Data Foundations for Functional Prediction in Enzyme Engineering

Thesis by

Yueming Long

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy in Chemistry

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2026

(Defended March 18, 2026)

Yueming Long
ORCID: 0009-0006-5112-7791

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor, Professor Frances Arnold. She has been both a personal and scientific role model throughout my PhD. Her example shaped how I think about doing research that matters, pursued with intellectual honesty and scientific rigor. I am especially grateful for the opportunities she created for me to learn from outstanding scientists, and for the many conversations in which she consistently pushed us toward the questions most relevant to the field. I am also deeply grateful to my thesis committee members, each of whom has been an essential part of my PhD experience. Professor Hosea Nelson has been a constant source of energy and encouragement. I had the privilege of serving as his teaching assistant for three quarters across multiple classes, and I learned a tremendous amount from watching how he teaches. He brings clarity and excitement to the classroom, sparks students' curiosity, and shows up for them with generosity and support when it matters. Professor Shu-ou Shan shaped my early training as a scientist. I had the pleasure of rotating in her lab during my first quarter, and I remain grateful for her guidance. She is exceptionally detail-oriented and thorough, with a rare ability to attend to every aspect of a project while keeping the core scientific question in focus. Working with her taught me how to think carefully about research problems and how to build work that is both rigorous and complete. Finally, I want to thank Dr. Bruce Wittmann, who first drew me into computational research and scientific tooling, and who helped plant the initial vision for building a database for enzyme engineering. He has been a wonderful friend and scientist, and he welcomed me into the field in a way that deeply shaped my trajectory. I am especially thankful for his influence on how I think about the minimal requirements needed to establish data foundations for functional prediction. Thank you for everything you have shared, and for including me in your work in ways that allowed me to learn and grow as a scientist.

I had the great pleasure of working alongside Dr. Sabine Brinkmann-Chen and Dr. Kathleen Sicinski throughout my time at Caltech. They have been mentors to me both scientifically and personally. Through them, I got into exercise, yoga, and healthier habits, which helped me tremendously—especially during stressful periods. They are also simply wonderful to talk to, and knowing they were in the lab each day made Caltech feel like home.

I would like to dedicate this section to Professor Ariane Mora, with whom I have worked most closely over the past two years. She is among the most curious and genuinely excited scientists I have ever met, and I feel honored to have been part of her scientific journey. My thesis would not exist in its current form without her. I am deeply grateful for her mentorship, her drive, and the doors she opened

simply by being the kind of scientist who relentlessly pursues the questions that matter most. She also has a remarkable ability to bring together the right people and resources to make ambitious work possible. I have no doubt I will soon be visiting her in Austria and seeing her thriving, successful lab firsthand.

I also want to acknowledge Professor Anuvab Das and Dr. Jennifer Kennemur, two close collaborators on the chemistry and directed evolution projects that shaped my PhD. They taught me, with patience and precision, the practical and conceptual differences between kinetic and thermodynamic resolution, and how to execute each experiment thoughtfully and rigorously. Beyond their scientific insight, they were kind, understanding, and exactly the kind of collaborators anyone would hope for. Professor Anuvab Das is now at Nanyang Technological University, a place I have long wanted to visit. I feel lucky to have been his first mentee, and I believe wholeheartedly that he will be an exceptional professor.

I would also like to thank Jason Yang and Ryen O'Meara, my cohort-mates and friends throughout this journey. Although we each pursued very different research paths, we supported one another at every stage, sharing advice, perspective, and encouragement when it was most needed. Knowing I was not alone in this experience meant a great deal to me, and I am deeply grateful for the community.

Speaking of community, I would like to thank all of my friends and collaborators: Dr. Kadina Johnston, Dr. Francesca Zhoufan-Li, Dr. Martin Power, Dr. Chenghao Liu, Théophile Lambert, Jarrid Rector-Brooks, Severin Stalter, Dr. Julia Reisenbauer, Dr. Ravi Lal, Dr. Edwin Alfonzo, Dr. Shilong Gao, Dr. Ziqi Li, Dr. Hayden Carder, Dr. Ziyan Zhang, Dr. Casey Ritts, Dr. Ziyang Qin, Dr. Daniel Roth, and Deirdre Hanely. I have been fortunate to interact and work with each of you, and I am grateful not only for your help and friendship, but also for the joy of thinking together and building projects as a team.

I would also like to thank my mentees over the years. Thank you for your patience with me as I learned how to mentor well, and for giving me the space and trust to grow into a better mentor.

Additionally, I would like to thank Dr. Cheryl Nakashima and Rosie, whose work on the administrative and maintenance side keeps the lab running smoothly. Every day, they work hard to ensure the lab is clean, organized, and well supported, and that the countless logistical details and paperwork are handled with care. I am deeply grateful for everything they do behind the scenes.

My family has been instrumental in making it possible for me to complete this work. My parents, Yingfeng Long and Xiangyang Wu, are both PhDs and strongly influenced my decision to pursue an advanced degree. I grew up spending time with them in their office—while they worked on research, I did my homework—and that environment made reading, learning, and research feel like a natural part of life. I am also deeply grateful to my in-laws, Thomas Kefer and Ilse Kefer, who have been a constant source of support and encouragement, especially during the difficult moments when I struggled to keep going. I cannot fully express how much I appreciate all of them. As a family, they gave me everything they could so that I could explore what I love and become the person and scientist I hoped to be.

Finally, I want to thank Paul Kefer, my partner of ten years, and the person who has been my steady center through every season of this PhD. You have carried me through days of joy and days of doubt, and you have given me the kind of courage that makes the ordinary moments possible. No one knows me the way you do, and I don't think anyone knows you the way I do. I am proud of the life we have built together, and grateful that you've been there for every step of who I have grown into. This thesis belongs to both of us.

ABSTRACT

Despite transformative progress in protein structure prediction and de novo design, accurate prediction of enzyme function remains elusive. The central barrier is not a lack of sequences or structures, but a lack of learnable, experimentally grounded labels: in enzyme engineering, function is defined by assay context and deployment constraints, yet most available sequence–function datasets are sparse, biased toward successes, incompletely genotyped, and inconsistently annotated. As a result, models that appear strong on curated benchmarks often fail to guide real engineering decisions, especially for new-to-nature transformations and out-of-distribution substrates. This thesis advances a data-centered strategy for functional prediction by building a practical pipeline that turns routine protein engineering experiments into prediction-ready sequence–function records. Chapter 2 establishes the importance of assay-defined, application-relevant labels through a stereodivergent biocatalytic platform for azaspirocycle synthesis, where directed evolution of carbene transferases enables access to azaspiro[2.y]alkane scaffolds with tunable stereochemical outcomes, quantified by deployment-aligned measurements including turnover and enantio- and diastereoselectivity. Chapter 3 develops LevSeq, a high-throughput and cost-effective genotyping and analysis workflow that assigns full-length variant identities at screening scale using dual barcoding and nanopore long reads and reports practical quality-control metrics to prevent silent failure modes and preserve reliable sequence assignment across both hits and non-hits. Chapter 4 introduces EnzEngDB, a curated database and data model that standardizes sequences, reaction context, assay conditions, units, quality metrics, and provenance so datasets can be queried, filtered, and aggregated across campaigns, while retaining negative and neutral outcomes as informative constraints. Chapter 5 closes the loop by using the resulting data substrate to support decision-making in *de novo* enzyme validation, emphasizing that predictive ranking is currently most reliable within reactions with dense standardized coverage and hypothesize that generalization to unseen reactions will require systematic expansion of assay-defined labels across diverse transformations. Together, these chapters provide the tooling and framework needed to make sequence–function data compounding rather than ephemeral, enabling reliable functional prediction and iterative, model-guided enzyme design.

PUBLISHED CONTENT AND CONTRIBUTIONS

‡ Denotes equal contribution

1. Yang, J., Li, F.-Z., **Long, Y.**, Arnold, F.H., 2025. Illuminating the universe of enzyme catalysis in the era of artificial intelligence. Cell Syst. doi:10.1016/j.cels.2025.101372

Y.L. contributed to the conceptual mapping of sequence–function space and carried out the research to identify chemically relevant reactions for the figure. Y.L. also assisted with editing and proofreading the manuscript.

2. Kennemur, J.L. ‡, **Long, Y.** ‡, Ko, C.J., Das, A., Arnold, F.H., (2025). Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]alkanes. In J. Am. Chem. Soc. 147, 27165–27171. doi:10.1021/jacs.5c07015

Y.L. performed initial screening, sequencing, directed evolution, and synthesis of relevant chemical substrates and standards. Y.L. conducted lineage validation, investigated high-substrate-loading (high-concentration) reaction conditions, and purified enantioenriched products. Y.L. and J. L. K. contributed equally to this work and took primary responsibility for writing the manuscript, with feedback from all authors.

3. **Long, Y.** ‡, Mora, A. ‡, Li, F.-Z., Gürsoy, E., Johnston, K.E., Arnold, F.H., 2025. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. ACS Synth. Biol. 14, 230–238. doi:10.1021/acssynbio.4c00625

Y.L. conceptualized the project and performed initial testing of the barcoded-primer design. Y.L. wrote the software and developed laboratory protocols to enable broad adoption of the method, validated the accompanying codebase, and built cross-architecture software distributions using both pip and Docker. Y.L. and A. N. M. contributed equally to this work and took primary responsibility for writing the manuscript, with feedback from all authors.

4. **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., 2026. Enzyme Engineering Database (EnzEngDB): a platform for sharing and interpreting sequence–function relationships across protein engineering campaigns. Nucleic Acids Res. 54, D564–D571. doi:10.1093/nar/gkaf1142

Y.L. conceptualized the project and led the content design and visual design of the website components. Y.L. developed the API-based extraction tool for mining data from published enzyme engineering papers and manually validated all extracted entries. Y.L. also manually validated all entries derived from in-house datasets and wrote the manuscript with feedback from other authors.

PUBLISHED CONTENT NOT INCLUDED IN THESIS

‡ Denotes equal contribution

1. Das, A‡.; **Long, Y‡.**; Maar, R. R.; Roberts, J. M.; Arnold, F. H. Expanding Biocatalysis for Organosilane Functionalization: Enantioselective Nitrene Transfer to Benzylic Si–C–H Bonds. ACS Catal. 2024, 14, 148–152.

Y.L. performed screening, sequencing and directed evolution. Y.L. conducted lineage validation. Y.L. and A.D. contributed equally to this work.

2. Athavale, S. V.; Gao, S.; Das, A.; Mallojjala, S. C.; Alfonzo, E.; **Long, Y.**; Hirschi, J. S.; Arnold, F. H. Enzymatic Nitrogen Insertion into Unactivated C–H Bonds. J. Am. Chem. Soc. 2022, 144, 19097–19105.

Y.L. performed substrate scope testing and characterization for this project and help proofread the manuscript.

3. Mora, A.; Reisenbauer, J. C.; Schmid, H.; Miyazaki, I.; **Long, Y.**; Yang, J.; O'Meara, R.; Arnold, F. H. Enzyme-Tk: Searching the Sequence Space to Find New Enzymatic Starting Points 25 for Bioremediation. In preparation.

Y.L. performed initial literature research and cloning of the QHH variant, Y.L. assisted in sequencing method update specific to the project and proofread the manuscript.

4. O'Meara, R. L.; Mora, A. N.; Quinn, E. C.; **Long, Y.**; Maar, R. R.; Arnold, F. H. Title: Enzymatic Mineralization of Siloxanes by the Sulfite Reductase CysI. In preparation.

Y.L. performed LCMS method development and troubleshooting for this project.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

NOMENCLATURE

Active Learning: A machine-learning workflow where a model is trained on existing data, proposes which new datapoints to measure next, is retrained after those measurements, and iterates to improve outcomes with minimal experiments.

ALDE (Active Learning-Assisted Directed Evolution): A directed-evolution strategy that uses active learning (often Bayesian optimization) to choose which variants to build and test in successive rounds.

Assay-Defined Label: A quantitative measurement produced by a specific assay under specified conditions; in this thesis, "function" is treated as an assay-defined label rather than an intrinsic annotation of sequence.

Assay Context: The experimental conditions that define what a label means (e.g., substrate identity, reaction format, time, detection method, temperature, cosolvent, loading).

Benchmark Task: A standardized prediction/evaluation setup (with defined data splits and metrics) used to compare models on a clearly specified practical task.

Benjamini–Hochberg (BH) Correction: A multiple-testing procedure that controls the false discovery rate when many hypothesis tests are performed (e.g., mutation calls across positions and wells).

Binomial Test: A hypothesis test used here to decide whether an observed nucleotide frequency exceeds what would be expected from the sequencing error rate.

Carbene Transferase: An engineered (often heme-dependent) enzyme that catalyzes transfer of a carbene equivalent from a diazo reagent to a substrate.

Chemical Reaction Context: The chemically explicit representation of the transformation being measured, treated as part of the model input and essential for comparing records across campaigns.

Chemoselectivity: Preference for forming one type of product (or one pathway) over others when multiple reactions are possible.

Consensus Sequence: A single inferred sequence for a well/sample derived from multiple reads, intended to represent the true variant while averaging down sequencing errors.

Coverage (Sequencing Coverage): The number of reads supporting a sequence call for a given well; low coverage reduces confidence in variant identity.

Cross-Reaction Generalization: A model's ability to predict performance for reactions (or reaction conditions) not seen during training, rather than interpolating within a single reaction.

CSV (Comma-Separated Values): A simple tabular file format used for standardized data exchange and upload into EnzEngDB.

Data Provenance: Information describing where a record came from (source paper/experiment), how it was generated, and what transformations/filters were applied.

Dataset Shift: A change between training and testing distributions (e.g., different substrates, assay formats, scaffolds, or reaction conditions), often causing performance degradation.

De Novo Design: Designing protein sequences (often with structure prediction/generation) rather than starting from a natural scaffold and improving by mutation.

Demultiplexing: Computationally assigning pooled sequencing reads back to their original samples (here, plate and well) using barcode sequences.

Deployment-Aligned Measurement: A label chosen to reflect real operating constraints (e.g., high substrate loading, aqueous conditions, whole-cell or lysate format) rather than a convenient proxy.

Directed Evolution (DE): Iterative cycles of mutagenesis, expression, screening/selection, and fixation of improved variants to optimize an assay-defined objective.

Diastereomeric Ratio (dr): The ratio of diastereomer products formed in a reaction; a measure of diastereoselectivity.

Diastereoselectivity: Preference for forming one diastereomer over another in product formation.

Dual Barcoding: Using distinct barcode sequences on both ends (or on two primers) to encode well and plate identity for high-throughput pooling.

Enantiomeric Ratio (er): The ratio of enantiomer products formed; a measure of enantioselectivity.

Enantioselectivity: Preference for forming one enantiomer over the other.

Engineering Campaign: A coordinated set of experiments (often multi-round) aimed at improving a protein for a defined assay objective; produces structured local explorations of sequence space.

Epistasis: Non-additive interactions between mutations where the combined effect differs from the sum of individual effects.

epPCR (Error-Prone PCR): A mutagenesis method that introduces random nucleotide substitutions during PCR amplification, generating diverse variant libraries.

ESM-2 Embedding: A learned numerical representation of a protein sequence produced by a pretrained protein language model, used as input features for downstream prediction.

Fitness (Protein Fitness): How well a protein performs a specified assay-defined function under defined conditions; the quantitative target of optimization and prediction.

Fitness Landscape: The mapping from sequence to fitness (label), conceptualized as a high-dimensional surface f(sequence) = fitness.

Flongle: A low-throughput Oxford Nanopore flow cell format used for smaller sequencing runs.

FDR (False Discovery Rate): The expected fraction of false positives among accepted discoveries; controlled here during mutation calling across many tests.

Full-Length Genotyping: Determining the complete coding sequence (or full amino-acid sequence) of each screened variant rather than sequencing only targeted sites.

GC-FID (Gas Chromatography with Flame Ionization Detection): An analytical method used to quantify products (often with calibration curves and internal standards).

GCMS (Gas Chromatography–Mass Spectrometry): An analytical method combining chromatographic separation with mass detection for product identification/quantification.

Genotype Assignment: Linking each experimental measurement (well/variant) to the correct underlying sequence identity.

Hamming Distance: The number of positions at which two sequences differ (for equal-length sequences), used to quantify mutational distance within a lineage.

Hit Rate: The fraction of tested variants exceeding a defined activity threshold in a given assay context.

HPLC (High-Performance Liquid Chromatography): A chromatography method used for separation and quantification; can be coupled to chiral stationary phases for er/dr.

kcat, KM, kcat/KM: Standard kinetic parameters describing catalytic turnover rate, substrate affinity, and catalytic efficiency, respectively.

LCMS (Liquid Chromatography–Mass Spectrometry): An analytical method combining LC separation with MS detection for product identification/quantification.

LevSeq (Long-read Every Variant Sequencing): A sequencing + analysis workflow that assigns full-gene variant identities at screening scale using dual barcoding and nanopore long reads, with explicit quality-control outputs.

Library Bias: Systematic skew in the variant population (e.g., mutation rate, stop codons, recombination, overrepresented clones) that can waste screening effort and distort learned relationships.

Lyophilized Lysate: Freeze-dried cell lysate used as a practical catalyst format that can be reconstituted like a chemical reagent.

Masked-Token Prediction: A pretraining objective where some tokens in a sequence are masked and the model learns to predict them from context.

MinION: A benchtop nanopore sequencer used for in-house long-read sequencing runs.

minimap2: A long-read aligner used to map nanopore reads to a reference sequence for downstream variant calling.

Mixed Well: A well containing more than one variant (e.g., contamination or multiple colonies), often detected by statistically significant evidence for multiple mutations at the same position.

MLPE (Machine Learning–Guided Protein Engineering): Protein engineering workflows that use ML to prioritize variants, propose mutations, or guide exploration of sequence space.

Mol*: A web-based molecular visualization engine used for interactive 3D structure viewing in EnzEngDB.

Nanopore Sequencing: A sequencing approach that reads DNA by measuring ionic current changes as nucleic acids pass through a nanopore; enables long reads but has higher per-read error rates than many short-read methods.

Negative Outcome: A measured result indicating lack of desired activity (or deleterious performance) under the assay definition; valuable as a constraint for learning and calibration.

Neutral Outcome: A measured result showing little or no change relative to baseline; informative for identifying plateaus and non-impactful mutations.

One-Step Colony PCR: PCR performed directly from a small amount of colony culture to amplify the target gene (here, with barcoded primers) without plasmid prep.

Oxford Nanopore: Oxford Nanopore Technologies that produces nanopore sequencing devices (e.g., MinION) and library preparation kits.

Pairwise Ranking Loss: A training objective that learns to order pairs of sequences correctly (relative ranking) rather than predicting absolute activity values.

Plate-Scale Screening: Screening variants arrayed in microplates (e.g., 96-well) to enable high-throughput measurement and systematic linking of genotype and phenotype.

Plotly Dash: A Python web framework used to build interactive dashboards (used for EnzEngDB's interface).

Primer Barcode: A short DNA tag appended to primers so that amplified products can be traced back to specific wells/plates after pooling.

Protein Scaffold: A protein family or backbone used as the starting framework for engineering (e.g., protoglobins).

Protoglobin (Pgb): A small, thermostable heme protein scaffold used extensively here for carbene/nitrene transfer engineering (e.g., ApePgb, TamPgb, ParPgb variants).

Quality Control (QC): Checks and summary metrics that detect failure modes (e.g., low coverage, mixed wells, implausible variants) and allow downstream filtering by confidence.

Reaction SMILES: A machine-readable encoding of substrates and products as a SMILES reaction string, used to represent chemical transformations consistently.

RDKit: An open-source cheminformatics toolkit used for validating/standardizing SMILES and computing fingerprints and reaction similarities.

Read Depth: The number of reads assigned to a sample/well; higher depth generally increases the reliability of consensus and mutation calls.

SFC-MS (Supercritical Fluid Chromatography–Mass Spectrometry): A chromatographic method (often fast and chiral-capable) coupled to MS for product analysis.

Sequence Identity: The fraction of positions that match between two aligned sequences; used to define relatedness among homologs/variants.

Sequence–Function Dataset: A collection of sequence–function records suitable for learning, evaluation, and comparison when context and units are consistent.

Smith–Waterman Alignment: A local alignment algorithm used here in a bespoke implementation to efficiently detect barcode sequences at read ends.

SMILES (Simplified Molecular Input Line Entry System): A text representation of chemical structures used for canonical, machine-readable encoding of substrates/products.

Spearman Correlation (ρ): A rank-based correlation metric used to evaluate whether predicted ordering agrees with measured ordering (useful for prioritization tasks).

SSM (Site-Saturation Mutagenesis): A mutagenesis approach where a specific residue position is diversified to many or all amino acids to map local effects.

Stereodivergence: The ability to access different stereoisomer products (or opposite enantiomers) via different engineered enzyme lineages/variants.

Tanimoto Similarity: A similarity metric (often applied to molecular fingerprints) used to compare reactions/structures; higher values indicate more similar fingerprints.

TBamp / LBamp: Terrific Broth or Luria Broth media supplemented with ampicillin, used for selection/propagation of plasmid-bearing E. coli cultures.

TTN (Total Turnover Number): Product molecules formed per catalyst molecule (often per enzyme active site/cofactor), used as an activity/efficiency metric.

Variant Calling: Inferring the mutations present in a well/sample relative to a reference sequence, given read alignments and an error model.

Whole-Cell Reaction Format: Running catalysis in intact cells expressing the enzyme, rather than using purified protein or lysate; affects context and label meaning.

EnzEngDB (Enzyme Engineering Database): A curated database + analysis platform that standardizes sequences, reaction context (SMILES), quantitative metrics, units, QC, and provenance so records can be queried and compared across campaigns, including negative/neutral outcomes.

Chapter 1

FROM SCREENING TO PREDICTION-READY SEQUENCE–FUNCTION DATA

This chapter incorporates figure from the following publication: 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.

1.1 Overview

Protein engineering sits at the intersection of a practical need and an information paradox. On the one hand, sequence space is vast, experimental iteration is expensive, and the ability to predict function would immediately change how enzymes are discovered, optimized, and deployed. On the other hand, the last decade has delivered an abundance of biological information: sequence databases have grown to extraordinary scale [1], and structure prediction has become fast and widely accessible [2]. The paradox is that even with far more sequences and far more structures, it is still difficult to predict whether a proposed enzyme will catalyze a desired reaction with useful activity and selectivity under real experimental conditions [3].

This persistent gap is sometimes framed as a failure of models, but in practice it is more often a failure of data to be learnable. In the lab, "function" is measured as an assay outcome that reflects a specific substrate, a specific reaction format, and a specific operational context [4]. In many datasets, those details are missing, inconsistent, or unlinked from the exact genotype that produced the measurement. When that happens, prediction becomes brittle: models can interpolate within narrow training distributions yet fail when asked to guide decisions across substrate classes, across scaffolds, or across experimental formats. The result is a common

experience in enzyme engineering: computational suggestions are useful as ideas, but the decisive evidence still comes from screening.

This thesis argues that function prediction in enzyme engineering is limited upstream by three bottlenecks: the definition of function as an assay-defined label, the ability to attach those labels to complete and correct genotypes at screening scale, and the standardization required to make records reusable across campaigns. These bottlenecks are conceptually simple, but they are systematically under-invested in because they sit between disciplines. They are partly chemistry problems (what is the reaction and what metrics matter), partly assay design problems (what is measured), partly sequencing problems (what genotype is linked to each well), and partly data engineering problems (how records are represented so they can be reused). The core premise of this thesis is that solving these interfaces turns routine protein engineering work into learnable sequence–function data, and that this data foundation is what makes downstream prediction practical rather than aspirational (Figure 1-1).

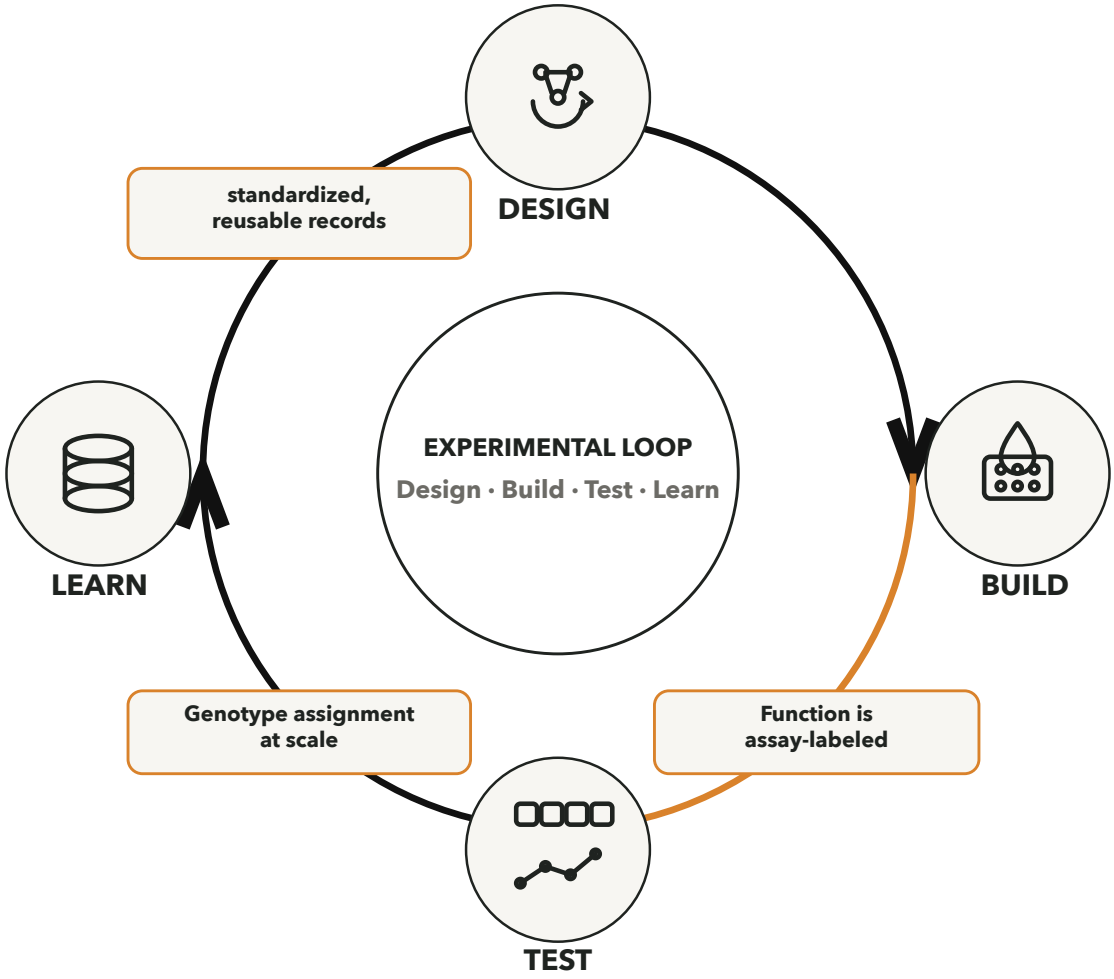


Figure 1-1. Data bottlenecks in the design–build–test–learn loop limit practical enzyme function prediction. Although sequence databases and structure prediction have expanded rapidly, prediction in enzyme engineering is constrained upstream by how function is defined and measured in a specific assay context, how those measurements are linked to complete and correct genotypes at screening scale, and how records are standardized so results remain quantitative and reusable across campaigns. Chapters 2–5 address these interfaces to make dataset-quality capture routine and to enable predictors that meaningfully shift experimental testing toward success.

A second premise is that enzyme engineering already generates the right kind of supervision, but it is rarely captured in a way that compounds. Directed evolution produces structured explorations of local sequence neighborhoods with measured outcomes, including failures, plateaus, tradeoffs, and occasional leaps [5]. These patterns are exactly what learning systems

need to identify mutational rules and predict promising enzymes. However, engineering campaigns are usually optimized to deliver a catalyst, not to deliver a dataset. This thesis aims to align those goals by making dataset-quality capture routine and low-friction.

Historically, the field has progressed by strengthening each link in the experimental loop: building larger and better targeted libraries, increasing throughput of screening, and developing assays that report more directly on a desired chemical outcome. These advances have produced remarkable catalytic capabilities, but they have also raised expectations for prediction. If the field can build and test thousands of variants per week, then a predictor that is only marginally better than random sampling will not justify its complexity; it must meaningfully shift the distribution of tested variants toward success. Achieving that level of leverage depends less on model novelty than on whether the underlying data are coherent, quantitative, and reusable [6–8].

This chapter frames the thesis around that central claim. Chapter 2 focuses on generating quantitative, stereochemically resolved, deployment-aligned labels. Chapter 3 establishes full-length genotype assignment with explicit quality control at screening scale. Chapter 4 standardizes and consolidates records in a database representation that preserves chemistry context and retains negative and neutral outcomes. Chapter 5 then uses these foundations to support prioritization of experimental validation of designed enzymes.

1.2 Why sequence and structure are not enough

Sequence and structure are necessary, but they do not fully determine catalytic function in an engineering sense. Enzymatic catalysis depends on a physical mechanism that couples binding, positioning, electrostatics, conformational dynamics, and chemical steps that often involve transient states and competing pathways [9–11]. Two proteins can share the same fold and even similar active-site motifs while displaying dramatically different rates, selectivities, and substrate scopes. Conversely, enzymes with weak overall similarity can converge on similar reactivities through different solutions.

Structure prediction has removed a major barrier to mechanistic reasoning, but it has also highlighted the limits of what structure alone can provide. A static model can suggest plausible binding modes and identify residues likely to contact a substrate, but it rarely reveals which microstates dominate under reaction conditions, how conformational equilibria shift with mutations, or how the active site reorganizes over a catalytic cycle [12–15]. Many catalytic properties that matter in engineering, such as chemoselectivity against competing pathways, stereochemical fidelity over conversion, or tolerance to high substrate loading, depend on subtle energetic differences that are not robustly inferred from a single structural snapshot.

This becomes more pronounced as engineering moves away from biology's native distribution. When a reaction is new-to-nature or a substrate is far from natural metabolites, the relevant interactions have not been filtered by evolution, and catalytic solutions are more idiosyncratic[16–19]. In these settings, sequence similarity offers limited guidance, and structural plausibility is an unreliable proxy for productive turnover. Even when a model is structurally accurate, it does not automatically predict whether the enzyme will express well, load its cofactor efficiently, remain stable under reaction conditions, or avoid off-pathway reaction [20–23].

A useful way to phrase this limitation is that sequences and structures describe what a protein is, whereas catalysis measures what a protein does under a particular set of constraints. The gap between "is" and "does" can be narrow for conserved natural reactions, but it widens quickly when the objective is engineered. For many campaigns, the decisive differences between variants are not large structural rearrangements, but small changes that bias conformational sampling, reshape electrostatics, or suppress a competing pathway. These effects may be invisible to coarse structural reasoning yet dominate measured outcomes.

For prediction, these realities imply that the most valuable data are not merely sequences with categorical annotations, but sequences paired with quantitative outcomes measured under defined conditions [4]. The datasets that matter for engineering are those that reflect the actual objective function: product formation and selectivity measured in the relevant format,

with enough context to interpret differences across experiments (Figure 1-2). Chapter 2 is built around this idea by treating a stereochemically resolved catalytic objective as the primary label rather than as secondary characterization.

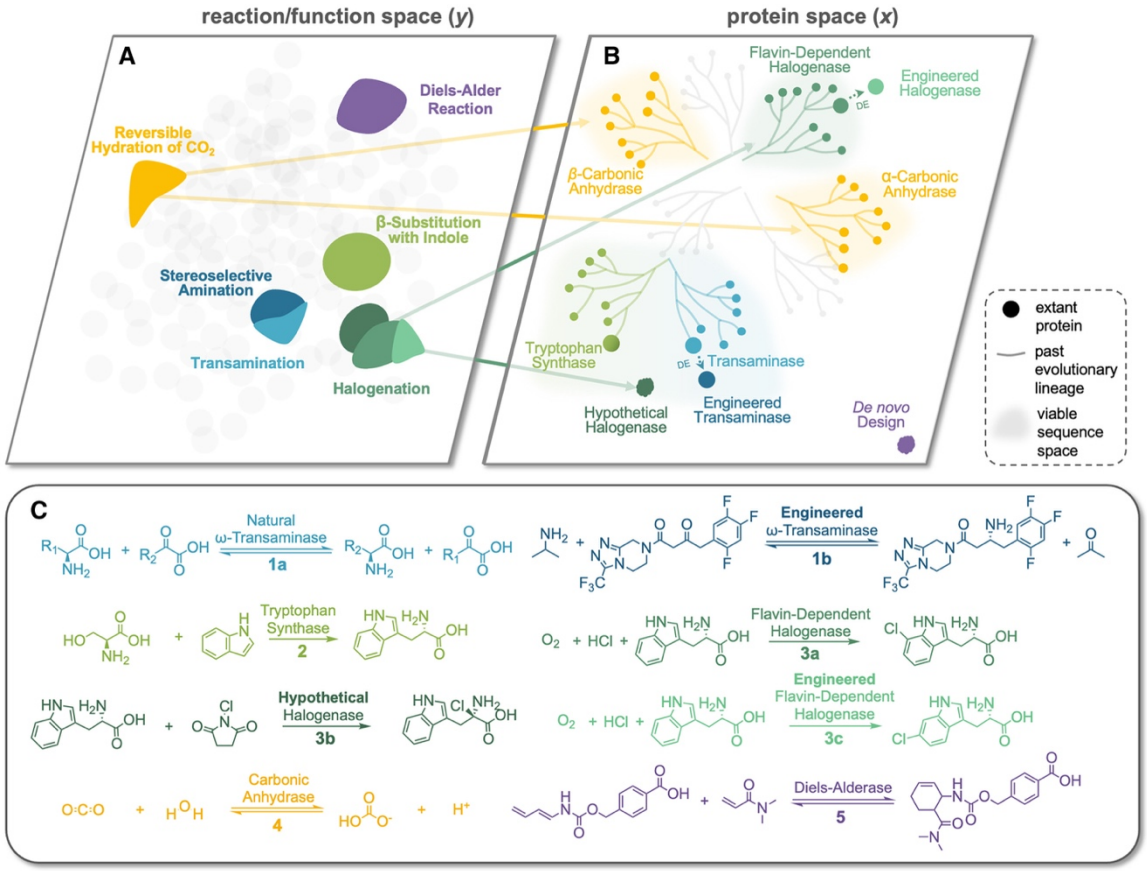


Figure 1-2. Conceptual framework for learning a mapping between protein space (x) and reaction/function space (y), illustrating incremental exploration by directed evolution and larger functional 'jumps' enabled by machine learning. (A and B) Learning a mapping between (A) function space (y) and (B) protein space (x) using ML. (C) Example of a chemical reaction associated with each function, color coded.

