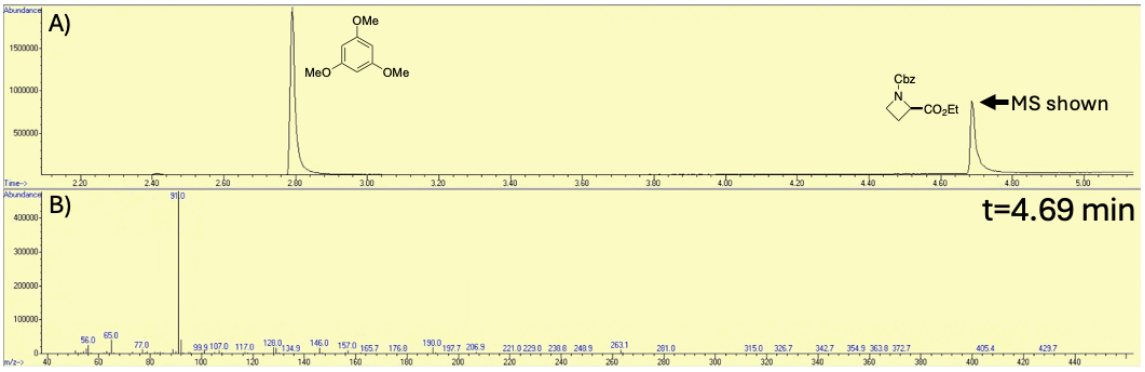
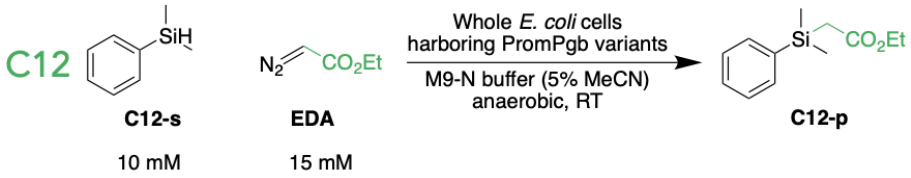


Figure D-45. (A) GC-MS trace for an authentic standard of **C11-p**. **(B)** Selected mass spectrum from the authentic standard peak (t = 4.69 min), showing representative mass peaks for the ring expansion product **C11-p**.

Reaction C12:

Scheme D-12. Reaction conditions for reaction **C12** and expected products.



For library screening, reaction **C12** was analyzed by GC-FID. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C12** was tested with round 1 and round 2 of predictions. Activity data for reaction **C12** with round 1 of predictions were not used to update the model before generating the second round of predictions in this work.

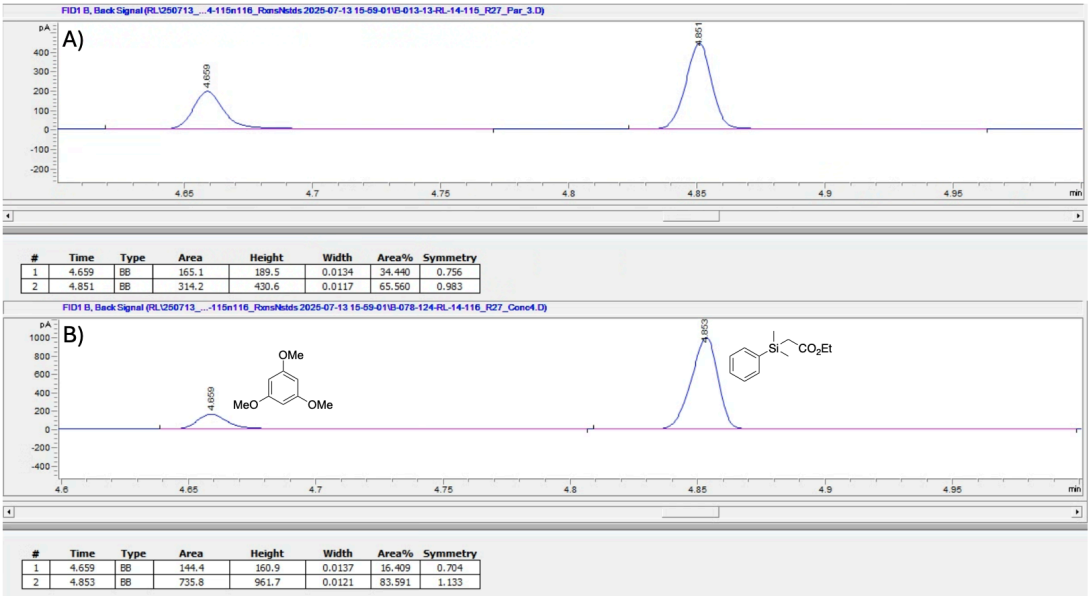


Figure D-46. (A) Representative GC-FID trace for reaction **C12** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of **C12-p** with 1,3,5-trimethoxybenzene as an authentic standard.

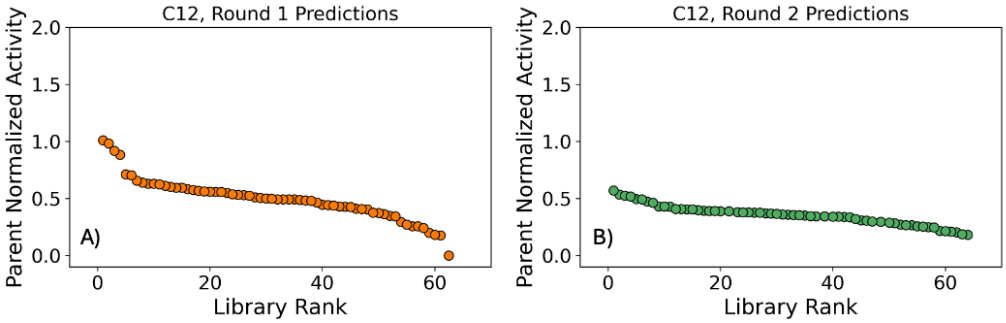
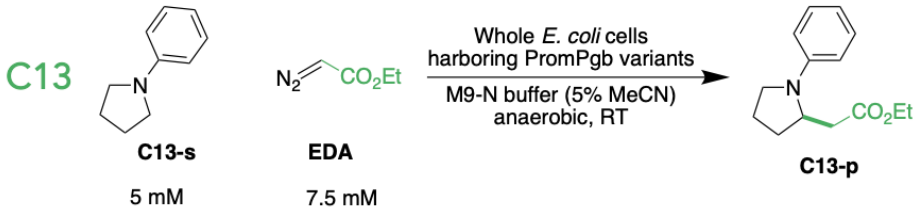


Figure D-47. Retention of function plots for reaction **C12** activity for variants in the **(A)** first round of predictions and **(B)** the second round of predictions. Parent normalized activities are computed from the total formation of alkylation product for each variant relative to PromPgb.

Reaction C13:

Scheme D-13. Reaction conditions for reaction **C13** and expected products.



For library screening, reaction **C13** was analyzed by GC-FID. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C13** was tested with round 1 and round 2 of predictions. While collecting data with round 1 predictions for model updating, a pooled GC-MS analysis was used for reaction **C13**, yielding qualitative reaction data. Reaction extracts from reactions **C13**, **C14**, **C15**, **C16**, and **C17** were pooled and analyzed by GC-MS. As a result, low-performing variants may have been over diluted and analytes may have been brought below the limit of detection. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation. Thirty-seven variants were found to have activity for reaction **C13** for model updating in this work. Later, reaction **C13** was evaluated with round 1 and round 2 of predictions using a quantitative GC-FID method, yielding the data shown in **Figure D-48**.

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	11.6	BV	600.2	476.3	0.0195	73.789	1.334
2	11.641	VB	201.9	140.1	0.0216	24.817	0.66
3	14.299	BB	11.3	7.7	0.022	1.394	0.51

#	Time	Type	Area	Height	Width	Area%	Symmetry
1	11.638	BB	146.2	94	0.023	25.982	0.521
2	14.286	BB	416.5	417.5	0.0156	74.018	1.028

Figure D-48. (A) Representative GC-FID trace for reaction **C13** with PromPgb. **(B)** GC-FID trace for a sample of the authentic standard of **C13-p** with 1,3,5-trimethoxybenzene as an authentic standard.

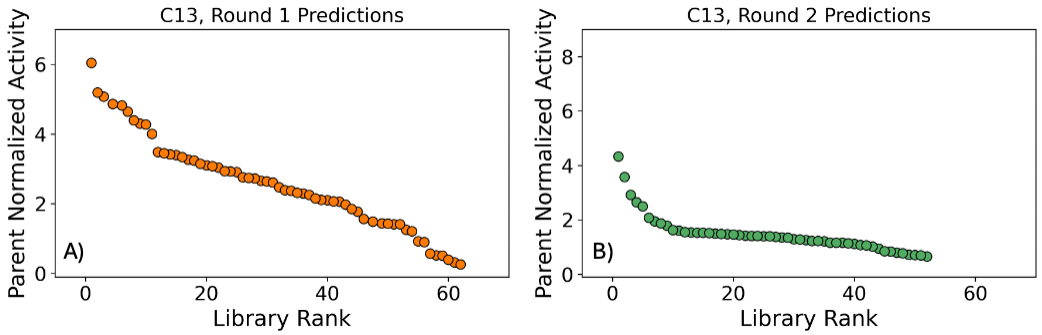
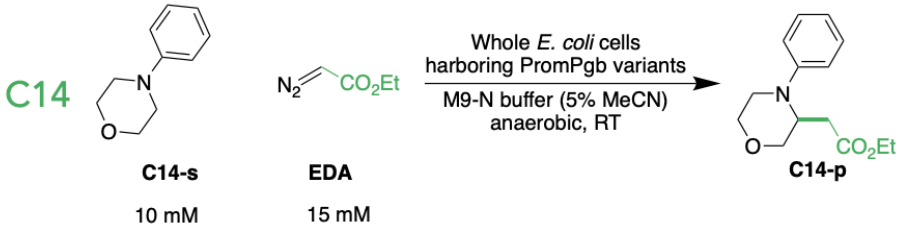


Figure D-49. Retention of function plots for reaction **C13** activity for variants in the **(A)** first round of predictions and **(B)** the second round of predictions. Parent normalized activities are computed from the total formation of alkylation product for each variant relative to PromPgb.

Reaction C14:

Scheme D-14. Reaction conditions for reaction **C14** and expected products.



For library screening, reaction **C14** was analyzed by GC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **C14** was tested with round 1 and round 2 of predictions. PromPgb was incapable of catalyzing reaction **C14** under these conditions. While collecting data with round 1 predictions for model updating, a pooled GC-MS analysis was used for reaction **C14**, yielding qualitative reaction data. Reaction extracts from reactions **C13**, **C14**, **C15**, **C16**, and **C17** were pooled and analyzed by GC-MS. As a result, low-performing variants may have been over diluted and analytes may have been brought below the limit of detection. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation. Six variants were found to have activity for reaction **C14** for model updating in this work. Later, reaction **C14** was evaluated with round 1 and round 2 of predictions using an un-pooled GC-MS strategy, uncovering 42 variants which could perform this transformation.

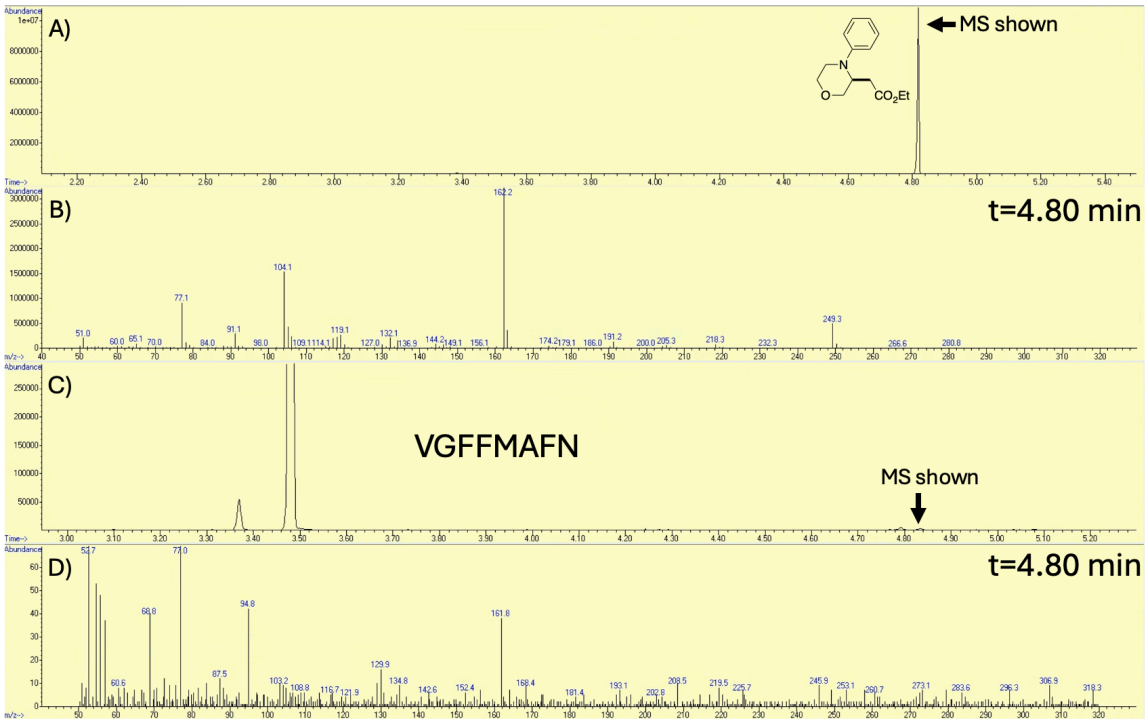
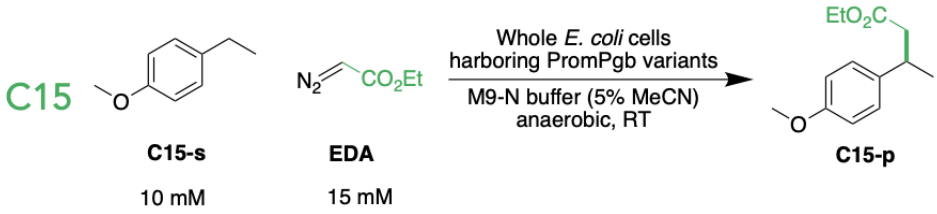


Figure D-50. (A) GC-MS trace for an authentic standard of **C14-p**. **(B)** Selected mass spectrum from the authentic standard peak (t = 4.80 min), showing representative mass peaks for the alkylation product **C14-p**. **(C)** GC-MS trace for reaction **C14** run with variant VFGGMAFN. **C14-p** can be observed in trace quantities. **(D)** Selected mass spectrum from the observed trace product (t = 4.80 min) showing the characteristic m/z = 162 mass for **C14-p**.

Reaction C15:

Scheme D-15. Reaction conditions for reaction **C15** and expected products.



For library screening, reaction **C15** was analyzed by GC-MS. Authentic standards for this reaction were purchased from a commercial source. Reaction **C15** was tested with round 1 and round 2 of predictions. In the course of collecting data with round 1 predictions for model updating, a pooled GC-MS analysis was used for reaction **C15**, yielding qualitative reaction data. Reaction extracts from reactions **C13**, **C14**, **C15**, **C16**, and **C17** were pooled and analyzed by GC-MS. As a result, low-performing variants may have been over diluted and analytes may have been brought below the limit of detection. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation. No variants were found to have activity for reaction **C15** for model updating in this work. Later, reaction **C15** was evaluated with round 1 and round 2 of predictions using an un-pooled GC-MS strategy, after which there were still no variants found to catalyze this transformation.

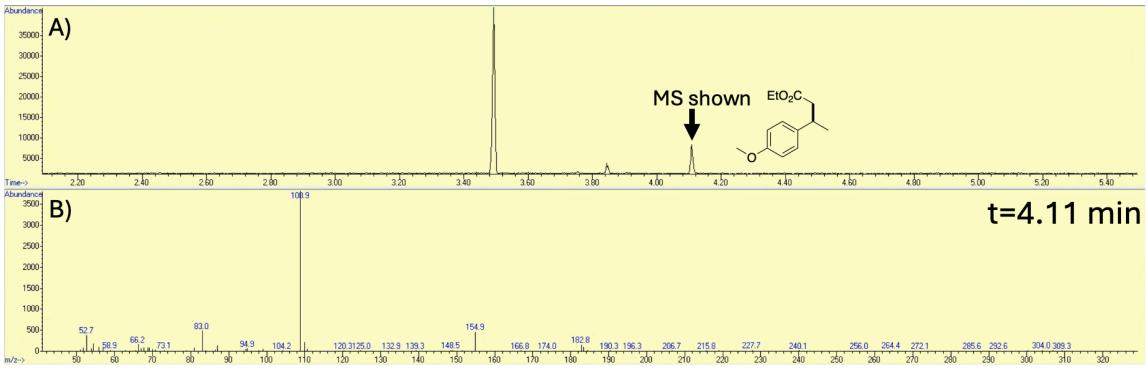
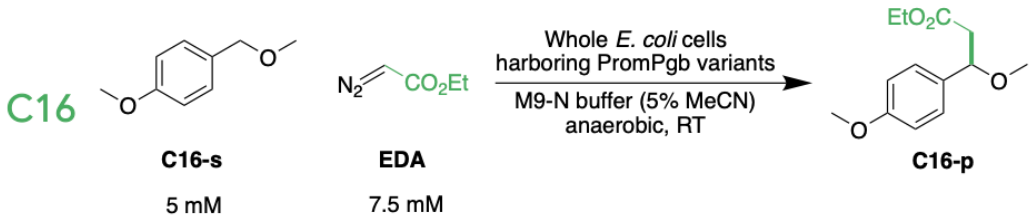


Figure D-51. (A) GC-MS trace for an authentic standard of **C15-p**. **(B)** Selected mass spectrum from the authentic standard peak (t = 4.11 min), showing representative mass peaks for the alkylation product **C15-p**.

Reaction C16:

Scheme D-16. Reaction conditions for reaction **C16** and expected products.



For library screening, reaction **C16** was analyzed by GC-MS. Authentic standards for this reaction were purchased from a commercial source. Reaction **C16** was tested with round 1 and round 2 of predictions. PromPgb was incapable of catalyzing reaction **C16** under these conditions. In the course of collecting data with round 1 predictions for model updating, a pooled GC-MS analysis was used for reaction **C16**, yielding qualitative reaction data. Reaction extracts from reactions **C13**, **C14**, **C15**, **C16**, and **C17** were pooled and analyzed by GC-MS. As a result, low-performing variants may have been over diluted and analytes may have been brought below the limit of detection. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation. No variants were found to have activity for reaction **C16** for model updating in this work. Later, reaction **C16** was evaluated with round 1 and

round 2 of predictions using an un-pooled GC-MS strategy, uncovering 54 variants which could perform this transformation.