1.3 Enzyme engineering as iterative measurement

Directed evolution and related protein engineering workflows are, at their core, a repeated loop of hypothesis, diversification, screening, and selection. The loop can be executed with many different design philosophies, from random mutagenesis to structure-guided libraries

to modern generative approaches. Regardless of how variants are proposed, the loop is ultimately grounded in measured performance, and the most successful campaigns are those that make measurement fast, reliable, and aligned with the objective [24].

In this framework, prediction should not be thought of as a replacement for experiments, but to make experiments more informative and more efficient. A predictor that cannot be connected to the exact assay conditions and the exact genotype identities that arise in real campaigns will not close the loop. This is why benchmarks and curated datasets can be sometimes misleading: they often smooth away the messiness that dominates real screening, plate effects, batch variability, expression variability, detection limits, and practical constraints on assay time and throughput [25,26].

A key observation motivating this thesis is that directed evolution (Figure 1-3) already generates unusually valuable supervision for learning. Variants are not sampled uniformly; they are explored in local neighborhoods shaped by human choices and selection pressures [27]. It can learn which mutations are consistently beneficial, which are conditional on background, which trade activity for selectivity, and which open or close substrate scope [28]. It can also learn when a sequence is likely to be unproductive under a given assay definition, which matters as much as identifying hits when throughput is limited. Additionally, directed evolution campaigns are typically pursued only when the target function is considered meaningful and application relevant, which makes the resulting sequence–function data especially valuable to collect.

Importantly, the role of prediction differs by regime: within a campaign it guides local optimization, whereas for de novo designs it primarily serves as a conservative filter and prioritization tool, ideally paired with calibrated uncertainty and rapid experimental validation [29].

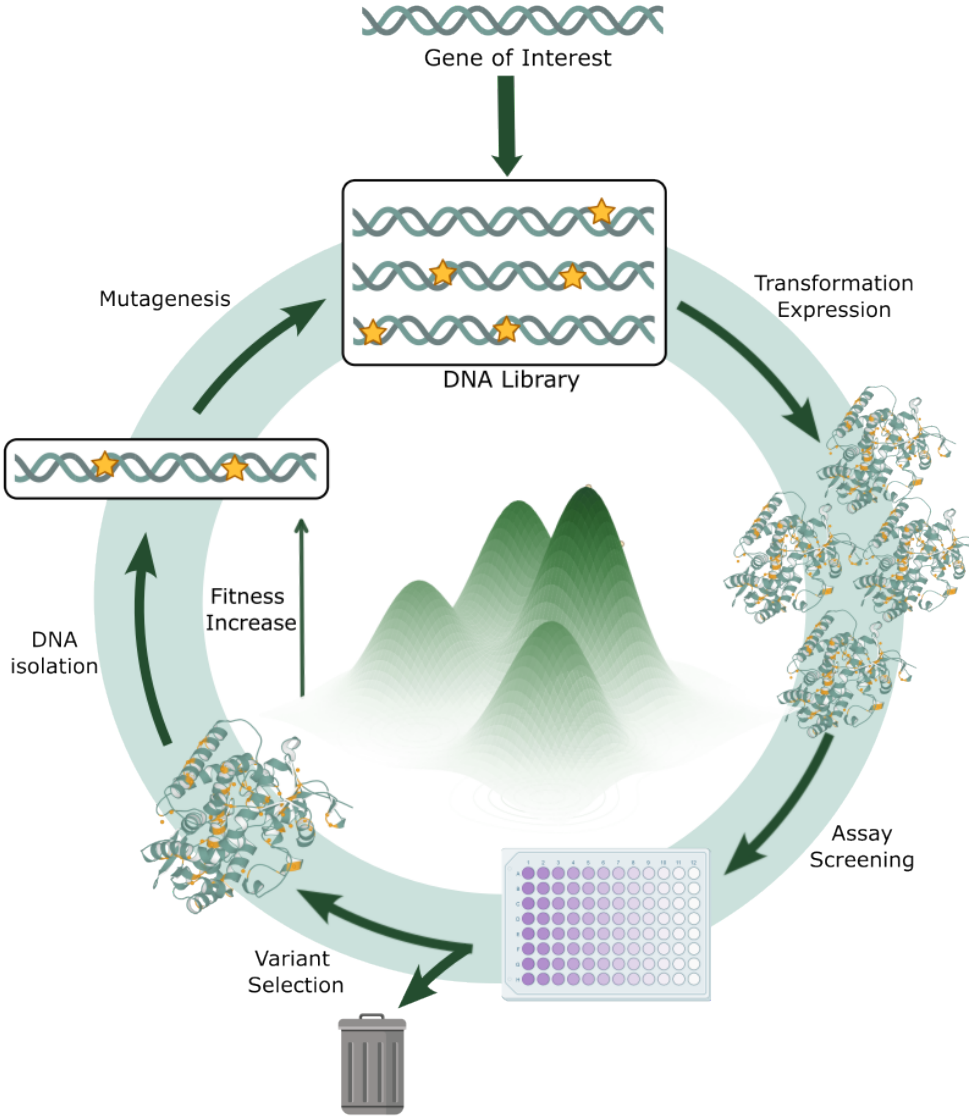


Figure 1-3. Directed evolution generates structured, assay-defined supervision for learning. Iterative rounds of library diversification, expression, screening, and selection explore local sequence neighborhoods and move variants up an assay-defined fitness landscape. The resulting trajectories produce measured outcomes that reveal beneficial, conditional, and deleterious mutations under a fixed experimental context.

At the same time, the iterative measurement framing highlights why naive dataset assembly often fails. In a real campaign, measurements are made across rounds, across plates, and across conditions that may shift as the project evolves. Assays may be tightened, substrates may be changed, and detection may improve. Without careful label definition and context

capture, the resulting dataset can become an accidental mixture of objectives. For a project, this is survivable because the experimenter knows the story. For a model, it is damaging because the model sees a heterogeneous mapping without the explanatory metadata that would make it conditional [30].

Over time, this creates a mismatch between how enzyme engineering is performed (measurement-rich, iterative, local exploration) and how prediction datasets are typically assembled (sparse, curated, often centered on positives). This thesis is designed to close that mismatch by treating the experiment-to-record pipeline as part of the engineering problem. Chapter 3 addresses the genotype side of this pipeline, and Chapter 4 addresses the record structure side, so that iterative measurement becomes an iterative dataset rather than a series of isolated results.

1.4 Function in a prediction-ready dataset

The word function is overloaded. In databases and annotations, function often means membership in an enzyme class, assignment of a reaction, or the presence of a conserved motif. In protein engineering, function is an operational quantity that includes performance metrics and practical constraints. An enzyme that is correctly annotated but only active at low substrate loading, only in purified format, or only under conditions incompatible with the intended process is not functional in the engineering sense [31,32].

This thesis defines function as an assay-defined label: a quantitative measurement (or set of measurements) produced by a specified assay under specified conditions. The essential point is that labels are conditional on context, and the context must be represented with enough clarity that measurements can be compared, aggregated, and learned from. For catalysis, that context includes at minimum the substrate identity, the reaction format (whole cell, lysate, purified), the time point or kinetic window, and the detection method. It also includes practical details that often dominate outcomes, such as substrate loading, cosolvent fraction, and whether the measured number reflects a single endpoint or an integrated conversion over time.

Equally important is the idea that the label must reflect what the campaign ultimately cares about. Many screening assays prioritize throughput and therefore use proxy readouts, fluorescence surrogates, coupled assays, growth selections, or colorimetric signals, that correlate with product formation but are not identical to it. Proxy labels are not inherently bad. They can be extraordinarily powerful when they are stable, calibrated, and mechanistically linked to the true objective [33]. The risk is that proxies often contain hidden confounders: expression level, well evaporation, scattering artifacts, or metabolic effects in whole cells can influence the readout without reflecting catalytic chemistry. A model trained on such labels will learn the proxy, including its confounders, and may optimize for artifacts rather than for the desired transformation unless the label definition is explicit and the dataset includes the necessary controls and context.

A prediction-ready label also benefits from being high-specificity. In catalysis, many failure modes collapse into low signal if measured coarsely. Labels that resolve stereochemical outcomes or product distributions provide sharper constraints. They capture not only whether the enzyme reacts, but how it reacts, and they often reveal mechanistic coupling that would be invisible in a scalar activity readout. In many engineering programs, stereoselectivity is not an accessory metric; it is the metric. When selectivity is treated as a first-class label, it shapes the kinds of sequence–function relationships a model must learn and therefore shapes what prediction means [34].

This thesis therefore emphasizes labels that preserve selectivity information and that align with deployment decisions, because these are the labels that enzyme function prediction ultimately must learn to reproduce. Chapter 2 uses a stereodivergent catalytic objective as a concrete demonstration of this philosophy and treats label specificity as an enabling layer for prediction.

1.5 The hidden bottleneck: genotype assignment at screening scale

A prediction-ready dataset requires that each assay outcome be linked to the correct sequence, ideally the full-length variant sequence rather than a partial barcode or a list of

intended mutations. In practice, genotype assignment is often the least glamorous and most fragile part of a screening workflow. It is also the part that most directly determines whether results are reusable.

When genotyping is performed sporadically, or only for top hits, or only for a subset of positions, the resulting dataset becomes biased and incomplete (Figure 1-4) [35]. It may be adequate to advance a project, because a few best variants can be confirmed and carried forward, but it is inadequate as training data. The model's job is not to explain only the best variants; it is to learn the mapping across the landscape, including the many variants that fail, plateau, or trade off activity against selectivity. Those "non-hits" define the boundary conditions of function and excluding them makes prediction less calibrated and less useful for prioritization.

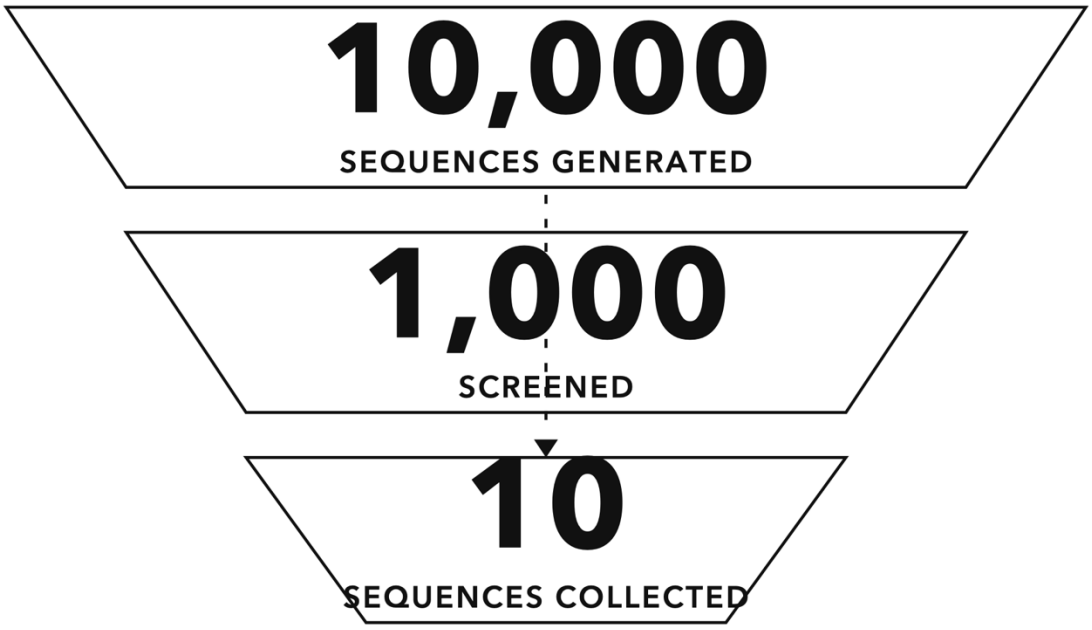


Figure 1-4. Selective genotyping biases sequence–function datasets toward hits. Large libraries are screened, but only a small subset is genotyped, excluding the non-hits and tradeoffs needed for calibrated prediction.

Genotype assignment must also be reliable, not merely present. Contamination, well swaps, mixed populations, recombination artifacts, and sequencing errors can create silent failure

modes where a measurement is attributed to the wrong sequence[36]. Such errors are catastrophic for learning because they inject label noise that is structured and difficult to detect after the fact. In an individual project, such mismatches sometimes reveal themselves when a variant fails to reproduce. In a dataset, they can persist undetected and degrade models in ways that are hard to debug.

For these reasons, genotyping workflows must include quality control outputs that are designed for reuse, not just for internal troubleshooting. QC should report whether a sequence call is confident, whether reads indicate mixed populations, whether barcodes are consistent with plate mapping, and whether variant identities are plausible given library design [37]. These checks turn genotyping from a confirmatory step into a data integrity layer. They also make the dataset self-aware by allowing later users to filter or weight records based on confidence rather than treating every sequence–label pair as equally trustworthy.

Chapter 3 introduces LevSeq as the solution to this bottleneck. It is positioned not merely as a sequencing method, but as a genotyping-and-QC layer designed to make plate-scale screens produce reusable, learnable sequence–function records with explicit provenance.

1.6 Why standardization and chemistry context matter

Even when assays are quantitative and sequences are complete, the resulting records are often trapped in inconsistent formats that prevent reuse. Spreadsheets differ in column names, units, and conventions. Substrates are recorded as names that are ambiguous or non-canonical. Reaction conditions are described in prose, scattered across notebooks, or omitted. Measurements may be normalized in one project and absolute in another. Without standardization, aggregation becomes a manual and error-prone process, and the dataset cannot grow into a coherent resource [4,7,38–40].

For enzymatic catalysis, chemistry context is not optional metadata; it is part of the input. A sequence–function record is only interpretable if the reaction being measured is represented in a way that can be compared across records. At minimum, this requires canonical substrate

identity and a clear description of the transformation and metric. It also requires enough information to interpret differences across records: reaction format, temperature, pH, cofactors, and whether the number reflects endpoint conversion, initial rate, total turnover, or selectivity at a specified conversion range. When these attributes are unstructured, later users cannot determine what is comparable, and models may learn spurious correlations that reflect dataset assembly rather than chemistry.

Standardization is also how engineered reaction data become interoperable. Existing public resources were largely built around natural enzymes and canonical biological transformations. They are invaluable, but they often underrepresent engineered scope expansions and new-to-nature chemistry. As the field pushes into these areas, engineering campaigns become the primary source of supervision. Standardization provides the mechanism by which those campaigns can contribute to a shared data substrate rather than remaining isolated case studies.

This thesis treats standardization as a minimal representation problem: what is the smallest set of fields that must be structured so that a record is reusable and learnable, while remaining practical enough that it can be adopted broadly. Chapter 4 introduces EnzEngDB as the solution to this bottleneck, with an emphasis on representing sequence–function pairs with chemistry context in a standardized, searchable form that supports aggregation and reuse.

1.7 Negative and neutral outcomes as constraints, not waste

A central failure mode in the creation of learnable datasets is survivorship bias. Many workflows record only improved variants, only successful reactions, or only the final optimized series. This is understandable from the perspective of project progression, but it is harmful for prediction. A model trained only on positives cannot learn the boundary of feasibility, cannot calibrate confidence, and will often assign high scores to implausible designs because it has not been trained on the failures that define what does not work [35,41,42].

Negative outcomes are also heterogeneous. A variant can fail because it does not express, because it misfolds, because it binds substrate non-productively, because it favors a side reaction, or because it loses stereocontrol while retaining activity. Collapsing all of these into "inactive" discards information that would be valuable both mechanistically and for learning. Even if a dataset does not explicitly annotate failure modes, retaining negative and neutral outcomes still improves learning because it teaches models what sequence neighborhoods do not yield the desired label under the given context.

Neutral outcomes are similarly valuable. In an engineering landscape, plateaus and trade-offs contain information about coupling between residues and outcomes. A dataset that includes neutral and negative results allows a model to learn not just what increases activity, but what changes do nothing, what changes break the enzyme, and what changes erode selectivity. These are precisely the constraints needed to make predictions robust and to prevent overconfident extrapolation.

Retaining these outcomes requires the upstream layers to be solid. If genotypes are incomplete, negatives cannot be interpreted. If labels are poorly defined, neutral results may simply reflect assay noise. If records are not standardized, failures cannot be aggregated across campaigns. This is why the thesis pipeline treats negatives and neutrals not as an optional add-on, but as a downstream beneficiary of better labels, better genotyping, and better record structure. Chapter 4's database layer is designed to make retaining and using these outcomes practical.

1.8 Thesis strategy: a pipeline for learnable sequence–function data

The strategy of this thesis is to build the upstream foundations that make functional prediction possible and useful. The pipeline begins with assays that generate quantitative, high-specificity labels aligned with practical deployment objectives. It then ensures that those labels can be attached to full-length genotypes routinely and accurately at screening scale, with explicit quality control that prevents silent data corruption. Finally, it consolidates

these records into a standardized database representation that preserves chemistry context and supports aggregation across projects while retaining negative and neutral outcomes.

This strategy is intentionally aligned with how protein engineering is already done. It does not require transforming every campaign into an academic data curation project. Instead, it aims to make a small set of practices routine: define labels explicitly, link every measured well to a verified genotype, record minimal context in a structured form, and keep the full distribution of outcomes. When these steps are built into the workflow, prediction becomes a natural next layer rather than a separate effort requiring bespoke dataset assembly.

With these foundations in place, the same dataset can support multiple uses. It can support local learning within a scaffold, helping choose the next round of mutations. It can support cross-scaffold comparisons when the reaction context is standardized. It can support prioritization of de novo designs by asking whether a candidate sequence resembles regions of sequence–function space that are known to be productive for a given transformation. It can also support uncertainty-aware decision-making by enabling models to detect when a design is outside the domain supported by data, which is often more important than assigning a single point estimate.

Chapter 5 is intentionally positioned as the “decision layer.” It uses the standardized data substrate produced by the earlier chapters to rank and prioritize designs for experimental validation, while keeping the core scientific contribution grounded in wet-lab testing. This reflects a practical thesis philosophy: prediction becomes compelling when it is embedded in a workflow that produces new high-quality data and when it is validated by experiments that matter.

1.9 Roadmap of the thesis

Chapter 2 demonstrates how a directed evolution campaign can be framed explicitly as the generation of prediction-ready functional labels, emphasizing quantitative outcomes and stereochemical resolution under conditions that reflect practical deployment formats. The

published work embedded in that chapter provides the experimental backbone, and the thesis wrapper makes the connection to the broader data foundations argument. The chapter is positioned as the "label specificity" layer: it illustrates why function prediction ultimately requires labels that capture not only whether an enzyme reacts, but how it reacts, and whether it performs under constraints that matter in real use.

Chapter 3 introduces LevSeq as the genotyping and quality control layer that makes plate-scale screening compatible with full-variant sequence resolution. The chapter focuses on how genotype assignment, when made routine and QC-aware, converts screening from a project-local activity into a generator of reusable records. It emphasizes that the limiting factor is not merely sequencing throughput, but the ability to map reads to wells, assign full sequences, and attach confidence metrics so that records can be trusted and reused without revalidation.

Chapter 4 presents EnzEngDB as the standardization and consolidation layer, designed to store sequence–function pairs with chemistry context in a way that supports aggregation, search, and analysis while elevating negative and neutral results as informative constraints. The chapter argues that database structure is not administrative overhead but a prerequisite for learnability and interoperability, especially for engineered and new-to-nature reactions where canonical public resources are sparse.

Chapter 5 then uses these foundations to support decisions, with a focus on experimentally validating de novo enzyme designs and incorporating a thesis-native analysis component that constructs a lightweight scoring head informed by EnzEngDB data. The intent is to demonstrate how prediction becomes actionable once the upstream data foundations are in place, and how experimental validation closes the loop by generating new, high-quality supervision.

Chapter 2

ENZYMATIC STEREODIVERGENT SYNTHESIS OF AZASPIRO[2.Y]ALKANES

This chapter incorporates material from the following publication, for which I am a co-first author: 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 (31), 27165–27171. The reprinted manuscript text is retained largely in its published form (including the collective voice "we"), and the surrounding Chapter 2 sections (2.1–2.2 and 2.5) provide thesis-specific framing and transitions.

Abstract

Azaspiro[2.y]alkanes are increasingly valuable scaffolds in pharmaceutical drug discovery; however, an asymmetric catalytic method for their synthesis remains unknown. Here, we present a stereodivergent carbene transferase platform for the cyclopropanation of unsaturated exocyclic N-heterocycles to provide structurally diverse and pharmaceutically relevant azaspiro[2.y]alkanes in high yield (21–>99% yield) and excellent diastereoselectivity and enantioselectivity (diastereomeric ratio (dr) values from 52.5:47.5 to >99.5:0.5; enantiomeric ratio (er) values from 51:49 to >99.5:0.5). These engineered protoglobin-based enzymes operate on gram scale, with no organic cosolvent, at substrate concentrations up to 150 mM (25 g/L) using lyophilized E. coli lysate as the catalyst. This platform represents a practical, scalable, and stereodivergent biocatalytic synthesis of azaspiro[2.y]alkanes, using low-cost engineered protoglobins and their native iron-heme cofactors.

2.1 Chapter Overview

This chapter embeds a first-author communication as a case study in application-driven enzyme engineering. The study reports directed evolution of protoglobins for cyclopropanation of unsaturated exocyclic heterocycles to form azaspiro[2.y]alkanes with high yield and strong stereocontrol, including stereodivergent access to multiple stereoisomers across a substrate panel.

In the context of this thesis, the key contribution of this chapter is methodological: it demonstrates how enzyme "function" can be defined through a specific assay context and captured as quantitative, deployment-relevant labels (activity and turnover, stereoselectivity, and operational tolerance under practical reaction formats). These labels motivate the need for scalable genotyping and standardized data infrastructure developed in later chapters.

2.2 Motivation and framing of the chapter

Rigid, three-dimensional scaffolds such as azaspiro[2.y]alkanes are increasingly valuable in medicinal chemistry because they expand accessible vectors for interaction while often improving physicochemical properties relative to more planar motifs. Yet despite their utility, routes to enantioenriched azaspiro[2.y]alkanes often require multi-step synthesis, specialized chiral catalysts, or downstream resolution, which increases cost and slows iteration during discovery.

Biocatalysis offers a complementary approach: enzyme active sites can impose stereocontrol in water and can be optimized by directed evolution to balance selectivity, activity, and operational practicality. In this chapter, engineered protoglobins catalyze cyclopropanation of unsaturated exocyclic heterocycles to generate azaspiro[2.y]alkanes with high yield and high enantio- and diastereoselectivity, including stereodivergent access to multiple stereoisomers across substrate classes. The work also emphasizes deployable catalyst formats and reaction conditions that matter for practical use, including activity in whole cells

or lysate-derived preparations and performance at high substrate concentration in fully aqueous media.

This chapter serves a second purpose in the thesis beyond demonstrating a useful transformation: it concretizes what is meant by enzyme "function" for prediction. Here, function is treated as an assay-defined quantity rather than an intrinsic annotation of sequence. The functional labels are the measurable outputs of a fixed experimental context, including product yield and total turnover, stereochemical outcomes (enantioselectivity and diastereoselectivity), and operational characteristics that determine deployability (catalyst format, substrate loading tolerance, and compatibility with aqueous conditions). These deployment-aligned labels are the starting point for the rest of the thesis: later chapters focus on scaling genotype assignment and quality control so such labels can be collected routinely across large libraries, and on standardizing metadata so results can be aggregated and compared across campaigns and laboratories.

2.3 Published work: Enzymatic Stereodivergent Synthesis of Azaspiro[2.y]Alkanes

2.3.1 Introduction

Azaspiro[x.y]alkanes are emerging scaffolds in pharmaceutical drug discovery, having been shown to enhance key pharmacokinetic and physicochemical properties, including potency, lipophilicity, target selectivity, and metabolic stability, among others (Figure 2-1A) [43,44]. These advantages are largely attributed to their rigid yet three-dimensional structure, featuring an expanded set of spatial vectors for chemical interactions with target proteins, compared to planar alternatives [45,46].

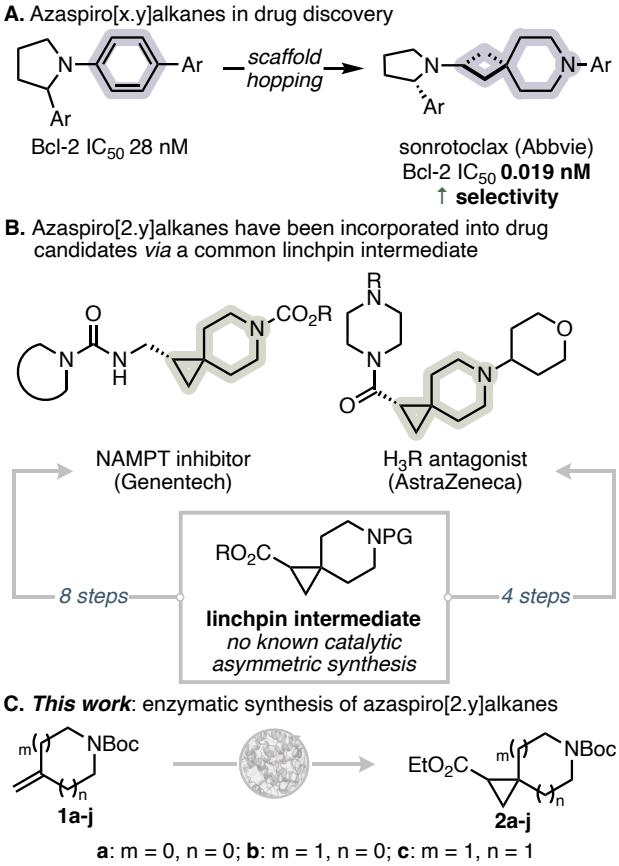


Figure 2-1. Azaspiro[x.y] alkanes overview (A) Azaspiro[x.y]alkanes are rigid, three-dimensional alterna-tives to aromatic linkers. (B) Azaspiro[2.y]alkanes have been incorporated into drug candidates spanning a range of thera-peutic areas. (C) This work describes an enzymatic route to azaspiro[2.y]alkanes of varying ring size.

Within this class of molecular frameworks, azaspi-ro[2.y]alkanes, characterized by a cyclopropane moiety joined at a quaternary carbon of a N-heterocycle, have been incorporated into numerous pharmaceutical drug candidates spanning various therapeutic areas (Figure 2-1B). For example, Genentech showed that replacing an aromatic linker with an azaspiro[2.5]octane improved the binding potency of a nicotinamide phosphoribosyltransferase (NAMPT) inhibitor while minimizing off-target interactions with the CYP2C9 enzyme[47]. AstraZeneca similarly utilized an azaspiro[2.5]octane in a histamine-3 receptor (H3R) antag-onist and found improved target selectivity and in vivo efficacy, while also reducing off-target effects [48]. Beyond these examples, the growing utility

of azaspiro[2.y]alkanes is evidenced by an increasing number of reports describing their incorporation into drug candidates, including azaspi-ro[2.3]hexanes [49,50] azaspiro[2.4]heptanes[51–53], and azaspi-ro[2.5]octanes[54,55], among other scaffolds[56–60].

The linchpin intermediate toward these structurally-diverse drug candidates commonly features an oxidized functionality bound to the cyclopropane—typically an ethyl ester—which, along with the amine, serves as a ver-satile handle for synthetic diversification[61,62]. Conventional syntheses toward this intermediate involve a Horner-Wadsworth-Emmons olefination of the requisite ketone, followed by a Corey-Chaykovsky cyclopropanation[55,63,64]. Alternatively, Rh- and Cu-catalyzed cyclopropanations of the corresponding alkene with a acceptor-type diazo carbene precursor are known[48,65,66]. However, an asymmetric catalyt-ic method has hitherto not been reported and chiral chroma-tographic separation is required to access enantioenriched intermediates, rendering these syntheses both costly and time- and resource-intensive[62,67,68].

Biocatalysis offers a promising asymmetric catalytic solution by providing a highly-ordered stereochemical environment that can be tuned to enhance activity, selectivity, and scala-bility. Our group has previously engineered heme-dependent enzymes for new-to-nature carbene transfer reac-tions, including cyclopropanations with acceptor-type diazo compounds[69–72]. We therefore hypothesized that an engi-neered carbene transferase could provide an asymmetric catalytic route to this valuable linchpin intermediate and that directed evolution (DE) would enable access to stereo-divergent pathways (Figure 2-1C). Our goal was to develop a platform of engineered enzymes that exhibit stereodiver-gence across a broad application-driven substrate scope, establishing a general, practical route to enantioenriched azaspiro[2.y]alkanes for drug discovery and production.

2.3.2 Results and Discussion

Hit identification and evolution from protoglobin libraries

At the outset, we screened an in-house library of 60 pro-toglobins previously engineered to effect carbene transfor-mations. Protoglobins are small (~200 amino acids), highly

thermostable hemoproteins that are easily synthesized by host microbes like Escherichia coli[73,74]. Excitingly, we discovered an Aeropyrum pernix protoglobin (ApePgb) containing four mutations from wildtype (W59A Y60G V63P F145W) (denoted ApePgb-xHC-5311; where xHC = eXo-cyclic Heterocycle Cyclopropanation) that catalyzes the cyclopropanation of tert-butyl 3-methyleneazetidine-1-carboxylate (1a) with ethyl diazoacetate (EDA) to yield (R)-2a in 56% yield with good enantioselectivity (er = 82:18) in a whole-cell context (Figure 2-2). Fortuitously, within the same enzyme library, a protoglobin from Thermus amyloliquefaciens (TamPgb, 60% pairwise amino acid (AA) sequence identity to ApePgb) containing two mutations from wildtype (W59L Y60Q) (denoted TamPgb-xHC-5316) favored (S)-2a in 56% yield with an er = 62:38.

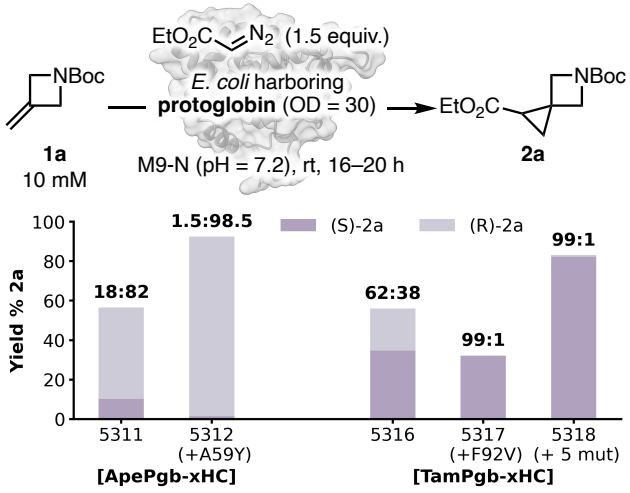


Figure 2-2. Enantiodivergent directed evolution of protoglobins toward (R)- and (S)-2a from 1a.

We considered ApePgb-xHC-5311 and TamPgb-xHC-5316 excellent starting points for the development of an enzyme controlled enantiodivergent pathway to 2a and conducted DE to improve both activity and selectivity in each stereo-chemical direction. Site saturation mutagenesis (SSM) on active site residues revealed ApePgb-xHC-5312, containing a mutation at site 59 (A59Y), which provides (R)-2a in 93% yield and excellent enantioselectivity (er = 98.5:1.5). Toward (S)-2a, a critical mutation at site 92 (F92V) pro-

vided TamPgb-xHC-5317, which yields (S)-2a with an er = 98:2, albeit with reduced yield (32%). From here, error-prone mutagenesis (epPCR) provided TamPgb-xHC-5318, containing five additional AA substitutions (D57G G61D K134R P137L K153R) that reinstated high activity (83% yield) while maintaining excellent enantioselectivity (er = 99:1).

Building the stereodivergent platform across ring sizes

Given the pharmaceutical relevance of azaspiro[2.y]alkanes of varying ring size and substitution, we sought to build upon our initial engineering campaign to provide a suite of engineered enzymes capable of achieving stereodivergence across multiple scaffolds. To do so, we tested variants TamPgb-xHC-5316–5318 and ApePgb-xHC-5311–5312, as well as select mutant libraries, on pyrrolidine- and piperidine-based substrates tert-butyl 3-methylenepyrrolidine-1-carboxylate (1b) and tert-butyl 4-methylenepiperidine-1-carboxylate (1c), respectively. We reasoned that 1b, a non-symmetric scaffold, would expand the platform's structural diversity, while 1c, a common motif in drug discovery, would bolster its utility.

Encouragingly, we observed promising activity and selectivity toward all six possible stereoisomers of 2b and 2c, with ApePgb-xHC-5312 providing isomer 2 of 2b in excellent yield and high diastereo- and enantioselectivity. We next conducted parallel DE campaigns, optimizing each enzyme lineage for selectivity and activity toward a given stereoisomer. We identified evolved enzymes capable of forming all possible stereoisomers of 2b and 2c with excellent stereoselectivity and exceptional activity (see Appendix A for evolution details). These engineering efforts culminated in a carbene transferase platform for the stereoselective synthesis of azaspiro[2.y]alkanes (Figrue 2-3). Site 92 remained a critical mutation for stereodivergence with both 1b and 1c, suggesting a conserved mechanistic role toward varying azaspiro[2.y]alkane frameworks (see Appendix A for details). Altogether, the sequences represented in this platform are highly conserved, with Ape and Tam variants averaging Hamming distances of 4.3 and 8.5 amino acids, respectively, underscoring the functional diversity accessible through minimal rounds of DE.