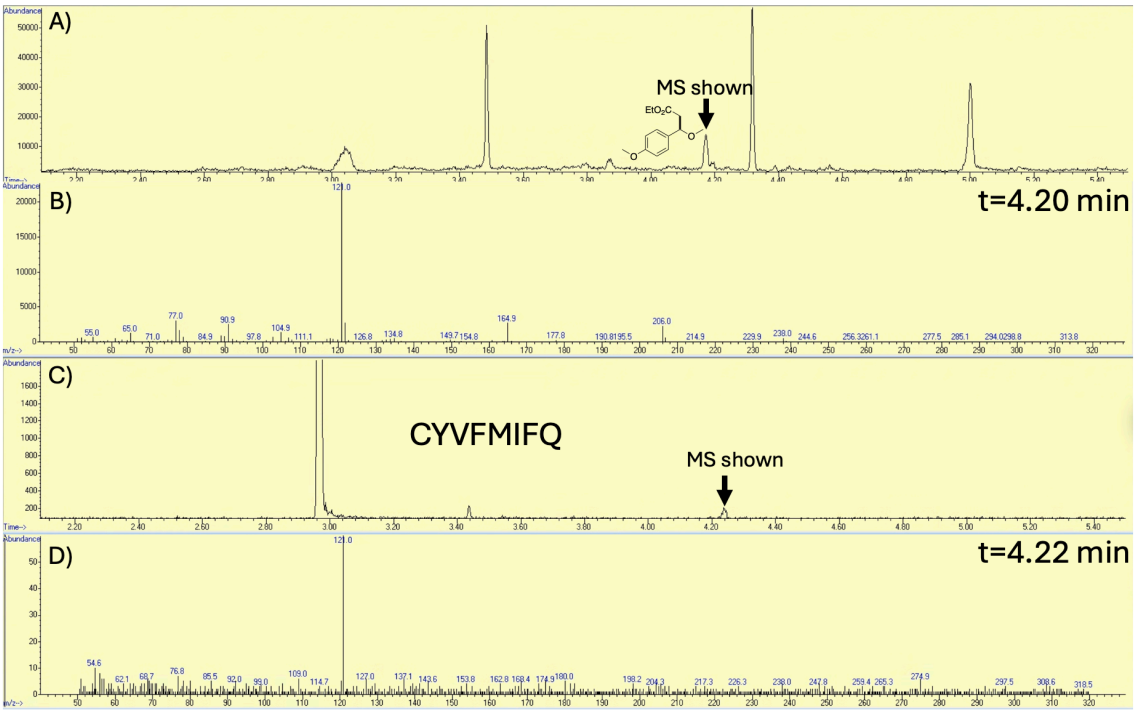
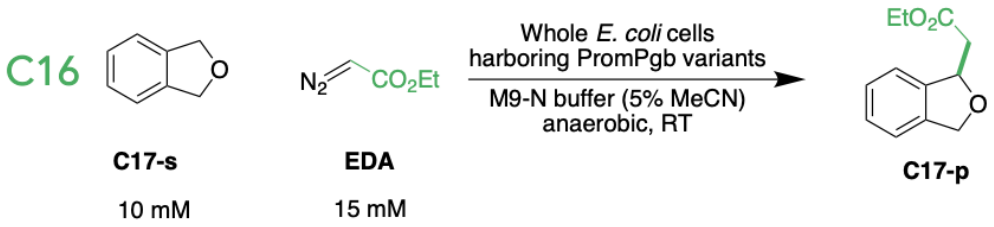


Figure D-52. (A) GC-MS trace for an authentic standard of **C16-p**. **(B)** Selected mass spectrum from the authentic standard peak (t = 4.20 min), showing representative mass peaks for the alkylation product **C16-p**. **(C)** GC-MS trace for reaction **C16** run with variant CYVFMIFQ. **C16-p** can be observed in trace quantities. **(D)** Selected mass spectrum from the observed trace product (t = 4.22 min) showing the characteristic m/z = 121 mass for **C16-p**.

Reaction C17:

Scheme D-17. Reaction conditions for reaction **C17** and expected products.



For library screening, reaction **C17** was analyzed by GC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory in a previous investigation. Full characterization data are reported in reference 11. In the course of collecting data with round 1 predictions for model updating, a pooled GC-MS analysis was used for reaction **C17**, yielding qualitative reaction data. Reaction extracts from reactions **C13**, **C14**, **C15**, **C16**, and **C17** were pooled and analyzed by GC-MS. As a result, low-performing variants may have been over diluted and analytes may have been brought below the limit of detection. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation. No variants were found to have activity for reaction **C17** for model updating in this work. Later, reaction **C17** was evaluated with round 1 and round 2 of predictions using an un-pooled GC-MS strategy, after which there were still no variants found to catalyze this transformation.

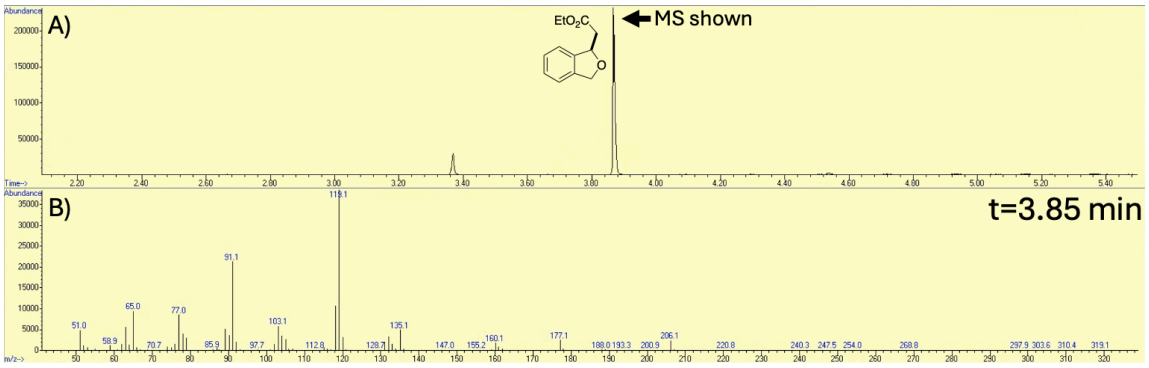
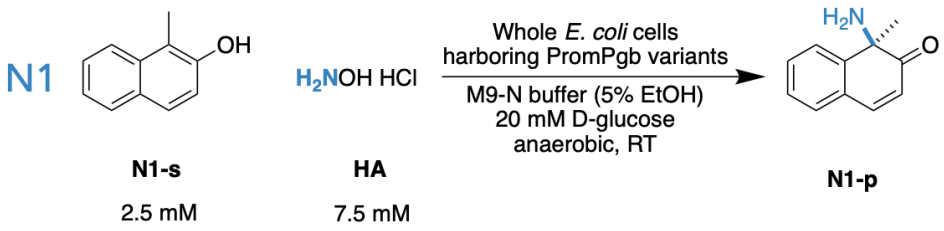


Figure D-53. (A) GC-MS trace for an authentic standard of **C17-p**. **(B)** Selected mass spectrum from the authentic standard peak (t = 3.85 min), showing representative mass peaks for the alkylation product **C17-p**.

Reaction N1:

Scheme D-18. Reaction conditions for reaction **N1** and expected products.



For library screening, reaction **N1** was analyzed by LC-MS. The authentic standard of **N1-p** was kindly provided by Dr. Kathleen M. Sicinski, who previously synthesized and characterized this compound in our laboratory. Full characterization data are reported in reference 12. Reaction **N1** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

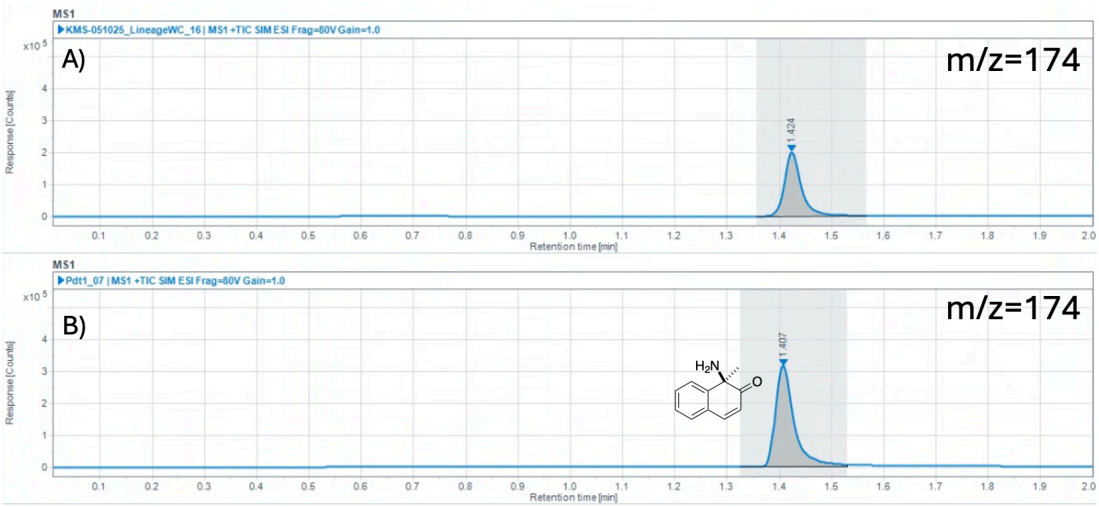


Figure D-54. (A) Representative LC-MS trace for reaction **N1** with PromPgb. **(B)** LC-MS trace for a sample of the authentic standard of **N1-p**. The Selective Ion Monitoring channel for the [M+H]+ ion of **N1-p** (m/z = 174) is shown.

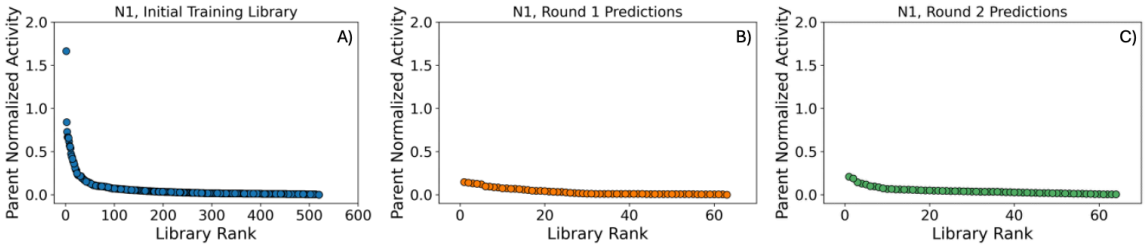
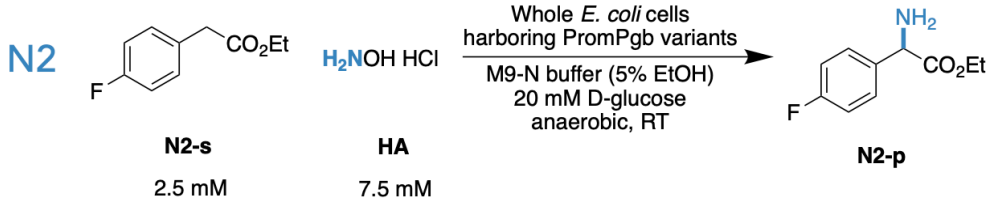


Figure D-55. Retention of function plots for reaction **N1** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of dearomatization product for each variant relative to PromPgb.

Reaction N2:

Scheme D-19. Reaction conditions for reaction **N2** and expected products.



For library screening, reaction **N2** was analyzed by LC-MS. The authentic standard of **N2-p** was kindly provided by Dr. Edwin Alfonzo, who previously synthesized and characterized this compound in our laboratory. Full characterization data are reported in reference 13. Reaction **N2** was tested with the initial training library, round 1 of predictions, and round 2 of predictions. PromPgb was incapable of forming **N2-p** under these conditions.

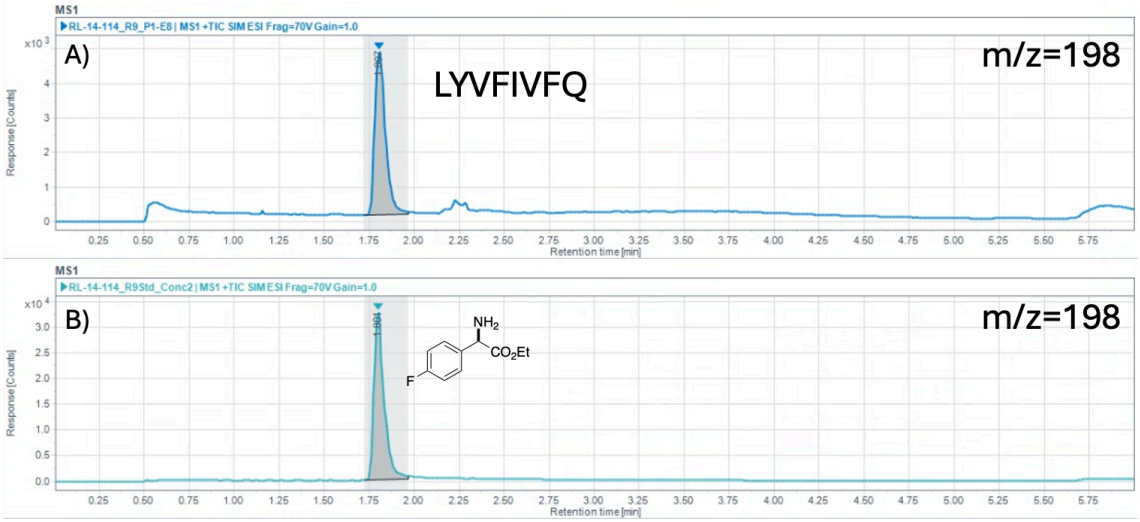


Figure D-56. (A) LC-MS trace for reaction **N2** with variant LYVFIVFQ from the first round of predictions. **N2-p** formation is observed. **(B)** LC-MS trace for a sample of the authentic standard of **N2-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N2-p** (m/z = 198) is shown.

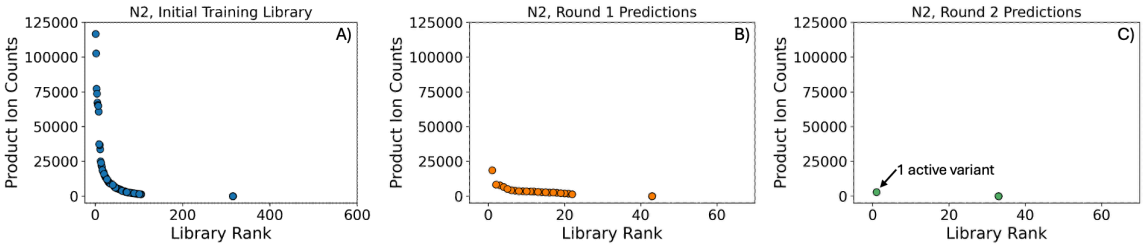
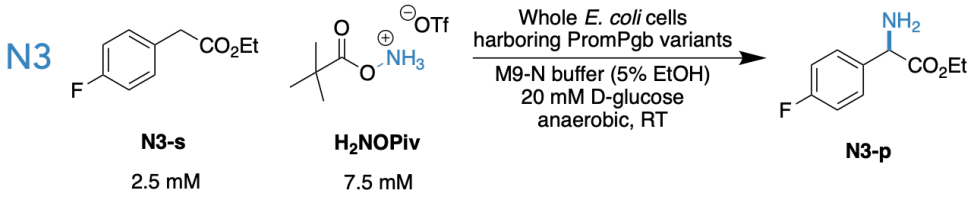


Figure D-57. Retention of function plots for reaction **N2** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Variants are ranked by measured ion counts corresponding to product **N2-p** in the SIM channel for the [M+H]+ ion (m/z = 198).

Reaction N3:

Scheme D-20. Reaction conditions for reaction **N3** and expected products.



For library screening, reaction **N3** was analyzed by LC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **N3** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

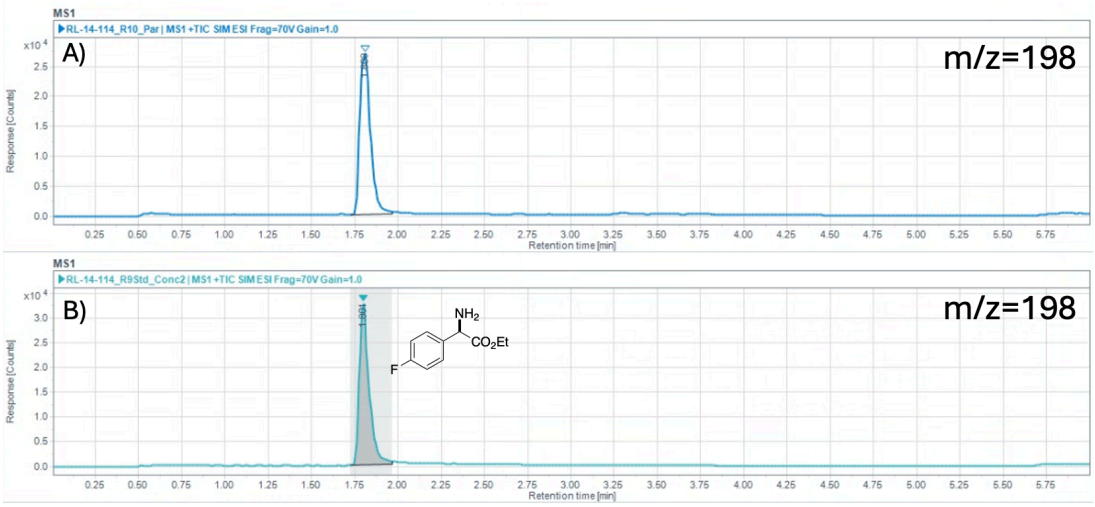


Figure D-58. (A) Representative LC-MS trace for reaction **N3** with PromPgb. **(B)** LC-MS trace for a sample of the authentic standard of **N3-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N3-p** (m/z = 198) is shown.

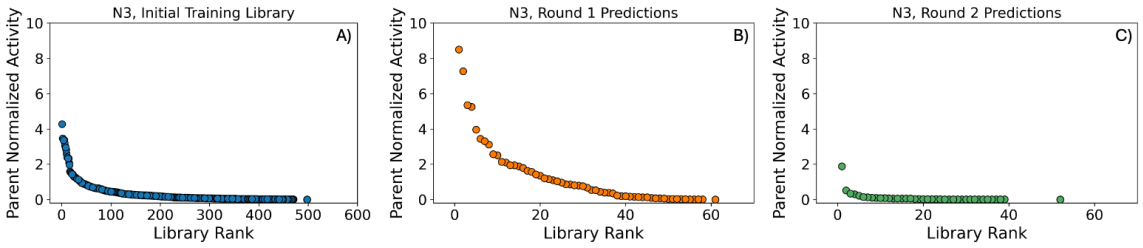
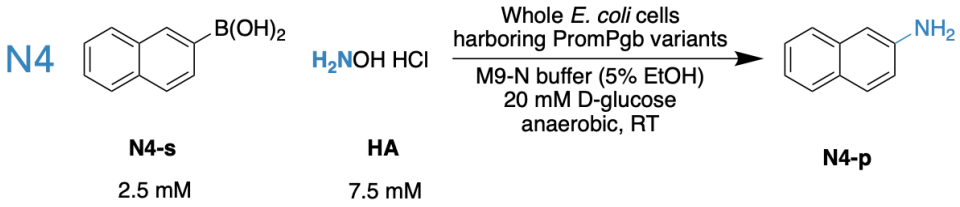


Figure D-59. Retention of function plots for reaction **N3** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of C–H amination product for each variant relative to PromPgb.

Reaction N4:

Scheme D-21. Reaction conditions for reaction **N4** and expected products.



For library screening, reaction **N4** was analyzed by LC-MS. Authentic standards for this reaction were purchased from a commercial source. Reaction **N4** was tested with the initial training library, round 1 of predictions, and round 2 of predictions.

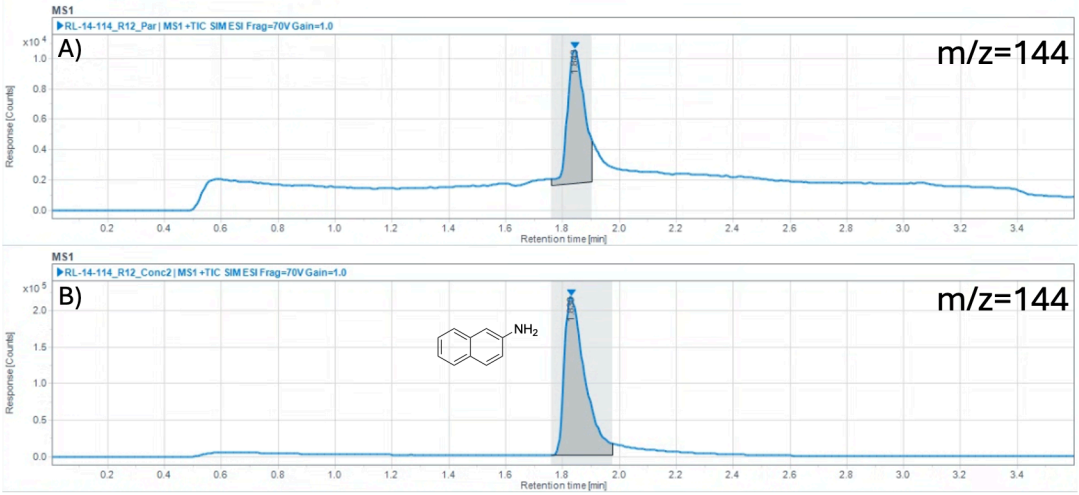


Figure D-60. (A) Representative LC-MS trace for reaction **N4** with PromPgb. **(B)** LC-MS trace for a sample of the authentic standard of **N4-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N4-p** (m/z = 144) is shown.