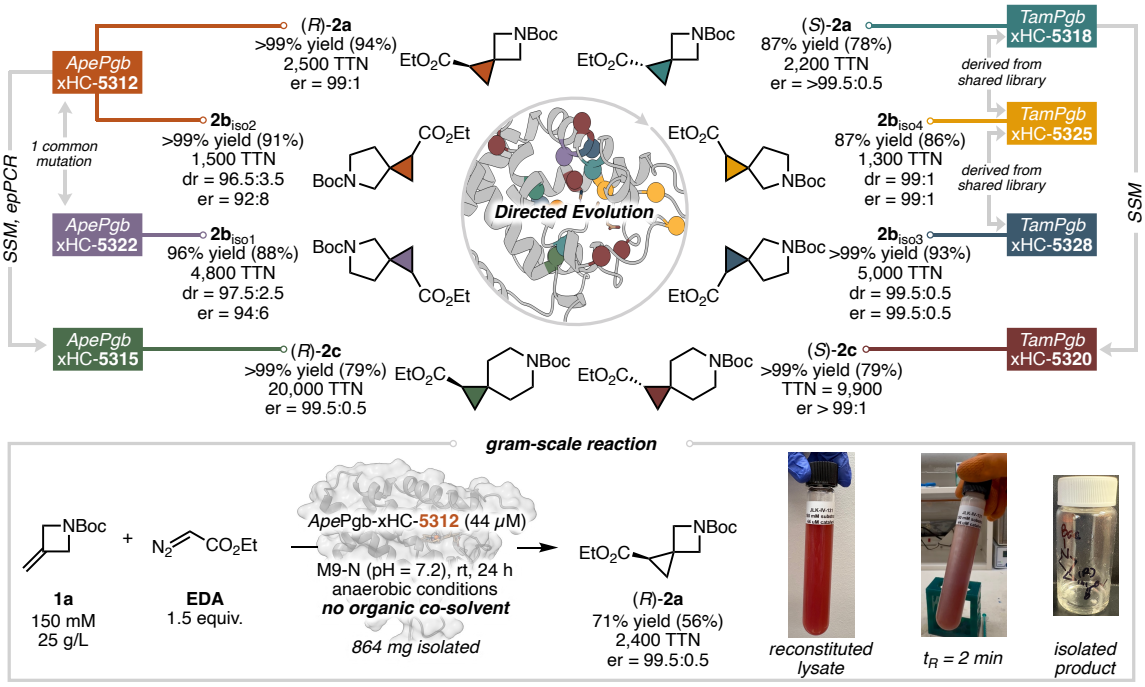


Figure 2-3. Engineered carbene transferase platform for synthesis of azaspiro[2.y]alkanes. Reactions were performed using reconstituted lyophilized lysate powder on 0.4 mmol scale (unless otherwise specified). Analytical yields were determined by comparison to an internal standard using GC-FID and a corresponding calibration curve. Isolated yields are designated in parentheses. TTNs are based off analytical yield. Stereoselectivities (er and dr) were determined on crude reaction mixtures using GC-FID equipped with a chiral stationary phase.

Practicality and scalability

Notably, all seven final variants are active catalysts as lyophilized lysate powders at 0.4 mmol scale and can be distributed similarly to any standard chemical reagent. Under these conditions, each variant exhibits TTNs between 1,000–20,000 and isolated yields between 78–94%. These protoglobin catalysts also retain activity at high substrate concentrations. Using a 1-L expression culture of ApePgb-xHC-5312, subjected to lysis and lyophilization, we observed 88% conversion of 150 mM (25 g/L) of 1a to yield 2a in 71% yield (56% isolated yield) (2400 TTN) with an er = 99:1. This reaction is performed in a fully aqueous environment without an organic co-solvent. This is amongst the highest reported substrate loading for a carbene transferase, without compromising stereoselectivity[75]. These results

highlight the practicality of our enzyme platform and its potential for large-scale applications without further protein engineering.

We next evaluated all seven final variants across a range of substrates, focusing on the synthesis of azaspi-ro[2.y]alkanes with established pharmaceutical relevance. First, we tested various N-protecting groups, finding that the platform is amenable to N-Cbz and N-Bz protecting groups, forming 2d and 2e in high yield and with excellent enantioselectivity, with distinct variants selectively produc-ing either the (R)- or (S)-enantiomer in each case (Figure 2-4).

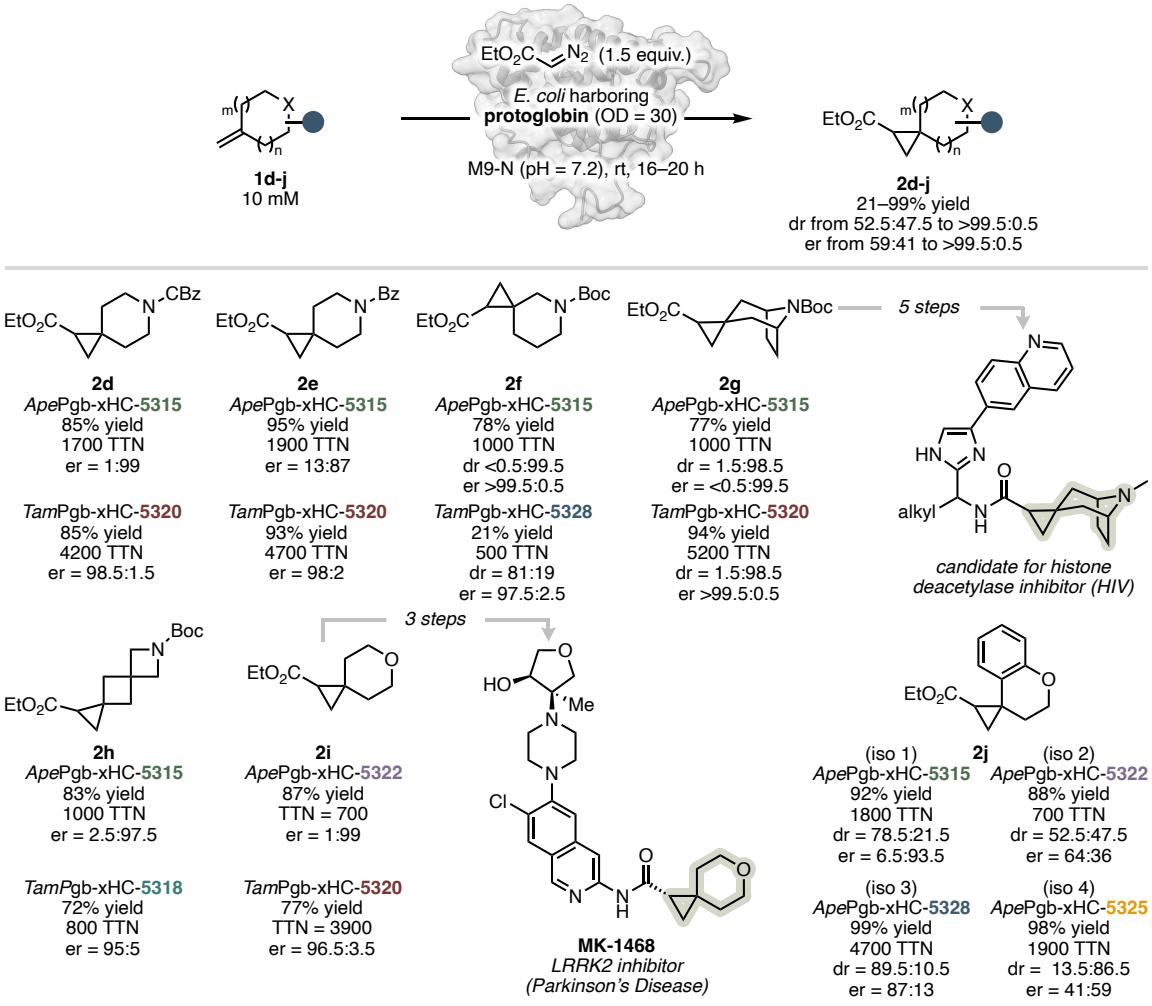


Figure 2-4. Engineered carbene transferase platform for synthesis of azaspiro[2.y]alkanes. Reactions were performed on 0.4 μmol scale using whole cells. Analytical yields were

determined by comparison to an internal standard using GC-FID and a corresponding calibration curve. Stereoselectivities (er and dr) were determined on crude reaction mixtures using either GC-FID, HPLC, or SFC-MS equipped with a chiral stationary phase.

We next tested a non-symmetric piperidine substrate (1f) to form 2f, a motif featured in numerous drug candidates, including anti-bacterial agents[76], MDM2 modulators, and DGAT1 inhibitors[59]. Excitingly, ApePgb-xHC-5315 furnished one isomer of this compound in high yield and with excellent enantio- and diastereoselectivity. In contrast, TamPgb-xHC-5328 yields another isomer in moderate yield and diastereoselectivity, with excellent enantioselec-tivity. Further, ApePgb-xHC-5322 and TamPgb-xHC-5325 provide excellent starting points to engineer variants selective for the remaining two possible stereoiomers (see Appendix A for details).

The platform similarly provides excellent starting points to selectively access multiple stereoisomers of 2g, a tro-pane derivative that has been incorporated into drug candi-dates targeting histone deacetylase and DGAT1. TamPgb-xHC-5320 and ApePgb-xHC-5315 provide two of the four possible stereoisomers in high yields and with excellent levels of diastereo- and enantioselectivity. Fur-ther, ApePgb-xHC-5312 is a promising starting point for the formation of a diastereomer highly disfavored in the Rh-catalyzed cyclopropanation of 1g (See Appendix A for details).

Additionally, 2h has shown promising potency as a muscarinic acetylcholine receptor (mAChR) antagonist[77]. Excitingly, ApePgb-xHC-5315 furnishes 2h in 83% yield with excellent enantioselectivity (er = 2.5:97.5). In con-trast, TamPgb-xHC-5318 delivers the antipode in 72% yield with an er of 95:5.

We additionally probed oxygen heterocycles and found that the platform is amenable to the synthesis of 2i, a pyran derivative, which has been incorporated into LRRK2 inhibitor MK-1468 by Merck[78] and studied by many others. We observed excellent activity and importantly, stereo-divergence: ApePgb-xHC-5322 provides the 2i with an er = 1:99 and TamPgb-xHC-5320 provides the antipode with an er = 96.5:3.5.

We next evaluated chromane derivative 1j, a scaffold markedly distinct from those used during platform evolution. Despite this structural divergence, the platform enables access to all four diastereomers of product 2j in high yield, with different variants selectively favoring each isomer. This result highlights the platform's substrate promiscuity, suggesting broad applicability to structurally diverse scaffolds beyond those explicitly represented in the evolution substrate set.

We describe a carbene transferase platform for the cyclopropanation of unsaturated exocyclic N-heterocycles to yield azaspiro[2.y]alkanes in high yields and with excellent levels of enantio- and diastereoselectivity. This stereodivergent platform shows excellent substrate generality, obviates the need for post synthetic resolution steps, and is amenable to gram scale synthesis; thus, improving applicability for both drug discovery and production efforts. The observation of stereodivergence on a substrate that is structurally distinct from those evolved on suggests that this enzyme platform contains biocatalysts capable of accessing a wide range of desired azaspiro[2.y]alkane stereoisomers. We envision this platform as a starting point for the selective synthesis of any desired stereoisomer of azaspiro[2.y]alkanes and anticipate its integration into drug discovery pipelines.

2.4 Conclusions and outlook

This chapter establishes a practical model for how enzyme function can be defined and measured in a way that is aligned with deployment: yields, stereoselectivity, and operational tolerance are captured as assay defined labels across engineered variants. This is the level of label specificity that function prediction ultimately requires. The natural next question is how to generate such labels not for dozens of variants, but for thousands, while maintaining confidence in the genotype assigned to each measurement. As directed evolution campaigns scale, variant identification and data integrity become limiting factors, especially when libraries screening include multiple fitness label. Chapter 3 therefore focuses on scalable genotyping by introducing LevSeq, a sequencing and analysis workflow that assigns full variant identities at high throughput and links them reliably to plate based functional readouts with routine quality control.

Chapter 3

LEVSEQ: RAPID GENERATION OF SEQUENCE-FUNCTION DATA FOR DIRECTED EVOLUTION AND MACHINE LEARNING.

This chapter incorporates material from the following publication, for which I am a co-first author: **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 (1), 230–238.

Abstract

Sequence-function data provides valuable information about the protein functional landscape, but is rarely obtained during directed evolution campaigns. Here, we present Long-read every variant Sequencing (LevSeq), a pipeline that combines a dual barcoding strategy with nanopore sequencing to rapidly generate sequence-function data for entire protein-coding genes. LevSeq integrates into existing protein engineering workflows and comes with open-source software for data analysis and visualization. The pipeline facilitates data-driven protein engineering by consolidating sequence-function data to inform directed evolution and provide the requisite data for machine learning-guided protein engineering (MLPE). LevSeq enables quality control of mutagenesis libraries prior to screening, which reduces time and resource costs. Simulation studies demonstrate LevSeq's ability to accurately detect variants under various experimental conditions. Finally, we show LevSeq's utility in engineering protoglobins for new-to-nature chemistry. Widespread adoption of LevSeq and sharing of the data will enhance our understanding of protein sequence-function landscapes and empower data-driven directed evolution.

3.1 Chapter Overview

This chapter establishes the genotyping and quality-control layer required to make plate-scale enzyme screens produce learnable sequence–function data. In Chapter 2, functional labels were defined by assay readouts; however, scaling campaigns exposes a limiting step that is often treated as ancillary: assigning full-length variant identities at throughput, and doing so with enough internal checks that genotype–phenotype pairs remain trustworthy as the number of clones, plates, and rounds grows. LevSeq (Long-read every variant Sequencing) is introduced here as a sequencing-and-analysis workflow that links plate-indexed measurements to full-gene sequences using long-read sequencing and a dual-barcoding strategy, enabling systematic coupling of genotypes to screening outcomes.

Beyond throughput, LevSeq is framed as a data integrity intervention: it produces explicit quality-control outputs that surface common failure modes (e.g., ambiguous identities, mixed wells, or insufficient evidence to call a consensus sequence) before such entries are silently absorbed into downstream analysis. In the thesis arc, this chapter therefore contributes more than "sequencing support", it converts screening from an experiment that yields isolated results into a workflow that yields auditable, sequence-resolved labels, at the scale and cadence of directed evolution.

3.2 Motivation and framing

Protein function prediction is ultimately constrained not by model capacity, but by the scarcity of datasets that are simultaneously large, sequence-resolved, and experimentally grounded. Directed evolution generates exactly the type of signal learning systems need, quantitative, deployment-relevant functional labels, but most campaigns do not routinely record full variant sequences for every screened variant. The consequence is structural: many experimental records remain trapped as "plate outcomes" that cannot be integrated across experiments, compared across rounds, or interpreted as points on a sequence–function landscape.

A second limitation is that the sequence–function datasets that do exist are often shaped by convenience rather than practice: deep mutational scans and restricted library designs provide controlled perturbations, but do not reflect the mutational structure of typical engineering workflows (whole-gene mutagenesis, recombination, and multi-round accumulation). The LevSeq framing is explicit that practical learning will require sequence–function data generated under the realities of enzyme engineering rather than only under idealized library regimes.

LevSeq addresses these bottlenecks by making full-variant genotyping routine at screening scale. It combines a dual barcoding strategy (to preserve plate context during multiplexing), nanopore long-read sequencing (to capture full coding regions in single reads), and a software workflow that performs demultiplexing, variant identification, and the explicit pairing of genotypes to functional readouts.

Critically, LevSeq is not framed as "sequencing for sequencing's sake," but as a means to produce data that are reusable and learnable: variant identities are assigned with built-in evidence and QC signals, so downstream analysis can separate high-confidence genotype–phenotype pairs from entries that require follow-up validation or exclusion. In the thesis logic, this chapter operationalizes the central principle introduced in Chapter 1: function becomes predictable only when assay-defined labels are attached to specific sequences with sufficient provenance and quality control to support aggregation across time, projects, and laboratories.

3.3 Published work: LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning.

3.3.1 Introduction

Directed evolution (DE) has been key to the discovery and engineering of biocatalysts for new-to-nature chemistry [18], development of sustainable bioprocesses for pharmaceutical synthesis [79,80], and engineering of proteins for environmental sensing [81] and bioremediation

[82], among many other applications. The power of directed evolution resides in the rapid evaluation of mutated proteins to traverse the fitness landscape toward those exhibiting improved fitness [27,82]. A typical directed evolution campaign involves the generation and screening of thousands of variants – a significant number but still only a tiny fraction of the possible sequence space [83]. To streamline directed evolution, machine learning (ML) can be employed [25,84,85] to guide sequence-function exploration towards variants with high fitness [29,86–88].

Traditional directed evolution (DE) approaches have generated datasets rich in activity labels but often lacking sequence information, as they focus on optimizing activity without sequencing all variants. Existing sequence-function datasets for protein evolution studies are primarily comprised of deep mutational scanning data covering all single mutations or combinatorial libraries targeting specific sites [35,89,90]. While valuable, these approaches are costly and capture only a fraction of the sequence diversity most useful for protein evolution [91]. To advance machine learning in protein engineering, we need a method for collecting, analyzing, and pairing sequence-function data from diverse mutagenesis approaches [92,93]. This method would work for random mutagenesis across whole genes, combinatorial libraries at sites distant in the primary sequence, and other targeted mutagenesis approaches.

Challenges that must be overcome to realize this vision include the high cost of sequencing entire gene [94] and the lack of a standardized format to create and distribute the data. There exist several approaches to sequence entire genes, yet many of these focus on validating construct quality and do not tie sequence to function [95,96]. Earlier methods used Illumina short-read sequencing for this purpose, whereas the methods since 2019 have started to use nanopore long-read sequencing [97–99]. The Arnold lab developed the Every Variant Sequencing (evSeq) method using Illumina short-read sequencing to capture the sequences of variants arrayed in 96-well plates [100]. Due to the short sequencing lengths (~250 base pairs), however, evSeq is not ideal for collecting full-gene-length gene sequences. In contrast, real-time sequencing technologies like nanopore sequencing can capture millions of long reads at a low cost [101], but nanopore sequencing is characterized by a high error rate

[97,102,103]. Previously published high-throughput nanopore sequencing methods, parSEQ [104] and SequenceGenie [97], have overcome this limitation by performing statistical analyses on consensus reads to detect true variants. UMIC-seq takes a different approach and clusters sequences rather than identifying individual variants, as the objective is to map evolutionary lineages [105]. Each of these methods uses a similar DNA-barcoding approach to demultiplex reads, which we have now coupled with the evSeq pipeline to enable collection of sequence-function data for directed evolution studies.

This work describes Long-read every variant Sequencing (LevSeq), which extends the previous evSeq method by utilizing the barcode strategy described in Currin et al. [97] for nanopore sequencing, enabling the evSeq pipeline to be utilized on full-length genes. LevSeq includes the following steps: 1) a colony polymerase chain reaction (PCR) to generate barcoded gene amplicons, 2) Oxford Nanopore sample preparation and sequencing to generate sequencing information, 3) demultiplexing and variant identification, 4) sequence-function data coupling accompanied by visualization and analyses, and 5) generation of data outputs that are amenable to downstream ML and compatible with existing databases. This method is rapid and robust under different mutagenesis conditions and enables researchers with no prior experience working with next-generation sequencing (NGS) data to perform analyses of mutagenesis libraries.

Importantly, LevSeq offers several advantages: a) the software is open source, easy to set up, and designed for directed evolution experiments; b) it requires as few as twenty reads to detect a variant in a well; c) results are available before the resource-intensive screening phase, enabling selection of specific variants for testing; d) fitness data are linked with sequence information to inform subsequent engineering steps. We demonstrate LevSeq in two protein engineering projects. First, we sequenced ~1,000 variants of an error-prone polymerase chain reaction (epPCR) random mutagenesis library and identified the top variants in a typical epPCR workflow by coupling sequence and function data. In the second demonstration, we applied LevSeq to variants sampled from a five-site combinatorial library, which yielded data for downstream ML packages to predict variants with increased activity

[106]. We show that LevSeq facilitates machine learning-guided protein engineering (MLPE) by collecting a small subset of sequence-function data from the studied system to be used as training or input data.

3.3.2 Materials and Methods

Design of backbone-specific barcoded primers

Universal binding sites of the pET-22b(+) cloning vector are first identified, and two sites upstream and downstream of the cloning site were chosen for primer design. All variants cloned using the pET-22b(+) vector can be sequenced using the primers designed for this research. Alternatively, primers can be designed for different cloning vectors, as long as the barcodes are attached to the upstream of the upstream primer and downstream of the downstream primer. (Appendix B Oligonucleotide Design).

Colony PCR for generating barcoded amplicons

PCR protocols are optimized for robust amplification of the full-length gene. Best performance is obtained using Taq polymerase and a touchdown PCR program. The PCR set up for each well includes 1 μL of overnight culture, 2 μL of 1 μM each barcoded primer mix, and 7 μL of PCR master mix. Using either 96-well or 384-well PCR thermocyclers, an initial 300 s denaturing step is performed followed by the touchdown PCR program detailed in the supplementary information. Depending on the length of the gene, one minute elongation time per 1 kb is recommended for optimal amplification. Amplicons were then pooled (Supplementary Information), analyzed, and purified with gel electrophoresis using the Zymoclean Gel DNA Recovery Kit (Zymo Research D4002).

LevSeq Sample preparation and sequencing

Purified amplicon samples were normalized and combined into one sample and prepared for sequencing using the Oxford Nanopore ligation sequencing kit (LSK-114). For the MinION and Flongle run setup, 0.02 Gb of base-called bases per 96 variants and super accurate base-

calling model are recommended as sequencing parameters. One Flongle flow cell is recommended for sequencing up to 1,600 variants, whereas the MinION flow cell can be washed and reused using the Oxford Nanopore wash kit until the number of pores decreases below 100. We recommend skipping the use of storage buffer as significant pore loss was observed after applying storage buffer to the flow cell (~300 pores lost).

Sanger sample preparation and sequencing

Overnight cultures from 96-well deep-well plates were transferred to 1.7 mL microcentrifuge tubes and centrifuged at 14,000 rpm for 1 minute. Cell pellets were processed for plasmid isolation using the Monarch Plasmid Miniprep Kit (NEB #T1110). The purified plasmids were submitted to Laragen for Sanger sequencing using the T7 primer.

Variant calling

The computational delineation of reads from a pooled sample is slow and thus we wrote a bespoke pairwise local alignment using the Smith-Waterman algorithm in C++ to efficiently detect barcodes at the 5’ and 3’ ends of each nanopore read. The 3’ end barcodes are aligned to the last 100 base pairs of each read, and the highest matching score above threshold 80 is used to assign each read to the 96-well plate of origin. Next, 5’ barcodes are aligned to the first 100 base pairs of each read, and the highest matching score is used to assign each read to a specific well within the assigned plate. The reads for each well are aligned using minimap2, version 2.1. The parameters for minimap2 are the standard long read parameters: “-ax map-ont”, a -B mismatch score of 2, a match score of 4, and a gap opening penalty of 10. These were chosen to deprioritize frame shift mutations, because they occur less frequently. If multiple reads with the same read ID are mapped to the same well, the read with the highest quality is retained. During variant calling several quality control files are produced: a multiple sequence alignment and a csv file for each well in each plate, which contains the p-values, p-adjusted values, and counts for each position in the sequence. The sequencing error for a nanopore device is comparatively high at approximately 10% and dependent on myriad factors such as the flow cell, age of the cell, run conditions, etc. As

such, for each well we calculate the error rate as the mean rate of non-reference nucleic acids per position. The probability that a mutation observed across a set of reads is a true mutation can be calculated using the binomial test. Namely, the null hypothesis, $\pi 0$, is that the observed sequence variation is due to the inherent sequencing error of the nanopore device, and the alternative hypothesis is that the observed nucleic acids are due to a mutation induced by SSM or epPCR. The number of trials, n, is the number of reads for a given well, the number of successes, k, is the number of a given nucleic acid or deletion, that is different to the reference sequence. Significance of the observed data is calculated using a one-sided test, testing for greater than expected error $\pi > \pi 0$, and calculating this for each A, T, G, C and deletion for each position for each well. The expected error rate used can be defined by the user; we use a default of 10%, or the mean error rate for the well. Multiple testing is corrected for by using the Benjamini-Hochberg test, with a false discovery rate of 0.05. For each well, for a given sequence, the number of tests that are corrected for is equivalent to the length of the sequence. Additionally, corrections for multiple testing are made across the wells that meet a mutation frequency threshold. If a well has a mutation frequency above a user-defined threshold, the well is checked for mutations and mixed wells. A well is classified as "mixed" if a position has more than one significant mutation by the FDR adjusted binomial test, using a threshold of $p < 0.05$ by default. Finally, post-variant calling we match the nucleotide variants to the amino acid changes.

Simulation study

For the simulation, the protoglobin used in case 1 and case 2 was chosen. This protein is 204 amino acids long. For epPCR, errors are introduced at the DNA level, and as such an error rate of 2% corresponds to approximately 12 nucleotide mutations. To test the effect of sequencing error, sequencing error was varied from 0 to 100% across the sequence by incrementing at 5% intervals with a constant read depth of 10 reads. For read depth, the number of reads varied from one read to 50 in increments of 1, with the sequencing error held constant at the reported nanopore sequencing error of 10%. To test the effect of sequence length, the sequence was trimmed to lengths between five and 200 at step sizes of 20.

Software stack

After the initial base-calling, which is a default option when using the MinION protocol, LevSeq is operating-system independent and runs entirely using docker or a web app. We modularized the application to comprise two components. The first is a command line and app that is used to de-multiplex and call variants. This is hosted locally, to reduce the need for large data transfer and processing. We opted to deploy this as an in-house tool as this component is stable and unlikely to change in future software updates. The code and docker image are available on GitHub (https://github.com/fhalab/LevSeq). It includes a multiple sequence alignment that shows the pileup from the bam files for quality control visualization. The output is an interactive HTML file that provides a per-plate view of the mutation, sequence count, and alignment probability. The second component is a web application where users upload screening data with coupled function. To calculate the combined "fitness", we normalize each user-provided feature and then compute the median across the normalized features. While median is the recommended summary statistic, as it is less susceptible to outliers, users can switch to calculating the mean across features.

Error-prone PCR random mutagenesis library generation for ParLQ

Error-prone mutagenesis libraries were prepared using a standard error-prone PCR protocol. We designed primers using the template given in Supplementary Information, Table S13. Different concentrations (100 mM, 200 mM, 300 mM, 400 mM) of MnCl2 were added to each PCR. Once PCRs finished, 1 μL of DpnI (NEB R0176S) was added to each of the reactions followed by incubation at 37 ◦C for 1 h to digest any residual template plasmid. DNA fragments with the desired size were excised from an agarose electrophoresis gel and then purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research D4002).

The expression plasmids containing an ampicillin resistance gene were constructed following standard Gibson assembly method. After 1 h of incubation at 50 ◦C, the reaction mixtures were used to transform T7 Express competent BL21 E. coli cells (NEB C2566H). Transformed cells were spread onto solid agar selection medium consisting of Luria broth

(RPI L24045000.0) supplemented with 0.1 mg/mL ampicillin (LBamp) and incubated at 30 ◦C until visible individual colonies are formed. To grow the error-prone libraries, 600 µL of LBamp were added into each well of 96-well deep-well plates (2-mL well volume). Individual colonies from the agar plates were then transferred into the wells with each well containing a single colony. The plates containing these overnight cultures were shaken at 220 rpm, 37 ◦C, and 80% humidity for 16 hours in an Infors Multitron HT shaking incubator. After overnight growth, 100 µL of overnight cultures were added to 100 µL of 50% glycerol solution to make glycerol stock plates, these plates can be used to store variants for future analysis.

Sequencing of ParLQ epPCR libraries

With the fresh overnight culture, sequencing libraries were prepared following the protocol described in Appendix B, LevSeq Library Preparation and Sequencing; the LevSeq software was run using all default parameters. Barcode-linked primer plates used are in Appendix; the barcode plates were paired to libraries as given in Appendix B.

Measuring cis and trans cyclopropane formation from 4-methoxystyrene

For expression of the variant libraries, 50 µL of the saturated overnight cultures were used to inoculate 900 µl of Terrific Broth with 0.1 mg/mL ampicillin (TBamp) in 96-well plates. These cultures were then grown at 37 ◦C, 220 rpm, and 80% humidity for 2.5 h in an Infors Multitron HT shaking incubator, after which they were placed on ice for 20 minutes. Following this, 25 µL of a 20 mM solution of isopropyl-β-d-thiogalactoside (IPTG; GoldBio # I2481C100) and 25 µL of a 40 mM solution of 5-aminolevulinic acid (ALA; thermo scientific # 103920050) in TBamp were added to each well to induce protein expression at a final concentration of 0.5 mM IPTG and 1 mM ALA. Expression proceeded in the same Infors shaker at 22 ◦C and 220 rpm for 18 h. Cells were harvested through centrifugation at 4,000g for 5 minutes, the supernatant was removed, and the pellets were resuspended in 380 µL of M9-N. In an anaerobic Coy chamber, 20 µL of 200 mM 4-methoxystyrene and 300 mM ethyldiazoacetate in acetonitrile were added into each well of the resuspended pellets.

The reaction plates were sealed using sticky aluminum foil and shaken at room temperature at 800 rpm on an IKA MTS 4 shaker for 20 h. Following the reaction, 800 µL of cyclohexane were added to each well, and the reactions were shaken and centrifuged at 5,000g for 10 minutes. The organic layer was transferred into GC screw vials (Agilent 5182715) and analyzed using GCMS (Agilent 7820A(G4350A)).

Individual measurement of cis and trans cyclopropane formation from 4-methoxystyrene

For expression of the Sanger sequence validated variants, 500 µL of the saturated overnight cultures were used to inoculate 50 mL of Terrific Broth with 0.1 mg/mL ampicillin (TBamp) in 125 mL Erlenmeyer flasks. These cultures were then grown at 37 ◦C, 220 rpm, and 80% humidity for 2.5 h in an INNOVA 42 shaking incubator, after which they were placed on ice for 20 minutes. Following this, 50 µL of a 500 mM solution of isopropyl-β-d-thiogalactoside (IPTG; GoldBio # I2481C100) and 50 µL of a 1000 mM solution of 5-aminolevulinic acid (ALA; Thermo Scientific # 103920050) in TBamp were added to each flask to induce protein expression at a final concentration of 0.5 mM IPTG and 1 mM ALA. Expression proceeded in the same INNOVA shaker at 22 ◦C and 220 rpm for 18 h. Cells were harvested through centrifugation at 4,000g for 5 minutes, the supernatant was removed, and the pellets were resuspended in 5 mL of M9-N, the optical density is measured at 600 nm wavelength and adjusted to 30 using M9-N. From the OD adjusted resuspension, 380 µL is transferred into 2 mL GC screw vial (Agilent 5191-8121). In an anaerobic Coy chamber, 20 µL of 200 mM 4-methoxystyrene and 300 mM ethyldiazoacetate in acetonitrile were added into each well of the resuspended pellets. The screw vials were capped and shaken at room temperature at 800 rpm on an IKA MTS 4 shaker for 20 h. Following the reaction, 800 µL of acetonitrile were added to each well, and the reactions were shaken and centrifuged at 5,000g for 10 minutes. The supernatant was transferred into a 400 µL GC screw vial insert (Agilent 5181-3377) and analyzed using LCMS (Agilent MSD/iQ G6160A).

3.3.3 Results and Discussion

A standardized workflow to sequence thousands of full-length variant genes

We use a dual backbone-specific barcoded primer system to streamline the sequencing process and maximize resource efficiency[107]. The primer sequences are designed for pET-22b(+) backbones and can be redesigned for other cloning vectors following standard design techniques (Appendix B Oligonucleotide Design). Compared to the original evSeq approach, LevSeq is not constrained by sequence length, can sequence any gene of interest in the cloning backbone, and has a short turnaround time (3–12 hours). The protocol commences with a one-step colony PCR that produces a full-length protein-coding DNA amplicon with unique barcode pairs at both ends (Figure 3-1 A). Using 96 unique forward barcodes for each well of a 96-well plate and 96 unique reverse barcodes for as many as 96 unique plates, it is theoretically possible to demultiplex and sequence 9,216 variants[108–111].

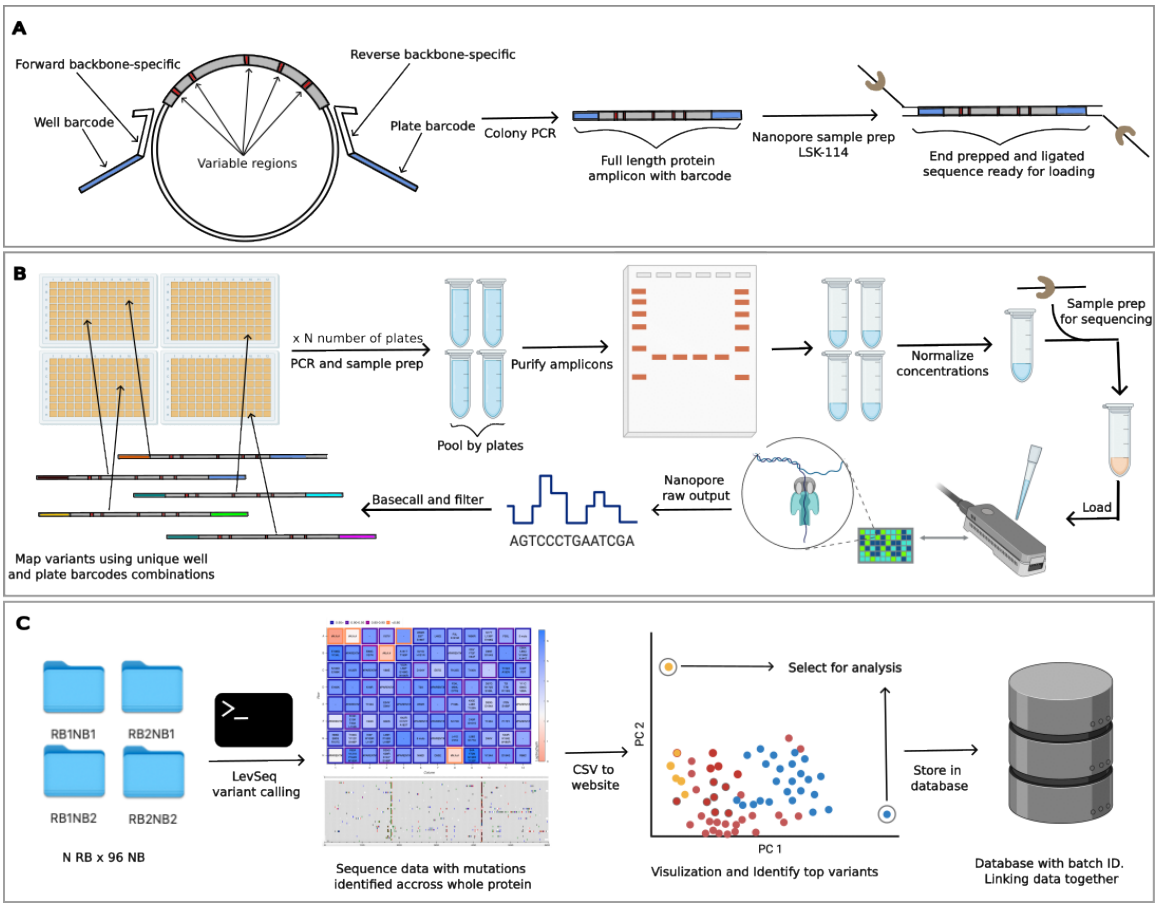


Figure 3-1. Overview of LevSeq library preparation, variant sequencing, and data visualization. A. The first step of LevSeq involves a one-step PCR using backbone-specific 5' and 3' end barcoded primers to amplify the full-length targeted gene. B. All PCR products from one 96-well plate are pooled for gel purification. The purified DNA samples from each plate are normalized by molarity and combined for nanopore sample preparation using the ligation sequencing kit. The sequencing run is performed in-house on a MinION sequencer, and the raw voltage signals are base-call-converted into nucleotides with the resulting fastq reads filtered by quality. C. Sequence function data pairing, visualization, and storage in format compatible with database.