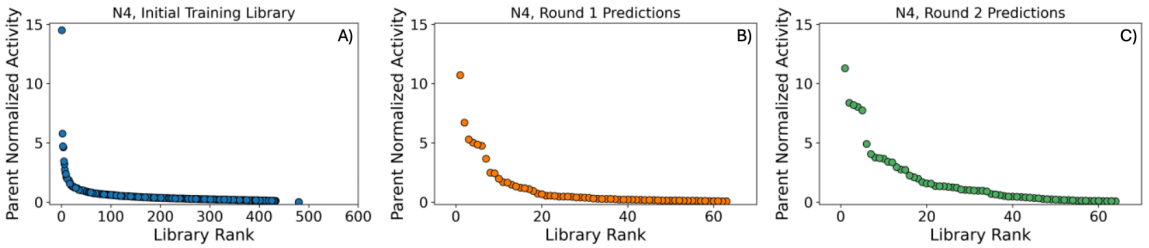
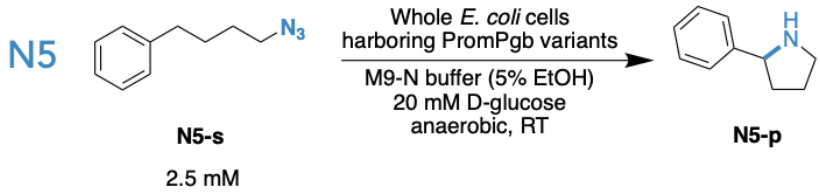


Figure D-61. Retention of function plots for reaction **N4** activity for variants in the **(A)** initial training library, **(B)** the first round of predictions, and **(C)** the second round of predictions. Parent normalized activities are computed from the total formation of C–H amination product for each variant relative to PromPgb.

Reaction N5:

Scheme D-22. Reaction conditions for reaction **N5** and expected products.



For library screening, reaction **N5** was analyzed by LC-MS. Authentic standards for this reaction were purchased from a commercial source. Reaction **N5** was tested with round 1 and round 2 of predictions.

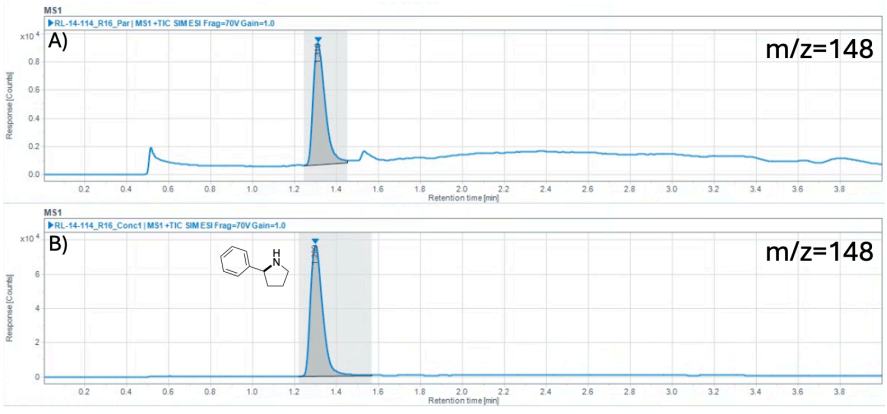


Figure D-62. (A) Representative LC-MS trace for reaction **N5** with PromPgb. **(B)** LC-MS trace for a sample of the authentic standard of **N5-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N5-p** (m/z = 148) is shown.

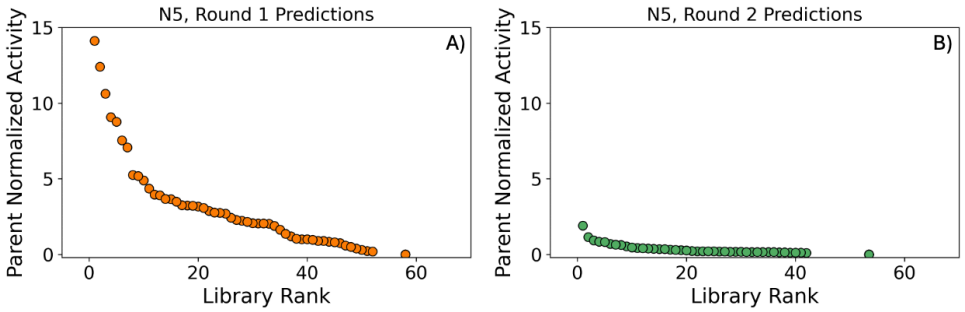
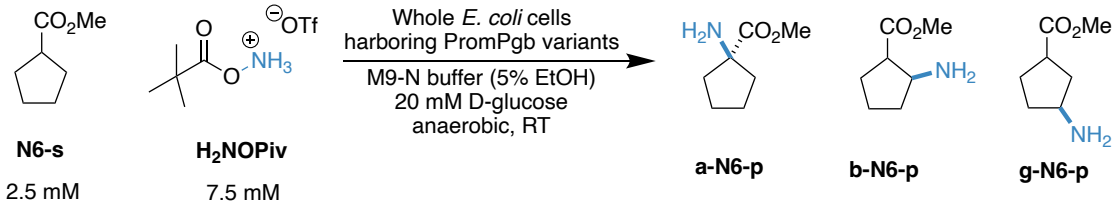


Figure D-63. Retention of function plots for reaction **N5** activity for variants in the **(A)** first round of predictions and **(B)** the second round of predictions. Parent normalized activities are computed from the total formation of intramolecular reaction product for each variant relative to PromPgb.

Reaction N6:

Scheme D-23. Reaction conditions for reaction **N6** and expected products.



For library screening, reaction **N6** was analyzed by LC-MS. Authentic standards for the possible α-, β-, γ-amination products for this reaction were purchased from a commercial source. Reaction **N6** was tested with round 1 and round 2 of predictions. PromPgb was incapable of catalyzing reaction **N6** under these conditions. In library screening, five variants were found in the first round of predictions which could generate trace quantities of **N6-p** isomers and no variants were found in the second round of predictions. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation.

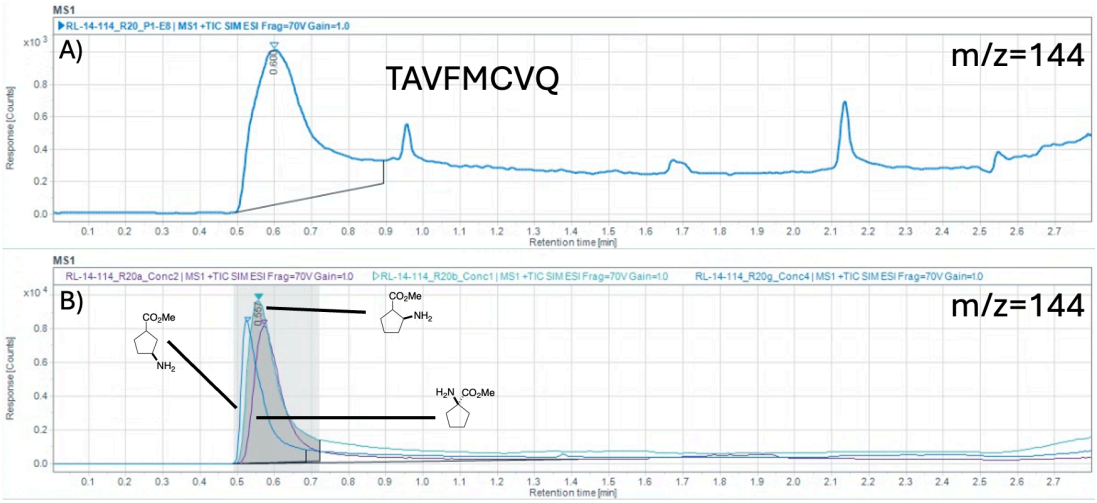
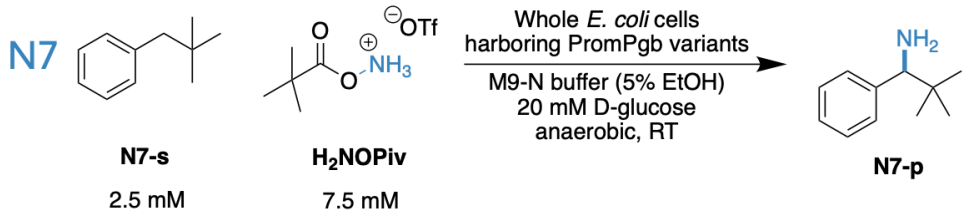


Figure D-64. (A) LC-MS trace for reaction **N6** with variant TAVFMCVQ from the first round of predictions. The three possible regioisomers of **N6-p** formation could not be well-separated on the LC-MS column and so were treated as a single eluting peak. **(B)** LC-MS traces for samples of the authentic standards of **a-N6-p**, **b-N6-p**, and **g-N6-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N6-p** (m/z = 144) is shown.

Reaction N7:

Scheme D-24. Reaction conditions for reaction **N7** and expected products.



For library screening, reaction **N7** was analyzed by LC-MS. Authentic standards for this reaction were purchased from a commercial source. Reaction **N7** was tested with round 1 and round 2 of predictions. PromPgb was incapable of catalyzing reaction **N7** under these conditions. In library screening, three variants were found in the first round of predictions which could generate trace quantities of **N7-p** isomers and no variants were found in the second round of predictions. For model updating, data from this reaction were treated as binary, where a value of 0 denoted no reaction, and a value of 1 indicated detectable product formation.

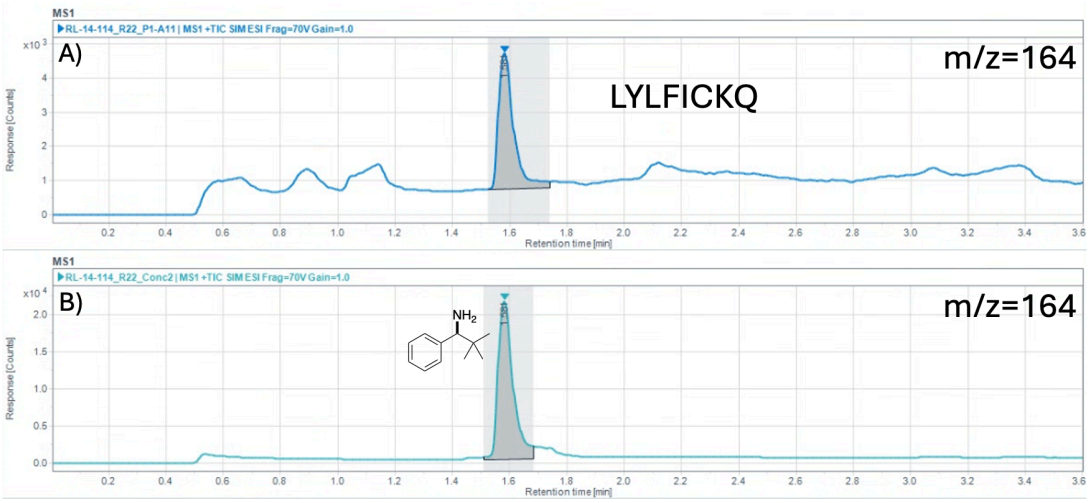
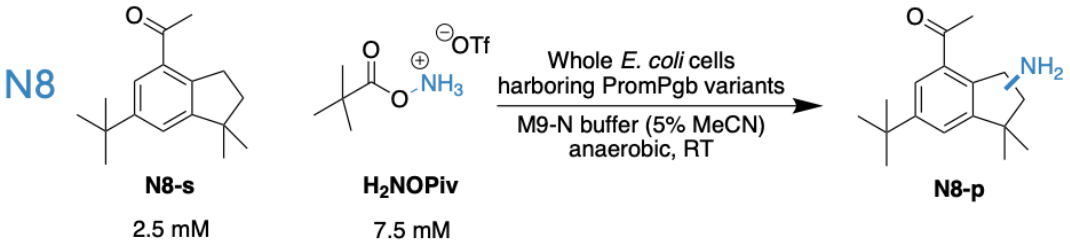


Figure D-65. (A) Representative LC-MS trace for reaction **N7** with variant LYLFICKQ from the first round of predictions. **(B)** LC-MS trace for a sample of the authentic standard of **N7-p**. The Selective Ion Monitoring (SIM) channel for the [M+H]+ ion of **N7-p** (m/z = 164) is shown.

Reaction N8:

Scheme D-25. Reaction conditions for reaction **N8** and expected products.



For library screening, reaction **N8** was analyzed by LC-MS. For reaction **N8**, the exact product structure was not determined; activity was quantified based on an LC–MS peak corresponding to the expected $[M+H]^+$ ion. Taken together, the data shown in **Figure D-65** support the occurrence of a protoglobin-catalyzed nitrene transfer reaction with **N8-s**. The product peak observed at t = 1.73 min in panels **Figure D-65A** and **Figure D-65B** is detected only in reactions containing whole cells harboring protoglobin variants, and its abundance depends on the identity of the protoglobin. This peak is not observed in the absence of whole cells (**Figure D-65C**) or in the presence of whole cells expressing TrpB,

a PLP-dependent enzyme serving as our negative control (**Figure D-65D**). Reaction **N8** was tested with round 1 of predictions.

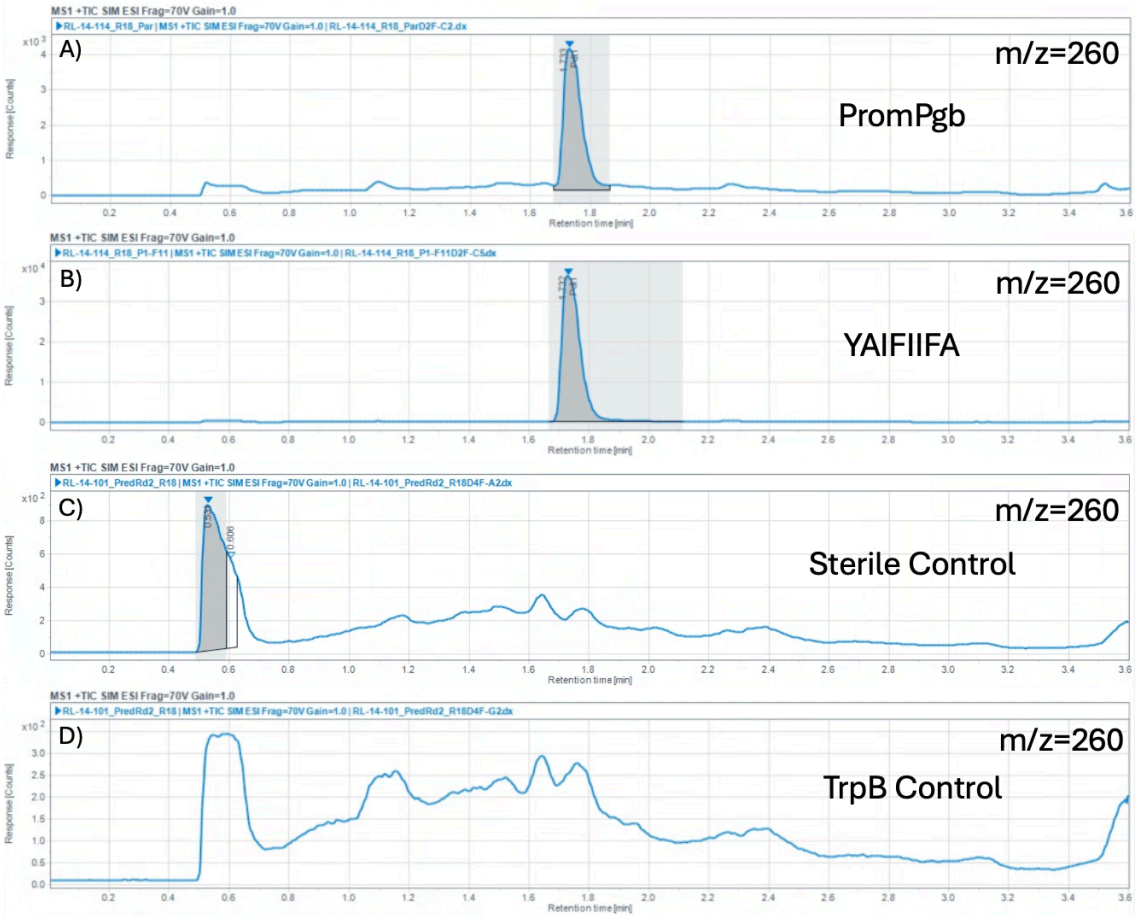


Figure D-66. (A) Representative LC-MS trace for reaction **N8** with PromPgb. **(B)** LC-MS trace for reaction **N8** with the high-performing variant YAIFIIFA. **(C)** Representative LC-MS trace for reaction **N8** in the absence of any whole cells. All reaction conditions are otherwise kept the same. **(D)** Representative LC-MS trace for reaction **N8** in the presence of whole cells expressing The Selective Ion Monitoring (SIM) channel for the $[M+H]^+$ ion of **N8-p** (m/z = 260) is shown for all chromatograms.

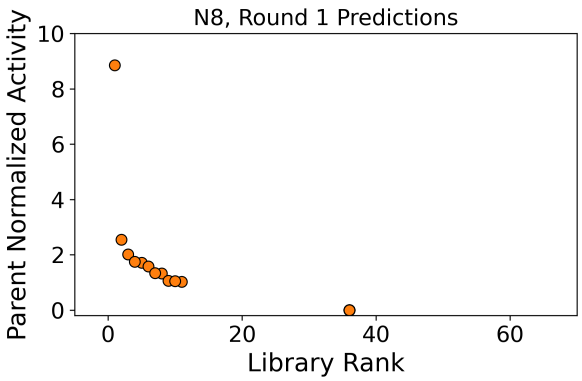
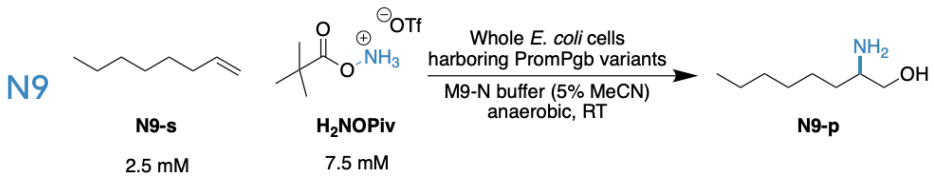


Figure D-67. Retention of function plot for reaction **N8** activity for variants in the first round of predictions.

Reaction N9:

Scheme D-26. Reaction conditions for reaction **N9** and expected products.



For library screening, reaction **N9** was analyzed by LC-MS. Authentic standards for this reaction were synthesized and characterized in our laboratory. Reaction **N9** was tested with round 1 and round 2 of predictions. During Round 1 data collection, four variants were initially assigned activity for reaction **N9**, and these measurements were incorporated into the model update. Subsequent reanalysis of these samples did not confirm product formation under the reported conditions. PromPgb was not capable of performing this reaction, and upon further screening no variants were found in either round of predictions which were capable of performing this transformation with detectable yield.

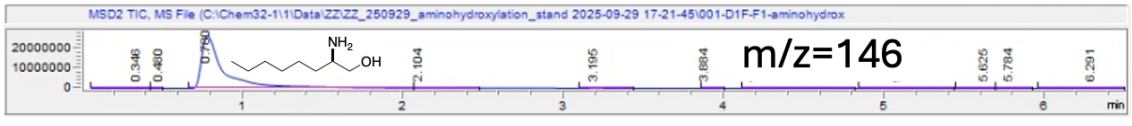


Figure D-68. LC-MS trace for a sample of the authentic standard of **N9-p**. The Selective Ion Monitoring channel for the $[M+H]^+$ ion of **N9-p** (m/z = 146) is shown.

D.6. Computed Summary Statistics

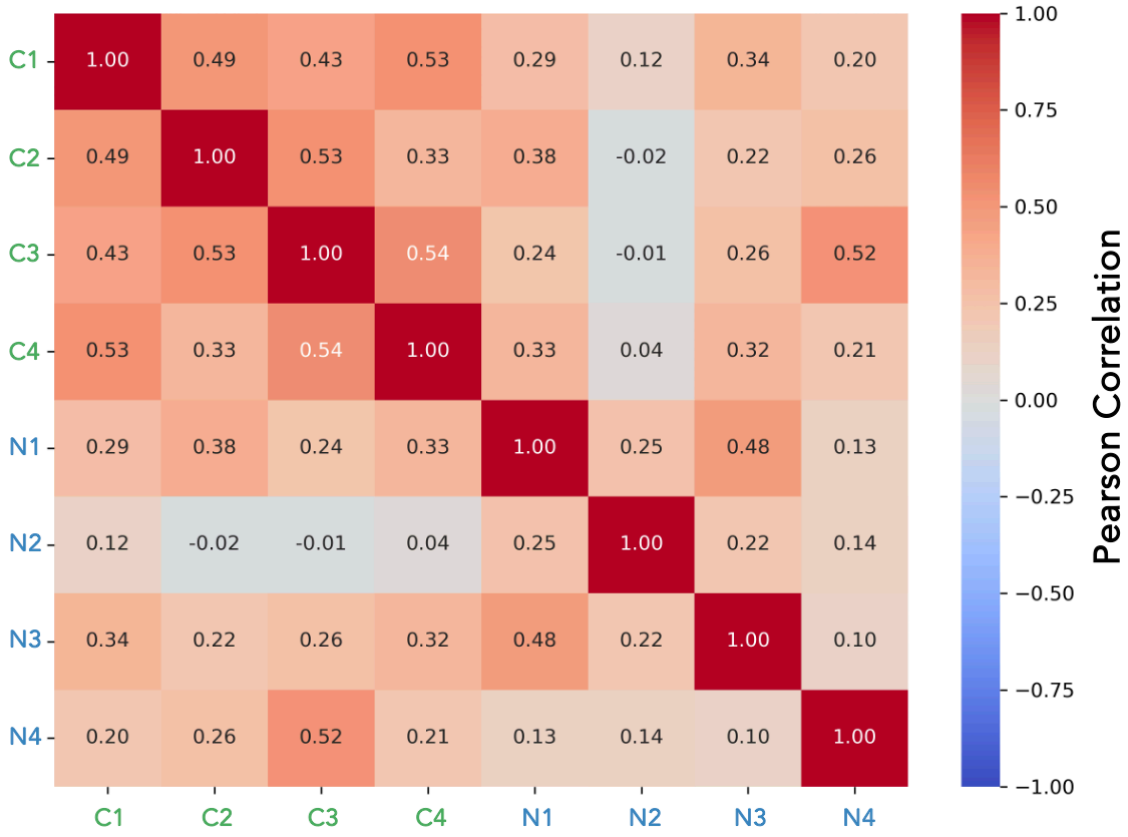


Figure S69. Computed pairwise Pearson correlation coefficients (r) between the fitness values for each of the eight ITRs, calculated using data from variants in the initial training library. Correlations were computed across all variants for which measurable activity was observed in both reactions. Positive values indicate that mutations tending to enhance activity in one reaction also enhance activity in the other, whereas low or near-zero correlations suggest reaction-specific mutational effects.

D.7. Machine Learning Details

D.7.1. Data Processing

At the start of each round of the computational pipeline, sequences were paired with a set of respective experimental yields for carbene and nitrene reactions. The reaction yields were processed and normalized to the parent (starting) sequence, such that for a given reaction, 1.0 corresponded to the same yield as parent. For variants displaying activity for reaction **N2**, fold-improvements in activity are normalized to the yield of parent for reaction **N3**, which shares the same product. Outliers with very high yield (i.e., greater than 4 times that of parent) were rounded down to 4.0, to encourage promiscuity rather than optimizing for a specific reaction. For reactions with trace activity that were not detected in parent, a variant was assigned 1 if could perform that reaction and 0 otherwise. The two fitness objectives used as predictive outputs for ML model training were the mean fitness values of all carbene and nitrene reactions, respectively, thus weighting improvements to carbene reactions relatively equally to improvements to nitrene reactions.

D.7.2. Surrogate Model Training

Most Bayesian optimization algorithms consist of two main components: (1) a probabilistic surrogate model of the objective function and (2) an acquisition function. The surrogate model predicts the objective function values at unobserved inputs, while the acquisition function quantifies the potential benefit of evaluating any given batch of inputs based on these predictions. In each iteration of the Bayesian optimization loop, a new batch of inputs is selected by maximizing the acquisition function. After evaluating the objective function at these new inputs, the surrogate model is updated, and the process repeats. Let $\boldsymbol{X}$ denote the input space (i.e., the space of feasible protein sequences) and let $f: \boldsymbol{X} \rightarrow \boldsymbol{R}$ denote the objective function (i.e., the metric we wish to optimize). In this work, we used deep ensembles as the surrogate model, which is based on their high performance and calibrated uncertainty seen in past work.[8]

Deep ensembles are constructed by training identical deep neural network architectures multiple times, each with different random initializations of the weight parameters. Here, we train the deep ensembles with bootstrapping; each ensemble consists of 20 models where 90% of the total training data is randomly seen during training, and each model is a multilayer perceptron with two hidden layers and a hidden dimension of 50. The input to each model is a one-hot encoding of the 8 residues being studied, with dimension 160, and the outputs are two values, corresponding to the predictions for the two normalized carbene and nitrene reaction yields, described above. We surmised that performing multi-task training by sharing weights would improve regularization and reduce overfitting in the model, as the two distinct objectives are correlated. For each model, we performed training with a batch size of 128 and a learning rate of 0.001 for 100 epochs. Model training time was less than one minute on a single GPU. These independently trained networks were then collectively used as if they were samples from a Bayesian posterior distribution over the objective function f.

D.7.3. Sample Acquisition

We used the expected improvement (EI) acquisition function, which is given by $\alpha_n(x) = E_n[\{f(x) - f_n^*\}^+]$, where $f_n^* = \max_{i=1,\ldots,n} f(x_i)$ and the expectation is computed with respect to the posterior distribution given $\mathcal{D}_n$.[15] For Gaussian posterior distributions and noise-free observations (where f_n^* is a constant rather than a random variable), the EI can be expressed in a closed form using the posterior mean and variance. In scenarios where these conditions do not hold, computing the EI often requires approximate calculation, typically through Monte Carlo sampling techniques.

In this case, we extended expected improvement to expected hypervolume improvement (eHVI), which corresponds to expected improvement in the pareto front "area" across the two objectives described above and is commonly used in multi-objective settings. We acquired a batch of new samples to maximize eHVI following a procedure similar to that used in MORBO[16] because this approach enables better scalability to large batch sizes, by exploiting the submodularity of the acquisition function. Thus, efficient approximation can

be achieved through a greedy optimization strategy, selecting each input in the batch sequentially. The main difference between our approach and the approach used in MORBO is that we used Thompson sampling with a frequentist ensemble of neural networks rather than a Gaussian process as the surrogate function. We used a reference fitness value of 0 for both objectives in the expected improvement acquisition function. The algorithm we used to obtain a batch of 96 proposed samples was implemented using the botorch python package:[17]

1. Sample a function from the ensemble of supervised models (Thompson-style sampling)
2. Evaluate the predicted increase in hypervolume (across two objectives) for all points in the design space, using the sampled function
3. Query the next point in the batch as the point that increases the hypervolume the most
4. Repeat from step one until the entire batch is sampled, but the increase in area is evaluated with respect to expected gains from previously queried points

To speed up calculations, we only evaluated the acquisition function for a subset of the design space by only considering amino acid substitutions that corresponded to variants with some activity in the initial round of data collection, reducing the design space to about 180 million variants, based on a somewhat arbitrary cutoff. Evaluating the acquisition function 96 times took around 3 hours on a single H100 GPU.

D.8. Quantitative Analysis of Select PromPgb Variants

Reactions **C1**, **C8**, **C13**, and **N5** were all validated under controlled conditions with small-scale protein expression of PromPgb and top-performing variants for each of these reactions.

D.8.1. Protocols for Small-Scale Reaction Setup in GC Vials

D.8.1.1. Small-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 6 mL of LB-Amp in a 15-mL culture tube. Cultures were incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 shaking incubator. A 1-mL aliquot of each of these overnight cultures was used to inoculate 50 mL of Terrific Broth with 100 μg/mL of ampicillin (TB-Amp) (0.5% v/v starter culture in expression culture) in 125-mL unbaffled Erlenmeyer flasks. The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaking incubator, at which point they are moved to room temperature for 25 minutes. Protein expression was then induced by direct addition of 50 μL of stock solutions containing 500 mM isopropyl-β-D-thiogalactoside (IPTG) and 1.0 M 5-aminolevulinic acid (ALA) such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16–18 hours in an Innova 42 shaker.

D.8.1.2. Small-Scale Biocatalytic Carbene Transfer Reaction Preparation

The corresponding 50-mL expression cultures were pelleted (4,000 × g for 15 minutes at 4 ºC) and resuspended in 5 mL of M9-N buffer. The optical density at 600 nm (OD_{600}) of this suspension was measured and adjusted to OD_{600} = 31.5 with the addition of more M9-N buffer. A 380-μL aliquot of the cell suspension was added to 2-mL GC vials (Agilent). These whole-cell suspensions were transferred into a vinyl Coy anaerobic chamber, at which point 10 μL of a solution of the reaction substrate in MeCN followed by 10 μL of a

solution of EDA in MeCN were added. The GC vials were tightly capped with screwcaps with a septum and were allowed to shake at room temperature for 16 hours.

Once complete, the reactions were transferred to 1.7-mL Eppendorf tubes and mixed with 600 μL of a 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). The layers were vigorously mixed, and the samples were centrifuged (14,000 × g for 10 minutes at RT). Afterwards, an aliquot of the organic layer was subjected to GC analysis.

D.8.1.3. Small-Scale Biocatalytic Nitrene Transfer Reaction Preparation

The corresponding 50-mL expression cultures were pelleted (4,000 × g for 15 minutes at 4 ºC) and resuspended in 5 mL of M9-N buffer containing 20 mM D-glucose. The optical density at 600 nm (OD_{600}) of this suspension was measured and adjusted to OD_{600} = 31.5 with the addition of more M9-N buffer. A 370 aliquot of the cell suspension was added to 2-mL GC vials (Agilent). These whole-cell suspensions were transferred into a vinyl Coy anaerobic chamber, at which point 20 μL of a solution of the reaction substrate in EtOH followed by 10 μL of a solution of nitrene precursor in H_2O were added. The GC vials were tightly capped with screwcaps with a septum and were allowed to shake at room temperature for 16 hours.

Once complete, the reactions were transferred to 1.7-mL Eppendorf tubes and mixed with 800 μL of an acetonitrile solution containing papaverine as an internal standard (50 μM concentration). The layers are vigorously mixed after which the samples were stored at -20 °C for 1 hour. The samples were then centrifuged (14,000 × g for 10 minutes at RT). Afterwards, an aliquot of the resulting clarified solution was subjected to LC-MS analysis.

D.8.2. Preparation of Calibration Curves for Analytical Yield Determination

All reactions with quantifiable yield were assayed either by gas chromatography equipped with flame ionization detection (GC-FID) or by liquid chromatography equipped with mass spectrometric detection (LC-MS). To quantify yields, calibration curves were prepared

from synthesized authentic standards or products isolated from reactions. Calibration curves were constructed by one of two methods.

D.8.2.1. GC-FID Calibration Curve Construction – Method A

Method A was used for products which could be separated and detected by GC-FID. For these compounds, a 10 mM stock of authentic standard or isolated product was prepared in a solution of 1:1 solution of ethyl acetate:cyclohexane with 1,3,5-trimethoxybenzene as an internal standard (1.0 mM concentration). This stock solution was diluted in the same solution of ethyl acetate:cyclohexane solution containing 1.0 mM standard to a range of concentrations within the range of the measured enzymatic product concentration. In instances where two diastereomers could be formed, calibration curve samples were generated for each diastereomer. Samples were analyzed by GC-FID. Calibration curves were then generated by plotting the concentration of the product against the ratio of its signal (P) to the signal of the internal standard (IS).

D.8.2.2. LC-MS Calibration Curve Construction – Method B

Method B was used for products which could be separated and detected by LC-MS. For these compounds, 380 μL of M9-N buffer were mixed with 20 μL of the authentic product dissolved in ethanol at several unique concentrations such that the sample contained the product at final concentrations within the range of measured enzymatic formation. This mixture was then processed according to the protocol for working up enzymatic reactions for LC-MS quantification (**Small-Scale Biocatalytic Nitrene Transfer Reaction Preparation**).

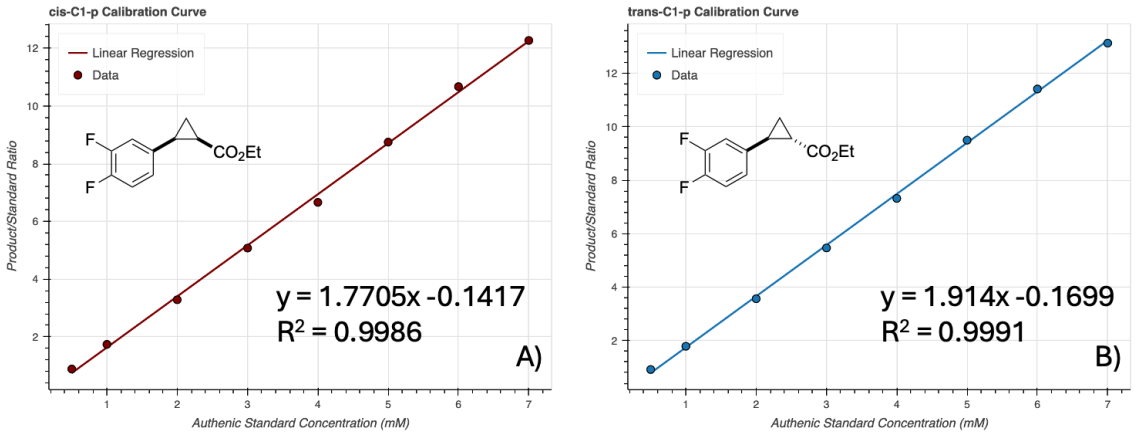


Figure D-70. Calibration curves for the products of reaction **C1**. **(A)** Achiral GC-FID calibration curve for ***cis*-C1-p**. **(B)** Achiral GC-FID calibration curve for ***trans*-C1-p**. Samples were generated using method A.

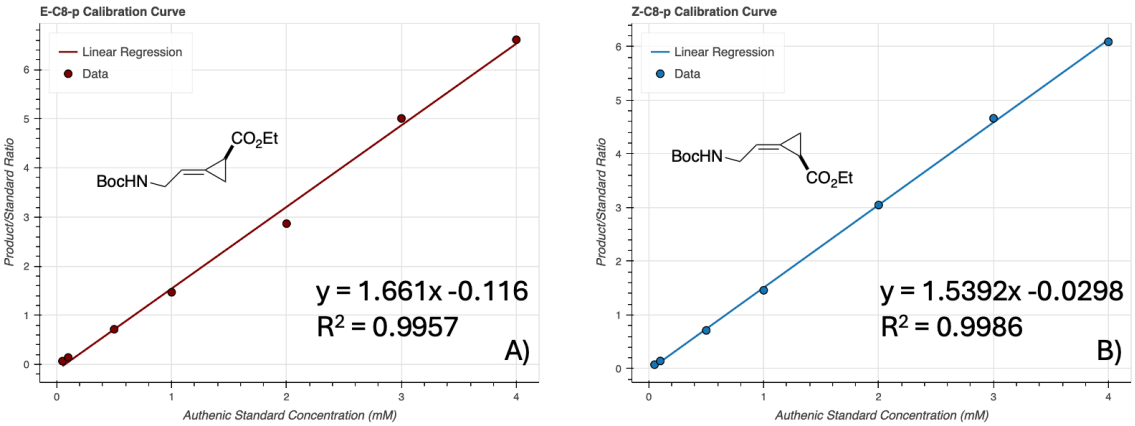


Figure D-71. Calibration curves for the products of reaction **C8**. **(A)** Achiral GC-FID calibration curve for ***E*-C8-p**. **(B)** Achiral GC-FID calibration curve for ***Z*-C8-p**. Samples were generated using method A.

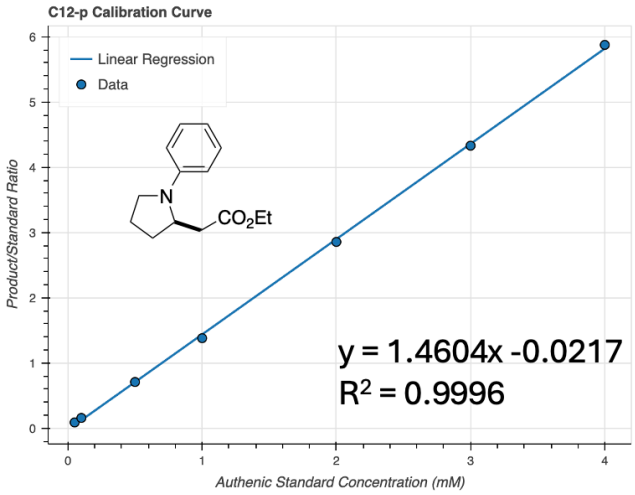


Figure D-72. Achiral GC-FID calibration curve for **C12-p**. Samples were generated using method A.

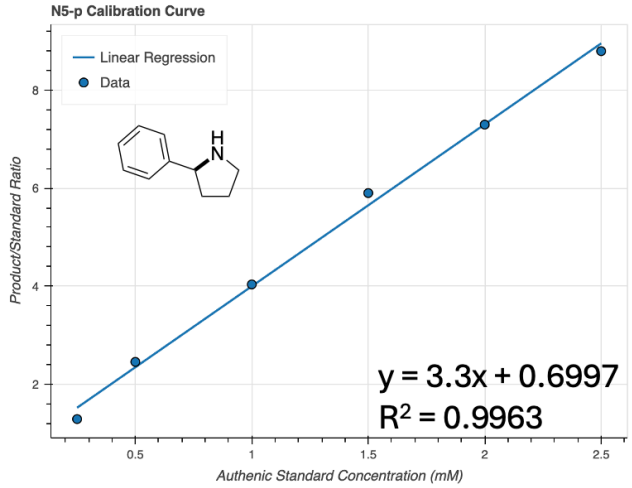
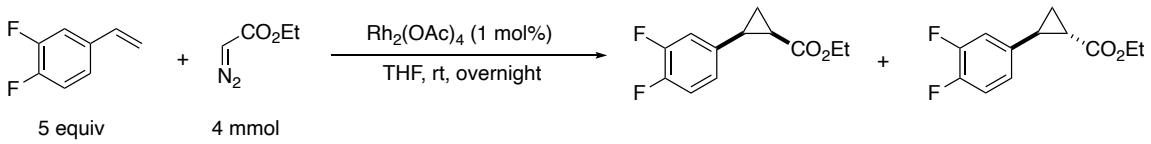


Figure D-73. Achiral GC-FID calibration curve for **N5-p**. Samples were generated using method B.

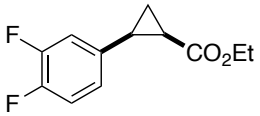
D.9. Preparation of Authentic Standards

*Synthesis of **cis-C1-p** and **trans-C1-p***



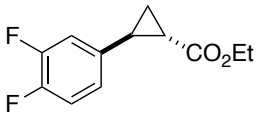
A solution of ethyl diazoacetate (1 equiv, 4 mmol, ≥13 wt.% DCM) in THF (1.2 ml) was added dropwise to a stirred mixture of 1,2-difluoro-4-vinylbenzene (5 equiv, 20 mmol) and rhodium (II) diacetate dimer (1.0 mol %, 0.04 mmol) in dry THF (2 ml) under an argon atmosphere. The reaction mixture was stirred for 16 h, then filtered through a short plug of alumina gel using diethyl ether as an eluent. The filtrate was concentrated *in vacuo*. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the cyclopropanation products as separate diastereomers.

(±)-*cis*-ethyl 2-(3,4-difluorophenyl)cyclopropane-1-carboxylate ***cis-C1-p***



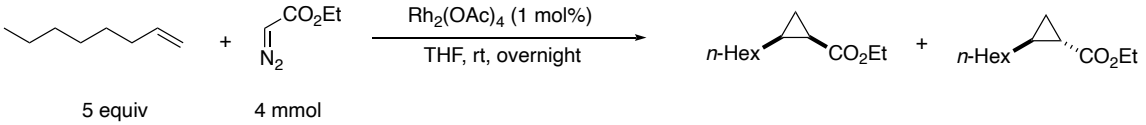
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.12 – 6.93 (m, 3H), 3.93 (q, *J* = 7.1 Hz, 2H), 2.58 – 2.43 (m, 1H), 2.08 (ddd, *J* = 9.2, 7.9, 5.6 Hz, 1H), 1.63 (dt, *J* = 7.4, 5.4 Hz, 1H), 1.35 (ddd, *J* = 8.7, 7.9, 5.2 Hz, 1H), 1.05 (t, *J* = 7.1 Hz, 3H).