A typical round of directed evolution with LevSeq begins with isolating colonies into a 96-well plate for overnight culture, followed by protein expression and screening. The LevSeq protocol is executed during the protein expression time, after overnight cultures of individually arrayed colonies in 96-well plates are grown to saturation.

A small amount of overnight culture is combined with PCR master mix and barcoded primers to generate barcoded gene amplicons from each well (Supplementary Protocols). After colony PCR, DNA samples are pooled and normalized (Figure 3-1 B). Samples are prepared for sequencing using the ligation kit provided by Oxford Nanopore before being loaded onto the MinION or Flongle flow cell. The real-time sequencing data, stored as raw and base-called files, are readily processed by the open-source software for comprehensive data analysis, visualization, and storage (Figure 3-1 C). Our software also provides a template for LC-MS instruments based on the sequencing results to screen only true variants, reducing the screening load and automatically coupling the sequence function data (Figure 3-2).

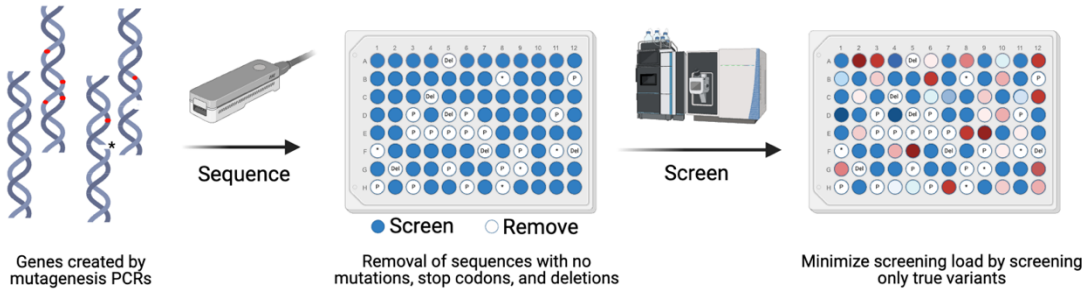


Figure 3-2. LevSeq reduces screening burden by enabling removal of sequences with no mutations, stop codons, and deletions.

LevSeq has an accurate variant caller that is robust to experimental designs

We developed a software suite for LevSeq that consists of two components to process and analyze every variant. The first is an operating system-independent docker image that performs efficient barcode de-multiplexing and runs on the sequencing computer [112]. The de-multiplexed plate and well data are parsed by a Python package to identify statistically significant variants and notify users of any poor-quality mappings or mixed wells. If a particular barcode is undesirable, users can edit the barcode sequences file provided in the software suite to incorporate any customized barcodes; no information beyond the barcode sequence is needed for the demultiplexing step. We validated the variant calling software by performing over 1,000 simulations to test the effect of experimental and sequencing

conditions on the variant calling accuracy. Simulated experimental variation, defined here as protein sequence length and epPCR error rate showed no effect on the efficacy of variant calling (Appendix B Figure B1). Sequencing variation, such as nanopore error rate, does not affect the ability to accurately call variants if more than 10 reads are assigned to a well, which is within the typical flow cell operating range (error < 20%), (Appendix B Figure B1). With simulated data, our pipeline showed that variants are accurately detected (>99%) with 10 reads41 to generate a consensus sequence [113,114] up to a sequencing error rate of 20%, which exceeds the expected error of nanopore devices [115] (Figure 3-3A). However, we note that experimental variation may reduce this accuracy and recommend a minimum sequence depth of 20 reads, as strand bias may impact calling accuracy. Previous methods have addressed high error rates through alternative approaches: SequenceGenie, for instance, leverages strand bias and Bayesian statistics, parSEQ achieves high accuracy by utilizing a minimum of 300 reads to generate a consensus sequence. Details for different plasmid library sequencing methods can be found in Supplementary Information section Alternative Methods and Recommendations.

To test the experimental variability of the LevSeq pipeline we compare LevSeq with the gold-standard Sanger sequencing (Laragen) by running both protocols on a 96-well plate. Laragen samples were individually miniprepped, with 79 of 94 (two empty wells) sequenced with high quality. The 96-well plate was also sequenced with LevSeq using two barcode plates to test for variance between barcodes. Barcode plate 26 sequenced all 94 samples, while barcode plate 25 sequenced 90, with four wells recording low coverage (< 20 reads); results of successfully sequenced variants matched across all three methods. Details for this sequencing experiment can be found in Appendix.

The second component of the software suite is a website that processes variant files for future reference and downstream ML. Users upload variant information from LevSeq along with associated fitness data. Top variants are returned to the user along with a visualization of the coupled sequence function data. This process enables the collection and consolidation of standardized data along with the associated screening conditions (e.g. chemical reaction,

stability, etc.). The deployment of this software is a primary differentiator between LevSeq and other nanopore variant calling methods, parSEQ and SequenceGenie. While SequenceGenie is also suitable for an individual lab, it requires users to build the docker image and has limited documentation, making it a challenge for the standard bench scientist to implement. ParSEQ is an extensive software suite and utilizes comprehensive cloud computing, making it ideal for larger scale operations that can leverage cloud infrastructure.

Given the similarities in the LevSeq, ParSEQ, and SequenceGenie experimental protocols, LevSeq's sequence function pairing, variant calling, and downstream software can be used with previously published protocols. While base-called fastq files from any of these methods can be directly supplied to LevSeq for variant calling and processing, we cannot guarantee demultiplexing accuracy with barcode designs that differ from LevSeq's fixed 24-mer DNA barcodes. Therefore, we recommend users with different experimental setups perform their own demultiplexing and provide the demultiplexed files in LevSeq's format for variant calling and visualization. LevSeq followed the evSeq approach and was designed to be easily deployable with minimal installation and a single command to run analyses and output data in an interoperable format [6,8,116,117]. Our hope is that it will become a useful resource for protein engineers who seek to create data-driven models of protein sequence-function landscapes.

Use case 1: Analyze random mutagenesis libraries and inform next steps in DE

To demonstrate the utility of LevSeq in a random mutagenesis experiment, we constructed and screened ten 96-well plates from an error-prone PCR library of Pyrobaculum arsenaticum protoglobin (ParPgb) LQ variants [118]. ParPgb is isolated from thermostable archaea and expresses well in Escherichia coli. Over the past decade, our laboratory has shown that protoglobins exhibit remarkable tolerance to mutations that alter their catalytic capabilities [69,74,119,120]. Three of the ten plates exhibited unusually low sequencing coverage, contributing to a large number of samples with insufficient coverage (Appendix B Figure B3). The insufficient coverage in this case resulted from improper PCR amplification. This can be mitigated by adjusting the PCR setup method (Appendix Protocols) and was not

observed in subsequent sequencing experiments (Appendix B Figures). The final dataset for ParPgb included 211 sequences with zero amino acid mutations and 539 sequences with up to five amino acid mutations from the parent. Single amino acid mutations were most prevalent, occurring in 210 out of 539 sequences (Figure 3-3B and 3-3C). The mutation distribution aligned with the expected outcome of the mutagenesis method. Following sequencing, we utilized the LevSeq toolkit to generate sequence-function data.

All ten plates were screened for activity and specificity for catalyzing the formation of the cis and trans cyclopropanation products of 4-methoxystyrene (Figure 3-3A). The software automatically returns the top-performing variants for each recorded fitness value, which in this case were top variants for both the cis and trans cyclopropanation products (see Methods for details on selection criteria). We found a single mutation conferred the highest activity for each desired stereoisomeric outcome: F70L for cis preference and F89S for trans (Figure 3D). Sequencing every variant also revealed sites with epistatic interactions. For example, the single mutation D72G improved the formation of both cis and trans products 1.5-fold, and the F89L mutation improved trans product formation nearly 3-fold. However, combining D72G and F89L resulted in activity similar to the parent, indicating higher-order interactions [28,121] between sites 72 and 89, which could be further investigated with a double-site saturation mutagenesis experiment. Detailed results for all individually tested variants are presented in Appendix Figure B5.

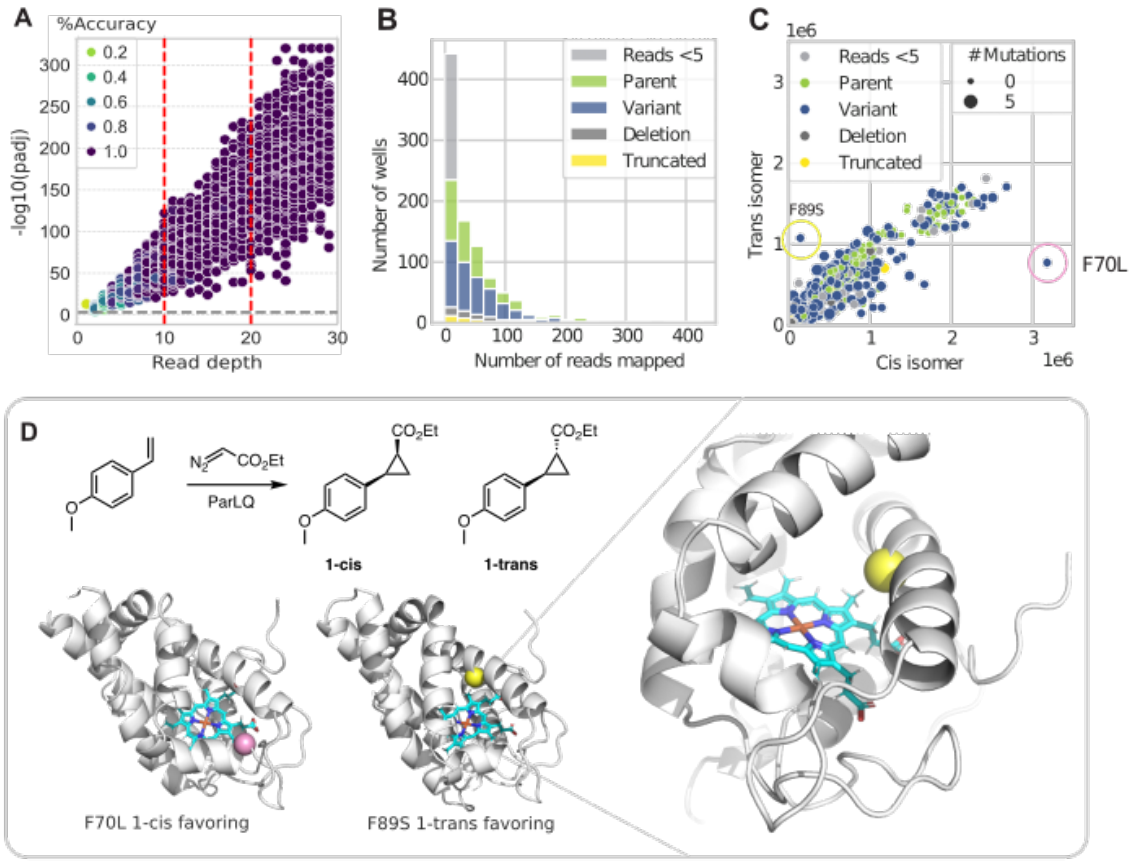


Figure 3-3. Sequence-function analysis and insights from random mutagenesis libraries of ParPgb LQ. A. Accuracy of detecting variants using simulation studies on ParPgb LQ, varying read depth from 1 to 20 using an epPCR mutation rate of 2% and a nanopore sequencing error rate of 10%. B. Sequencing coverage across 10 plates of ParPgb LQ epPCR variants. C. Ion counts clustered by reads across all ten plates from the epPCR experiment with outlier sequences highlighted. D. Enzyme-catalyzed reaction screened in the epPCR experiments is cyclopropanation of 4-methoxystyrene, leading to cis and trans cyclopropane products. The positions mutated for highlighted variants in C are colored on an AlphaFold3 structure of Pgb.

With a MinION flow cell, LevSeq can generate reads for a theoretical value of 2,500 96-well plates in a single flow cell, assuming 1,000 base pair lengths, as noted in previous nanopore sequencing methods. Sequencing bias, increased sequencing length, and low-quality samples will reduce the number of useful sequences obtained. However, the reusability of the flow cell makes LevSeq a more economical option compared to Sanger and short-read next-

generation sequencing. LevSeq is specifically designed to develop sequence-function datasets for research labs and as such the software has been designed for ease of installation and speed. A limitation of this is that the increased demultiplexing speed results in fewer reads assigned to each well compared with other methods [97,104]. When fewer than 20 reads are assigned to a well, a warning is issued to inform the user of the low read counts. For experiments that require a high degree of accuracy we recommend validating these wells using a second method such as Sanger. For LevSeq, we opted for this trade-off, which is beneficial when running experiments on a per lab basis where real-time data analysis is important for guiding the next step of an experiment. For industrial scale protein engineering procedures or in a sequencing facility, ParSEQ may be a more suitable pipeline.

Use case 2: Collecting sequence-function data to optimize variants using ML

To use in silico tools for protein optimization, one ideally starts with sequence-function data for a small subset of the studied system and these datasets serve as a foundation from which to make predictions and guide the engineering process. As one use case, LevSeq was used for active learning-assisted directed evolution (ALDE) by Yang et al. to sequence and analyze four 96-well plates of ParPgb LQ variants from a 5-site combinatorial library. From the four plates, 216 unique variants without stop codons were identified and screened for activity and specificity for catalyzing the formation of the cis and trans cyclopropanation products of 4-methoxystyrene, the same reaction as in case 1. The sequence data from LevSeq and corresponding labels were used as initial training data for a batch Bayesian optimization algorithm, forming the baseline distribution to capture the effect on function of different amino acids at specific residues. This model was then used to suggest 96 sequences for testing; the researchers ordered exact genes for the 96 designed variants. Through three active learning loops the yield of non-native cyclopropanation reaction increased from 12% to 99%, with a 14:1 cis:trans selectivity ratio. LevSeq enabled a critical step of collecting sequence-fitness datasets for model training in ML-assisted workflows. This foundation enhanced the effectiveness of subsequent rounds of ML guided directed evolution, leading to more successful outcomes.

In addition to collecting data for active learning, LevSeq can be used for experimental validation of suggested variants from various MLPE tools such as focused training MLDE (ftMLDE) , cluster learning-assisted directed evolution (CLADE) [122], and degenerate codon optimization for informed libraries (DeCOIL) [123].

LevSeq serves three primary functions for protein engineering: optimizing directed evolution workflows, gathering sequence information for specific MLPE projects, and consolidating sequence-function data for training future generalizable MLPE models. During exploration of the vast sequence-function space, experimentalists often encounter library bias; bias can lead to time and resources wasted on evaluating low-quality libraries. By providing sequence information prior to screening, LevSeq ensures that the gathered data are useful, regardless of whether the fitness results are positive, negative, or neutral.

In addition to its role in optimizing directed evolution workflows, the LevSeq pipeline can be used to generate high-quality datasets for MLPE. The cost-effective and efficient sequencing of variants from random mutagenesis studies using LevSeq helps overcome the bottleneck of limited data. Moreover, the scalability of LevSeq allows for the generation of datasets from a wide range of mutagenesis experiments, further expanding the scope of MLPE applications and facilitating advancements in protein engineering. By creating diverse and representative datasets that capture relevant sequence-function relationships, LevSeq can enable more robust and accurate models to be trained, ultimately leading to improved protein engineering and design.

3.4 Conclusions and outlook

LevSeq makes a routine but historically under-instrumented step in directed evolution—full-variant genotyping—fast, scalable, and operationally useful. By combining dual barcoding with nanopore long-read sequencing and an analysis workflow designed around plate-based experiments, LevSeq enables researchers to attach assay-derived fitness measurements to full-length gene identities at the same throughput as screening. This pairing turns directed evolution outputs from isolated plate results into sequence-resolved records that can be

compared across rounds, aggregated across projects, and retained as reusable sequence–function data.

A central contribution of LevSeq is that it treats sequencing as a quality-control tool rather than a post hoc validation step. By reporting coverage, warning on ambiguous or mixed wells, and surfacing library pathologies before resource-intensive assays, the workflow reduces wasted screening effort and increases confidence that measured labels correspond to the intended genotypes. Just as importantly, LevSeq preserves negative and neutral outcomes as first-class data: when variant identities are known, “non-hits” become informative constraints on the landscape rather than discarded experimental noise.

In the broader thesis arc, this chapter supplies the genotyping and provenance needed to make assay-defined labels learnable. The next step is to ensure that the resulting sequence–function pairs remain interpretable outside the originating experiment by standardizing what is recorded about conditions, measurements, and metadata, and by storing the data in formats that support retrieval and reuse. Chapter 4 therefore focuses on consolidating these records into database-ready, interoperable representations that enable cross-campaign analyses and provide a foundation for generalizable machine learning in protein engineering.

Chapter 4

ENZYME ENGINEERING DATABASE (ENZENGDB): A PLATFORM FOR SHARING AND INTERPRETING SEQUENCE–FUNCTION RELATIONSHIPS ACROSS PROTEIN ENGINEERING CAMPAIGNS.

This chapter incorporates material from the following publication, for which I am the first author: **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 Res. 2026, 54 (D1), D564–D571.

Abstract

The discovery and engineering of new enzymes is important across the bioeconomy, with diverse applications from foods to pharmaceuticals, sensors to agriculture. However, enzyme engineering, in particular machine learning-guided engineering, is hampered by a lack of data. Currently there exists no database designed to capture and interpret datasets created in this domain, nor are there easy analysis and visualisation tools. We developed the Enzyme Engineering Database to provide a centralized resource and an online analysis tool to consolidate sequence-function data from enzyme engineering campaigns, thereby making three contributions: (i) a database into which researchers can deposit public data, (ii) visualisation and analysis tools for protein engineers to analyse their own data or compare enzyme variants to other engineering campaigns, and (iii) a gold-standard dataset for benchmarking automated extraction along with the first large language model extraction pipeline specific for enzyme engineering campaigns. The Enzyme Engineering Database is accessible at http://enzengdb.org/.

4.1 Chapter Overview

This chapter establishes the data standardization and consolidation layer required to turn sequence–function measurements into an asset that can be reused beyond the experiment that generated them. Chapter 3 focused on making genotype assignment routine at screening scale; however, once full-length sequences can be collected, the next bottleneck is organizational rather than experimental: sequence–function data remain difficult to store, query, compare, and interpret across campaigns because they are scattered across spreadsheets, scripts, supplementary files, and inconsistent reporting conventions.

To address this, the Enzyme Engineering Database (EnzEngDB) is presented as both a deposition resource and an analysis environment purpose-built for protein engineering campaigns. Each record links (i) a full-length enzyme sequence, (ii) a reaction encoded as canonical SMILES, and (iii) quantitative performance values such as yield, turnover number (TTN), or selectivity. In addition to supporting deposition, EnzEngDB provides interactive visualization and querying tools that make campaign data explorable as sequence–function landscapes.

In the thesis arc, this chapter provides the step that follows LevSeq: sequence–function pairs become standardized records with consistent metadata and chemistry context, enabling aggregation and comparison across projects and time.

4.2 Motivation and framing

Machine learning-guided protein engineering is often limited by the availability of datasets that are both experimentally grounded and structured for reuse. Even when labs generate information-rich measurements during directed evolution and screening, results are frequently difficult to aggregate because full variant sequences are not routinely recorded at scale, and because datasets are rarely published in a format that supports systematic comparison across studies.

Existing public resources only partially address this gap. Major enzyme databases primarily center natural activities and EC-number organization, which leaves many engineered enzymes and new-to-nature transformations underrepresented. At the same time, curated resources tend to skew toward positive results, whereas engineering decisions and learning systems benefit from negative and neutral measurements as well as positive hits. General deposition portals can accept mutational datasets, but they are not designed as enzyme-engineering-specific workspaces with the chemistry context and interactive tools needed to interpret screening data as a landscape.

EnzEngDB is motivated by a practical requirement: reusable sequence–function data need a shared minimal representation that connects sequence, transformation, and quantitative performance in one place, while remaining compatible with the metrics and workflows used in enzyme engineering. The platform therefore couples a standardized schema and upload checks with interactive analysis and visualization tools that lower the barrier to adoption and make deposition useful within ongoing campaigns.

Finally, because the literature contains large volumes of enzyme engineering results in non-standard formats, the chapter frames database growth as a dual strategy: a manually curated benchmark and an assisted extraction workflow that accelerates candidate capture while preserving expert validation.

4.3 Published work: Enzyme Engineering Database (EnzEngDB)

4.3.1 Introduction

Enzymes are engineered using cycles of mutagenesis, or rational substitution, and selection or screening to enhance specific properties such as activity and selectivity [18,124,125]. Such engineering has enabled enzymes to catalyse non-natural reactions, such as carbene and nitrene transfers or reductive amination of non-native substrates, and to improve cold stability for laundry processing [23,71,126]. Enzyme engineering's numerous techniques include directed evolution (DE) campaigns, targeted mutation studies, and machine learning (ML)-

guided designs. Large, information-rich datasets emerge from the labour- and resource-intensive engineering workflows, and yet most data remain uninterpreted, unused, or unpublished due to the high cost of sequencing and lack of standardised analysis [91,127]. While recent high-throughput workflows such as LevSeq, ParSeq, SequenceGenie, and others [97,99,104,128] have enabled the capture of plate-scale sequence-function maps during screening, these datasets stay siloed in individual labs, and researchers routinely juggle multiple scripts and spreadsheets to analyse one round of experiments before planning the next [129].

Current public databases partly address the need for an enzyme engineering data resource, but important gaps remain. Public resources such as UniProtKB [116], BRENDA [7], Rhea [117], and KEGG consolidate enzyme information on primarily natural activities classified by enzyme commission (EC) numbers. As a result, many engineered reactions and variants, such as de novo designed enzymes or repurposed enzymes that catalyse new-to-nature chemical reactions, are missing [130–133]. These databases are also highly curated and therefore contain mostly positive data, whereas ML models learn best from both positive and negative examples. ProtaBank [6] helps bridge this gap by accepting any mutational dataset; however, it is designed as a general-purpose deposition portal rather than an enzyme-specific resource with analysis features. Beyond experimental data, the scientific literature constitutes a rich yet non-standardised source of enzyme engineering data. Large language model (LLMs) and agentic systems offer practical routes to automated and standardised data collection from literature [134]. Recent pipelines such as Enzyme Co-Scientist [135] and EnzyExtract [136] have parsed thousands of preprint papers to extract kinetic constants (kcat, KM, kcat/KM), assembling >90 ,000 curated records with precision scores ranging from 0.8 to 0.9. However, these pipelines focus solely on kinetic constants and overlook sequence-function maps, negative variants, and activity metrics such as yield, turnover number (TTN), and selectivity, all of which are central metrics in enzyme engineering campaigns [137–139]. This underscores the need for a unified platform linking enzyme sequence, chemical transformation, and quantitative performance data, including both positive and negative results to accelerate discovery and improve reproducibility.

To address this gap, we developed the Enzyme Engineering Database (EnzEngDB), an open-source database coupled with an interactive web analysis tool. EnzEngDB is designed specifically for enzymes where both the catalyst and the chemical transformation are well defined: each record contains a canonical SMILES representation of the reaction (substrates to products), the full-length enzyme sequence, and quantitative performance metrics such as yield, turnover number (TTN), or selectivity. In its current release, EnzEngDB includes datasets spanning directed evolution campaigns, targeted mutation studies, and ML-guided designs, with examples from both natural and new-to-nature reactions. Our primary focus remains on transformations not found in nature, reflecting both the unique expertise of the founding laboratory and the scarcity of such data in existing resources, but the database structure is intentionally broad and designed to accommodate enzyme engineering datasets of all types. This design choice ensures compatibility with a wide range of present and future applications while giving immediate priority to the most information-rich, underrepresented datasets.

4.3.2 Materials and methods

Manual data collection and cleaning

For our gold standard dataset, we selected peer-reviewed enzyme engineering studies from our laboratory from the past decade and have included four additional combinatorial or model-guided mutation datasets that cover diverse enzyme engineering strategies. Working with these publications lets curators locate details quickly and cross-check against lab records, giving the most reliable baseline for benchmarking. The curators were three postdoctoral researchers, five PhD candidates, and two masters students. Curators independently extracted four fields from each paper: (i) enzyme/variant identifier, (ii) full-length amino-acid sequence, (iii) model reaction encoded as canonical SMILES (substrates to products), and (iv) performance values, Appendix. Notably, curators struggled to locate consistent sequence annotations and performance values because the data were scattered across the main text, figures, and supplementary files. Conflicts were resolved by a third curator, who consulted the original figures and deposited the consensus record. These

manually extracted data ensured consistency between DNA and protein, parents and enzyme variants, and the validity of the SMILES using RDKit.

Automated literature mining and data standardisation

To scale curation beyond our manually compiled gold standard dataset, we built a four-step, LLM-based extraction pipeline, validated it on the gold standard papers, and then applied it to a wider literature set. We conducted a comprehensive PubMed search for papers related to enzyme engineering for unnatural chemistry to capture data that is not covered by existing resources, APPENDIX C2. This query retrieved 912 papers with DOIs. Abstracts were scored for relevance with the Gemini 2.5 Flash model, and those above 0.90 confidence yielded 327 candidate manuscripts, including 32 of our 34 gold-standard references; see APPENDIX C2 for download and scoring details. Full-text PDFs and supplements were downloaded from PubMed Central when possible. The pipeline identified engineered variants and their full-length sequences, sanitised the sequences, encoded each reported reaction as a canonical SMILES string, and pulled quantitative metrics such as yield or TTN into schema-compliant CSV and JSON files; see APPENDIX C3 for detailed information on the extraction method. Complete reaction, sequence, and performance records were recovered for 133 out of 327 papers, which were then manually validated; the rest lacked yield data or used chemical names the pipeline could not resolve. The same curators who prepared the gold standard set reviewed every output and ensured that the recorded sequences and reactions matched the source manuscripts (Figure 4-1).

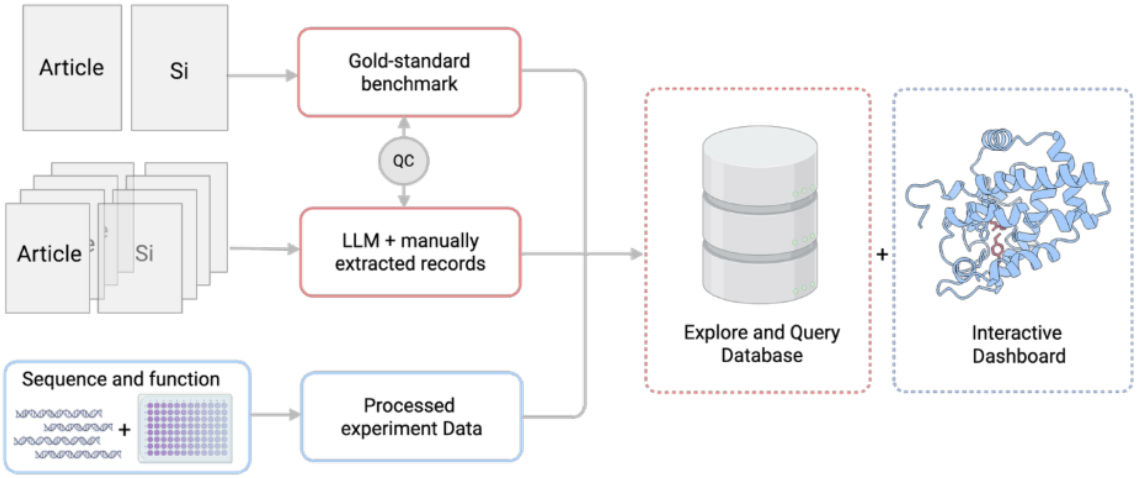


Figure 4-1. Integrated workflow for EnzEngDB population and visualization. Literature PDFs are processed via two routes (i) manual curation to create a gold-standard benchmark and (ii) an LLM-based extraction pipeline that generates auto-extracted records. Both outputs, together with CSV uploads from plate-scale experiments, converge into a unified, schema-compliant CSV collection. These curated records feed an interactive front-end that links sequence tables, fitness heat maps and 3-D structural views.

Development of Enzyme Engineering Database (EnzEngDB)

EnzEngDB was implemented in Python with the Plotly Dash web framework [https://plotly.com/dash] and packaged in a Docker container. All curated records and user-uploaded datasets are stored as CSV and .CIF files and are available for direct download on the website. Protein structures are displayed with Mol* through the Dash-Molstar wrapper [https://github.com/everburstSun/dash-molstar], and two-dimensional depictions of substrates and products are generated on the back end with RDKit. Sequence handling and alignment use BioPython (Appendix).

4.3.3 Results

Datasets consolidated in Enzyme Engineering Database (EnzEngDB)

The present EnzEngDB release already ingests plate-scale CSV formatted datasets produced by the LevSeq screening workflow. Additionally, any CSV file that conforms to the standard

column headers specified in APPENDIX C1 can be uploaded to the database. A dedicated Upload tab in the database and an accompanying Information tab perform automatic checks for required columns before ingestion, ensuring standardisation and lowering the barrier to community participation. To demonstrate the platform's capacity to store and analyse experimental results, we have deposited an error-prone PCR random mutagenesis dataset and a five-site combinatorial mutagenesis dataset, together representing sequence-function measurements equivalent to roughly eight 96-well plates of variants, each measured for two products.

We also manually curated a gold-standard dataset comprising 6,234 sequence-function entries across 40 papers. The 635 reactions include carbene, nitrene, and condensation reactions; and the 1,846 unique protein variants include heme proteins and tryptophan synthase (Figure 4-2A) These variants trace back to six protein scaffolds and span a wide range of non-natural substrates and transformations (Figure 4-2 A). Although their sequences show high similarity to entries in the PDB (31) and SwissProt, the chemistries they perform are distinct (Figure 4-2 B). Reaction space comparison, quantified with the Tanimoto similarity using RDKit structural fingerprints indicates that most transformations in EnzEngDB are different to those in EnzymeMAP, confirming that EnzEngDB captures largely unique biocatalytic reactions (Figure 4-2 C). To complement the directed evolution studies, we include data from other enzyme engineering approaches such as combinatorial studies and ML-guided designs, including three non-natural substrate datasets (included above) and four native reaction datasets (447,759 entries) (Appendix C).

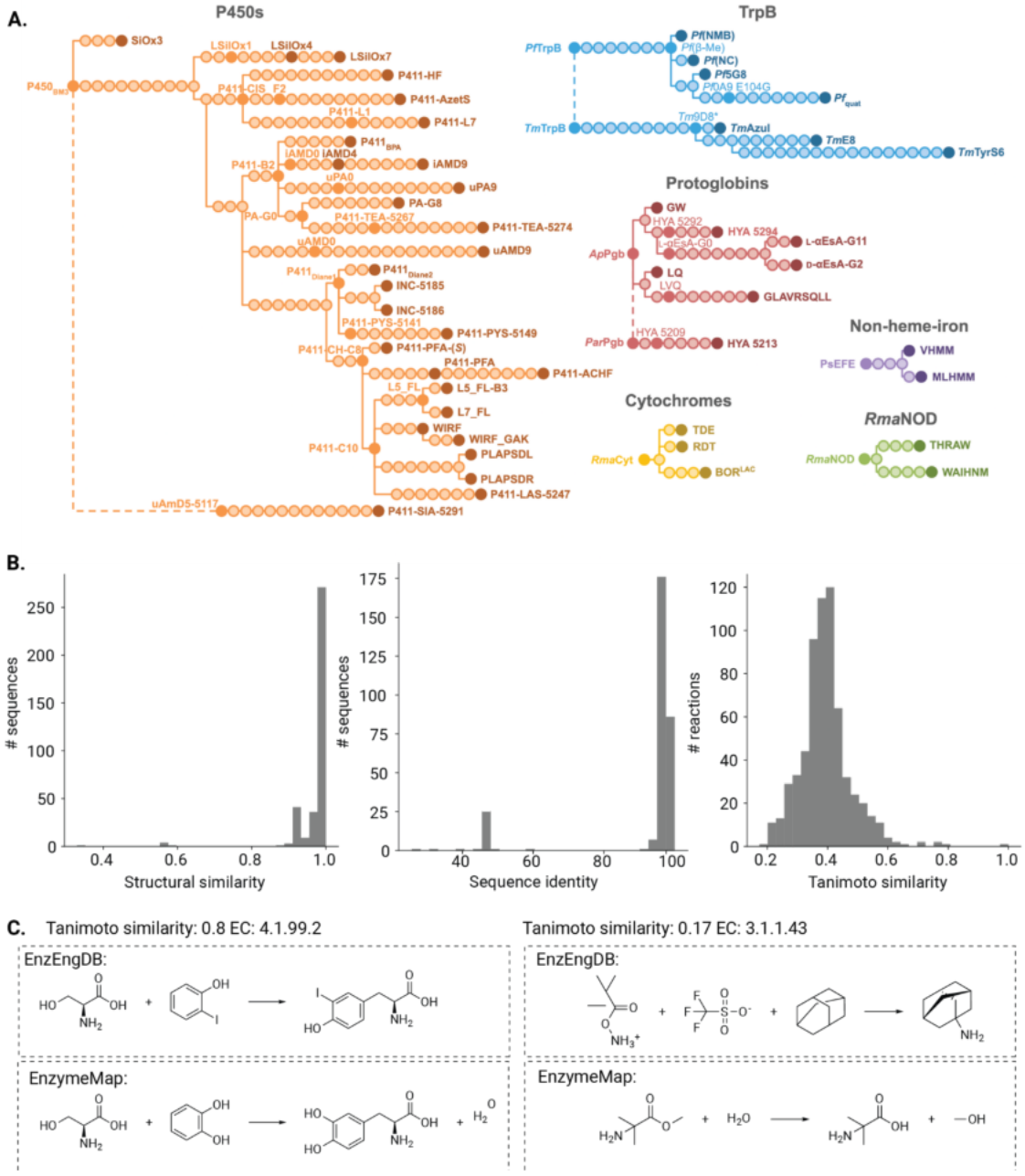


Figure 4-2. Data composition of the database. A. Schematic representation of the six enzyme families across the 366 enzymes from directed evolution campaigns captured in the EnzEngDB manual gold-standard dataset (See Appendix C). B. The similarities between the gold standard section of the EnzEngDB and the publicly available databases are shown: structural similarity was compared to PDB, sequence identity to SwissProt, and the reactions were compared to unmapped reactions in EnzymeMap. C. The most similar (that was not an exact match) and least similar reactions are shown alongside their

matching reaction and EC classes in EnzymeMap to show the difference in the reactions between the EnzEngDB manual gold-standard dataset and EnzymeMap.

As the manual extraction of an enzyme sequence and its corresponding function is time consuming, taking each curator approximately an hour per paper, we developed an automated extraction pipeline to expand our dataset. To evaluate the automated extraction pipeline, we reprocessed the PDFs and supplementary files for every gold-standard paper, omitting those without sequence information. The extracted parent entries for each campaign were mapped to the reference set under following criteria: sequences could differ by terminal tag truncation, and reactions were deemed equivalent when the Tanimoto similarity exceeded 0.80 (Appendix C). Under these criteria, 30% of the LLM-extractions passed (Appendix). Papers from the automated extraction (133 downloaded papers) were manually validated and 64% required manual intervention for reasons such as no sequence or standard chemical name provided (Appendix C). We found that the LLM extraction pipeline reduced the time to curate data, requiring correcting rather than searching the SI and PDF for information. Rerunning sequence extraction multiple times helps correct sequence-extraction errors, and providing full IUPAC names in the text reduces reaction-extraction errors. Common reaction extraction errors arise from missing stereochemistry information or a manual curator mistakenly extracting the wrong chemistry (Appendix C). Moreover, our attempts to extract chemical structures directly from figures revealed that current LLMs and software packages still struggle to capture structural information directly. Common LLM sequence extraction errors were character issues (e.g. M as W) however, we also identified cases where an author deposited the incorrect sequence in the SI or referenced an incorrect PDB ID. The effort required to reconcile these details shows how easily critical information can be missed or misinterpreted and underscores the need for a community-wide, standard reporting format that records variant lineage, reaction context and quantitative metrics at the time of publication. While automation reduced the average hands-on curation time from roughly an hour per paper to less than ten minutes per paper, expert validation remains essential.

The extraction pipeline and manual validation produced an additional 2,415 entries drawn from 102 papers, averaging 50,000 output tokens and approximately fifty cents per paper. These records cover 818 protein variants that catalyze 1,078 reactions.

Interactive experiment dashboard

The EnzEngDB website includes over 10k experimentally collected records. For each enzyme engineering campaign an interactive dashboard collects all essential content in a single view: the parent amino acid sequence, experimental metadata, the canonical SMILES representation of the model reaction and a manipulable protein structure (Figure 4-3). The interactive three-dimensional protein structure is provided alongside a sortable table of variants listing each mutation with the parent-normalized fitness value, and the fold-change ratio. Clicking on a row highlights the mutated residues on the 3D model and auto-zooms to their local environment, letting users investigate structural effects from sequence substitutions that previously required separate scripts or multiple programs (Figure 4-3). The structure is rendered in a 3D viewer that supports zoom and rotation for in-depth analysis. Filters can be used to organize the variant list by any metric, and the protein viewer will remain synchronized with selected row. The integrations of a protein viewer and spreadsheet-style analytics turn raw screening data into an exploratory workspace, reducing the time from data upload to actionable hypotheses.

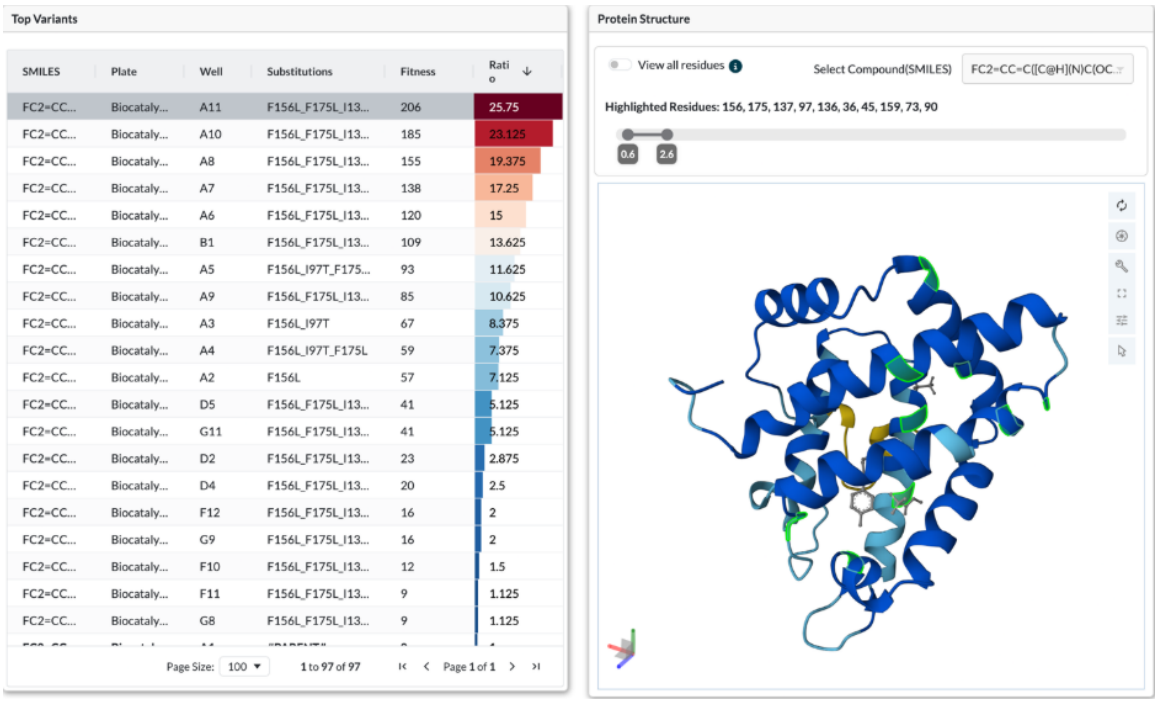


Figure 4-3. Interactive dashboard partial view: selected variant mutations (positions 156, 175, 137, 97, 136, 36, 45, 159, 73, 90) highlighted on the 3D protein structure.

Database browsing and queries

All campaigns in EnzEngDB can be queried and compared using the web interface. While the visualization dashboard is optimized for directed evolution datasets, where substitutions are highlighted on the protein structure and variant fitness can be compared to a parent, the database can also accommodate other enzyme engineering campaigns, provided each entry includes a reaction SMILES and a quantitative product measurement. From the dashboard, users can run a sequence similarity search to find enzymes with different functions or to identify mutations that improve function in other campaigns, with results filterable by experiment name, percentage sequence identity, reaction SMILES, or assay type. For every dataset with a specified parent, variant fitness is automatically reported as a fold-change relative to that reference, while the others are reported as the extracted values. These normalized values reveal gain- and loss-of-function mutations and can guide the choice of target sites for the next library. For instance, a similarity search on the protoglobin nitrene-

transfer lineage reveals an active-site substitution that enhances selectivity but has yet to be explored in other protoglobin projects, flagging a promising engineering target without relying on informal knowledge transfer (Figure 4-4).

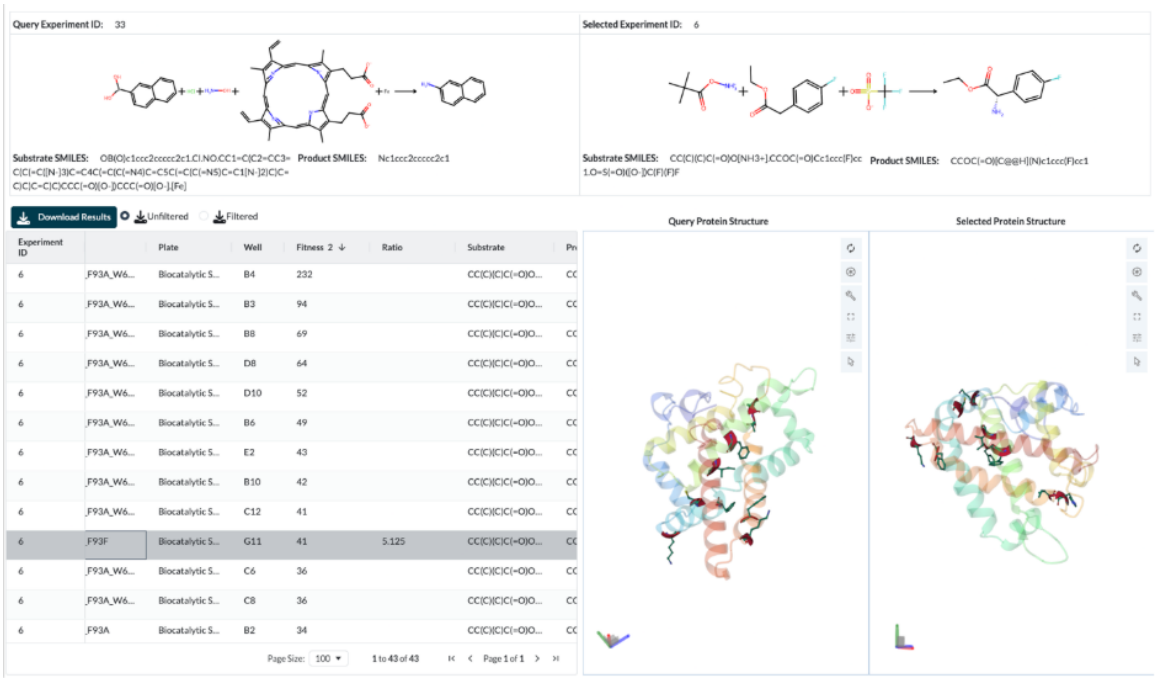


Figure 4-4. An example of a sequence-similarity search within the protoglobin nitrene-transfer lineage revealing an active-site mutation that enhances selectivity yet remains unexplored in other campaigns.

Private and Public Instance of EnzEngDB

To incentivize participation, the database includes both a private and a public instance. The private instance provides full visualization and formatting functions to help researchers store and analyze their own data locally, while the public instance is intended for deposition once datasets are published. The private instance can be locally served using one line of code deploying a docker container, so the data is stored and standardized for local use, once the owner makes the data public, the public database instance will ingest the standardized screening data. Every dataset in the public instance can include a DOI link, a unique ID and

a downloadable BibTeX file during download to facilitate proper attribution and citation for data usage.

4.3.4 Discussion

We developed the Enzyme Engineering Database (EnzEngDB) as a data deposition hub that incentivizes researchers to share their screening data by providing interactive visualizations, enabling faster insights into gain-of-function mutations, and reducing the analysis workload. In addition to the public repository, EnzEngDB includes a private instance that allows researchers to store, format, and visualize their own data locally prior to publication. Public datasets with assigned DOIs will be accompanied by BibTeX files to ensure appropriate attribution to the original authors.

The framework of EnzEngDB is organized around amino acid sequence, fitness value, and fitness type, allowing a wide range of assay outputs to be captured in a consistent format. Whether the data originate from kinetic measurements (e.g., Km or kcat), spectroscopic readouts, or other experimental assays, the database structure allows data to be stored, filtered, and processed in a consistent manner by representing them as fitness values. While not yet comprehensive, EnzEngDB should be regarded as an ongoing effort, with its scope expected to broaden as additional datasets are contributed and standardized ingestion pipelines are developed. Direct comparisons are recommended only within individual CSVs, where experimental conditions and formats are consistent. This design choice enables flexibility across diverse enzyme engineering campaigns while maintaining internal consistency.

We envision that EnzEngDB will convert siloed datasets into accessible resources, improving reproducibility, data access, and the creation of balanced datasets by including negative results. While these values may differ across campaigns, they are measured under consistent conditions within each dataset and are included in the downloadable CSV. For DE campaigns with a defined lineage, identical reaction conditions allow fold-change calculations, enabling standardized comparisons of mutation effects. We envision that this

data will support the design of machine learning (ML) models that can be used to guide engineering, reducing the experimental burden [106,140,141]. Ultimately, systematic data collection of diverse enzyme engineering campaigns will bridge the gap between resource-intensive experiments and efficient, data-driven engineering methods [142,143].

Alongside the platform we assembled a manually curated set of new-to-nature reactions and used it to benchmark an LLM-based extraction pipeline, establishing the first benchmark for future automated curation. While the present EnzEngDB version shows that enzyme engineering data can be consolidated and searched in one place, coverage remains limited. The curated dataset mostly comes from a single laboratory, and every automatically extracted paper required expert validation, which is evidence that LLM extraction alone cannot yet capture the full detail of directed evolution campaigns and underscores the value of manual curation by domain experts.

Community participation is essential for EnzEngDB to become the comprehensive, standardized resource the field needs. As the field increasingly moves toward machine learning applications, it is essential to ensure that the extensive experimental effort required to generate training data is appropriately acknowledged. Even relatively modest ML tasks, such as fine-tuning, can demand substantial volumes of high-quality measurements. EnzEngDB was designed in part to address this need by providing clear mechanisms for dataset sharing and attribution for experimentalists who upload data.

We encourage researchers to deposit complete datasets in the format compatible with EnzEngDB, including both positive and negative data, at the point of screening and publication, and we invite developers to test and improve extraction scripts against the benchmark set. EnzEngDB is currently compatible with LevSeq. However, we make our code open source and encourage community developed integrations with other sequencing methods via GitHub (https://github.com/ssec-jhu/levseq-dash). These contributions will turn EnzEngDB into the resource that can advance the communities understanding of biocatalysis and provide new methods in data-driven enzyme engineering and design.

4.4 Conclusion and outlook

EnzEngDB addresses a practical but foundational barrier to functional prediction in protein engineering: even when laboratories generate quantitative sequence–function measurements, the results are rarely captured in a form that is portable across projects, comparable across studies, and reusable for modeling. By organizing records around three linked primitives—full-length enzyme sequence, chemically explicit reaction representation, and quantitative performance metrics—EnzEngDB provides a consistent structure for storing and interpreting enzyme engineering outcomes as a sequence–function landscape rather than as isolated experimental anecdotes.

A key contribution of the platform is that it bridges the gap between what is measured at the bench and what can be learned from data. Directed evolution produces labels under diverse conditions, assay formats, and mutational strategies; EnzEngDB makes these results findable and interpretable by pairing measurements with the minimal contextual metadata needed to support comparison and aggregation. In doing so, it elevates outcomes that are often lost—particularly negative and neutral measurements—into durable constraints on the landscape, improving the fidelity of downstream inference and reducing redundant experimentation.

This chapter also clarifies what it will take for the community to scale sequence–function knowledge. Literature data are abundant but fragmented across text, figures, and supplementary files, and consistent extraction is limited by heterogeneous reporting conventions. EnzEngDB therefore motivates a dual approach: curated benchmarks that establish high-confidence ground truth, paired with assisted extraction workflows that accelerate candidate capture while preserving expert validation. The long-term value of the resource will depend not only on technical infrastructure, but on shared expectations for how engineered enzymes and their measurements are reported and deposited.

Within the thesis arc, EnzEngDB completes the infrastructure needed to make assay-defined labels learnable at scale. Chapter 2 defined function through deployment-aligned assays; Chapter 3 made full-variant genotyping and quality control routine; and Chapter 4 makes the

resulting sequence–function pairs standardized, searchable, and aggregatable. Together, these components convert protein engineering from a sequence of local optimizations into a data-generating process whose outputs can be reused for mechanistic interpretation, cross-campaign generalization, and model development. The next chapter builds on this foundation by using consolidated datasets to define modeling-ready prediction tasks and evaluate what functional prediction can achieve under realistic experimental constraints.

This chapter establishes a standardized substrate for learning: sequence–function records that preserve chemical context, quantitative labels, and provenance in a form that supports aggregation and retrieval. With this foundation in place, the next question is no longer whether we can store data, but whether we can use it to make better choices—in particular, to prioritize de novo designed enzymes for experimental testing under realistic assay constraints. Chapter 5 closes the loop by combining a multimodal generative design model with a lightweight, data-driven scoring head informed by EnzEngDB, then evaluating whether database-derived signals can rank candidates and predict outcomes when designs are synthesized and validated in the laboratory. Together, Chapters 4 and 5 connect data infrastructure to decision-making: from consolidated measurements to testable hypotheses about which sequences are most likely to function.

Chapter 5

FROM DATA FOUNDATIONS TO DECISIONS: RANKING AND VALIDATING DE NOVO DESIGNS.

Parts of the work described in this chapter are ongoing and are being prepared for publication. Results are presented here as progress to date and are subject to refinement as additional experiments are completed.

5.1 Chapter overview

Chapters 2–4 argue that enzyme function becomes learnable only when it is defined by an assay, tied to full genotypes, and recorded with enough context to make measurements comparable. Chapter 5 moves from those data foundations to the decision problem they are meant to support: how to choose what to build and test when the design space is vast and experimental validation is the limiting step. De novo enzyme design makes this tension explicit. Modern generative models can propose many candidate sequences, but only a small fraction can be cloned, expressed, and assayed with careful analytics. In this regime, the practical bottleneck is not generating candidates, but prioritizing them under a specific, deployment-relevant assay definition.

This chapter focuses on de novo designed carbene transferases as a test for decision-oriented prediction. The first part establishes the design and experimental context: what the designs are, how candidates were generated and screened, and how fitness is defined and measured in a consistent whole-cell assay. The chapter then motivates a single, well-labeled reaction setting, 4-methoxystyrene cyclopropanation, as a practical anchor for quantitative comparison across sequences.

The second part presents an analysis performed for this thesis: a lightweight ranking model trained on in-house sequence–function records and evaluated on *de novo* designs to ask a narrow but actionable question, whether predicted fitness correlates with measured fitness when fitness is defined as yield under a fixed assay. Importantly, this model was not used to filter candidates during the design campaign; rather, it serves as a retrospective probe of how far the current data structure supports prioritization. The results clarify what is predictable today within a defined reaction context, what breaks when reaction identity is ignored, and how this kind of prioritization can serve as a component of an active learning loop as reaction-specific data accumulate.

5.2 Motivation: *de novo* validation as a stress test for functional prediction

De novo enzyme design creates a separation between what is easy and what is hard in current protein engineering workflows. Generating candidate sequences has become inexpensive: models can propose hundreds to thousands of putative catalysts for a target transformation in the time it takes to validate a handful experimentally. In contrast, experimental confirmation remains rate-limiting. Each candidate must be built, expressed, screened under defined conditions, and quantified with an analytical readout that distinguishes true catalysis from background and cofactor-only activity. The practical problem is therefore not whether designs can be generated, but how to decide which designs are most worth testing when validation throughput is finite [144–146].

This decision problem makes *de novo* validation a stringent test for functional prediction. Unlike directed evolution, where variants are explored in a local neighborhood of an already functional parent and selection pressure enriches for incremental improvements [147], de novo libraries are broader and less constrained by historical function. Many sequences will be nonfunctional for reasons that are orthogonal to the intended chemistry (poor expression, misfolding, weak cofactor incorporation, unstable active-site geometry), and even among well-behaved proteins, catalytic competence can be highly conditional on assay format and reaction setup. In this setting, apparent predictive success on curated benchmarks is not sufficient; what matters is whether a model can support triage under a specific assay

definition, enriching for candidates that measurably outperform background in the format used for validation.

The motivation of this chapter is thus twofold. First, it grounds the chapter in the design-to-test interface for *de novo* carbene transferases: what the designs represent, how they are experimentally evaluated, and how fitness is defined in a way that is quantitative and comparable across candidates. Second, it uses this validation setting to examine a thesis-relevant question: given the kind of standardized, assay-defined sequence–function records developed in earlier chapters, how far can a lightweight model go toward prioritizing de novo candidates. Although no model was used to filter designs during the work described here, a retrospective ranking analysis provides a controlled way to assess whether predicted fitness aligns with measured fitness (product yield) and to identify the conditions under which prioritization is plausible versus where reaction context and label noise dominate.

In many new-to-nature transformations, prioritization begins in a cold-start regime where there is no established activity and only a small subset of *de novo* candidates can be tested. In that setting, the goal of early experiments is not only to identify hits, but to learn which sequence and assay factors separate background from measurable yield so that subsequent rounds become more targeted. This naturally motivates an iterative design–test–learn strategy, where each screening round updates what is known and informs the next set of candidates. The practical implication for the rest of this chapter is that prediction becomes most useful when it is coupled to complete record-keeping of all sequences and outcomes, including negatives, so that each round improves both the ongoing campaign and the reusable data substrate.

5.3 Design context: what was designed and how candidates were generated

In this chapter, *de novo* designs refer to computationally generated heme-dependent proteins intended to catalyze carbene transfer reactions. Rather than starting from a naturally occurring enzyme or an evolved lineage and exploring a local mutational neighborhood, the candidates considered here originate from a generative design process that produces new

sequences with corresponding structural hypotheses. The design goal is not a generic folded protein, but a scaffold that can support a heme cofactor and present an active-site environment compatible with organometallic-like carbene transfer chemistry.

Candidate sequences were produced using a multimodal design workflow that couples sequence generation with structural constraints. At a high level, the generator is conditioned to produce proteins that are (i) predicted to adopt a stable fold, (ii) capable of binding a heme cofactor in a defined geometry, and (iii) able to accommodate reactive intermediate states relevant to carbene transfer. This conditioning step is important for interpretation: the designs are not drawn uniformly from protein space, but from a constrained subset shaped by the model's priors and by explicit requirements imposed during generation. The output of this process is therefore best viewed as a set of hypotheses: candidate catalysts that satisfy computational filters, rather than variants already known to be catalytically competent.

Between generation and wet-lab validation, candidates pass through practical selection criteria that reflect the realities of experimental throughput. Designs are screened for basic feasibility (e.g., sequence sanity and constructability), and they are typically prioritized using computational confidence measures and structural heuristics that aim to reduce obvious failure modes (sequences with only Alanine) before committing to cloning and screening. These prefilters do not constitute functional prediction in the sense of forecasting catalytic outcomes under assay conditions; instead, they are intended to enrich for designs that are likely to express and fold and to maintain the intended cofactor-binding environment.

A key point for interpreting the validation results is what is held fixed versus what is allowed to vary in this study. Across candidates, the expression host, construct format, and whole-cell screening workflow are held constant, and designs are evaluated with the same analytical readout and background controls. What varies is the protein sequence and its resulting folded environment around the heme, which is the intended locus of design. This separation is deliberate: it allows differences in measured outcomes to be attributed primarily to sequence-dependent effects under a defined reaction context, rather than to changes in protocol. The

next section describes the validation workflow and the fitness label used for both screening and the thesis-only ranking analysis.

5.4 Experimental workflow and fitness definition

De novo candidates were evaluated using a standardized build–test–measure workflow designed to produce comparable, prediction-ready sequence–function records. In brief, each design was (i) synthesized or cloned into a common expression backbone, (ii) expressed in *E.* coli under a fixed induction protocol, (iii) assayed in whole cells using a defined carbene transfer reaction setup, and (iv) quantified by analytical chemistry to measure product formation relative to appropriate controls (See Appendix A). Keeping the validation pipeline consistent is not only a practical choice; it ensures that measured differences reflect a single operational definition of function.

Within this workflow, it is useful to separate two ways activity is handled. Operationally, primary screening often requires a simple rule for calling a design active versus inactive based on whether product formation rises above background, including hemin-only and no-enzyme controls. Such thresholds are valuable for triaging follow-up experiments, but they discard quantitative information once a candidate clears the bar. For the analysis in this chapter, fitness is instead treated as a continuous, assay-defined label.

Accordingly, the fitness label used for the quantitative analyses in this chapter is product yield (See Appendix B). Yield provides a direct quantitative measure aligned with the decision problem: among designs tested in the same format, which sequences most reliably drive substrate to product above background. Importantly, this definition embeds assay context. Yield depends on the whole-cell environment, cofactor availability, expression and folding, and reaction setup. It is therefore not a universal property of the sequence, but a measured outcome under a specified protocol. This framing is consistent with the thesis argument that learnable supervision in enzyme engineering must be grounded in explicit assay definitions and recorded with sufficient context to interpret comparability.

Finally, each measurement was captured as a structured sequence–function record that links the full genotype to the assay-defined label and the reaction context under which it was obtained. This is the connection point to Chapters 2–4: standardized recording of sequences, conditions, and quantitative outcomes makes it possible to later ask decision-oriented questions, such as whether a model trained on in-house records can rank *de novo* candidates by expected yield under the same assay definition, without rebuilding the dataset from scratch. In an early-stage campaigns for new reactions, preserving all tested sequences and outcomes, including near-background results, is what enables iterative improvement of prioritization as the campaign progresses.

5.5 Assay choice and label definition: why 4-methoxystyrene cyclopropanation

The design-to-validation pipeline in this chapter spans multiple potential carbene transfer reactions, but a single, well-defined assay setting is needed to make quantitative comparisons meaningful and to support the thesis-only ranking analysis. I therefore focus on 4-methoxystyrene cyclopropanation as the primary anchor reaction for defining fitness and evaluating prioritization. This choice is pragmatic rather than aesthetic. Among the available assays, this reaction provides the most consistent combination of analytical tractability, background separation, and in-house measurement density, making product formation a reliable, comparable label across many sequences.

Several practical considerations motivate this selection. First, product formation can be quantified robustly and repeatedly across screening runs. Second, the reaction exhibits a usable dynamic range under whole-cell conditions, with a measurable baseline and a spread of outcomes across sequences that supports ranking rather than collapsing most designs into an undifferentiated inactive class. Third, relative to other reaction settings, the laboratory has accumulated the most internal sequence–function coverage for this substrate and assay format, which reduces variance introduced by small sample sizes and enables a cleaner train–test separation for the retrospective model.

Finally, anchoring the chapter around this assay makes it possible to separate two distinct questions that are often conflated. The first is whether *de novo* designs can produce measurable carbene transfer activity in the chosen format, which is addressed by the validation results summarized in the next section. The second is whether a model trained on standardized in-house records can prioritize designs by expected conversion within that same assay definition. By fixing the reaction context and label definition here, the later ranking analysis tests prediction in a way that is interpretable and directly connected to the experimental decision problem.

5.6 Progress to date: validation of de novo carbene transferases

Validation of *de novo* carbene transferase designs has proceeded by testing generated candidates through a consistent whole-cell assay pipeline and quantifying yield as described above. Across campaigns, the primary outcome is a distribution of binary classification of active versus non-active as well as yields. Most designs cluster near background, a smaller fraction produce measurable product, and a small subset achieve substantially higher yield. This long-tailed outcome distribution is typical for *de novo* validation and reflects the fact that successful catalysis requires multiple requirements to be satisfied simultaneously [148,149], including expression and folding, cofactor compatibility, and an active-site configuration that supports productive carbene transfer under the assay conditions.

A key qualitative observation from progress to date is that apparent hit rates and yield distributions vary substantially across reaction settings, even when the experimental format is held constant (Figure 5-1). Some reactions yield a higher fraction of designs with measurable yield, while others show near-background outcomes for most candidates. This variability reinforces the central thesis theme that function is assay-defined and reaction-specific: sequence effects are evaluated relative to a particular substrate, carbene precursor, and whole-cell context, and performance measured in one setting does not automatically translate to another. Practically, this also means that mixing outcomes across reaction contexts without explicitly accounting for reaction identity can obscure signal and inflate label noise. For the remainder of the chapter, the 4-methoxystyrene cyclopropanation assay

provides quantitative comparison across many sequences. In this setting, the validated designs span a broad dynamic range of yield, enabling both a meaningful summary of experimental progress and a decision-oriented evaluation of prioritization.

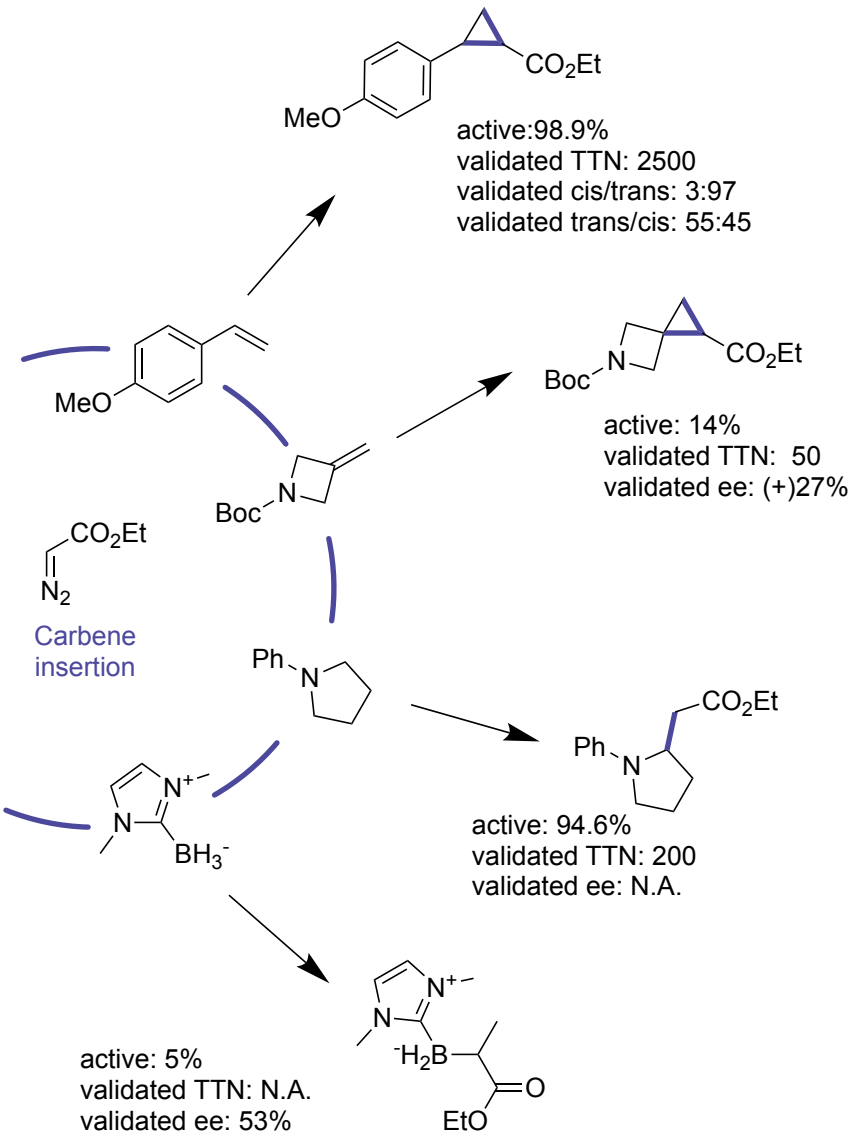


Figure 5-1. Different chemical reactions have different hit rate. Performance distributions differ across transformations, emphasizing that prediction must be reaction-scoped or explicitly context-conditioned.

Finally, these results motivate the specific computational question examined next. If standardized in-house sequence–function records contain learnable signal within a fixed assay definition, then a lightweight ranking model should preferentially score designs that later exhibit higher measured yield in the same context. The next section tests this idea directly by evaluating the correlation between predicted fitness and experimental yield for *de novo* designs, and by quantifying the extent to which explicit reaction context improves prioritization.

These progress-to-date results also highlight that *de novo* design and directed evolution are complementary rather than competing approaches. *De novo* design expands the set of starting points by proposing candidates in regions of sequence space that may be difficult to reach through incremental mutation of existing scaffolds. At the same time, the screening outcomes underscore that *de novo* candidates frequently require subsequent optimization. Once measurable activity is identified in a given assay context, iterative evolution remains essential for improving turnover, shaping selectivity, and increasing robustness under deployment-relevant conditions [145,149–151]. Early rounds are therefore valuable not only for finding hits, but for learning which design features and conditions most strongly separate background from measurable activity, informing subsequent rounds.

5.7 Thesis-only analysis: reaction-aware ranking and low-data adaptation for de novo designs

For *de novo* designed proteins, the practical question is which candidates should be built and tested first when validation throughput is limited. This section presents a thesis-only retrospective analysis that asks whether in-house sequence–function records contain enough learnable signal to support prioritization of newly generated designs within a fixed assay context. The analysis is anchored to 4-methoxystyrene cyclopropanation, where yield was recorded for essentially all tested designs, enabling rank-based evaluation against a continuous experimental label rather than a binary hit call.

I trained lightweight ranking models using a combined dataset of 9,600 sequence–function records assembled from internal campaigns and published work. Each record consists of a full-length amino-acid sequence, a reaction SMILES string encoding the transformation, and an assay-defined quantitative label (yield) with associated metadata (format, controls, and measurement method). Protein structure features were derived from predicted models generated using both AlphaFold and Chai, providing a consistent structural hypothesis for each sequence in the training set. Protein features consisted of ESM-2 sequence embeddings [152] together with a compact set of structure-derived features from AlphaFold 3 predictions [153]. Specifically, these included the predicted confidence score pLDDT, an estimate of active-site pocket volume, and the fraction of polar versus hydrophobic residues lining the predicted pocket surface. Models were trained with a pairwise ranking loss to learn relative activity ordering within an assay-defined reaction context [154,155].

After training, each model assigns a scalar score to *de novo* candidates, producing an ordered list intended to enrich for higher-yield designs under the same screening format. Predictive utility was evaluated by comparing predicted ranks to experimentally measured ranks derived from yield using Spearman correlation ρ, which directly tests ordering quality rather than absolute calibration (Figure 5-2) [156].