(±)-*trans*-ethyl 2-(3,4-difluorophenyl)cyclopropane-1-carboxylate ***trans-C1-p***



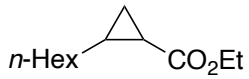
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.05 (dt, *J* = 10.2, 8.3 Hz, 1H), 6.93 – 6.79 (m, 2H), 4.17 (q, *J* = 7.3 Hz, 2H), 2.47 (ddd, *J* = 9.2, 6.4, 4.1 Hz, 1H), 1.84 (ddd, *J* = 8.5, 5.4, 4.2 Hz, 1H), 1.59 (ddd, *J* = 9.2, 5.3, 4.7 Hz, 1H), 1.28 (t, *J* = 7.2 Hz, 3H), 1.26 – 1.21 (m, 1H).

Synthesis of ***cis-C3-p*** *and* ***trans-C3-p***



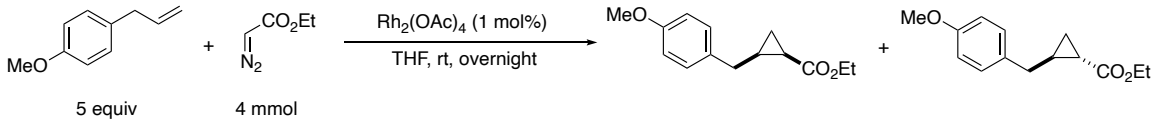
A solution of ethyl diazoacetate (1 equiv, 4 mmol, ≥13 wt.% DCM) in THF (1.2 ml) was added dropwise to a stirred mixture of 1-octene (5 equiv, 20 mmol) and rhodium (II) diacetate dimer (1.0 mol %, 0.04 mmol) in dry THF (2 ml) under an argon atmosphere. The reaction mixture was stirred for 16 h, then filtered through a short plug of alumina gel using diethyl ether as an eluent. The filtrate was concentrated *in vacuo*. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the cyclopropanation products as a mixture of diastereomers.

ethyl 2-hexylcyclopropanecarboxylate ***C3-p***



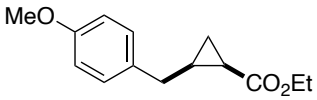
^{1}H NMR (400 MHz, $CDCl_3$) δ 4.06 (q, *J* = 7.1 Hz, 2H), 1.59 (ddd, *J* = 8.9, 7.8, 5.5 Hz, 1H), 1.19 (td, *J* = 8.6, 4.2 Hz, 13H), 0.92 (td, *J* = 8.1, 4.4 Hz, 1H), 0.88 – 0.83 (m, 1H), 0.83 – 0.78 (m, 3H).

Synthesis of ***cis-C4-p*** *and* ***trans-C4-p***



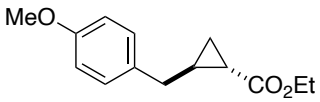
A solution of ethyl diazoacetate (1 equiv, 4 mmol, ≥13 wt.% DCM) in THF (1.2 ml) was added dropwise to a stirred mixture of 2-allylanisole (5 equiv, 20 mmol) and rhodium (II) diacetate dimer (1.0 mol %, 0.04 mmol) in dry THF (2 ml) under an argon atmosphere. The reaction mixture was stirred for 16 h, then filtered through a short plug of alumina gel using diethyl ether as an eluent. The filtrate was concentrated *in vacuo*. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the cyclopropanation products as separate diastereomers.

(±)-*cis*-ethyl 2-(4-methoxybenzyl)cyclopropanecarboxylate ***cis-C4-p***



^{1}H NMR (400 MHz, $CDCl_3$) δ 7.17 – 7.08 (m, 2H), 6.88 – 6.78 (m, 2H), 4.11 (qd, *J* = 7.2, 1.0 Hz, 2H), 3.79 (s, 3H), 2.70 (dd, *J* = 14.8, 6.4 Hz, 1H), 2.52 (dd, *J* = 14.8, 7.1 Hz, 1H), 1.71 – 1.61 (m, 1H), 1.49 (dt, *J* = 8.5, 4.4 Hz, 1H), 1.31 – 1.16 (m, 4H), 0.81 (ddd, *J* = 8.2, 6.4, 4.2 Hz, 1H).

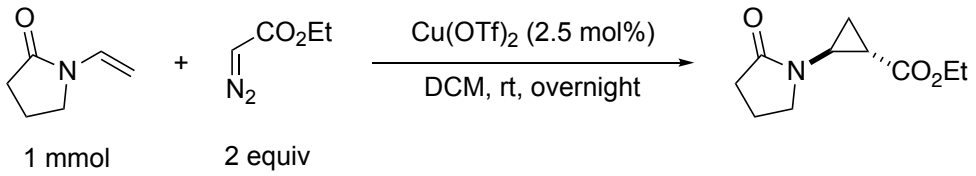
(±)-*trans*-ethyl 2-(4-methoxybenzyl)cyclopropanecarboxylate ***trans-C4-p***



^{1}H NMR (400 MHz, $CDCl_3$) δ 7.16 – 7.08 (m, 2H), 6.85 – 6.79 (m, 2H), 4.13 (q, *J* = 7.1 Hz, 2H), 3.78 (s, 3H), 2.86 (dd, *J* = 14.9, 6.9 Hz, 1H), 2.77 (dd, *J* = 14.9, 7.6 Hz, 1H), 1.76

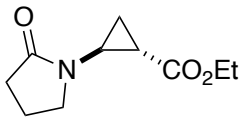
(ddd, *J* = 8.8, 7.5, 5.8 Hz, 1H), 1.55 – 1.43 (m, 1H), 1.23 (d, *J* = 7.2 Hz, 4H), 1.14 – 0.99 (m, 2H).

Synthesis of ***trans-C6-p***



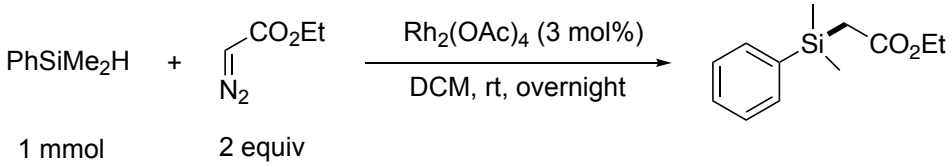
A solution of ethyl diazoacetate (2 equiv, 2 mmol, ≥13 wt.% DCM) in DCM (2 ml) was added dropwise to a stirred mixture of *N*-vinyl-pyrrolidone (1 mmol) and copper(II) trifluoromethanesulfonate (2.5 mol %, 0.025mmol) in dry DCM (5 ml) under an argon atmosphere. The reaction mixture was stirred for 16 h, then filtered through a short plug of silica gel using EtOAc as an eluent. The filtrate was concentrated *in vacuo*. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the cyclopropanation product.

Ethyl 2-(2-oxopyrrolidin-1-yl)cyclopropane-1-carboxylate ***trans*-C6-p**



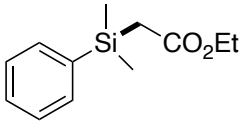
Yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 4.13 (qt, *J* = 7.1, 1.2 Hz, 2H), 3.30 (dd, *J* = 7.4, 6.6 Hz, 2H), 3.21 – 3.10 (m, 1H), 2.38 (t, *J* = 8.1 Hz, 2H), 1.99 (dddd, *J* = 14.9, 8.3, 7.3, 0.9 Hz, 2H), 1.83 (dddd, *J* = 9.1, 5.9, 3.1, 0.7 Hz, 1H), 1.50 – 1.34 (m, 2H), 1.25 (td, *J* = 7.1, 0.9 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 176.2, 172.3, 61.0, 47.3, 34.1, 31.7, 19.8, 18.1, 14.2, 14.1.

Synthesis of ***C12-p***



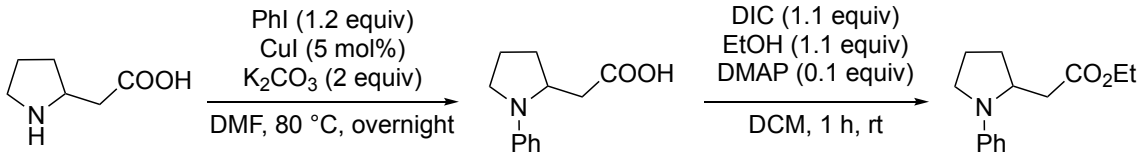
A solution of ethyl diazoacetate (2 equiv, 2 mmol, ≥13 wt.% DCM) in DCM (2 ml) was added dropwise to a stirred mixture of dimethylphenylsilane (1 mmol) and rhodium(II) acetate (3 mol%, 0.03 mmol) in dry DCM (5 ml) under an argon atmosphere. The reaction mixture was stirred for 16 h, then filtered through a short plug of silica gel using EtOAc as an eluent. The filtrate was concentrated *in vacuo*. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the product.

Ethyl 2-(dimethyl(phenyl)silyl)acetate **C12-p**



Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.59 – 7.48 (m, 2H), 7.44 – 7.32 (m, 3H), 4.04 (q, *J* = 7.1 Hz, 2H), 2.11 (s, 2H), 1.16 (t, *J* = 7.1 Hz, 3H), 0.41 (s, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.7, 137.1, 133.6, 129.6, 128.0, 60.1, 26.4, 14.4, -2.6.

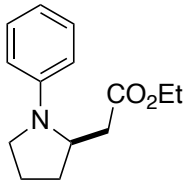
Synthesis of ***C13-p***



Step 1: To a sealed tube flushed with nitrogen were added pyrrolidine (5 mmol), potassium carbonate (2 equiv, 10 mmol), copper (I) iodide (0.25 mmol, 5 mol%), iodobenzene (1.2 equiv, 6 mmol) and DMF (7.5 ml). The mixture was heated at 90°C for 48 hours, then cooled to room temperature. Water was added, and the pH value was adjusted to <3 with concentrated HCl. The aqueous phase was extracted four times with ethyl acetate. The combined organic layers were washed with brine, dried over magnesium sulfate, filtered and concentrated *in vacuo*. Purification by silica gel chromatography (0 to 100% EtOAc/hexane gradient) afforded the product.

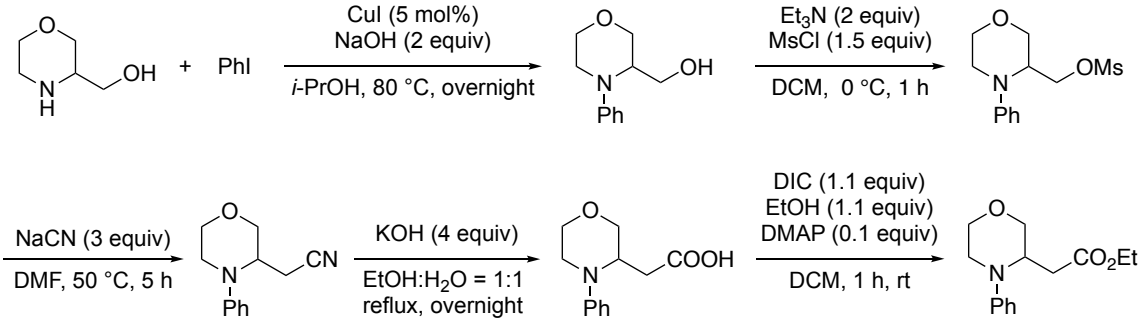
Step 2: A round-bottom flask was charged with carboxylic acid (1.0 equiv), ethanol (1.1 equiv), and DMAP (0.1 equiv). Dichloromethane was added (0.4 M), and the mixture was stirred vigorously. DIC (1.1 equiv) was then added dropwise via syringe, and the reaction mixture was stirred until consumption of the acid was complete, as determined by TLC. The mixture was filtered through a fritted funnel and rinsed with CH_2Cl_2/Et_2O. The solvent was removed under reduced pressure, and purification by column chromatography afforded the corresponding ester.

Ethyl 2-(1-phenylpyrrolidin-2-yl)acetate **C13-p**



Yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.26 – 7.22 (m, 2H), 6.71 – 6.65 (m, 1H), 6.65 – 6.55 (m, 2H), 4.20 – 4.15 (m, 3H), 3.46 – 3.39 (m, 1H), 3.22 – 3.13 (m, 1H), 2.79 (dd, *J* = 15.0, 2.9 Hz, 1H), 2.26 – 2.17 (m, 1H), 2.17 – 1.94 (m, 4H), 1.28 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.1, 146.5, 129.5, 129.4, 121.6, 116.0, 111.9, 55.4, 47.9, 37.8, 31.0, 23.1, 14.3.

*Synthesis of **C14-p***



Step 1: A seal tube with a magnetic stirring bar was charged with CuI (1 mmol, 0.05 equiv.), powder NaOH (40 mmol, 2 equiv), and morpholin-3- ylmethanol (20 mmol, 1 equiv). To the tube were added isopropyl alcohol (5 mL, 0.8 M) and iodobenzene (24 mmol, 1.2 equiv) via syringe. The reaction was heated to 80 °C and left to stir overnight. The following day the reaction was reduced *in vacuo* and reconstituted in DCM and water. The mixture was introduced into a separatory funnel and the organic layer was separated. The water layer was extracted twice with DCM, and the combined organic layers were washed with brine

and dried over $MgSO_4$. The solution was reduced *in vacuo* and purified via flash chromatography, affording the arylation product as an oil.

Step 2: A round-bottom flask with a magnetic stirring bar was charged with (4-phenylmorpholin-3-yl)methanol (10 mmol, 1 equiv) and placed under an argon atmosphere. Thereafter, DCM (0.2 M) was added via syringe and the reaction was stirred and cooled to 0 °C (ice/water bath). Triethylamine (2 equiv, 2.0 mmol) and methanesulfonyl chloride (MsCl) (1.5 equiv, 1.5 mmol) were added dropwise in that order. The reaction was left to warm to room temperature and stirred for 1 hour at that temperature. The reaction was then introduced into a separatory funnel, washed with saturated $NaHCO_3$, brine, and dried over $MgSO_4$. The resultant oil was used immediately in the following reaction.

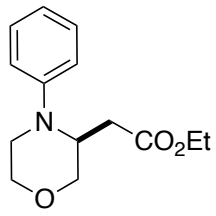
Step 3: A round-bottom flask with a magnetic stirring bar and the resultant oil from the previous reaction was charged with DMF (0.2 M) and NaCN (30 mmol, 3 equiv.). The reaction was sealed, blanketed with an argon atmosphere, and heated to 60 °C for 3 h. The reaction was cooled to room temperature and diluted with ethyl acetate and water. The organic layer was washed several times with brine, dried over $MgSO_4$, and reduced *in vacuo*. The resultant residue was purified using flash chromatography to afford the product as an oil.

Step 4: To a solution of 2-(4-phenylmorpholin-3-yl)acetonitrile (5 mmol) in ethanol (5 mL) and water (5 mL) was added potassium hydroxide (20 mmol) and the resulting mixture was refluxed overnight. The ethanol was removed under reduced pressure, and then the solution was cooled to below 10 °C. and acidified with concentrated HCl to pH 1. The mixture was extracted with EtOAc and the combined organic extracts were dried over anhydrous sodium sulfate and concentrated under vacuum to afford the carboxylic acid as a yellow oil.

Step 5: A round-bottom flask was charged with carboxylic acid (1.0 equiv, 3 mmol), ethanol (1.1 equiv), and DMAP (0.1 equiv). Dichloromethane was added (0.4 M), and the mixture was stirred vigorously. DIC (1.1 equiv) was then added dropwise via syringe, and the reaction mixture was stirred until consumption of the acid was complete, as determined by TLC. The mixture was filtered through a fritted funnel and rinsed with

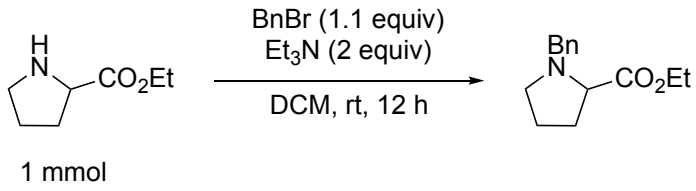
CH_2Cl_2/Et_2O. The solvent was removed under reduced pressure, and purification by column chromatography afforded the corresponding ester.

Ethyl 2-(4-phenylmorpholin-3-yl)acetate **C14-p**



Yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 – 7.25 (m, 2H), 6.92 – 6.82 (m, 3H), 4.19 – 4.14 (m, 1H), 4.09 – 3.98 (m, 3H), 3.92 (dt, *J* = 11.6, 1.5 Hz, 1H), 3.85 (ddd, *J* = 11.5, 2.8, 1.6 Hz, 1H), 3.72 (td, *J* = 11.3, 3.4 Hz, 1H), 3.19 (ddd, *J* = 12.7, 3.4, 1.8 Hz, 1H), 3.10 (ddd, *J* = 12.2, 11.3, 3.7 Hz, 1H), 2.86 (dd, *J* = 15.8, 10.2 Hz, 1H), 2.36 (ddd, *J* = 15.8, 3.4, 1.6 Hz, 1H), 1.21 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.6, 149.3, 129.8, 120.1, 115.9, 69.8, 67.3, 60.9, 52.5, 43.4, 30.5, 14.5.

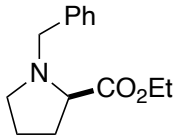
*Synthesis of **C10-p***



Triethylamine (2 equiv, 2 mmol) was added to a solution of pyrrolidine (1 mmol) in 10 mL DCM at room temperature. Benzyl bromide (1.1 equiv, 1.1 mmol) was added dropwise to the reaction mixture and the resulting solution was refluxed for 12 hrs. After cooling, sodium hydroxide (10 mL of 1 N solution) was added to the reaction mixture, and the product was extracted with DCM (2 x 30 mL). All the organic layers were combined, washed with brine, dried over $MgSO_4$, and concentrated under reduced pressure to afford

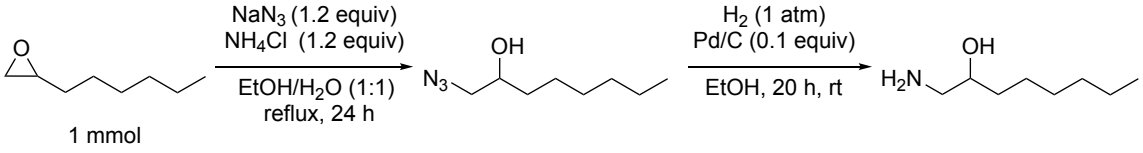
the crude product. The resulting residue was purified by column chromatography on silica gel in hexanes/EtOAc to afford the product.

Ethyl benzyl-prolinate **C10-p**



^{1}H NMR (400 MHz, $CDCl_3$) δ 7.31 – 7.21 (m, 4H), 7.21 – 7.14 (m, 1H), 4.06 (qd, *J* = 7.1, 1.9 Hz, 2H), 3.86 (d, *J* = 12.7 Hz, 1H), 3.49 (d, *J* = 12.7 Hz, 1H), 3.17 (dd, *J* = 8.8, 6.3 Hz, 1H), 2.98 (ddd, *J* = 9.0, 7.6, 3.1 Hz, 1H), 2.32 (td, *J* = 8.8, 7.8 Hz, 1H), 2.13 – 2.01 (m, 1H), 1.95 – 1.77 (m, 2H), 1.77 – 1.65 (m, 1H), 1.19 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 174.3, 138.5, 129.3, 128.3, 127.2, 60.6, 58.8, 53.3, 29.4, 23.0, 14.4.

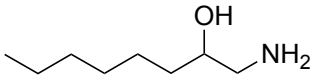
*Synthesis of **N9-p***



NaN_3 (1.2 equiv, 1.2 mmol), NH_4Cl (1.2 equiv, 1.2 mmol) and a 1,2-epoxy-*n*-alkane (1 mmol) were dissolved in sufficient EtOH/H_2O (1:1 v:v) to give a homogeneous solution. The solution was refluxed for 24 h, then the EtOH was removed *in vacuo.* The residual aq. soln. was extracted with Et_2O (3 × 10 ml) and the combined organic extracts washed with water and dried ($MgSO_4$). After filtration and concentration *in vacuo*, purification by flash column chromatography (hexanes/EtOAc) gave a colorless oil. The azidoalcohol and Pd/C catalyst (0.1 equiv, w:w) were placed in a flask with EtOH (3 ml). The mixture was degassed by three freeze-pump-thaw cycles from an H_2 atmosphere. The mixture was then stirred under H_2 (balloon, 1 atm.) at room temperature for 20 h. After filtration through

Celite®, the EtOH was removed *in vacuo* to give the product as a colorless oil. No further purification was necessary.