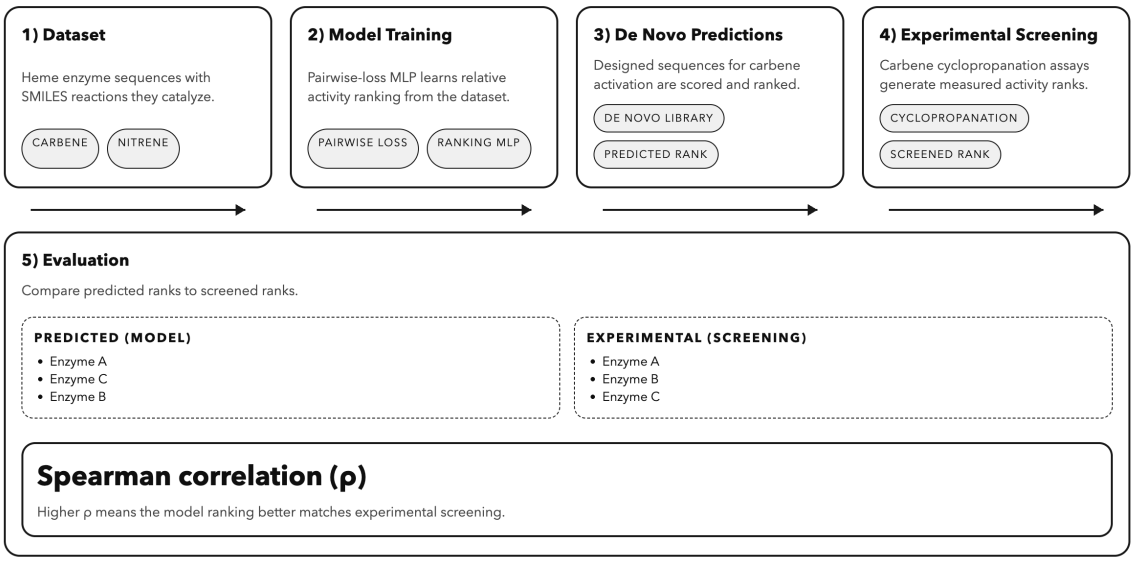


Figure 5-2. Workflow for lightweight prioritization and rank-based evaluation of de novo designs. In-house sequence–function records for heme enzymes are used to train a pairwise ranking model that learns relative activity order within an assay-defined reaction context. The trained model scores and ranks a de novo design library for a specified carbene-transfer reaction, candidates are tested by whole-cell screening, and predictive utility is quantified by comparing predicted ranks to experimentally measured ranks using Spearman correlation (ρ).

To quantify the role of chemistry context and the practical value of small amounts of target-reaction data, I compared three prioritization settings. The first model uses only protein sequence and structure features. The second model augments these protein features with an explicit representation of the reaction, encoded from reaction SMILES as Morgan fingerprints, so that reaction identity is provided as a structured input rather than an implicit label. The third model uses the same protein and reaction features as the second model but is further fine-tuned on a small labeled subset of de novo designs. Specifically, from a 96-member de novo library, 24 designs were used for fine-tuning and the remaining 72 designs were held out for evaluation. Fine-tuning and evaluation were performed within the same assay-defined reaction setting, mirroring the early stage of a new reaction campaign in which only a limited initial batch can be experimentally characterized.

This analysis is not part of the planned manuscript claims and was not used to filter candidates during the design campaign. It is included here because it provides a concrete demonstration of the thesis argument: prediction becomes decision-relevant when the data are structured as assay-defined, comparable labels linked to full genotypes and explicit.

The central observation is that prioritization improves as reaction context and limited target-reaction data are incorporated. As shown in Figure 5-3, the protein-only model shows minimal agreement with measured ranks (Spearman $\rho = 0.045$). Adding reaction fingerprints improves agreement ($\rho = 0.158$), and 24-shot fine-tuning yields the strongest agreement on held-out designs ($\rho = 0.408$). These trends are consistent with the idea that yield depends jointly on protein sequence and reaction setting, and that modest amounts of target-reaction data can calibrate a transferred model for a new design library. Together, the results support

a practical decision framework: incorporate reaction-aware representations to avoid pooling incompatible labels, and update prioritization as soon as an initial batch of measurements becomes available [157,158].

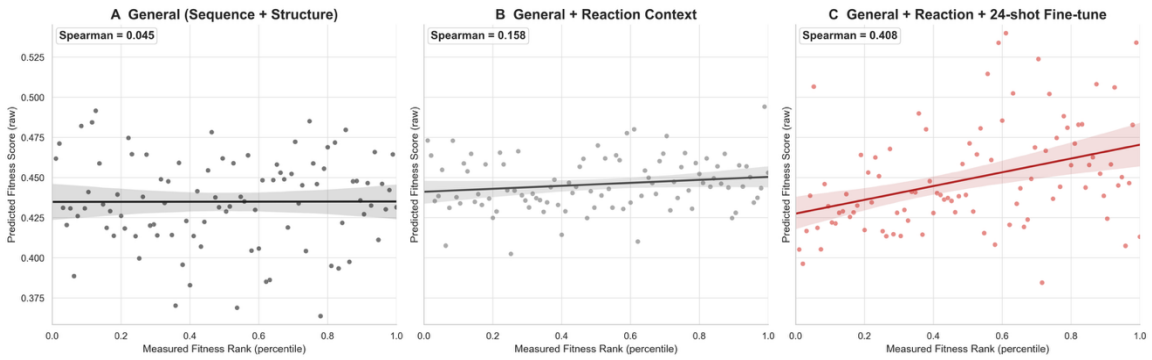


Figure 5-3. Reaction context and low-data fine-tuning improve rank agreement for de novo designs. Spearman correlation between predicted fitness scores and measured fitness rank percentiles for three ranking models evaluated on the 4-methoxystyrene cyclopropanation de novo library. (A) Protein-only model using sequence and structure features. (B) Protein features plus reaction context encoded from reaction SMILES as Morgan fingerprints. (C) Same as (B) with 24-shot fine-tuning on a labeled subset of de novo designs and evaluation on held-out designs. Reported values indicate Spearman ρ.

Taken together, these results delineate what is currently predictable and what is not yet supported by the present data regime. The most reliable use case is within-reaction prioritization for reactions where internally consistent sequence–function records already exist, with additional gains achievable by fine-tuning on a small initial batch of target measurements. In contrast, collapsing heterogeneous reactions into a single context-free label space degrades performance because beneficial mutations and failure modes differ across transformations and assay baselines differ. In this sense, the ranking experiment motivates a data strategy rather than a model strategy: broader predictive scope will require standardized, context-rich records across reaction space so that reaction-aware transfer and rapid fine-tuning are feasible.

5.8 Implications: what is predictable now, and what is not

Taken together, these results delineate what is currently predictable and what is not yet supported by the present data regime. The clearest decision value arises within a fixed assay setting, where internally consistent sequence–function records exist and a model can be used to rank candidates for the next round of testing. In this framing, prediction is not a one-time filter but a component of an iterative design–test–learn process. A practical strategy for new reactions is to screen an initial batch, record all sequences and outcomes, and then update the prioritization model so that subsequent batches are chosen to be maximally informative or higher performing. The improvement observed with 24-shot fine-tuning is consistent with this active learning interpretation: even a small amount of target-reaction data can materially improve ranking for the remaining candidates when reaction context is represented explicitly.

The same results also clarify what is not yet supported. Collapsing heterogeneous reactions into a single context-free label space degrades performance because sequence effects, failure modes, and assay baselines differ across transformations. In this sense, the ranking experiment motivates a data strategy rather than a model strategy. To make active learning and transfer reliable in new settings, datasets must preserve reaction identity and key assay context in a standardized form and must include complete outcome distributions, including negatives, so that each screening round both advances the campaign and becomes usable supervision for the next.

Many more expressive modeling choices could plausibly improve ranking beyond the lightweight baseline used here. Examples include uncertainty-aware Bayesian models for batch selection, models that jointly learn protein and reaction representations end-to-end, and selection objectives that directly optimize top-K enrichment rather than global rank correlation. I did not pursue an extensive model sweep in this thesis because the goal is not to identify the best possible architecture, but to show what becomes possible once labels and context are recorded in a reusable form. The empirical pattern in Figure 5-3 suggests that incorporating reaction information and enabling rapid fine-tuning are higher leverage than increasing model complexity in the absence of comparable data.

5.9 Outlook and thesis contribution: why data tooling is the prerequisite for decision-grade prediction

The results in this chapter outline a realistic progression from prioritization within a fixed assay setting toward broader prediction across reaction space. In the current regime, the model is most useful when it is trained and evaluated within a single assay-defined reaction context. That strength also exposes a core limitation: reliable prioritization still depends on having quantitative data for the specific reaction of interest. For many new-to-nature transformations, this prerequisite is not met in an even more fundamental way. Early campaigns often begin without any established activity, and the first goal is simply to obtain measurable product above background. De novo design can generate plausible starting points for these settings, but experimental throughput limits validation to a small subset of candidates. In these early-stage campaigns, the bottleneck is not optimizing within a reaction with dense labels, but deciding which few designs to test while protocols are still being stabilized and most outcomes may cluster near background.

This regime naturally favors an iterative design–test–learn strategy. The most effective use of modeling is not as a one-time filter, but as an updateable prioritization rule that improves as soon as initial measurements exist. In practice, this is an active learning setting: screen an initial batch, record all sequences and outcomes including negatives, update the model, and select the next batch to maximize expected information gain or expected improvement under the same assay definition. The fine-tuning result in Figure 5-3 provides a concrete example of this logic, showing that modest amounts of target-reaction data can materially improve ranking for the remaining candidates when reaction context is represented explicitly. In practical terms, the 24-shot result gives a scale estimate: on the order of a few dozen quantitative measurements in a new reaction context can be sufficient to calibrate a transferred, reaction-aware ranker so that it improves the ordering of the remaining candidates, which is consistent with the goal of reducing wasted experiments rather than eliminating experimentation altogether. Under this use case, the model is intended to prioritize new designs within the same assay definition by enriching the top-ranked subset

for higher-yield candidates compared to random selection, with performance improving as soon as early-round measurements are incorporated through fine-tuning.

Transfer learning offers a plausible route to reduce dependence on large, reaction-specific datasets, but it does not remove the need for target-reaction data [159–162]. Pretrained sequence and structure representations can be reused across tasks, and models can be initialized from related reactions rather than trained from scratch. In practice, this can lower the number of labeled examples required before predictions become decision-relevant. However, transfer only helps when the target reaction is meaningfully connected to the source data through shared mechanisms and compatible assay definitions. If reaction identity and key context are not explicit, transfer can become negative transfer, where models carry over correlations that reflect dataset composition rather than chemistry. Even under favorable conditions, some amount of target-reaction data remains necessary to establish baselines, calibrate scores, and reveal reaction-specific failure modes that no amount of pretraining can infer.

A practical near-term strategy is therefore to combine reaction-aware modeling with deliberate early measurement designed to support rapid updating. This includes collecting small but structured datasets for new reactions, measuring full outcome distributions rather than only hits, and standardizing how labels are computed so that datasets remain comparable across campaigns. With such a substrate, transfer learning can be used in a grounded way: initialize from multi-reaction training, fine-tune on a modest number of target-reaction measurements, and report uncertainty when asked to extrapolate beyond the supported domain. Over time, discrete reaction identifiers can be supplemented by structured reaction descriptors, such as substrate identity and coarse condition classes, enabling models to share information across related contexts without collapsing everything into a single global objective.

In this light, the thesis contribution of this chapter is not the specific ranking head, but the demonstration that decision-grade prediction depends on the structure and comparability of the underlying records. The tooling developed in earlier chapters makes it possible to accumulate assay-defined, genotype-linked data that support both within-context

prioritization and meaningful transfer. Chapter 5 then shows how such data can be used in a way that matches engineering constraints, where the goal is to prioritize what to test next and to improve that prioritization over successive rounds, rather than to assign a context-free notion of function. Better architectures alone will not solve the prioritization problem. Progress will come from building datasets that make reaction-aware transfer and rapid fine-tuning reliable, and from capturing the minimal target-reaction data required to turn transferred representations into dependable experimental decisions [38,163,164].

To support reuse and extension, the code used for the ranking and fine-tuning analyses in this chapter is available on GitHub (repository: github.com/YuemingLong/EnzRanker_Kit). The repository includes scripts for generating protein representations, assembling training examples from the standardized records introduced in Chapter 4, training the ranking models with reaction fingerprints, and performing few-shot fine-tuning on a small set of experimentally measured yields. With access to either in-house screening results or the database records described in Chapter 4 plus a small number of new measurements for a target reaction, readers can reproduce the analyses in Figure 5-3 and adapt the workflow to their own design campaigns.

Viewed across the thesis, the priority is to treat each engineering campaign as a learning system and a data-generating instrument. The path to more transferable prediction is to standardize how experiments are recorded so that early measurements in a new reaction immediately serve two roles: guiding the next round of testing and becoming compatible with prior datasets for fine-tuning and uncertainty calibration. Concretely, this means capturing complete genotypes, recording full outcome distributions including negatives, and stabilizing a minimal set of context fields that define comparability across labs and time. Under that regime, larger models and transfer learning become amplifiers rather than substitutes, because they can reuse accumulated supervision while still grounding decisions in the small amount of target-reaction data required for calibration.

Appendix A

SUPPORTING MATERIAL FOR CHAPTER 2

A.1 General Procedures

A.1.1 Materials

Unless otherwise specified, all reagents and solvents were obtained from known commercial suppliers (Sigma-Aldrich, Combi Blocks, Synthonix, VWR, Fischer, Ambeed, Enamine) and used without further purification. 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 treating with an appropriate TLC staining solution followed by heating if necessary.

Luria-Bertani (LB) and Terrific Broth (TB) media were used for bacterial cell growth and protein expression. Distilled water (dH_2O) was used for growth media preparation. Media were supplemented with ampicillin (amp) 0.1 mg/mL, and TB was additionally supplemented with glycerol according to the manufacturer's instructions. All media solutions were autoclaved for sterilization prior to use.

M9-N buffer was used for enzymatic reactions and was prepared from a 10X stock solution made by dissolving Na_2HPO_4 (68 g), KH_2PO_4 (30 g), and NaCl (5 g) in 1 L of double distilled water (ddH_2O) followed by sterilization using an autoclave. The 1X M9-N buffer was then prepared by diluting 100 mL of the 10X stock solution with 900 mL of ddH_2O, followed by the addition of 1 mL $CaCl_2$ (0.1 M) and 2 mL $MgSO_4$ (1.0 M). No further pH adjustment was performed.

The TAE buffer used in DNA gel electrophoresis was prepared from a 50X stock solution made by dissolving Tris free base (242g), disodium EDTA (18.61 g), and glacial acetic acid (57.2 mL) in 1 L of ddH_2O. The 1X TAE buffer was then prepared by diluting 40 mL of the 50X stock solution with 1960 mL of dH2O.

Oligonucleotides were obtained from Integrated DNA Technologies (IDT DNA). Unless otherwise stated, PCRs were run using the Phusion® High Fidelity PCR Kit (New England Biolabs). Gibson assembly mix1 was prepared by combining isothermal master mix and enzymes T5 exonuclease, Phusion® DNA polymerase, and Taq DNA ligase (New England Biolabs).

A.1.2 Instrumentation

All NMR spectra were obtained at the Caltech Liquid NMR Facility. ^{1}H and ^{13}C NMR spectra were recorded on a Bruker Prodigy 400 MHz (and 101 MHz) instrument in $CDCl_3$. ^{1}H and ^{13}C spectra are referred to residual $CDCl_3$ solvent signals at δ 7.26 and 77.0 ppm, respectively. High-performance liquid-chromatography mass spectroscopy (HPLC-MS) for reaction screening analysis was carried out using an Agilent 1200 series instrument equipped with a C18 column (Agilent Poroshell 120 EC-C18, 4.6 x 50 mm, 2.7 μm). Water and MeCN with 0.1 % HPLC grade glacial acetic acid were used as eluents. Gas chromatography (GC) was performed on an Agilent Technologies 7820A 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 in Section A.13. Sonication was performed using a Qsonica Q500 sonicator. Chromatographic purification was accomplished by flash chromatography using a Biotage Isolera One instrument with Sfär High-Capacity Duo columns using AMD Silica Gel 60, 230–400 mesh. Centrifugation was carried out in an Allegra 25R tabletop centrifuge equipped with a TS-5.1-500 swinging bucket rotor.

A.1.3 Cloning, Mutagenesis, and Plasmid Isolation

T7 Express competent E. coli (BL21) (New England Biolabs) cells were utilized for all experiments. Expression vector pET-22b(+) (Novagen) was used for cloning and expressing all variants. Plasmids were isolated using either a Monarch® Plasmid Miniprep Kit or a QIAprep® Spin MiniPrep Kit.

Site-saturation mutagenesis (SSM) was performed using a modified QuikChange™ mutagenesis protocol using primers bearing NNK degenerate codons.2 Upon completion of PCRs, the remaining template plasmid DNA was digested with the DpnI restriction enzyme at 37 ºC for 1.5 hours. The PCR products were purified by gel electrophoresis (1% agarose in 1X TAE buffer containing 1X SYBR Gold nucleic acid Stain) and extracted using the Zymoclean™ Gel DNA Recovery Kit. Fragments were then assembled into a circular plasmid using the Gibson assembly protocol, incubating at 50 ºC for 1 hour).1 Without further purification, 2.5 µL of the Gibson product were used to transform 22.5 µL of competent E. coli cells. Transformed cells were supplemented with 477.5 µL SOC medium and immediately plated on LB-amp (100 mg/mL) agar plates. Plates were incubated at 30 °C for 16–18 hours or 37 °C for 12–14 hours.

Random mutations were introduced using error-prone PCR (epPCR). PCRs were run using Taq polymerase (New England Biolabs) in the presence of 300, 400, 500, and 600 µM MnCl2 with the primers shown in Table A1.3 Primers 005 and 006 were used to amplify the DNA region encoding the full-length protein of interest. Primers 007 and 008 were used to amplify the backbone (pET-22b(+) vector) fragment containing a gene encoding for ampicillin resistance using Phusion polymerase and the template plasmid DNA. PCR products were isolated and assembled in the same manner as SSM libraries, and the resultant Gibson products were similarly used to transform competent E. coli cells and plated on LB-amp/agar. Libraries generated from 300, 400, 500, and 600 µM $MnCl_2$ were screened, and the library with the desired error rate was further evaluated.

Table A1. Primers used in error-prone PCRs.

Primers	Sequence (5' ⟶ 3')
005	GAA ATA ATT TTG TTT AAC TTT AAG AAG GAG ATA TAC ATA TG
006	GCC GGA TCT CAG TGG TGG TGG TGG TGG TGC TCG AG
007	CAT ATG TAT ATC TCC TCC TTA AAG TTA AAC AAA ATT ATT TC
007	CAT ATG TAT ATC TCC TCC TTA AAG TTA AAC AAA ATT ATT TC

A.1.4 Protein Expression and Reaction Screening in Plate

From agar plates bearing BL21 cells transformed with protoglobin variant libraries, 88 isolated single colonies were picked using sterile toothpicks and used to inoculate starter cultures (400 µL of LB-amp) in sterilized 96-well deep-well culture plates. Six wells on each plate were inoculated with single colonies of the parent enzyme (parent controls) and two wells were inoculated with a sterile toothpick (sterile controls). These starter culture plates were sealed with air-permeable tape and grown at 37 °C and 220 rpm with 80% humidity in a Multitron Infors shaker for 12–14 hours. Subsequently, 50 µL of the starter culture were used to inoculate expression cultures (900 µL TB-amp) in a separate, sterilized 96-well deep-well plate. Similarly, 50 µL of the overnight culture were added to 50 µL of sterilized 50% glycerol to prepare a glycerol stock replica, which was stored at -80 °C. Expression cultures were incubated at 37 °C and 220 rpm with 80% humidity for 2.5 hours. The plates were then cooled on ice for 30 minutes before inducing protein expression with isopropyl β-D-1-thiogalactopyranoside (IPTG, 0.5 mM) and δ-aminolevulinic acid (ALA, 1.0 mM). Expression was carried out at 22 °C and 220 rpm for 20–24 hours. The expression cultures were then centrifuged (4,000xg, 5 minutes, room temperature) and the supernatant was removed. The cell pellets were resuspended in 380 µL of M9-N (pH = 7.2) using an orbital shaker (900 rpm) and the plates were transferred into a Coy anaerobic chamber (~0–10 ppm O_2). Cell suspensions were supplemented with 20 µL of a combined substrate solution in acetonitrile (MeCN) ([alkene]final = 10 mM, [EDA]final = 15 mM; 5% MeCN). The plates

were sealed with an adhesive aluminum foil seal and incubated on an orbital shaker (700 rpm) in the Coy anaerobic chamber at room temperature overnight.

For workup and analysis, the plates were removed from the Coy chamber and HPLC-grade MeCN (800 µL/well) was added to each well. The resulting suspensions were mixed by shaking on an orbital shaker at 800 rpm for 20 minutes at room temperature and the plates were then centrifuged (5,000xg, 15 minutes, 4 °C) to remove proteins and cell debris from the supernatant. The supernatant (200 µL/well) was transferred to a polypropylene 96-well microtiter plate (Agilent) and sealed with an NAL-96 pierceable film (USA Scientific) and subjected to reverse-phase HPLC-MS analysis. The product MS peak was integrated using built-in analysis and wells showing signals higher than that of the parent wells were identified as "hits". The analyte corresponding to each hit was extracted into ethyl acetate (EtOAc) and subjected to GC analysis using a chiral stationary phase to determine the stereoselectivity. Wells displaying an optimal combination of activity and stereoselectivity were validated by resuspending cells from a 50 mL expression culture to OD_{600} = 30 in M9-N and screening this cell suspension alongside the parent. (see Section A.1.5 for details).

A.1.5 Protein Expression and Whole-Cell Validation Reactions

Following screening, wells displaying an optimal combination of activity and stereoselectivity relative to parent controls were streaked from the corresponding glycerol stock onto individual LB-amp agar plates. A single colony was picked and grown in 5 mL LB-amp overnight at 37 °C and 250 rpm. Subsequently, 0.5 mL of this starter culture was used to inoculate 50 mL of TB-amp in a 125-mL sterilized Erlenmeyer flask. The starter culture was also used to isolate variant plasmids using a miniprep kit and their DNA sequence was determined by Sanger sequencing (Laragen, Inc., Culver City, CA). The expression culture was incubated at 37 °C and shaken at 220 rpm until the optical cell density at 600 nm (OD_{600}) reached ~0.8–1.0. The expression culture was then cooled on ice for 30 minutes before inducing protein expression with IPTG (0.5 mM) and ALA (1.0 mM). Expression was carried out at 22 °C and 220 rpm for 20–24 hours. The expression culture was then transferred to a 5mL Falcon™ tube and centrifuged (4000xg, 5 minutes, room temperature).

The supernatant was removed, and the pellets were normalized to an OD600 = 30 in M9-N buffer. Subsequently, an aliquot of the normalized cell suspension (380 µL) was added to a 2.0 mL screw-cap vial (Agilent) and transferred into a Coy anaerobic chamber (~0–10 ppm O_2). For each variant, three technical replicates were performed. Cell suspensions were supplemented with 20 µL of a combined substrate solution in MeCN ([alkene]final = 10 mM, [EDA]final = 15 mM; 5% MeCN). The vials were capped and incubated on an orbital shaker (700 rpm) in the Coy anaerobic chamber at room temperature overnight. The vials were then removed from the Coy chamber and 800 µL of a 10 mM solution of 1,2-diphenylethane dissolved in 50:50 (v/v) EtOAc:cyclohexane were added to each vial. The resulting partition in each vial was mixed by shaking on an orbital shaker (900 rpm) for 20 minutes. The contents of each vial were then transferred to separate 1.7 mL microcentrifuge tubes, and centrifuged (14,000xg, 10 minutes, room temperature). The organic layer (300 µL/vial) was transferred to a 400 µL glass insert within a 2.0 mL screw cap vial for GC-FID analysis equipped with a chiral (for stereoselectivity) and achiral (for activity) stationary phase. The starting material, internal standard (1,2-diphenylethane), and product peaks were integrated using built-in ChemStation analysis software (Agilent). The variant displaying the most optimal combination of activity and selectivity compared to parent was carried forward in evolution.

A.1.6 Determination of Heme Concentration

Heme concentration was determined using the hemochrome assay with clarified cell lysate. Lysate was obtained by sonication (6 minutes, 1 second on, 2 seconds off, 35% amplitude, on wet ice) of whole-cell resuspended in M9-N buffer to OD600 = 30. The cell debris was removed by centrifugation (14,000g, 10 minutes, 4 °C) and the supernatant was stored on ice until further use. To a cuvette, 500 µL of lysate and 500 µL of freshly prepared solution I [0.2 M NaOH, 40% (v/v) pyridine, 0.5 mM $K_3Fe(CN)_6$] were added. The UV-Vis spectrum (380–620 nm) of the oxidized Fe(III) state was recorded immediately. Sodium dithionite (10 µL of 0.5 M solution in 0.5 M NaOH) was added, the mixture was thoroughly mixed through pipetting, and the UV-Vis spectrum of the reduced Fe(II) state was promptly recorded. The

heme concentration was calculated using Beer's law, with the extinction coefficient $\varepsilon[557_{reduced}–540_{oxidized}] = 23.98\ mM^{-1}cm^{-1}$ for heme, considering dilution factors.

A.1.7 Protein Purification

E. coli transformed with pET22b(+) constructs encoding various protoglobin variants were grown overnight in 50 mL LB-amp. Subsequently, 20 mL of this starter culture were used to inoculate 1 L of TB-amp in a 2-L sterilized Erlenmeyer flask. The expression culture was incubated at 37 °C and shaken at 140 rpm until OD600 reached ~0.9–1.1. The expression culture was then cooled on ice for 30 minutes before inducing protein expression with IPTG (0.5 mM) and ALA (1.0 mM). Expression was carried out at 22 °C, 180 rpm for 20–22 hours with a shaking radius of 25 mm. Once expression was finished, the cultures were centrifuged (4,000xg, 5 minutes, 4 °C) and the cell pellet was frozen at -20 °C until further processing.

After freeze-thawing thrice, the cell pellet was resuspended in 60 mL of KPi buffer and 60 mg of lysozyme lysate powder from chicken egg white, 10 mg of DNase I powder from bovine pancreas, and 60 µL of 2 M magnesium chloride solution (filtered) were added. The resuspended solution was incubated at 37 °C for 1 hour and was subsequently centrifuged (14,000xg, 15 minutes, 4 °C). The supernatant was applied to a 5-mL Ni^{2+}-NTA column equilibrated with 50 mM KPi (pH 8.0) at 4 °C. Once all of the lysate had been applied, the column was washed with 50–100 mL of buffer A (KPi buffer with 20 mM imidazole, pH 8.0) until the flow-through was clear. The protein-of-interest was eluted with 5 to 10 mL buffer B (KPi buffer with 500 mM imidazole, pH 8.0). Fractions were pooled and concentrated to >2.5 mL with 10 kDa MWCO filter. The heme concentration was measured using the hemochrome assay (Section 1.6). The concentrated, purified protein (0.4 – 1 mM stock) was flash-frozen and stored at -80 °C.

A.1.8 Lyophilized Lysate Preparation and Storage

E. coli transformed with pET22b(+) constructs encoding various protoglobin variants were grown overnight in 5mL LB medium supplemented with ampicillin (LB-amp).

Subsequently, 20 mL of this starter culture were used to inoculate 1 L of TB-amp in a 2-L sterilized Erlenmeyer flask. The expression culture was incubated at 37 °C and shaken at 140 rpm until the OD600 reached ~0.9–1.1. The expression culture was then cooled on ice for 30 minutes before inducing protein expression with IPTG (0.5 mM) and ALA (1.0 mM). Expression was carried out at 22 °C, 180 rpm for 20–22 hours with a shaking radius of 25 mm. Once expression was finished, the cultures were centrifuged (4000g, 5 minutes, 4 °C) and the wet cell pellet was weighed. Five times the wet cell weight amount of M9-N was used to resuspend the pellet. Lysate was obtained by sonication (6 minutes total, 1 second on, 2 seconds off, 35% amplitude, on wet ice). The cell debris was removed by centrifugation (14,000xg, 10 minutes, 4 °C). The cell lysates were combined, flash-frozen using liquid nitrogen, and lyophilized to obtain a dry powder that can be stored at room temperature on benchtop.

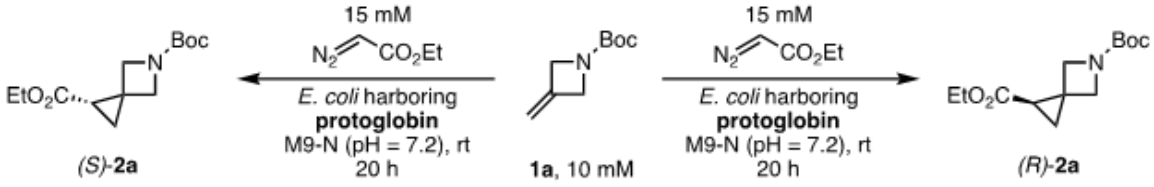
A.1.9 Reaction Setup with Lyophilized Lysate

The required amount of lyophilized lysate powder was weighed and dissolved in dH2O. This reconstituted solution is ready for direct use in reaction setups. Reconstituted lysate solution are transferred into GC 2.0 mL screw-cap vial (Agilent) and transferred into a Coy anaerobic chamber (~0–30 ppm O2). For each variant, three technical replicates were performed. Cell suspensions were supplemented with combined substrate solution in MeCN, lysate and reaction volume varies according to reaction set up. The vials were capped and incubated on an orbital shaker (700 rpm) in the Coy anaerobic chamber at room temperature overnight. The vials were then removed from the Coy chamber and 400 µL of either 50:50 (v/v) EtOAc:cyclohexane or a 10 mM solution of 1,2-diphenylethane dissolved in 50:50 (v/v) EtOAc:cyclohexane were added to each vial. The resulting partition in each vial was mixed by shaking on an orbital shaker (900 rpm) for 20 minutes. The contents of each vial were then transferred to separate 1.7 mL microcentrifuge tubes, and centrifuged (14,000xg, 10 minutes, room temperature). The organic layer (300 µL/vial) was transferred to a 400 µL glass insert within a 2.0 mL screw cap vial for GC-FID analysis equipped with a chiral (for stereoselectivity) and achiral (for activity) stationary phase. The starting material, internal

standard (1,2-diphenylethane), and product peaks were integrated using built-in ChemStation analysis software (Agilent).

A.2 Discovery of Initial Activity

A 6member subset of the Arnold lab collection of previously-evolved protoglobin variants from 12 different organisms was screened for product formation in whole-cell reactions in a 96-well deep-well plate (see Section A.1.4). We initially screened for starting activity with tert-butyl 3-methyleneazetidine-1-carboxylate (1a) in the presence of EDA using HPLC-MS to analyze product formation (Scheme A1). The enantiomeric ratio (er) of the top six variants was subsequently measured using GC with a chiral stationary phase. Fortuitously, in the presence of several engineered protoglobins, the cyclopropanation of 1a proceeded to form the desired compound (2a) (Figure A1). The activity and stereoselectivity of mutated variants corresponding to wells B06, B08, B09, D08, E07, and E09 was quantified according to the procedure described in Section 1.5. While most variants favor (R)-2a or provide the product as a near racemate, TamPgb-xHC-5316 (TamPgb W59L Y60Q) provides (S)-2a, establishing a starting point for the development of a stereodivergent pathway (Table A2). Therefore, ApePgb-xHC-5311 and TamPgb-xHC-5316 were chosen as the starting variants for a stereodivergent directed evolution campaign of 1a (see Sections A.4.1 and A.4.2, respectively).



Scheme A1. Initial activity discovery was conducted on 1a in the presence of EDA using a subset of the Arnold lab collection of previously-evolved protoglobin variants. Activity was observed in each stereochemical direction, yielding (S)-2a and (R)-2a, dependent on the mutated enzyme employed.

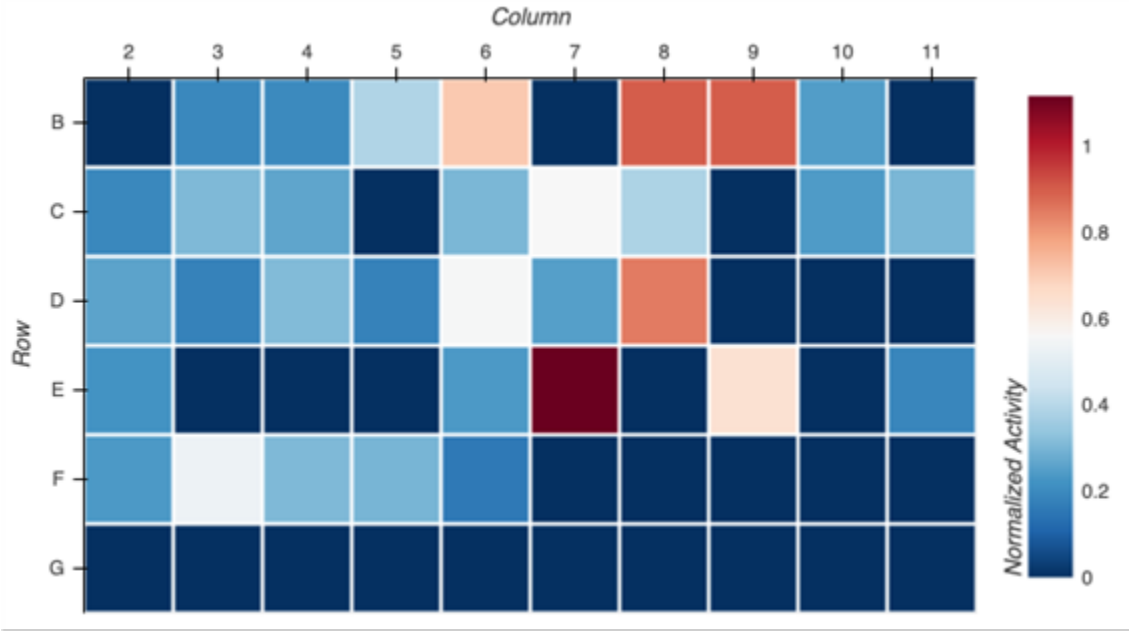


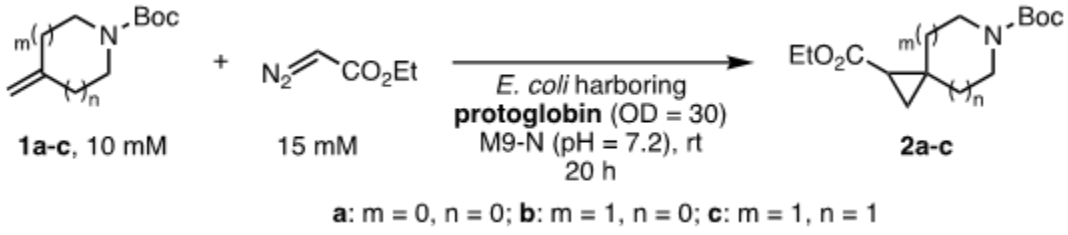
Figure A1. Heat map visualizing the concentration of 2a formed in an initial activity screening of 60 protoglobin variants previously engineered in our laboratory.

Table A2. Yields and stereoselectivities of 2a formed in the presence of engineered protoglobin variants. Yields were determined by comparison to an internal standard (1,2-diphenylethane) using GC and a corresponding calibration curve (see Section A.11). Enantioselectivities were determined using GC equipped with a chiral stationary phase.

Well	Variant	Yield (%)	er ((*S*)-**2a**:(*R*)-**2a**)
B06	*Ape*Pgb AGAGW	93	53:47
B08	*Ape*Pgb AGAMW	84	47.5:52.5
B09	*Ape*Pgb AGACW	57	38.5:61.5
D08	*Ape*Pgb AGPW	47	18:82
E07	*Pme*Pgb LQ	39	51:49
E09	*Tam*Pgb LQ	58	62:38

A.3 Control Experiments

Control experiments were performed to assess potential background activity using combinations of hemin, bovine serum albumin (BSA), and sodium dithionite (Na_2S_2O4) in the absence of a hemoprotein. The standard screening conditions are depicted in Scheme A2 and the control experiments are detailed in Table A3. Reactions with TamPgb-xHC-5316 and ApePgb-xHC-5311 were run in parallel for direct comparison. The results for substrates 1a, 1b, and 1c are shown in Sections A.3.1, A.3.2, and A.3.3, respectively. In all cases, TamPgb-xHC-5316 and ApePgb-xHC-5311 form the corresponding products 2a, 2b, and 2c in significant quantities; however, no significant product formation was observed under control conditions.



Scheme A2. Standard enzymatic screening conditions for substrates 1a-c to form 2a-c

Table A3. Experimental conditions for background activity assessment.

Entry	Changes to Standard Conditions
1	none, *Tam*Pgb-xHC-**5316** used
2	none, *Ape*Pgb-xHC-**5311** used
3	no enzyme, 50 μM hemin
4	no enzyme, 1 mg/mL BSA
5	no enzyme, 300 mM sodium dithionite
6	no enzyme, 50 μM hemin, 300 mM sodium dithionite
7	no enzyme, 50 μM hemin, 1 mg/mL BSA
8	no enzyme, 300 mM sodium dithionite, 1 mg/mL BSA
9	no enzyme, 50 μM hemin, 300 mM sodium dithionite, 1 mg/mL BSA
10	buffer only

A.3.1 Control Experiments on 1a

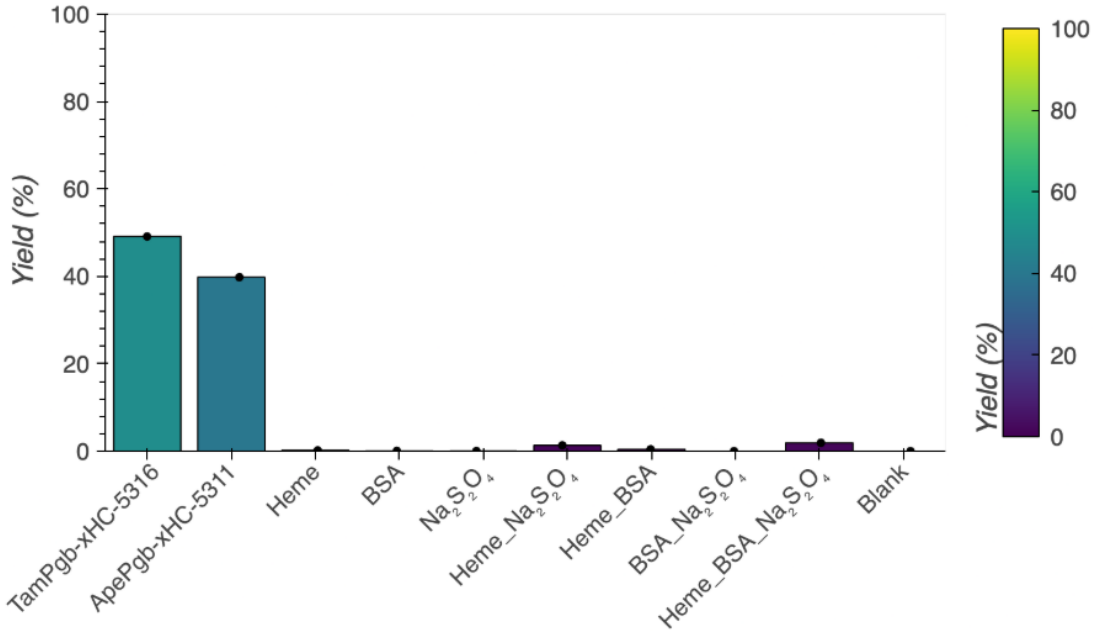


Figure A2. Bar plot depicting the enzymatic activity of TamPgb-xHC-5316, ApePgb-xHC-5311, and control conditions with substrate 1a. Yields of 2a were determined by comparison to an internal standard (1,2-diphenylethane) using GC and a corresponding calibration curve (see Section A.11). TamPgb-xHC-5316 and ApePgb-xHC-5311 exhibit distinct activity peaks, while no significant peaks are observed in any of the control conditions.

A.3.2 Control experiments on 1b

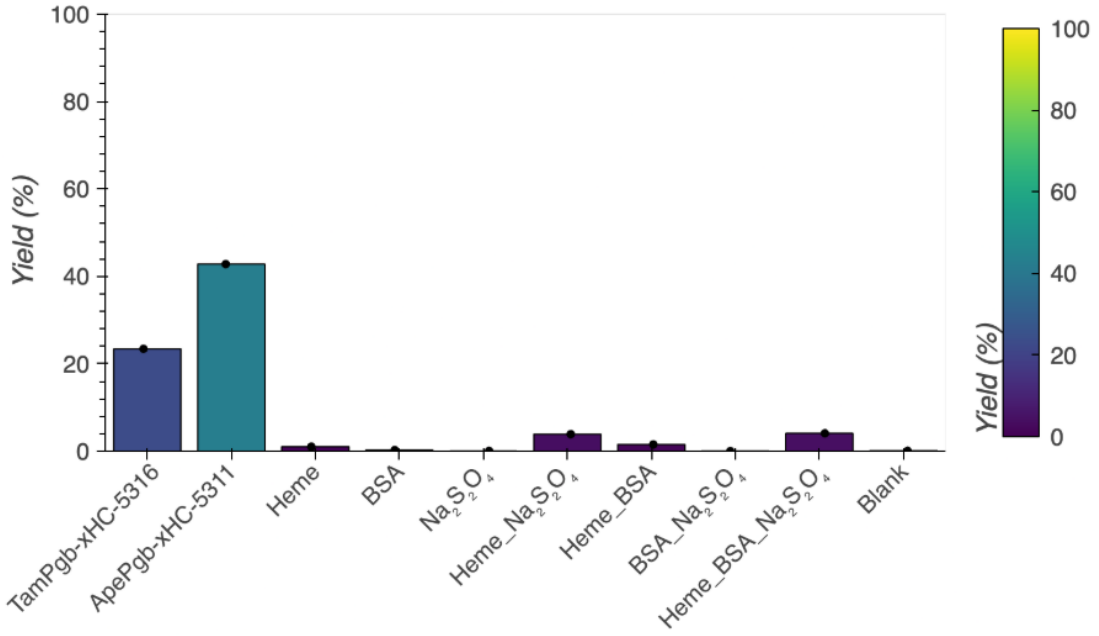


Figure A3. Bar plot depicting the enzymatic activity of TamPgb-xHC-5316, ApePgb-xHC-5311, and control conditions with substrate 1b. Yields of 2b were determined by comparison to an internal standard (1,2-diphenylethane) using GC and a corresponding calibration curve (see Section A.11). TamPgb-xHC-5316 and ApePgb-xHC-5311 exhibit distinct activity peaks, while no significant peaks are observed in any of the control conditions.

A.3.3 Control Experiments on 1c

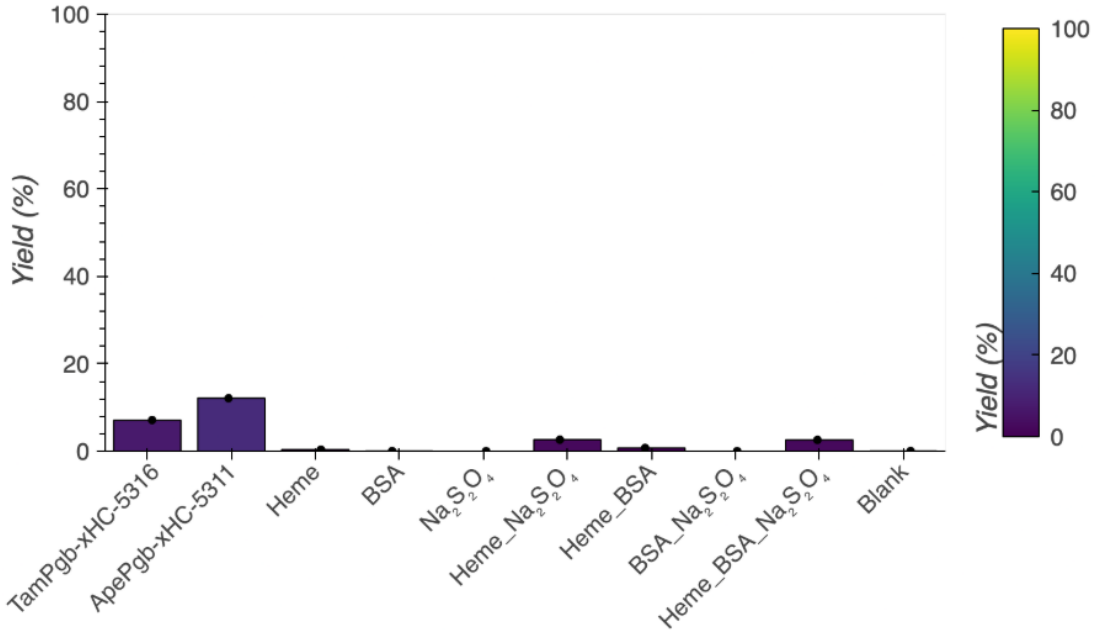


Figure A4. Bar plot depicting the enzymatic activity of TamPgb-xHC-5316, ApePgb-xHC-5311, and control conditions on substrate 1c. Yields of 2c were determined by comparison to an internal standard (1,2-diphenylethane) using GC and a corresponding calibration curve (see Section A.11). TamPgb-xHC-5316 and ApePgb-xHC-5311 exhibit distinct activity peaks, while no significant peaks are observed in any of the control conditions.

A.4 Directed Evolution Lineage Trajectories and Plots

The yield and stereoselectivity data presented in this section were collected according to the procedure detailed in Section A.1.5. All variants for a given substrate campaign were run in parallel. For each variant, two biological replicates were performed by inoculating two separate starter cultures from distinct single colonies, with three technical replicates carried out for each biological replicate (6 data points for each variant). Yields of 2a-c were determined by comparison to an internal standard (1,2-diphenylethane) using GC and a corresponding calibration curve (see Section A.11 for calibration curves and Section A.12 for GC-FID reaction traces). Enantio- and diastereoselectivities were measured using GC with a chiral stationary phase (see Section A.13 for traces).

A.4.1 Evolution trajectory of ApePgb-xHC-5311 to ApePgb-xHC-5312 for the synthesis of (R)-2a from 1a

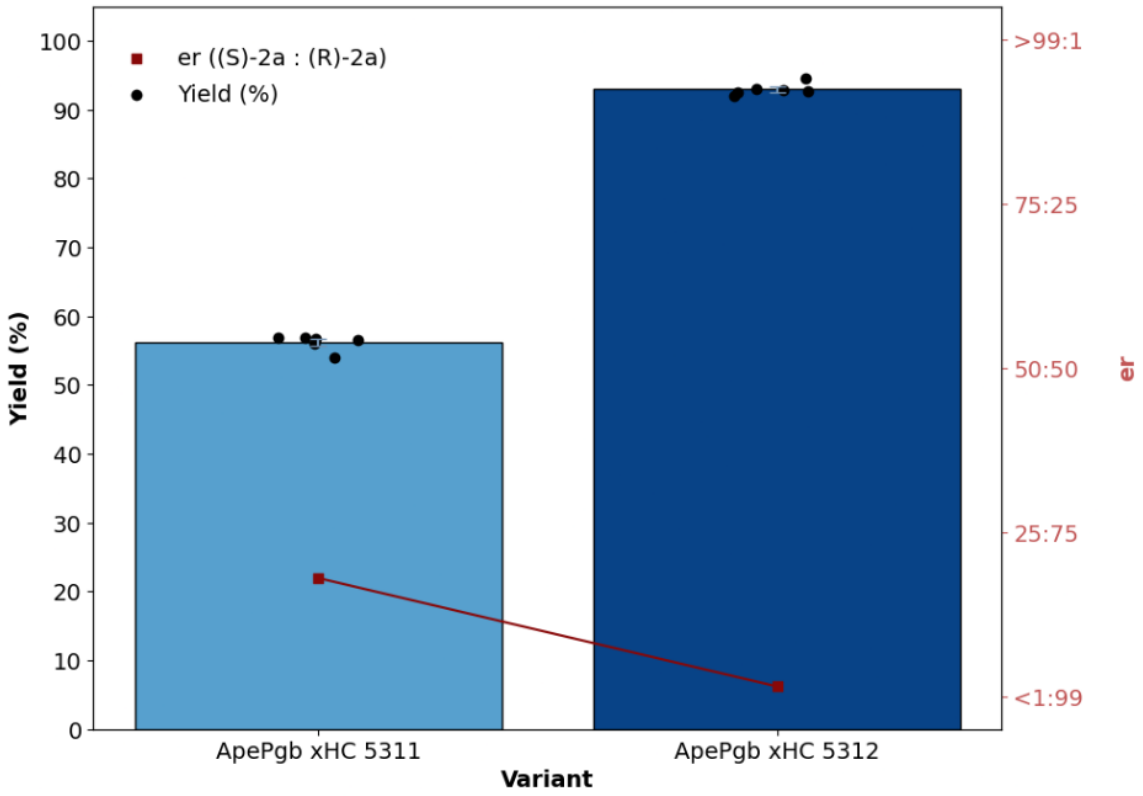


Figure A5. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of (R)-2a from 1a

Table A4. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of (R)-2a from 1a

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Ape*Pgb-xHC-**5311**	**W59A Y60G V63P F145W**
1	SSM	*Ape*Pgb-xHC-**5312**	**W59Y** Y60G V63P F145W

A.4.2 Evolution trajectory of TamPgb-xHC-5316 to TamPgb-xHC-5318 for the synthesis of (S)-2a from 1a

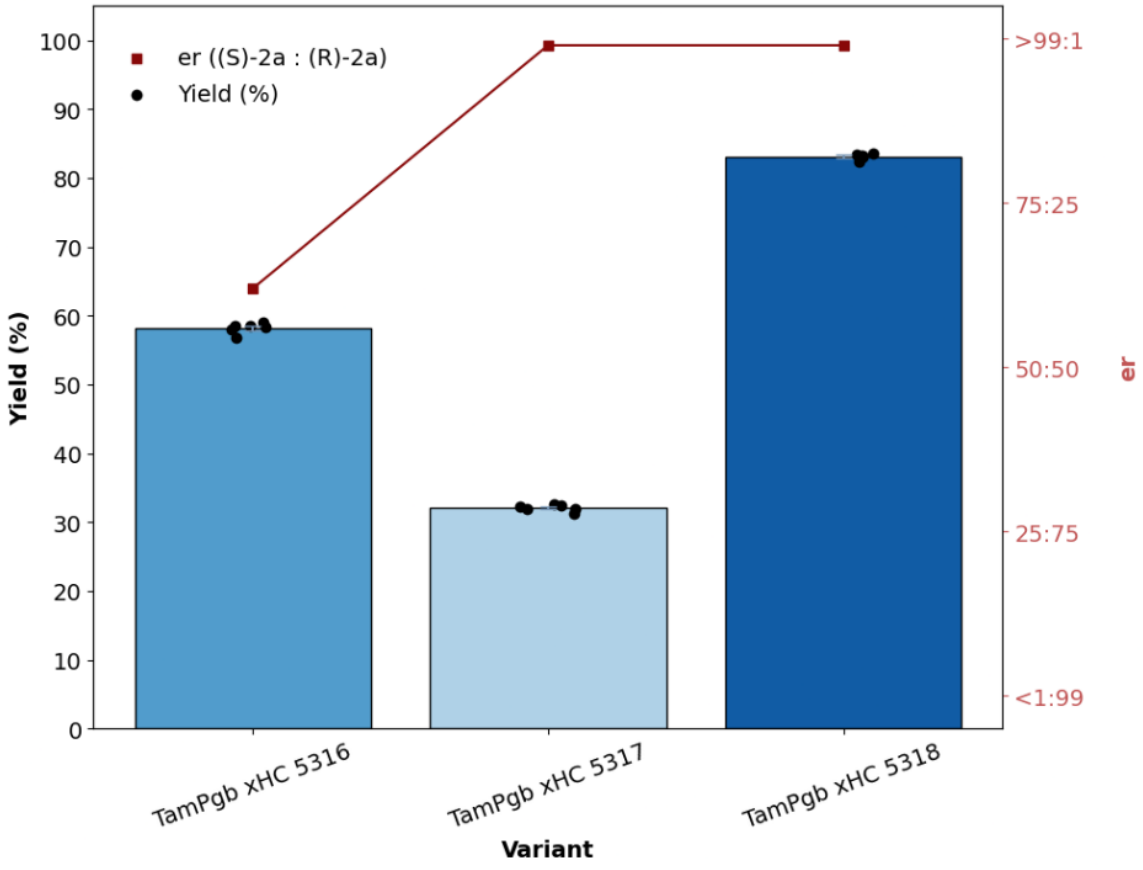


Figure A6. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of (S)-2a from 1a

Table A5. Table S5: Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of (S)-2a from 1a

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Tam*Pgb-xHC-**5316**	**W59L Y60Q**
1	SSM	*Tam*Pgb-xHC-**5317**	W59L Y60Q **F92V**
2	epPCR	*Tam*Pgb-xHC-**5318**	**L14L D57G** W59L Y60Q **G61D** F92V **R118R K124R P137L K153R**

A.4.3 Evolution trajectory of ApePgb-xHC-5321 to ApePgb-xHC-5322 for the synthesis of isomer 1 of 2b from 1b

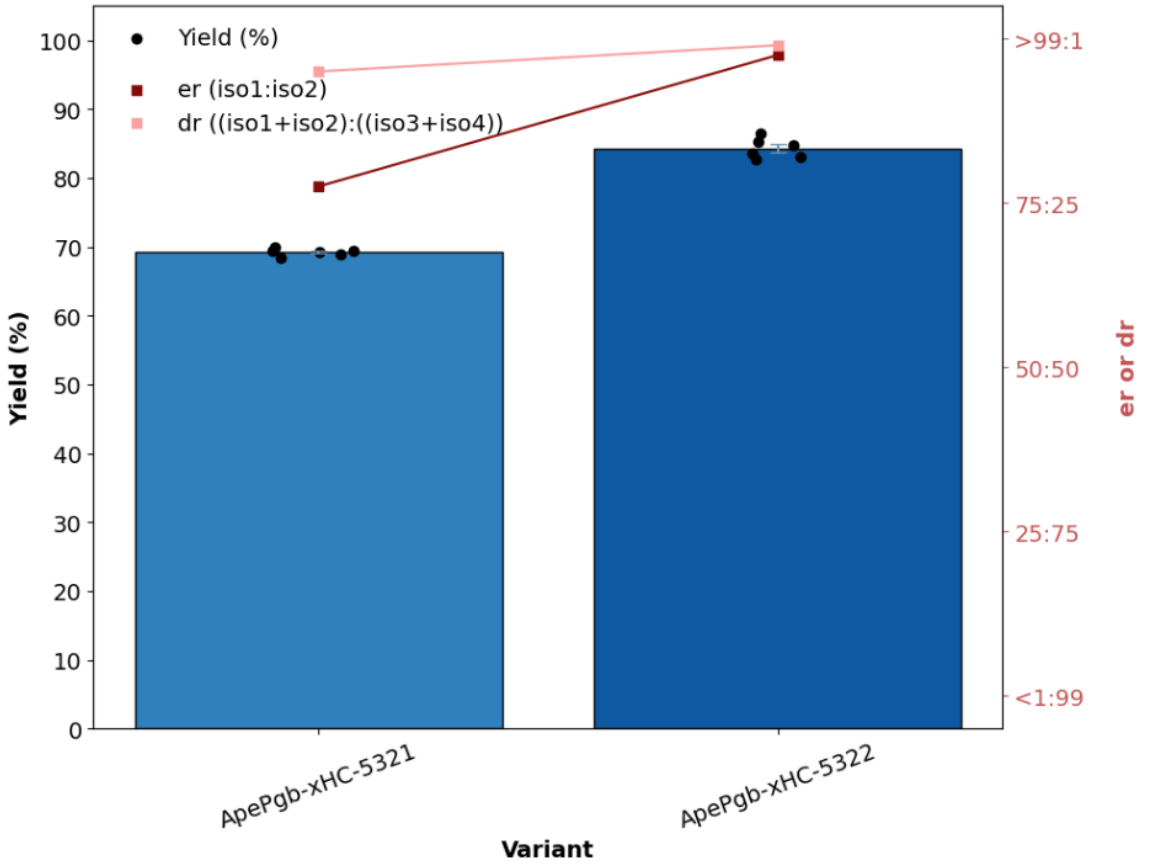


Figure A7. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of isomer 1 of 2b from 1b

Table A6. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of isomer 1 of 2b from 1b

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Ape*Pgb-xHC-**5321**	**Y60P**
1	SSM	*Ape*Pgb-xHC-**5322**	Y60P F145C

A.4.4 Evolution trajectory of ApePgb-xHC-5311 to ApePgb-xHC-5312 for the synthesis of isomer 2 of 2b from 1b

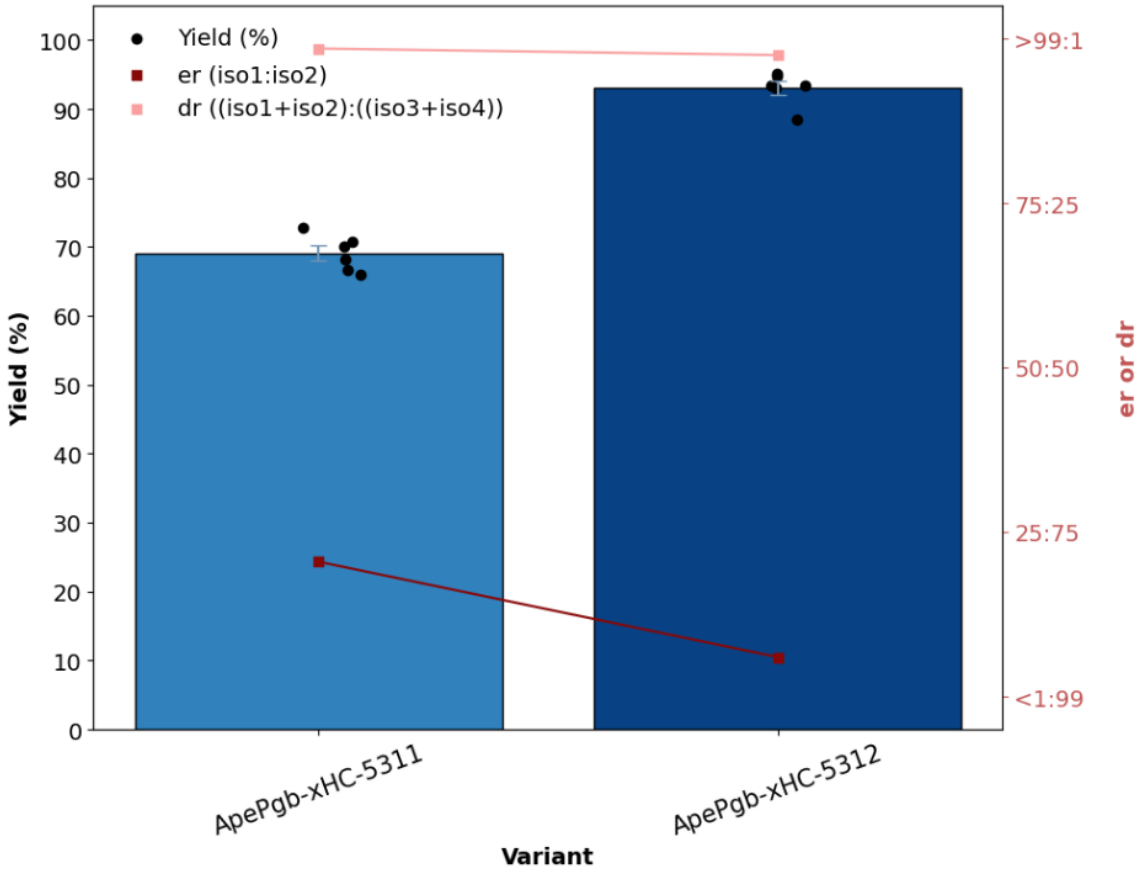


Figure A8. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of isomer 2 of 2b from 1b

Table A7. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of isomer 2 of 2b from 1b

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Ape*Pgb-xHC-**5311**	**W59A Y60G V63P F145W**
1	SSM	*Ape*Pgb-xHC-**5312**	**W59Y** Y60G V63P F145W

A.4.5 Evolution trajectory of TamPgb-xHC-5316 to TamPgb-xHC-5328 for the synthesis of isomer 3 of 2b from 1b

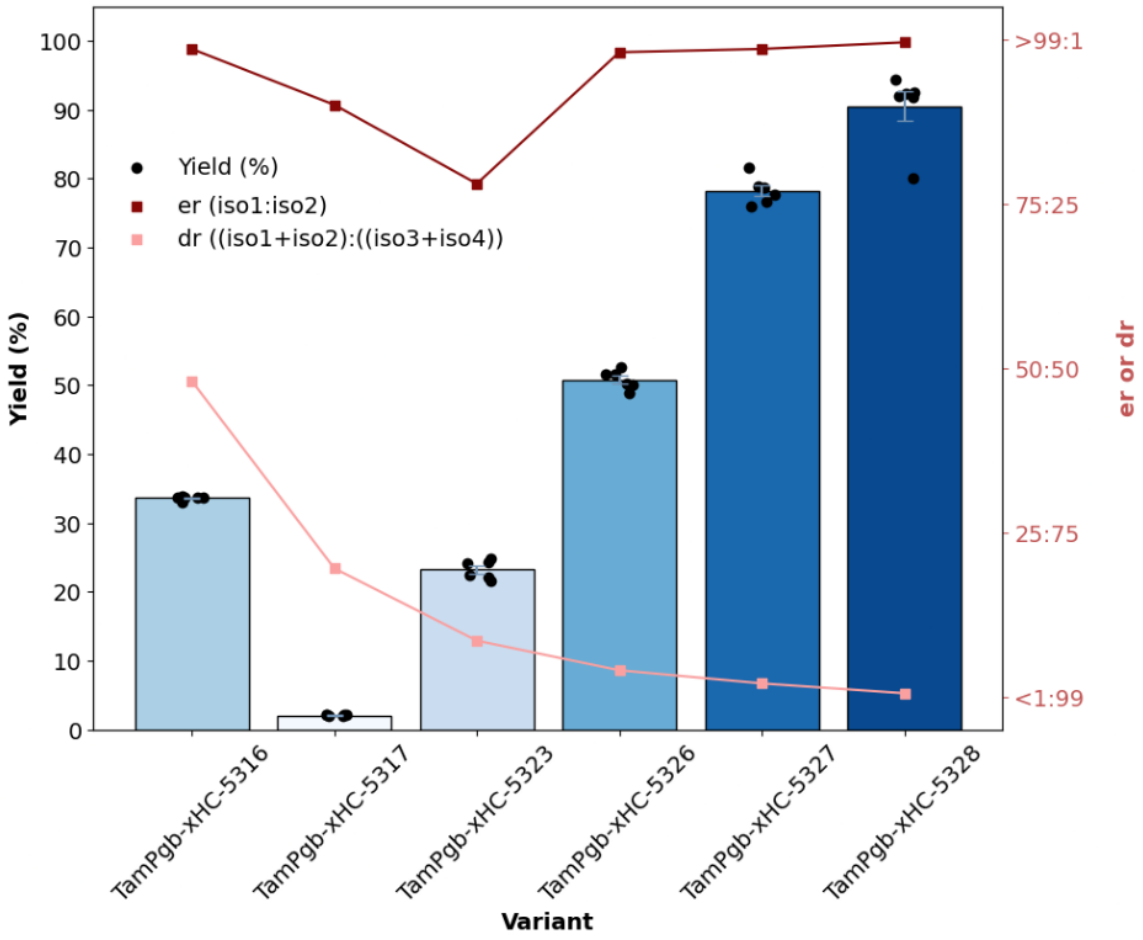


Figure A9. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of isomer 3 of 2b from 1b

Table A8. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of isomer 3 of 2b from 1b

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Tam*Pgb-xHC-**5316**	**W59L Y60Q**
1	SSM	*Tam*Pgb-xHC-**5317**	W59L Y60Q **F92V**
2	epPCR	*Tam*Pgb-xHC-**5323**	W59L Y60Q F92V **D98Y**
3	SSM	*Tam*Pgb-xHC-**5326**	W59L Y60Q **V88L** F92V D98Y
4	SSM	*Tam*Pgb-xHC-**5327**	W59L Y60Q V88L F92V D98Y **F144H**
5	epPCR	*Tam*Pgb-xHC-**5328**	W59L Y60Q **T74P** V88L F92V D98Y **T131A** F144H

A.4.6 Evolution trajectory of TamPgb-xHC-5316 to TamPgb-xHC-5325 for the synthesis of isomer 4 of 2b from 1b

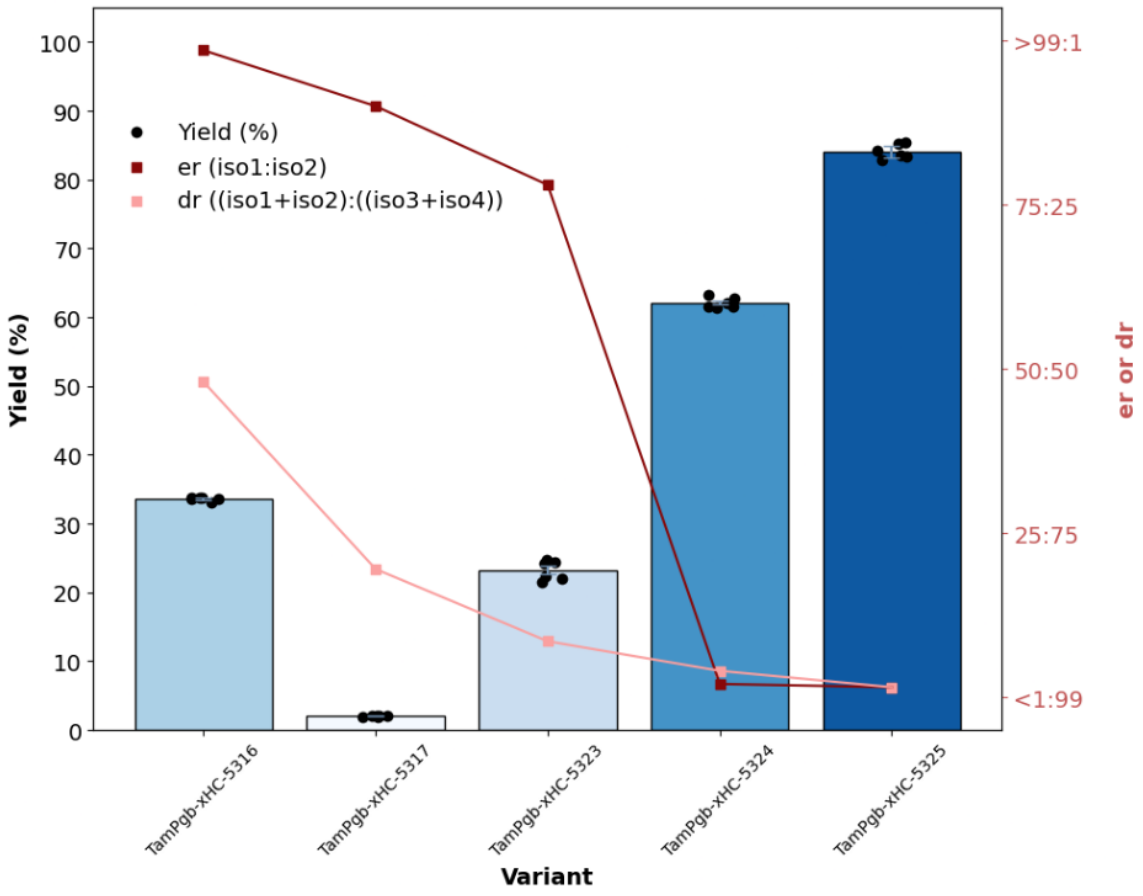


Figure A10. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of isomer 4 of 2b from 1b

Table A9. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of isomer 4 of 2b from 1b

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Tam*Pgb-xHC-**5316**	**W59L Y60Q**
1	SSM	*Tam*Pgb-xHC-**5317**	W59L Y60Q **F92V**
2	epPCR	*Tam*Pgb-xHC-**5323**	W59L Y60Q F92V **D98Y**
3	SSM	*Tam*Pgb-xHC-**5324**	W59L Y60Q **F92I** D98Y
4	SSM	*Tam*Pgb-xHC-**5325**	W59L **Y60R** F92I D98Y

A.4.7 Evolution trajectory of ApePgb-xHC-5311 to ApePgb-xHC-5315 for the synthesis of (R)-2c from 1c