1-Aminooctan-2-ol **N8-p**



Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 3.81 – 3.71 (m, 1H), 3.38 (dd, *J* = 12.4, 3.3 Hz, 1H), 3.25 (dd, *J* = 12.4, 7.4 Hz, 1H), 2.05 (s, 1H), 1.46 (ddt, *J* = 15.8, 11.7, 5.1 Hz, 3H), 1.36 – 1.23 (m, 7H), 0.92 – 0.85 (m, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 71.0, 57.2, 34.4, 31.8, 29.3, 25.5, 22.7, 14.2.

D.10. Preparative-Scale Enzymatic Reactions

D.10.1. Protocols of Preparative-Scale Enzymatic Reactions

D.10.1.1. Liter-Scale Protein Expression

Single colonies from LB-Agar plates were picked using a sterile pipette tip and were used to inoculate 50 mL of LB-Amp in a 125-mL Erlenmeyer flask. Cultures were incubated at 37 ºC with shaking at 220 rpm overnight in an Innova 4000 shaking incubator. A 25-mL aliquot of each of these overnight cultures was used to inoculate 1 L of Terrific Broth with 100 μg/mL of ampicillin (TB-Amp) (0.5% v/v starter culture in expression culture) in a 3.2-L unbaffled Erlenmeyer flask. The expression cultures were incubated at 37 ºC and 220 rpm for 2.5 hours in an Innova 42 shaking incubator, at which point they were moved to room temperature for 25 minutes. Protein expression was then induced by direct addition of 50 μL of stock solutions containing 500 mM isopropyl-β-D-thiogalactoside (IPTG) and 1.0 M 5-aminolevulinic acid (ALA) such that the final concentrations were 0.5 mM and 1.0 mM, respectively. The cultures were shaken at 22 ºC and 140 rpm for 16–18 hours in an Innova 42 shaker.

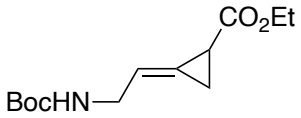
D.10.1.2. Large-Scale Biocatalytic Carbene Transfer

The corresponding l-L expression cultures were pelleted (4,000 × g for 30 minutes at 4 ºC) and resuspended in 100 mL of M9-N buffer. The optical density at 600 nm (OD_{600}) of this suspension was measured and adjusted to OD_{600} = 31.5 with the addition of more M9-N buffer. A 38-mL aliquot of the cell suspension was added to a 60-mL test tube with screw-cap seal. These whole-cell suspensions were transferred into a vinyl Coy anaerobic chamber, at which point 1 mL of a solution of the reaction substrate in MeCN was added. Subsequently, 1 mL of a solution of EDA in MeCN was added dropwise to the reaction vessel to prevent rapid release of N_2 gas upon diazo activation. The GC vials were tightly capped with screw caps equipped with septum and were allowed to shake at room temperature for 16 hours.

Once complete, the reaction was split into two 20-mL aliquots across two 50-mL falcon tubes. The suspensions are extracted three times with ethyl acetate as follows: 20 mL of ethyl acetate were added to each tube and the phases were mixed by hand-shaking the tubes. Upon thorough mixing, the phases were separated by centrifugation (5000 × g, 5 minutes, RT) and the organic phase was siphoned off. The combined organics were dried over sodium sulfate; once dry, the drying agent was decanted, and the volatiles removed under vacuum. The crude residue was purified by silica gel column chromatography to yield the titled compounds.

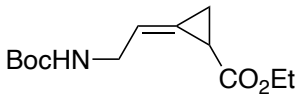
D.10.1.3. Characterization of Isolated Products

Ethyl (*E*)-2-(2-((*tert*-butoxycarbonyl)amino)ethylidene)cyclopropane-1-carboxylate ***E*-C8-p**



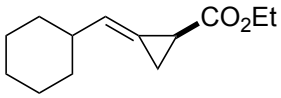
Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.91 (ddt, *J* = 5.7, 3.7, 2.1 Hz, 1H), 4.67 (s, 1H), 4.16 (qd, *J* = 7.2, 1.2 Hz, 2H), 3.85 (d, *J* = 6.2 Hz, 2H), 2.29 (dddd, *J* = 6.9, 4.6, 2.2, 1.1 Hz, 1H), 1.73 (td, *J* = 5.5, 2.3 Hz, 1H), 1.61 (tt, *J* = 6.6, 2.2 Hz, 1H), 1.44 (s, 9H), 1.27 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.2, 155.9, 123.5, 116.4, 79.5, 61.0, 41.6, 36.7, 28.5, 24.8, 23.4, 17.4, 14.3, 10.6. HRMS (FI+) *m/z*: $[M]^{\cdot+}$ calcd. for $C_{13}H_{21}NO_4$ 255.1465, found: 255.1463.

Ethyl (*Z*)-2-(2-((*tert*-butoxycarbonyl)amino)ethylidene)cyclopropane-1-carboxylate ***Z*-C8-p**



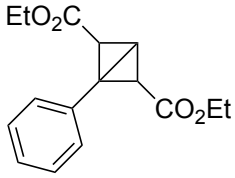
Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.91 (ddq, *J* = 5.7, 4.1, 2.2 Hz, 1H), 4.75 (s, 1H), 4.13 (qd, *J* = 7.1, 0.8 Hz, 2H), 3.91 (d, *J* = 7.3 Hz, 2H), 2.26 – 2.16 (m, 1H), 1.81 (ddd, *J* = 8.9, 4.5, 2.4 Hz, 1H), 1.63 (ddd, *J* = 8.6, 7.5, 2.2 Hz, 1H), 1.44 (s, 9H), 1.26 (d, *J* = 7.1 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.2, 123.4, 116.9, 79.5, 60.9, 41.5, 28.5, 16.9, 14.3, 10.9. HRMS (FI+) *m/z*: $[M]^{\cdot+}$ calcd. for $C_{13}H_{21}NO_4$ 255.1465, found: 255.1468.

Ethyl (*E*)-2-(cyclohexylmethylene)cyclopropane-1-carboxylate ***E*-C9-p**



Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.87 – 5.69 (m, 1H), 4.12 (q, J = 7.1 Hz, 2H), 2.26 – 2.11 (m, 2H), 1.83 – 1.60 (m, 7H), 1.34 – 1.15 (m, 8H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.7, 125.3, 119.2, 60.3, 39.8, 32.4, 32.2, 26.0, 25.8, 25.8, 16.4, 14.0, 11.4.

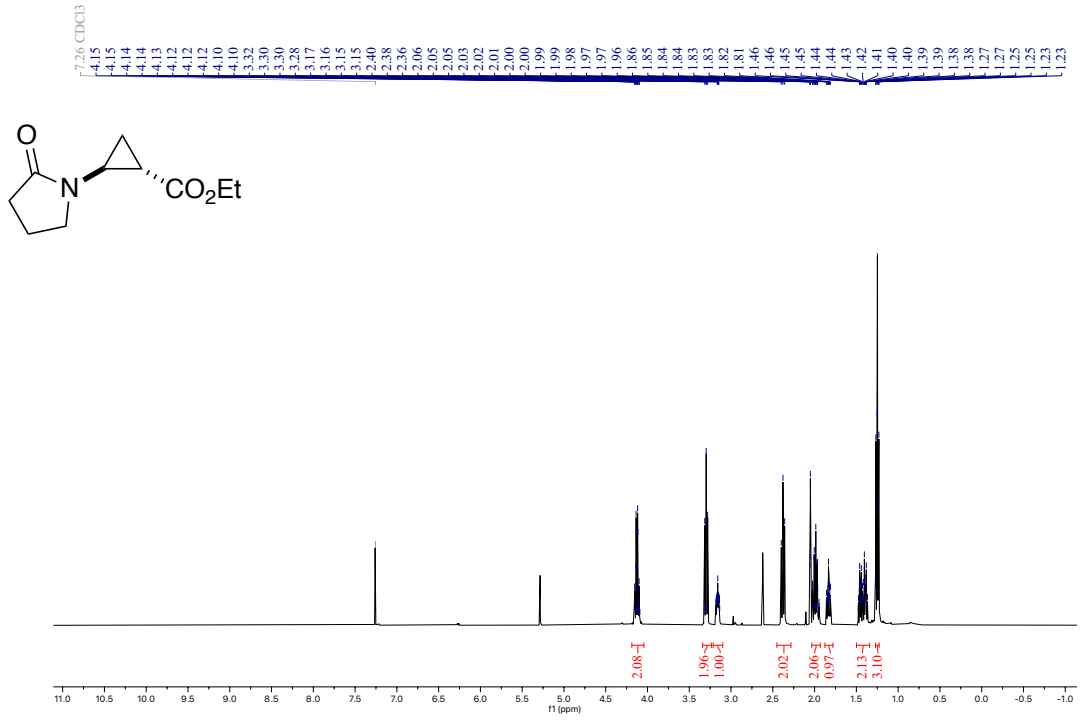
Diethyl 1-phenylbicyclo[1.1.0]butane-2,4-dicarboxylate **C7-p**



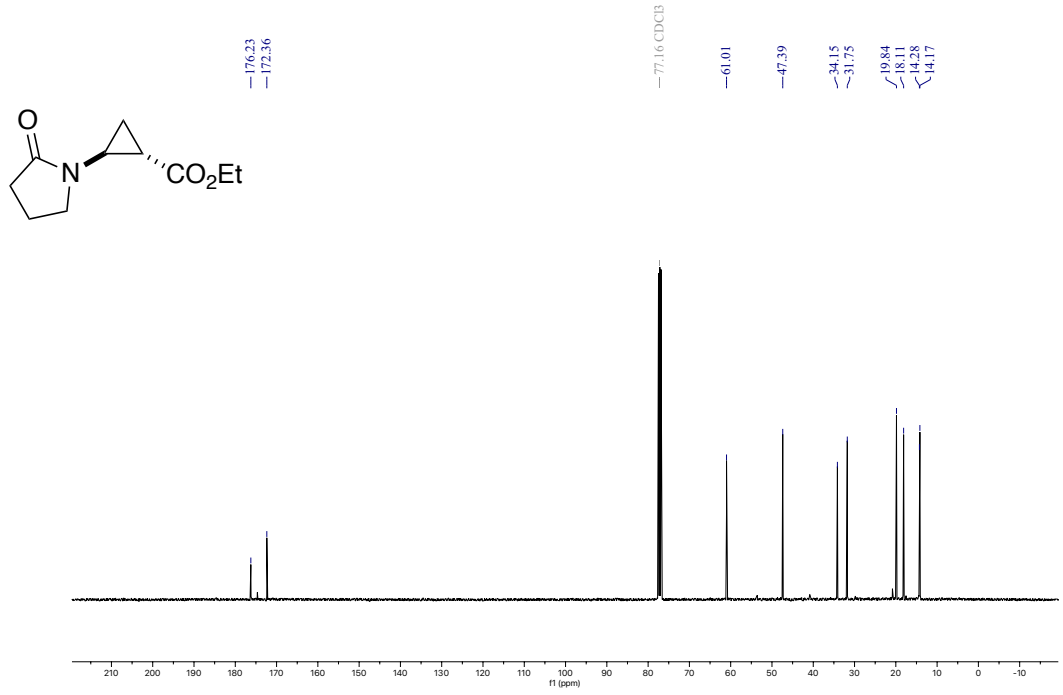
Colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.56 – 7.49 (m, 2H), 7.33 – 7.26 (m, 3H), 4.08 – 3.93 (m, 4H), 3.40 (s, 1H), 1.81 (s, 2H), 1.57 (s, 2H), 1.08 (t, J = 7.1 Hz, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 167.3, 131.7, 129.3, 128.1, 127.9, 60.8, 40.5, 25.6, 15.1, 14.1.

D.11. NMR Spctra of Authentic Standards and Isolated Products

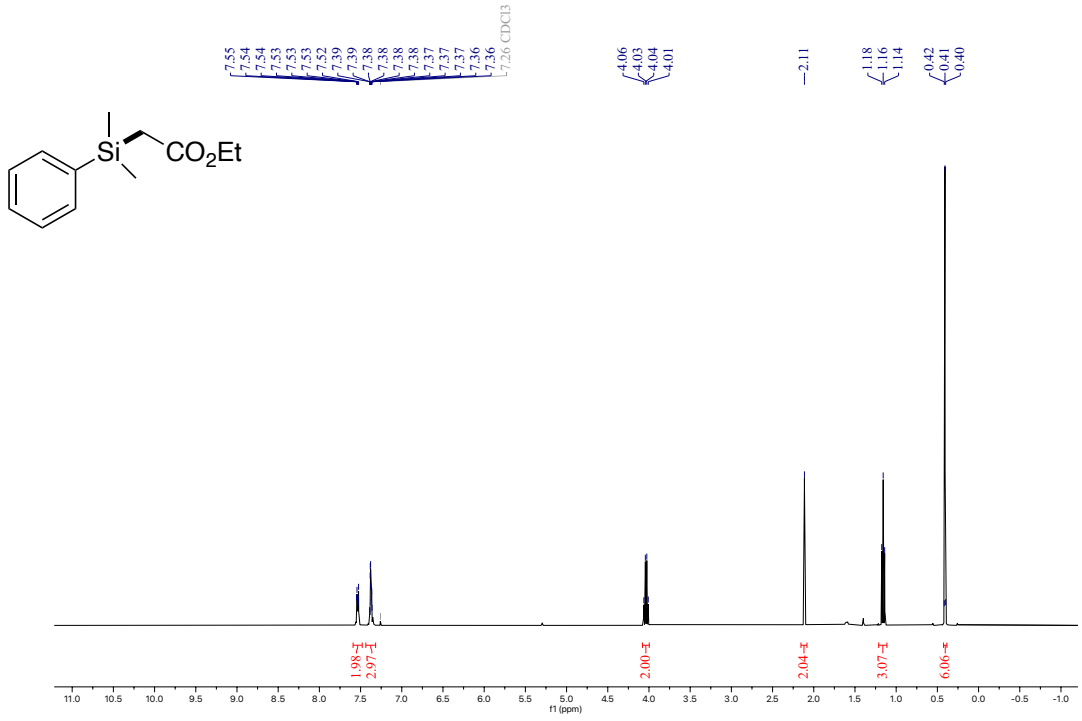
^{1}H NMR (400 MHz, $CDCl_3$) of ***trans*-C6-p**



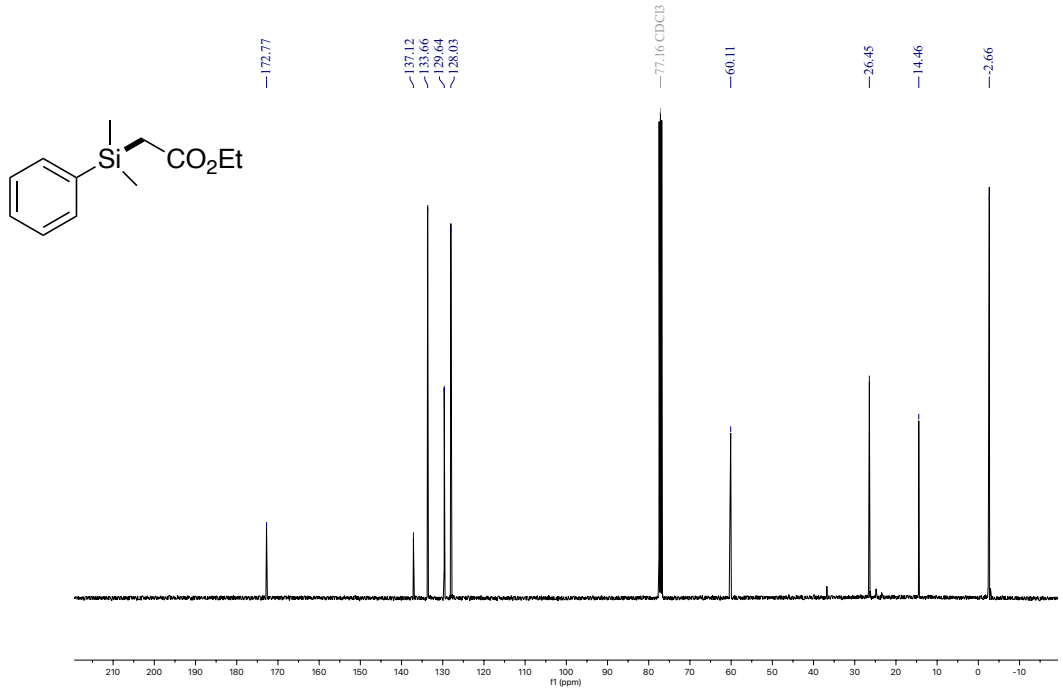
^{13}C NMR (101 MHz, $CDCl_3$) of ***trans*-C6-p**



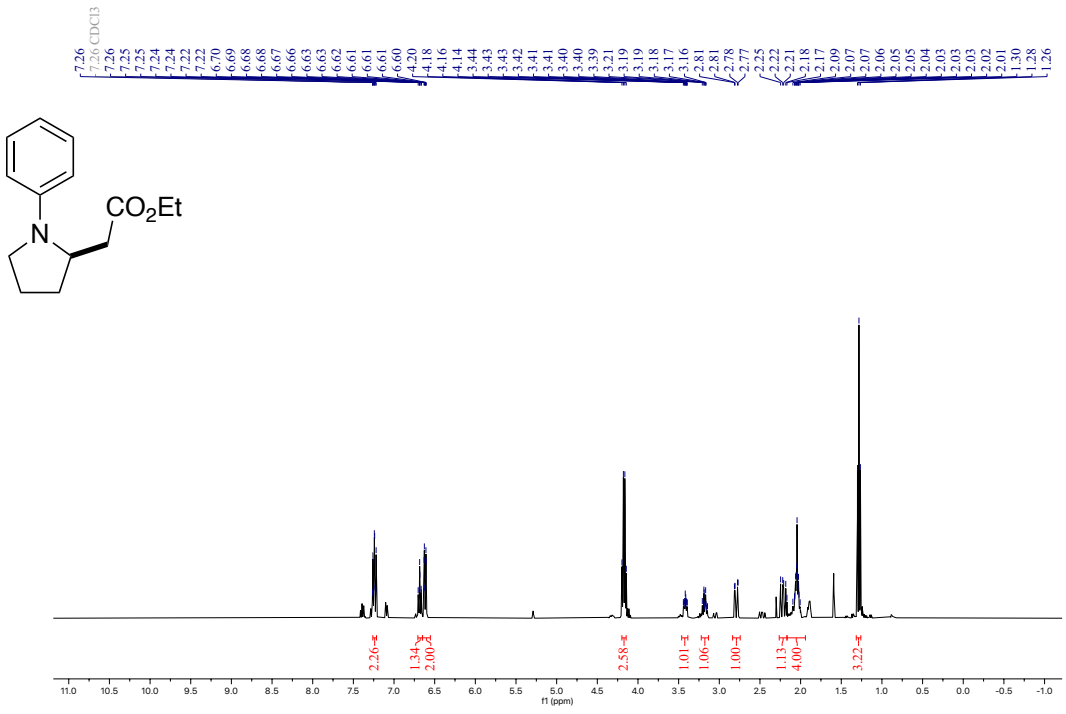
^{1}H NMR (400 MHz, $CDCl_3$) of **C12-p**



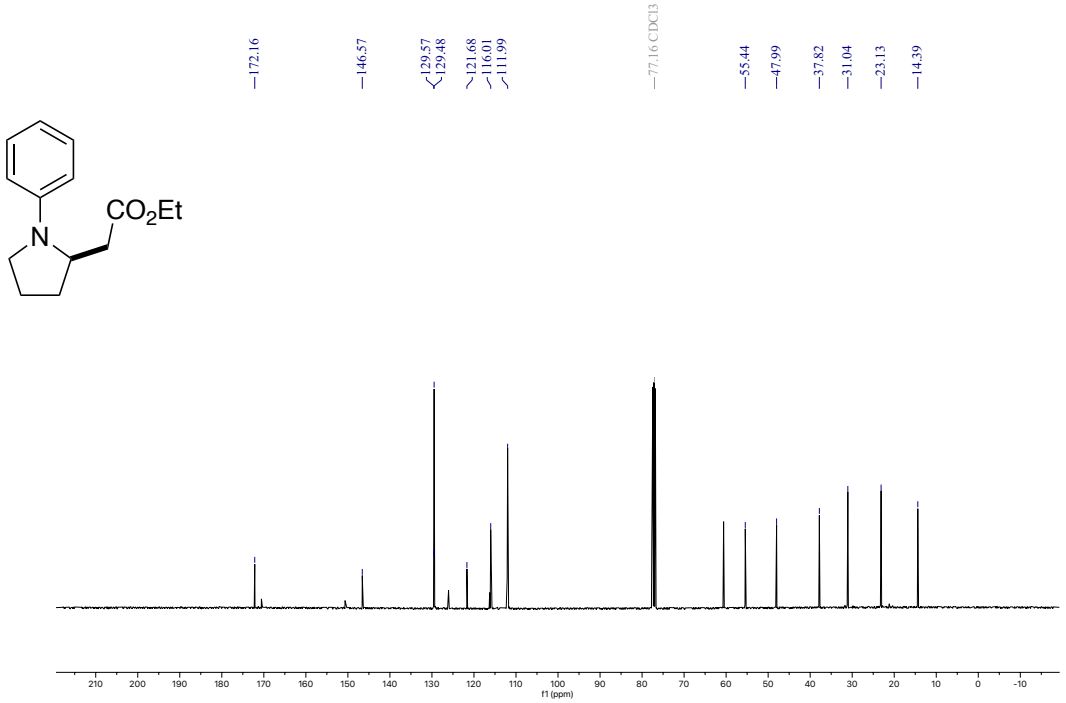
^{13}C NMR (101 MHz, $CDCl_3$) of **C12-p**



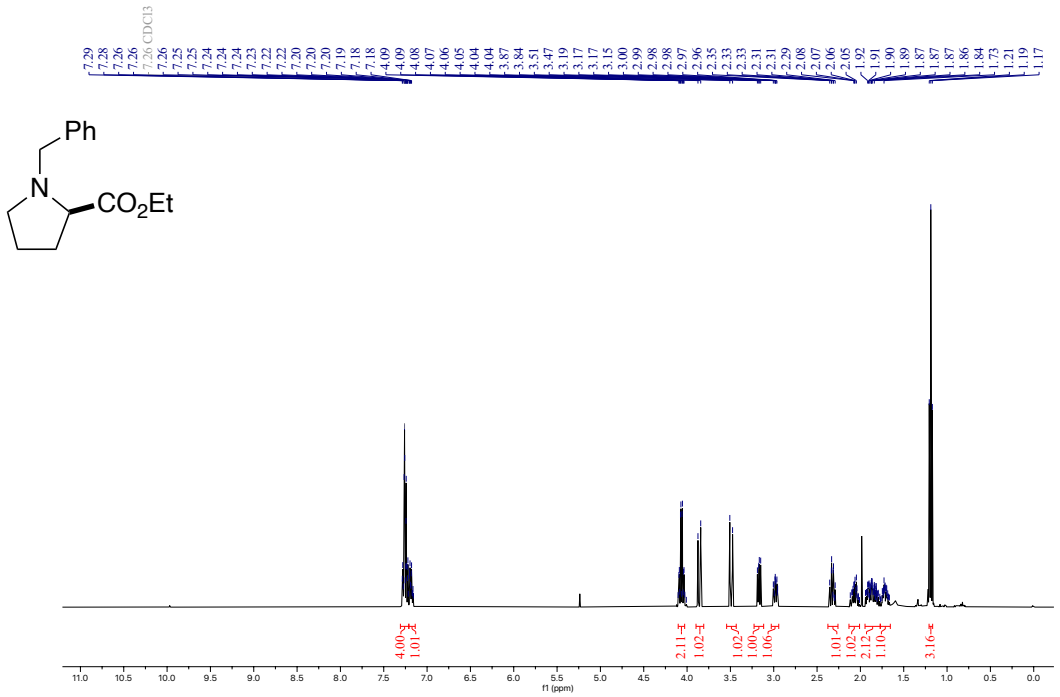
^{1}H NMR (400 MHz, $CDCl_3$) of **C13-p**



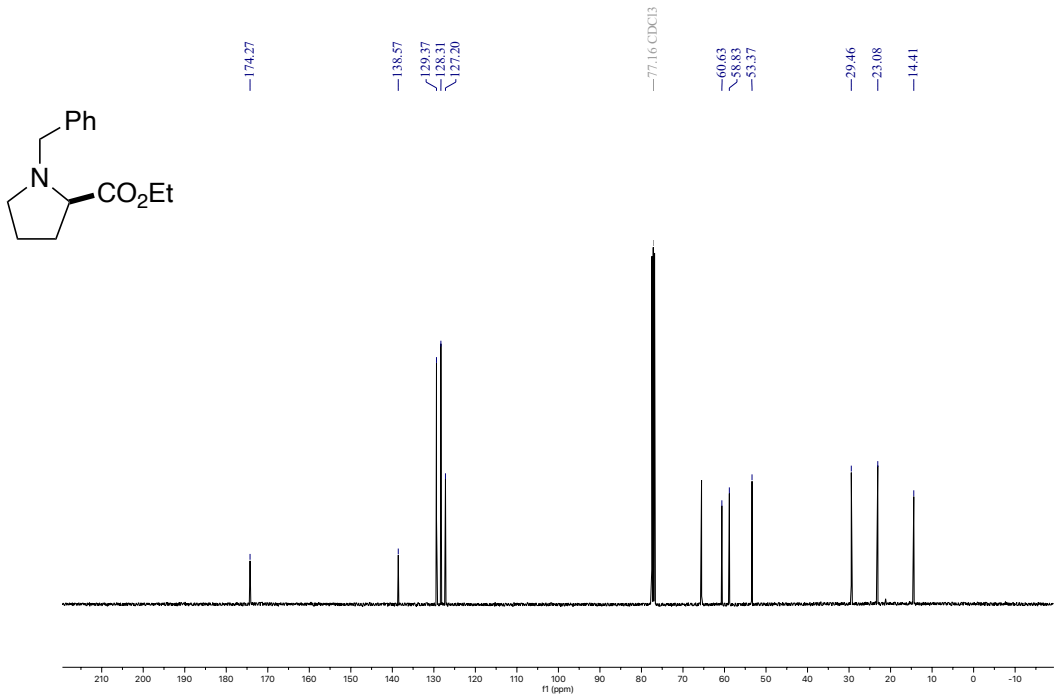
^{13}C NMR (101 MHz, $CDCl_3$) of **C13-p**



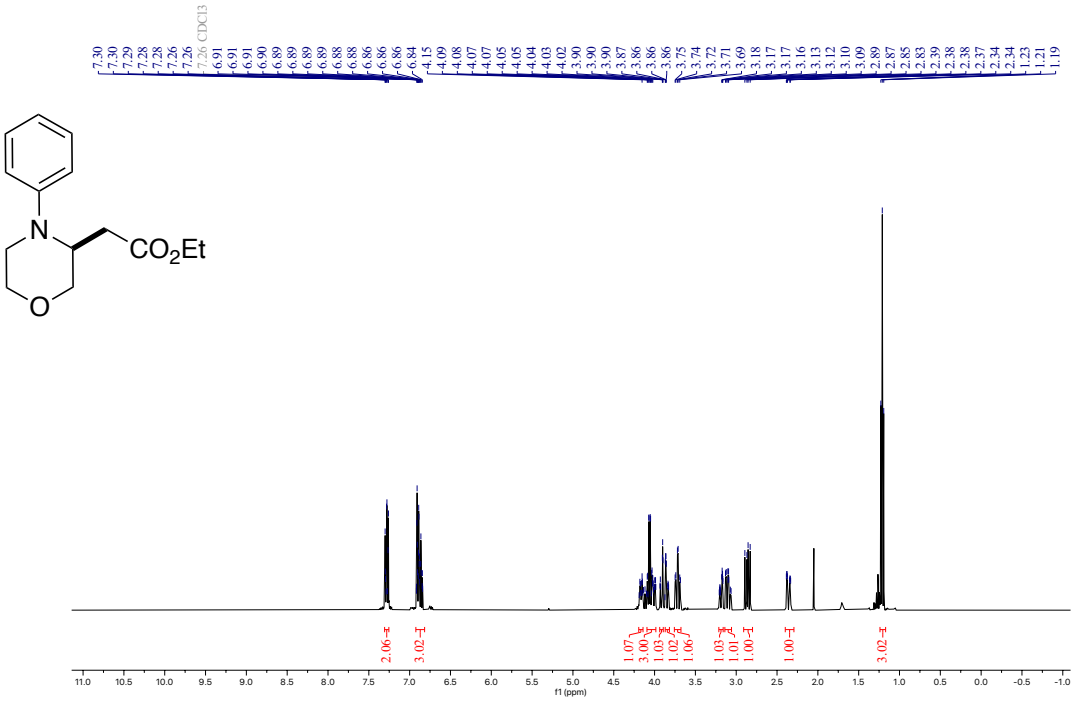
^{1}H NMR (400 MHz, $CDCl_3$) of **C10-p**



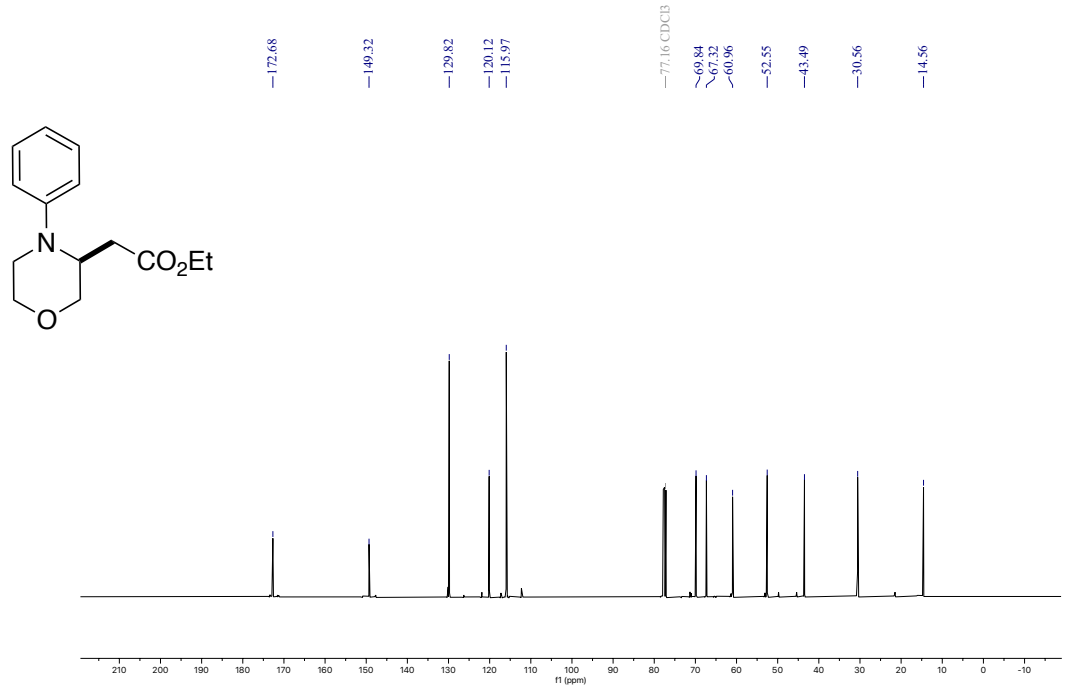
^{13}C NMR (101 MHz, $CDCl_3$) of **C10-p**



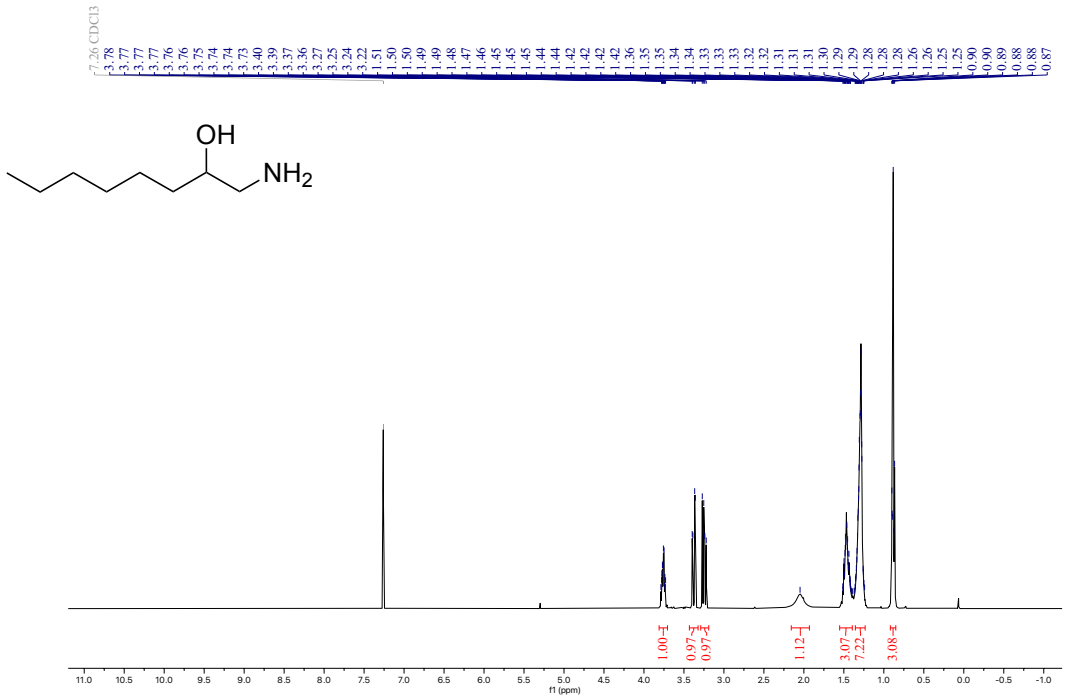
^{1}H NMR (400 MHz, $CDCl_3$) of **C14-p**



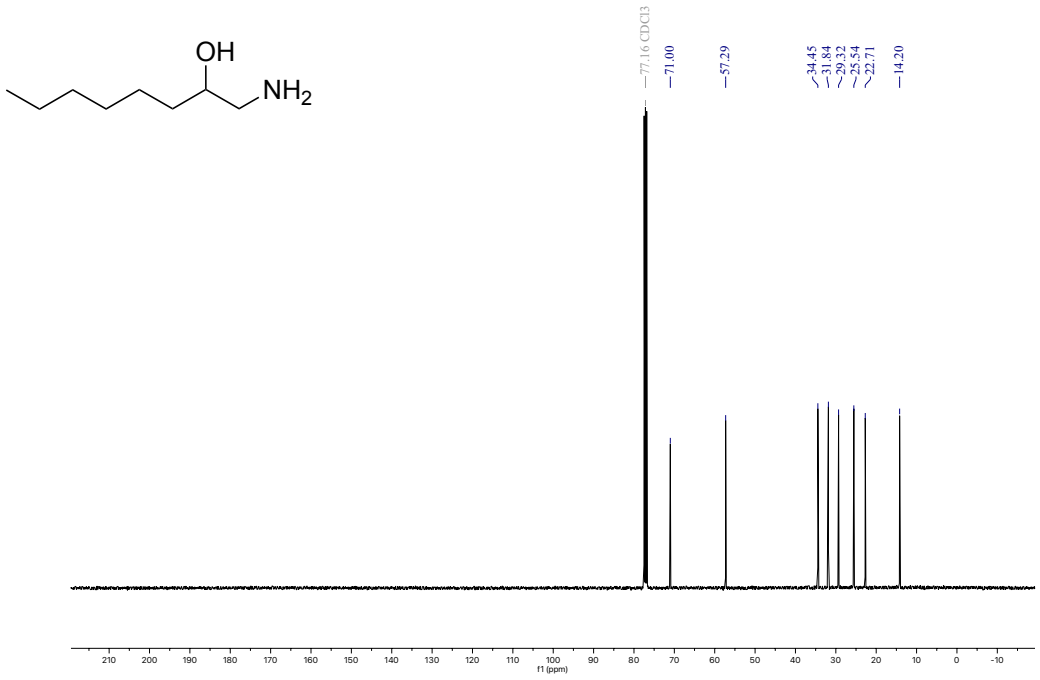
^{13}C NMR (101 MHz, $CDCl_3$) of **C14-p**



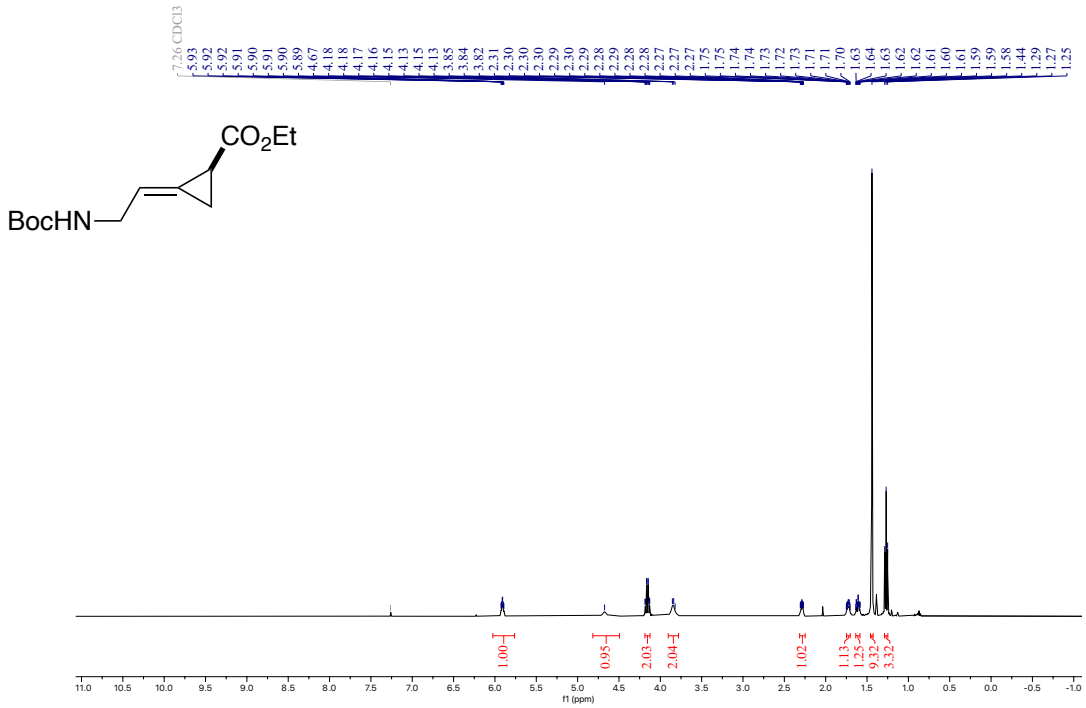
^{1}H NMR (400 MHz, $CDCl_3$) of **N9-p**



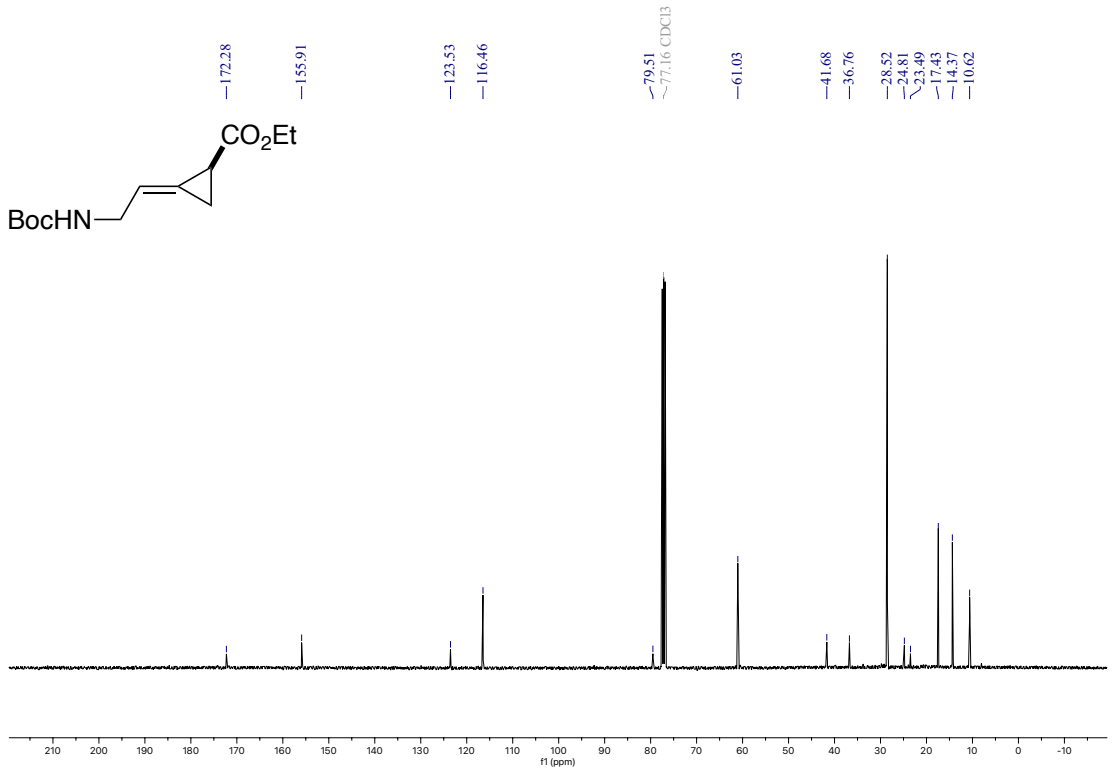
^{13}C NMR (101 MHz, $CDCl_3$) of **N9-p**



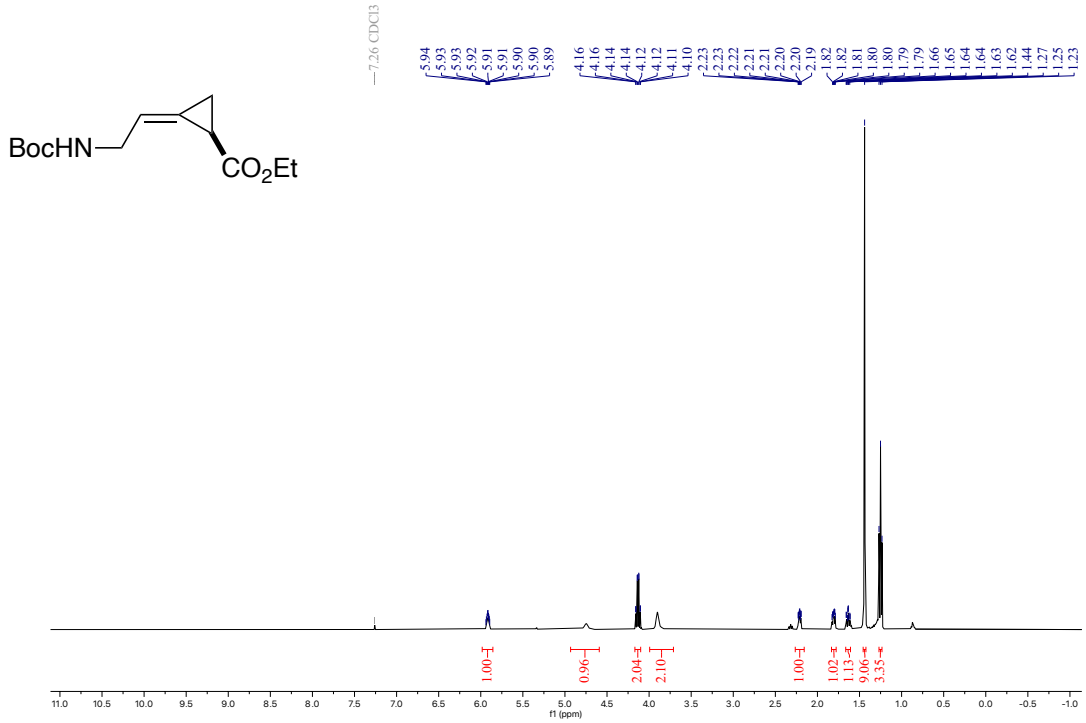
^{1}H NMR (400 MHz, $CDCl_3$) of ***E*-C8-p**



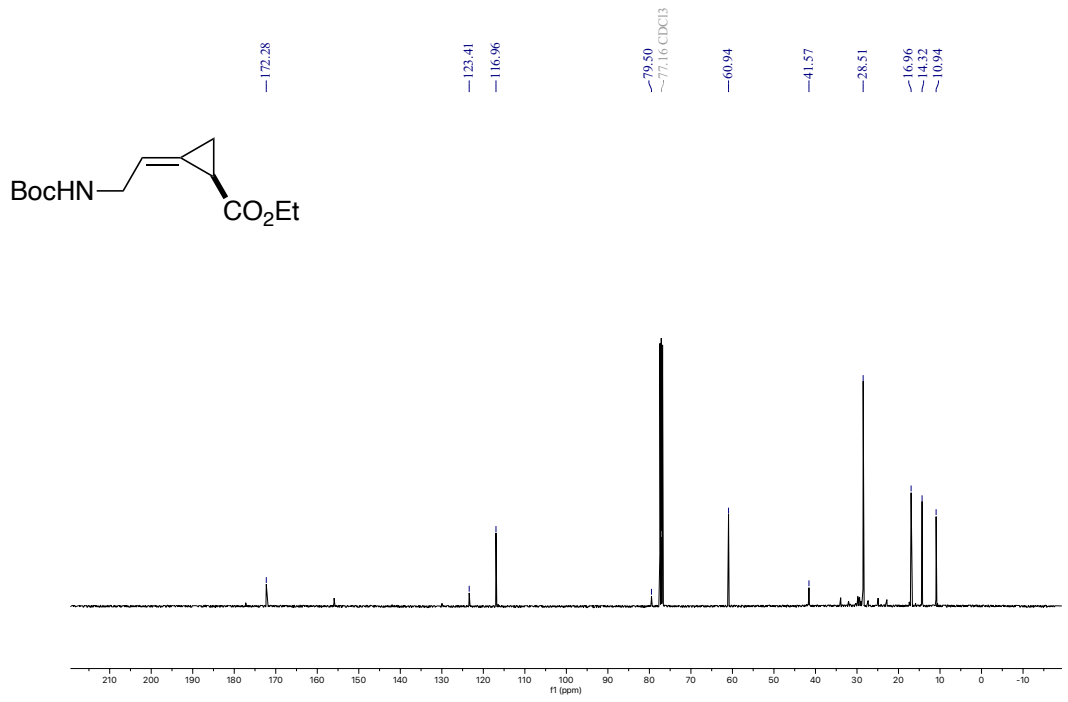
^{13}C NMR (101 MHz, $CDCl_3$) of ***E*-C8-p**



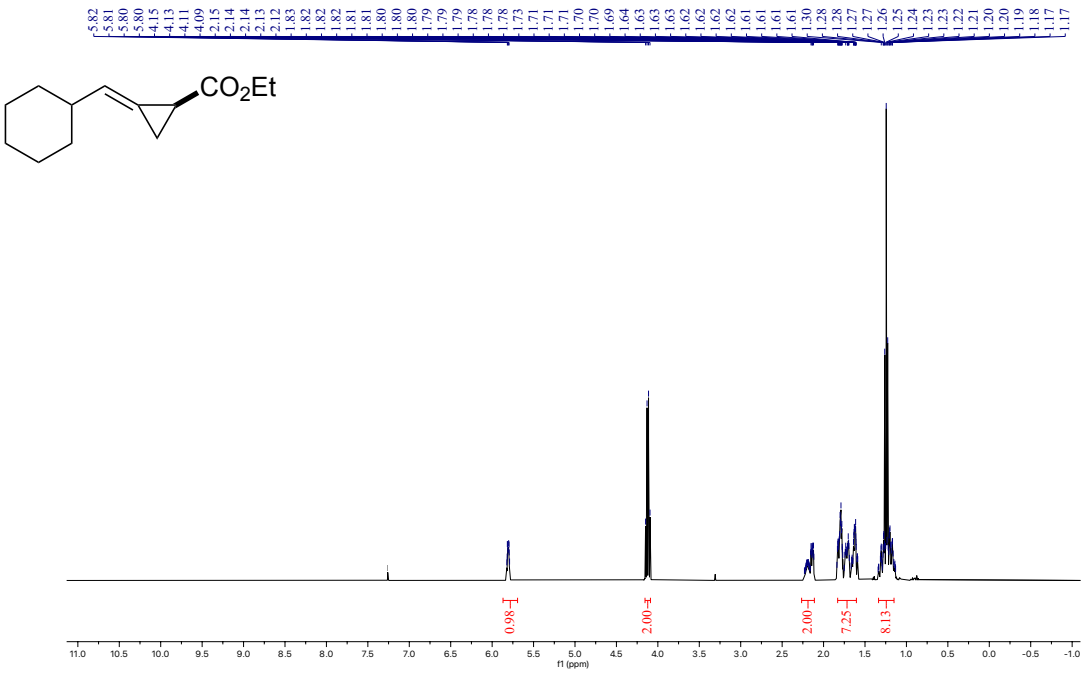
^{1}H NMR (400 MHz, $CDCl_3$) of ***Z*-C8-p**



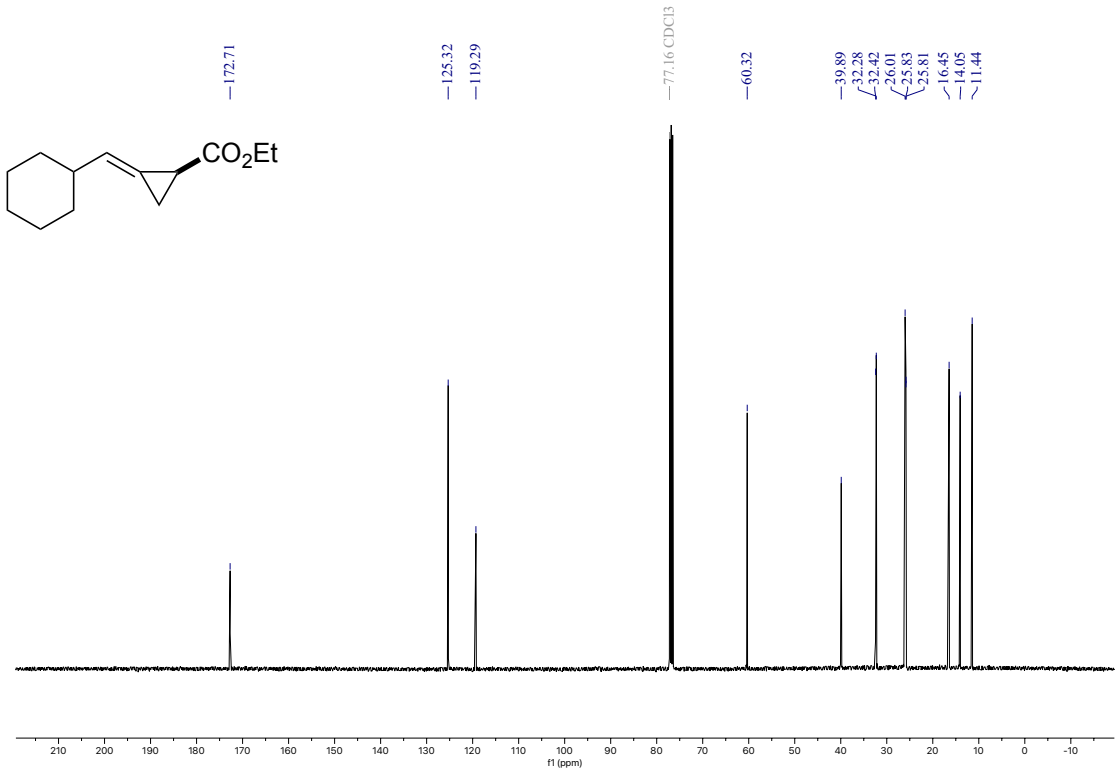
^{13}C NMR (101 MHz, $CDCl_3$) of ***Z*-C8-p**



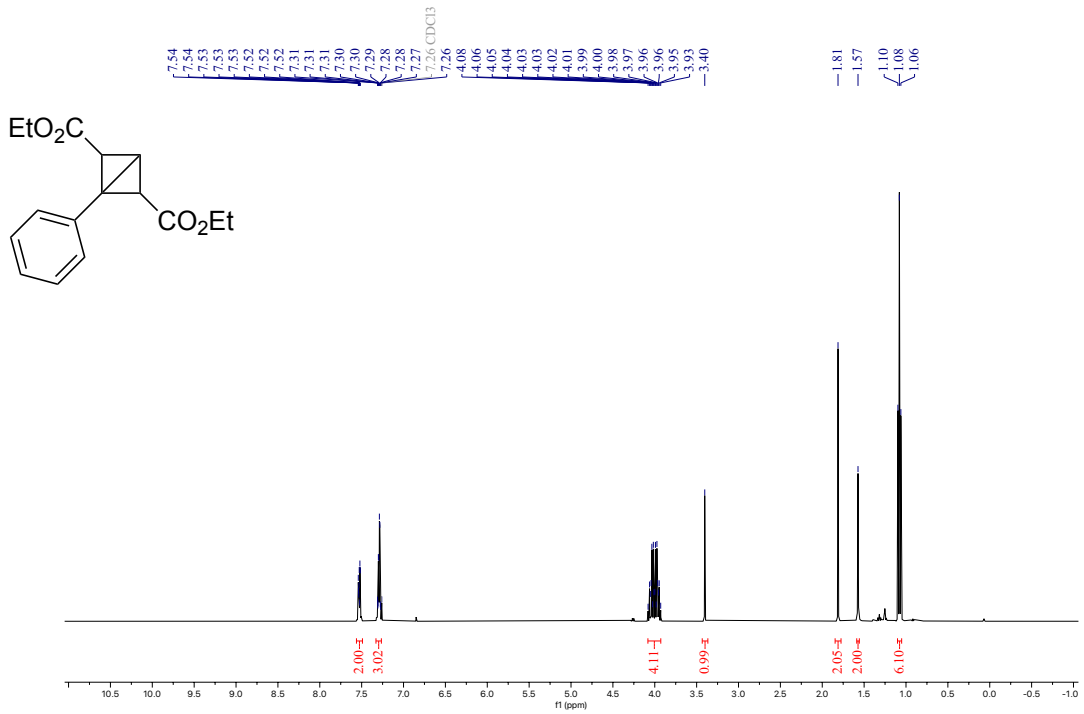
^{1}H NMR (400 MHz, $CDCl_3$) of ***E*-C9-p**



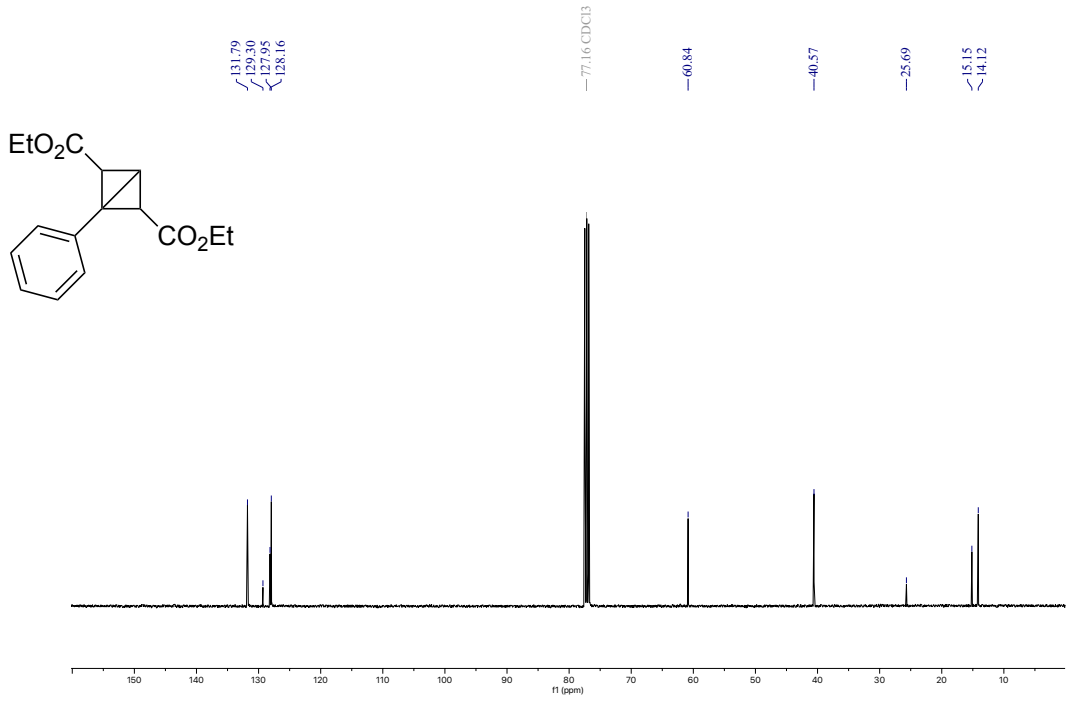
^{13}C NMR (101 MHz, $CDCl_3$) of ***E*-C9-p**



^{1}H NMR (400 MHz, $CDCl_3$) of **C7-p**



^{13}C NMR (101 MHz, $CDCl_3$) of **C7-p**



D.12. References

16. Studier, F.W.; Moffatt, B.A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol*. **1986**, *189*, 113-130.
17. Nov, Y. When Second Best Is Good Enough: Another Probabilistic Look at Saturation Mutagenesis. *Appl. Environ. Microbiol.* **2012**, *78*, 258-262.
18. Chester, N.; Marshak, D.R. Dimethyl sulfoxide-mediated primer Tm reduction: a method for analyzing the role of renaturation temperature in the polymerase chain reaction. *Analytical Biochemistry* **1993**, *209*, 284-290.
19. Gibson, D.G.; Young, L.; Chuang, R.-Y.; Venter, J.C.; Hutchinson III, C.A.; Smith, H.O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* **2009**, *6*, 343-345.
20. Long, Y.; Mora, A.; Li, F.-Z.; Gürsoy, E.; Johnston, K.E.; Arnold, F.H. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *ACS Synth. Biol.* **2025**, *14*, 230-238.
21. Abramson, J.; Adler, J.; Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493-500.
22. Kennemur, J.L.; Long, Y.; Ko, C.J.; Das, A.; Arnold, F.H. Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]alkanes. *J. Am. Chem. Soc.* **2025**, *147*, 27165-27171.
23. Yang, J.; Lal, R.G.; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.; Arnold, F.H. Active learning-assisted directed evolution. *Nat. Commun*. **2025**, *16*, 714.
24. Bloomer, B.J.; Joyner, I.A.; Garcia-Borràs, M.; Hu, D.B.; Garçon, M.; Quest, A.; Montero, C.U.; Yu, I.F.; Clark, D.S.; Hartwig, J.F. Enantio- and Diastereodivergent Cyclopropanation of Allenes by Directed Evolution of an Iridium-Containing Cytochrome. *J. Am. Chem. Soc.* **2024**, *146*, 1819-1824.
25. Miller, D.C.; Lal, R.G.; Marchetti, L.A.; Arnold, F.H. Biocatalytic One-Carbon Ring Expansion of Aziridines to Azetidines via a Highly Enantioselective [1,2]-Stevens Rearrangement. *J. Am. Chem. Soc.* **2022**, *144*, 4739-4745.
26. Zhang, R.K.; Chen, K.; Huang, X.; Wohlschlager, L.; Renata, H.; Arnold, F.H. Enzymatic assembly of carbon–carbon bonds via iron-catalysed sp^3 C–H functionalization. *Nature* **2019**, *565*, 67-72.
27. Sicinski, K.M.; Ritts, C.B.; Lin, E.; Long, Y.; Arnold, F.H. Transformation of Aromatic Feedstocks into Value-Added Products via Biocatalytic Primary Amination. *Unpublished*.
28. Alfonzo, E.; Hanley, D.; Li, Z.-Q.; Sicinski, K.M.; Gao, S.; Arnold, F.H. Biocatalytic Synthesis of α-Amino Esters via Nitrene C–H Insertion. *J. Am. Chem. Soc.* **2024**, *146*, 27267-27273.
29. Wilson, A.G.; Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *NeurIPS* **2020**, *34*, 4697.
30. Letham, B.; Karrer, B.; Ottoni, G.; Bakshy, E. Constrained Bayesian Optimization with Noisy Experiments. *arXiv* **2018**, arXiv:1706.07094.
31. Daulton, S.; Eriksson, D.; Balandat, M.; Bakshy, E. *arXiv* **2022**, arXiv:2109.10964.

32. Balandat, M.; Karrer, B.; Jiang, D.R.; Daulton, S.; Letham, B.; Wilson, A.G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *arXiv* **2019**, arXiv:1910.06403.