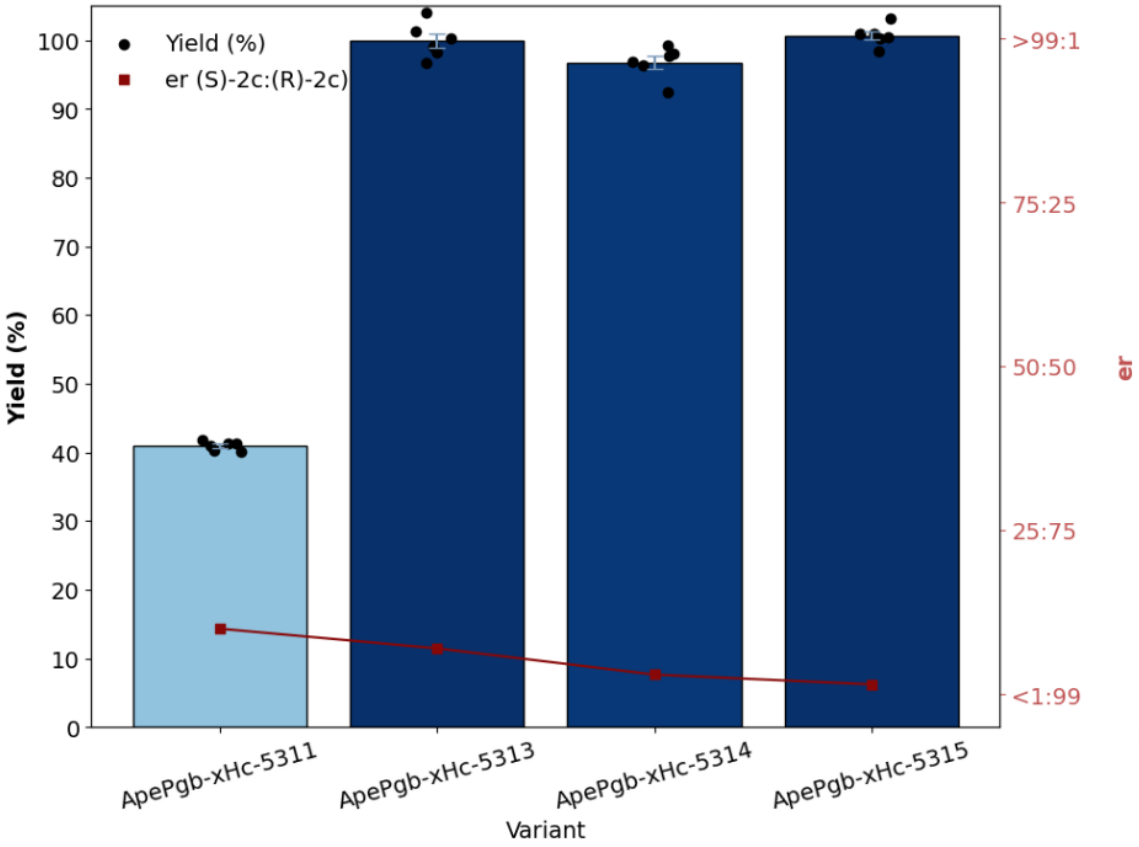


Figure A11. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of (R)-2c from 1c

Table A10. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of (R)-2c from 1c

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Ape*Pgb-xHC-**5311**	**W59A Y60G V63P F145W**
1	SSM	*Ape*Pgb-xHC-**5312**	**W59Y** Y60G V63P F145W
2	SSM	*Ape*Pgb-xHC-**5313**	**W59D** Y60G V63P F145W
3	epPCR	*Ape*Pgb-xHC-**5314**	W59D Y60G V63P F145W **W185S**
4	SSM	*Ape*Pgb-xHC-**5315**	W59D Y60G V63P F145W **I149Q** W185S

A.4.8 Evolution trajectory of TamPgb-xHC-5316 to TamPgb-xHC-5320 for the synthesis of (S)-2c from 1c

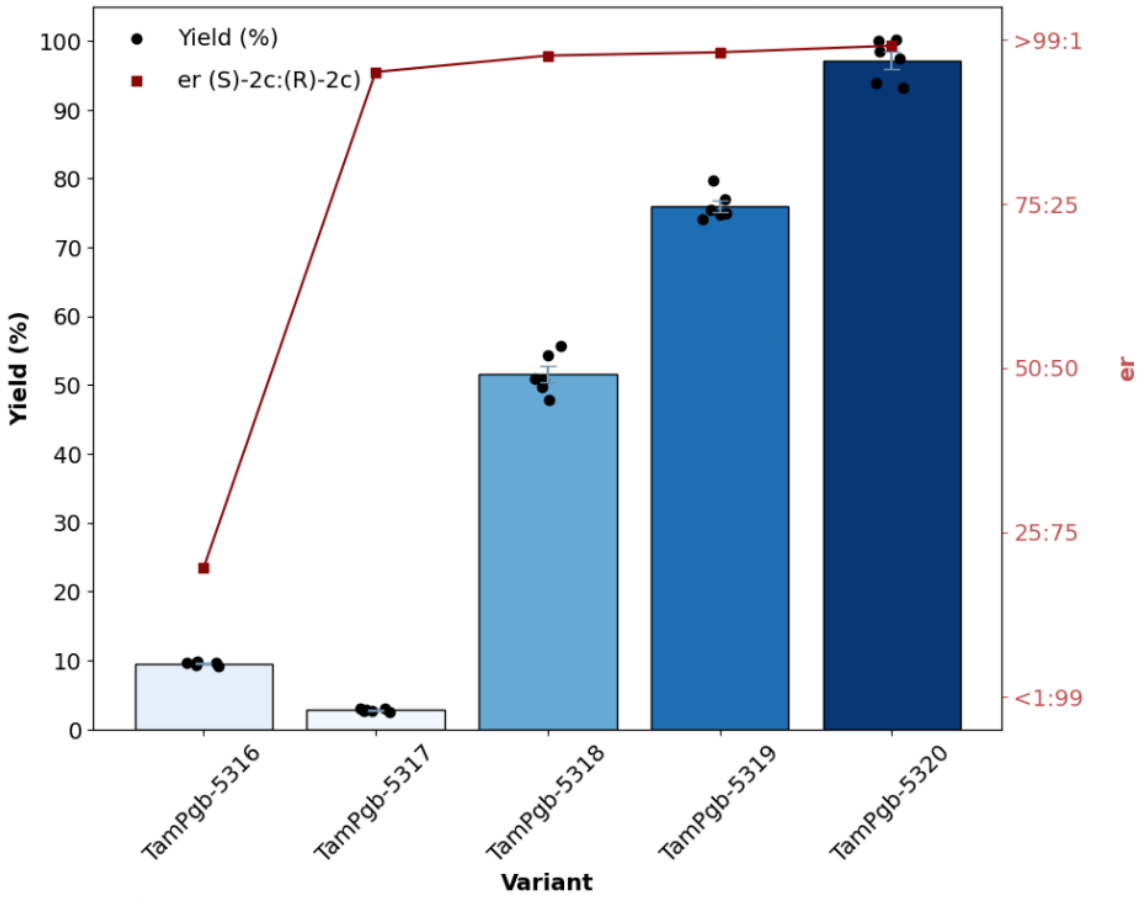


Figure A12. Bar plot visualizing the yield and enantioselectivity of variants evolved for the synthesis of (S)-2c from 1c

Table A11. Evolution strategy and incorporated mutations for each round of evolution for variants evolved for the synthesis of (S)-2c from 1c

Round	Method	Variant Name	Mutations from *Tam*Pgb WT
0	--	*Tam*Pgb-xHC-**5316**	**W59L Y60Q**
1	SSM	*Tam*Pgb-xHC-**5317**	W59L Y60Q **F92V**
2	epPCR	*Tam*Pgb-xHC-**5318**	**L14L D57G** W59L Y60Q **G61D** F92V **R118R K124R P137L K153R**
3	SSM	*Tam*Pgb-xHC-**5319**	L14L D57G W59L Y60Q G61D **V63G** F92V R118R K124R P137L K153R
4	SSM	*Tam*Pgb-xHC-**5320**	L14L D57G W59L Y60Q G61D V63G F92V R118R K124R P137L **L141M I148V** K153R

A.5 DNA Sequences of Evolved Enzymes

A.5.1 Off-Target Mutations Present in pET-22b(+) Vector

All variants described herein contain a 1- and 3-base pair deletion in non-coding regions of the pET-22b(+) vector. When indexed to begin at the ColE1 origin of replication (i.e., starting at TTGAGATCCTTTTTTT), the 3-bp deletion in both TamPgb and ApePgb variants occurs at 825_827delTAT. According to the same index, the 1-bp deletion in TamPgb and ApePgb occurs at 4788delC and 4791delC, respectively (position difference attributed to a codon insertion in the gene of interest of the TamPgb homologue).

Further, all variants described herein contain a single base pair mutation in the F1 origin of replication coding region. According to the index described above, the mutation in TamPgb and ApePgb is: 4778T>C and 4781T>C, respectively. With the exception of TamPgb-xHC-5328, all variants described herein contain an additional single base pair mutation in the F1 origin of replication coding region; the mutation in TamPgb and ApePgb is: 4357C>T and 4360C>T, respectively.

For the following sequences, base-pair mutations from wildtype are represented in red font.

A.5.2 DNA Sequence of ApePgb-xHC-5311 (AGPW)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 2173C>T.

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
GCGGGTGGTTGGCCGGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA
CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC

CATATCCCACTTCGTTATCTTATCGCATGGATCTATCCTATCACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCACTGA

A.5.3 DNA Sequence of ApePgb-xHC-5312 (YGPW)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 2173C>T.

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
TATGGTGGTTGGCCGGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATAC
AGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGATT
CTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTTG
GGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCCC
ATATCCCACTTCGTTATCTTATCGCATGGATCTATCCTATCACCGCCACTATCAAGCCA
TTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGGT
TCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATGA
CTGGCTCGAGCACCACCACCACCACCAC

A.5.4 DNA Sequence of ApePgb-xHC-5313 (DGPW)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 2173C>T.

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
GATGGTGGTTGGCCGGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA

CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC
CATATCCCACTTCGTTATCTTATCGCATGGATCTATCCTATCACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCAC

A.5.5 DNA Sequence of ApePgb-xHC-5314 (DGPWS)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 2173C>T.

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
GATGGTGGTTGGCCGGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA
CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC
CATATCCCACTTCGTTATCTTATCGCATGGATCTATCCTATCACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCAGTTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCAC

A.5.6 DNA Sequence of ApePgb-xHC-5314 (DGPWQS)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 2173C>T.

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
GATGGTGGTTGGCCGGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA
CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC
CATATCCCACTTCGTTATCTTATCGCATGGATCTATCCTCAGACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCAGTTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCAC

A.5.7 DNA Sequence of TamPgb-xHC-5316 (LQ)

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTTTCGGACAGTGGATTCT
GGATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.8 DNA Sequence of TamPgb-xHC-5317 (LQV)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTGTGGGACAGTGGATTCT
GGATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.9 DNA Sequence of TamPgb-xHC-5318 (G3)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCCTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGGTTT
ACTGCAGGACTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTGTGGGACAGTGGATTCT
GGATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGACATCATCGCACAAAAAGAAACCAAACTGACGGCGTGACCTCCGTCCCACAT
ATTCTGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAGACCCTT
CCTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTT
CAAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGAC
TTTCTCGAGCACCACCACCACCACCAC

A.5.10 DNA Sequence of TamPgb-xHC-5319 (G3G)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCCTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGGTTT
ACTGCAGGACTTTGGTGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTGTGGGACAGTGGATTCT
GGATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGACATCATCGCACAAAAAGAAACCAAACTGACGGCGTGACCTCCGTCCCACAT
ATTCTGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAGACCCTT
CCTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTT
CAAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGAC
TTTCTCGAGCACCACCACCACCACCAC

A.5.11 DNA Sequence of TamPgb-xHC-5320 (G3GMV)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCCTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGGTTT
ACTGCAGGACTTTGGTGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTGTGGGACAGTGGATTCT
GGATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGACATCATCGCACAAAAAGAAACCAAACTGACGGCGTGACCTCCGTCCCACAT
ATTCTGTTGCGCTATATGATCAGTTTTATTTATCCCGTGACAGCCACCATCAGACCCTT
CCTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTT
CAAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGAC
TTTCTCGAGCACCACCACCACCACCAC

A.5.12 DNA Sequence of ApePgb-xHC-5321 (P)

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
TGGCCGGGTTGGGTAGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA
CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC
CATATCCCACTTCGTTATCTTATCGCATTTATCTATCCTATCACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCAC

A.5.13 DNA Sequence of ApePgb-xHC-5322 (PC)

ATGACTCCCTCGGACATCCCGGGATATGATTATGGGCGTGTCGAGAAGTCACCCATCA
CGGACCTTGAGTTTGACCTTCTGAAGAAGACTGTCATGTTAGGTGAAAAGGACGTAAT
GTACTTGAAAAAGGCGTGTGACGTTCTGAAAGATCAAGTTGATGAGATCCTTGACTTG
TGGCCGGGTTGGGTAGCATCAAATGAGCATTTGATTTATTACTTCTCCAATCCGGATA
CAGGAGAGCCTATTAAGGAATACCTGGAACGTGTACGCGCTCGCTTTGGAGCCTGGAT
TCTGGACACTACCTGCCGCGACTATAACCGTGAATGGTTAGACTACCAGTACGAAGTT
GGGCTTCGTCATCACCGTTCAAAGAAAGGGGTCACAGACGGAGTACGCACCGTGCCC
CATATCCCACTTCGTTATCTTATCGCATGTATCTATCCTATCACCGCCACTATCAAGCC
ATTTTTGGCTAAGAAAGGTGGCTCTCCGGAAGACATCGAAGGGATGTACAACGCTTGG
TTCAAGTCTGTAGTTTTACAAGTTGCCATCTGGTCACACCCTTATACTAAGGAGAATG
ACTGGCTCGAGCACCACCACCACCACCAC

A.5.14 DNA Sequence of TamPgb-xHC-5323 (LQVY)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTGTGGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.15 DNA Sequence of TamPgb-xHC-5324 (LQIY)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTATTGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.16 DNA Sequence of TamPgb-xHC-5325 (LRIY)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 4705G>A.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCGTGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTGTTCGCCGCCGTATTGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.17 DNA Sequence of TamPgb-xHC-5326 (LQLVY)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTCTTCGCCGCCGTGTGGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTTTTATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.18 DNA Sequence of TamPgb-xHC-5327 (LQLVYH)

Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 146G>T.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCACTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTCTTCGCCGCCGTGTGGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGACCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTCATATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC
AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.5.19 DNA Sequence of TamPgb-xHC-5328 (LQPLVYAH)

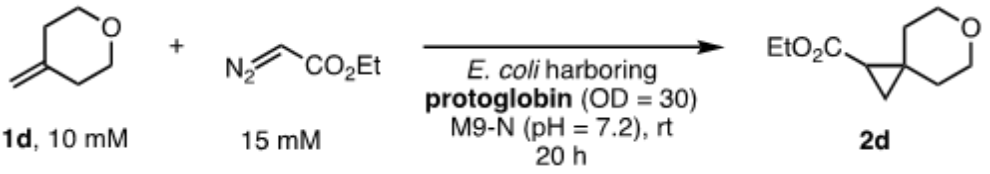
Note: In addition to the off-target pET-22b(+) vector mutations described in Section A.5.1, this variant contains a single base pair mutation in a non-coding region. According to the index described in Section A.5.1, the mutation is: 4608T>C.

ATGCCGCAAATTCCCGGTTACACTTATGGAGATCCCGCCTTACCTCCCAGTCCCGTTTC
TTTAGAGGAATTGGAGCGTCTGAAGGCTTGCTTGCTGTGGACGGAGGAGGATGATAA
AGCATTAGAGCAGGCCGGTAAAGTATTAGAGGACCAAGTAGAAGAAGTGTTAGATTT
ACTGCAGGGCTTTGTCGGGAGCCATCCTCATTTACTTCACTATTTCCCTGATCCTCAGG
GGAACCCGATCCCCGACTATCTTGAGCGTCTTCGCCGCCGTGTGGGACAGTGGATTCT
GTATACCTGCTTTCGTCCCAAAGACGAAACCTGGCTTCGTTATCAGCATGAGATTGGC
TTACGTCATCATCGCACAAAAAAAAACCAAACTGACGGCGTGGCCTCCGTCCCACATA
TTCCGTTGCGCTATTTGATCAGTCATATTTATCCCATTACAGCCACCATCAAACCCTTC
CTGACTAAGAAGGGCCACAATCCCGAGGAGGTGGAGCGTATGTATCAGGCATGGTTC

AAGGCAGTTGTATTGCAAGTAGCACTTTGGTCCTATCCTTACACTCAGCCTGGCGACTT
TCTCGAGCACCACCACCACCACCAC

A.6 Yields and Enantioselectivities for Enzymatic Reactions of 1d to 2d

Enzymatic reactions of tetrahydro-4-methyl-2H-pyran (1d) to yield ethyl 6-oxaspiro[2.5]octane-1-carboxylate (2d) were performed using whole-cell catalysts following the general procedure described in Section A.1.5 (Scheme A2). The yields of 2d shown in Table A12 were quantified by comparison to an internal standard (1,2-diphenylethane) using GC-FID and the corresponding calibration curve (Section A.11). The GC-FID crude reaction traces used for yield determination are shown in Section A.12.4. Enantioselectivities were measured using GC with a chiral stationary phase (see Section A.13.4 for traces).



Scheme A3. Standard enzymatic conditions for the reaction of 1d to form 2d

Table A12. Yields and enantioselectivities of compound 2d with enzyme variants in whole cell

Product	Variant	Yield (%)	er
2d	*Ape*Pgb-xHC-**5314**	34	4:96
2d	*Tam*Pgb-xHC-**5319**	10	96:4
2d	*Ape*Pgb-xHC-**5320**	23	94.5:5.5

A.7 Enzymatic Reactions Performed with Lyophilized Lysate

Reactions using lyophilized lysate were performed according to the general procedure described in Section A.1.9. The yields of 2a,b shown in Table A13 were quantified by comparison to an internal standard (1,2-diphenylethane) using GC-FID and the corresponding calibration curve (Section A.11). The GC-FID crude reaction traces used for yield determination are shown in Section A.12 Enantio- and diastereoselectivities were measured using GC with a chiral stationary phase (see Section A.13 for traces).

Table A13. Yields and enantioselectivities of compounds 2a,b using lyophilized lysate

Product	**Variant**	**Yield (%)**	**dr** (major:minor)	**er** (major:minor)
(*R*)-**2a**	*Ape*Pgb-xHC-**5312**	>99	NA	97:3
(*S*)-**2a**	*Tam*Pgb-xHC-**5318**	>99	NA	99:1
2b, iso1	*Ape*Pgb-xHC-**5322**	>99	96:4	88.5:11.5
2b, iso2	*Ape*Pgb-xHC-**5312**	95	95:5	88:12
2b, iso3	*Tam*Pgb-xHC-**5328**	>99	99:1	99.5:0.5
2b, iso4	*Tam*Pgb-xHC-**5325**	>99	98:2	99:1

A.8. Time Course Study with 500 Mm of Substrate

A.8.1 Time course study with 500 mM of 1a using ApePgb-xHC-5312 (YGPW)

ApePgb-xHC-5312 in the form of a lyophilized lysate powder was added to 2.0 mL screw cap vials and resuspended in 320 µL ddH2O to achieve a final concentration of 125 µM. The vials were transferred into a Coy anaerobic chamber (~0–10 ppm O2) and 39 µL of substrate 1a, as well as 41 µL of EDA were added to the solution (note: an organic co-solvent is not used). The mixtures were shaken for a fixed amount of time before quenching by addition of 400 µL of a 1:1 (v:v) solution of EtOAc:cyclohexane (for each time point, three technical replicates were conducted). The quenched reactions were shaken vigorously on an orbital shaker at 900 rpm for 30 minutes, then transferred to microcentrifuge tubes and centrifuged for 15 minutes at 14,000xg. The organic layer (300 µL/vial) was transferred to a 400 µL glass insert within a 2.0 mL screw cap vial and subjected to GC-FID analysis. The starting material and product peaks were integrated using built-in ChemStation analysis software (Agilent). The % conversion ((2a/(2a + 1a))*100) is plotted as a function of time (Figure A13).

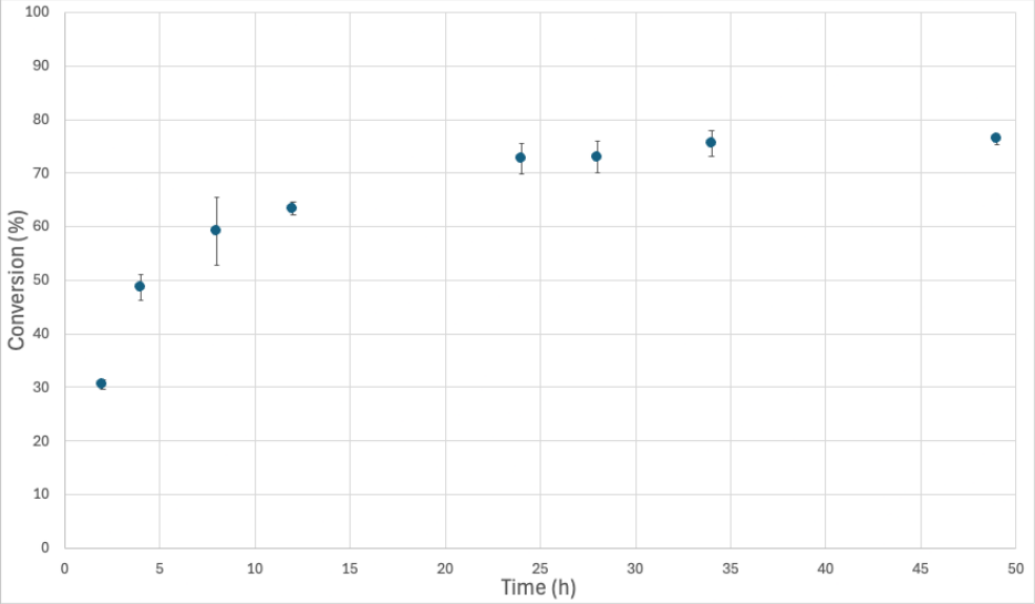


Figure A13. Conversion (%) to 2a plotted as a function of time in the presence of ApePgb-xHC-5312 with 500 µM substrate 1a at designated time points.

A.8.2 Time course study with 500 mM of 1c using ApePgb-xHC-5315

ApePgb-xHC-5315 in the form of a lyophilized lysate powder was added to a 2.0 mL screw cap vial and resuspended in 320 µL ddH_2O to achieve a final concentration of 125 µM. The vial was transferred into a Coy anaerobic chamber (~0–10 ppm O2) and 39 µL of substrate 1c, as well as 41 µL of EDA was added to the solution (note: an organic co-solvent is not required). The mixture is shaken for a fixed amount of time before quenching by addition of 400 µL of a 1:1 (v:v) solution of EtOAc:cyclohexane. The quenched reaction was shaken vigorously on an orbital shaker at 900 rpm for 30 minutes, then transferred to a microcentrifuge tube and centrifuged for 15 minutes at 14,000 rpm. The organic layer (300 µL/vial) was transferred to a 400 µL glass insert within a 2.0 mL screw cap vial and subjected to GC-FID analysis. The starting material and product peaks were integrated using built-in ChemStation analysis software (Agilent). The % conversion ((2c/(2c + 1c))*100) is plotted as a function of time (Figure A14).

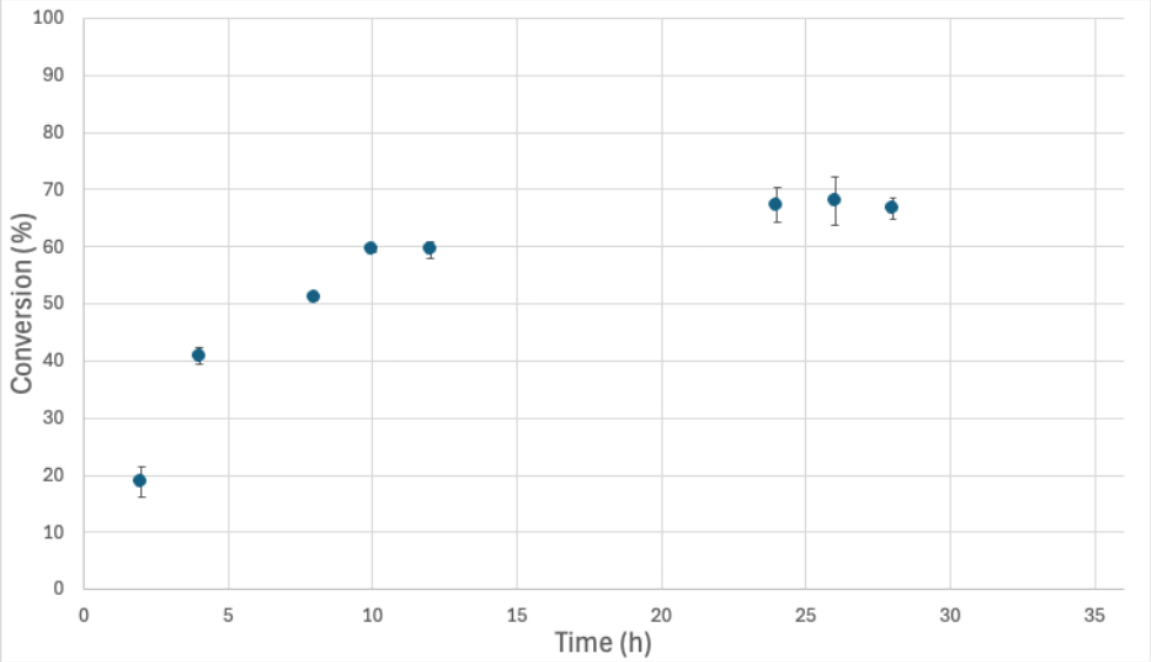


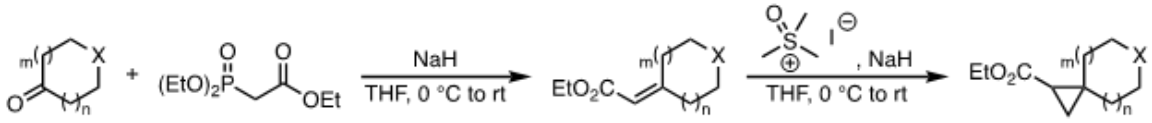
Figure A14. Conversion (%) to 2c plotted as a function of time in the presence of ApePgb-xHC-5315 with 500 µM substrate 1c at designated time points.

A.9 Starting Materials

The following starting materials were obtained commercially and used without prior purification: tert-butyl 3-methyleneazetidine-1-carboxylate (1a) (Combi Blocks), tert-butyl 3-methylenepyrrolidine-1-carboxylate (1b) (Combi Blocks), tert-butyl 4-methylenepiperidine-1-carboxylate (1c) (Synthonix), and tetrahydro-4-methyl-2H-pyran (1d) (Combi Blocks).

A.10 Synthesis and Characterization of Authentic Standards

The following authentic standards were obtained commercially and used without prior purification: 5-(tert-butyl) 1-ethyl 5-azaspiro[2.4]heptane-1,5-dicarboxylate (2b) (Combi Blocks), 6-(tert-butyl) 1-ethyl 6-azaspiro[2.5]octane-1,6-dicarboxylate (2c) (Combi Blocks), and ethyl 6-oxaspiro[2.5]octane-1-carboxylate (2d) (Synthonix). 5-(tert-butyl) 1-ethyl 5-azaspiro[2.3]hexane-1,5-dicarboxylate (2a) was synthesized according to the general procedures provided below.



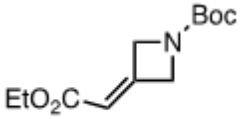
Scheme A4. The general synthetic route for the synthesis of the authentic standards presented herein consists of a Horner-Wadsworth-Emmons olefination of the corresponding ketone followed by a Corey-Chaykovasky cyclopropanation.

A.10.1 General procedure for Horner-Wadsworth-Emmons (HWE) olefination

In a flame-dried Schlenk flask under an argon atmosphere and equipped with a magnetic stir bar, NaH (1.5 equiv.) was taken up in dry THF (0.5 M with respect to ketone). The suspension was cooled to 0 °C using an ice bath and triethyl phosphonoacetate (1.5 equiv.) was added batchwise. The resulting solution was maintained at this temperature for 1 hour. The corresponding ketone (1.0 equiv.) was then added as a 1M solution in dry THF, and the reaction was stirred overnight, allowing the solution to ambiently warm to room temperature.

The reaction progress was analyzed by TLC and quenched by the addition of saturated aq. NH4Cl (3 mL/mmol). The mixture was then transferred to a separatory funnel. The phases were separated, and the organic layer was washed with dH_2O (2 x 2 mL/mmol) and brine (1 x 2 mL/mmol). The combined aqueous layer was washed with diethyl ether (3 x 2 mL/mmol) and the combined organic layers were subsequently dried with $MgSO_4$, filtered, and evaporated under reduced pressure. The crude material was then purified using automated column chromatography (EtOAc/Hexanes).

*tert-butyl 3-(2-ethoxy-2-oxoethylidene)azetidine-1-carboxylate (**S3a**)*



S3a

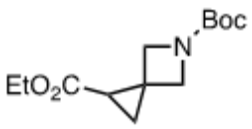
Compound S3a was prepared according to the General procedure for Horner-Wadsworth-Emmons (HWE) olefination using tert-butyl 3-oxoazetidine-1-carboxylate (20.0 mmol) and was obtained as a viscous oil in 65.0% yield (3.13 g, 13.0 mmol). ^{1}H NMR spectroscopic data is consistent with that found in literature.

A.10.2 General procedure for Corey-Chaykovasky cyclopropanation

In a flame-dried Schlenk flask under an argon atmosphere and equipped with a magnetic stir bar, NaH (1.5 equiv.) was taken up in dry THF (0.5 M with respect to ketone). The suspension was cooled to 0 °C using an ice bath and trimethylsulphoxonium iodide (1.5 equiv.) was added batchwise. The resulting solution was maintained at this temperature for 1 hour. The corresponding HWE product (1.0 equiv.) was then added as a 1M solution in dry THF, and the reaction was stirred overnight, allowing the solution to ambiently warm to room temperature. The reaction progress was analyzed by TLC and quenched by the addition of saturated aq. NH_4Cl (3 mL/mmol). The mixture was then transferred to a separatory funnel. The phases were separated, and the organic layer was washed with dH_2O (2 x 2 mL/mmol) and brine (1 x 2 mL/mmol). The combined aqueous layer was washed with diethyl ether (3 x 2 mL/mmol) and the combined organic layers were subsequently dried with

MgSO4, filtered, and evaporated under reduced pressure. The crude material was then purified using automated column chromatography (EtOAc/Hexanes).

5-(tert-butyl) 1-ethyl 5-azaspiro[2.3]hexane-1,5-dicarboxylate ((+)-2a)



(+)-**2a**

Compound (+)-2a was prepared according to the General procedure for Corey-Chaykovasky cyclopropanation using S3a (13.0 mmol) and was obtained as a viscous oil in 15.0% yield (498 mg, 1.95 mmol). 1H NMR spectroscopic data is consistent with that found in literature.

A.11 Calibration Curves for Starting Materials and Authentic Standards

A.11.1 General procedure for the generation of calibration curves

To prepare samples for the calibration curve, the authentic standard or starting material, as well as internal standard were dissolved in EtOAc to create 100 mM solutions. Six different product-to-standard (Pdt/Std) or starting material-to-standard (SM/Std) ratios were prepared: 2:1, 1:1, 0.5:1, 0.25:1, 0.125:1, and 0.0625:1, by mixing specific volumes of the 100 mM solutions. The total volume for each sample was adjusted to 500 µL with EtOAc. The prepared solutions were transferred to a 400 µL glass insert within a 2.0 mL screw cap vial and subjected to GC-FID analysis with an achiral stationary phase. The starting material, internal standard (1,2-diphenylethane), and product peaks were integrated using built-in ChemStation analysis software (Agilent). Calibration curves were generated by plotting the known concentrations of the product or starting material, based on prepared Pdt/Std or SM/Std ratios, as a function of the ratio of its FID signal (peak area) relative to that of the internal standard. The slope of the calibration curve was used to calculate the product yield and the remaining starting material in enzymatic reactions.

A.11.2 Calibration Curve of 1a

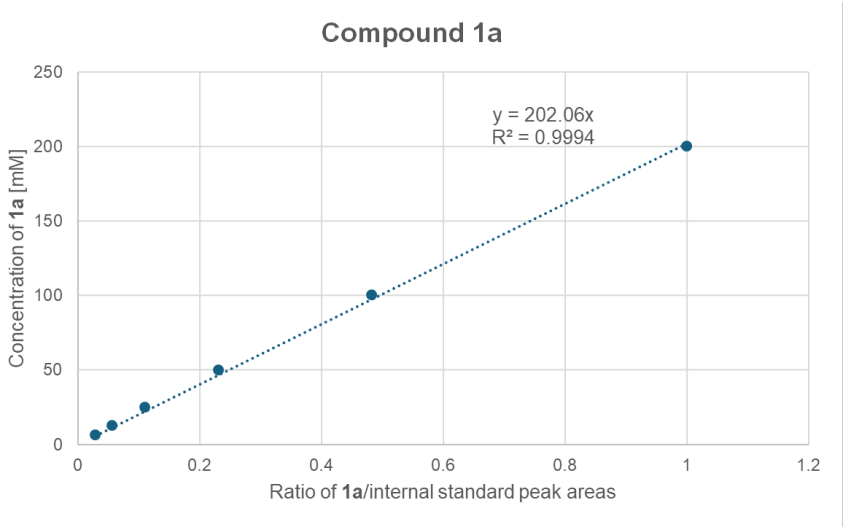


Figure A15. Calibration curve of compound 1a

A.11.3 Calibration Curve of 1b

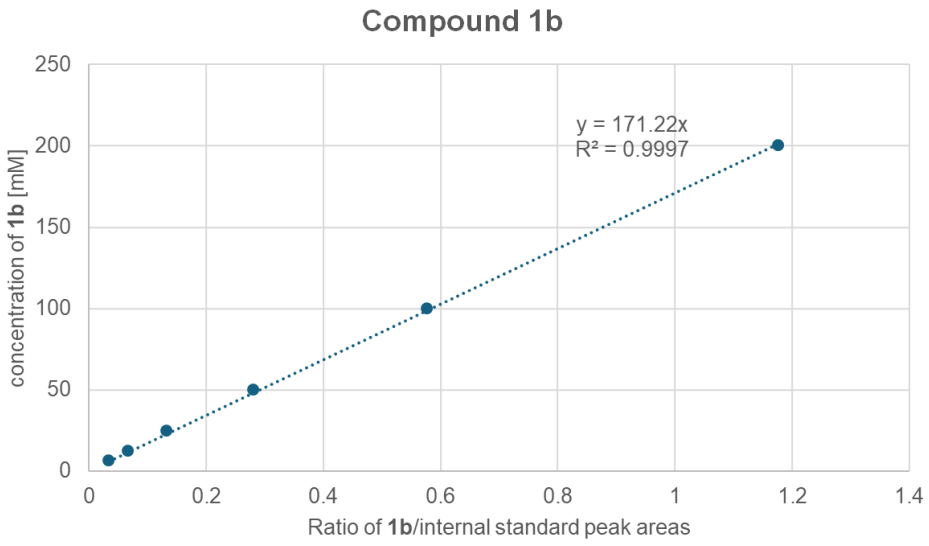


Figure A16. Calibration curve of compound 1b

A.11.4 Calibration Curve of 1c

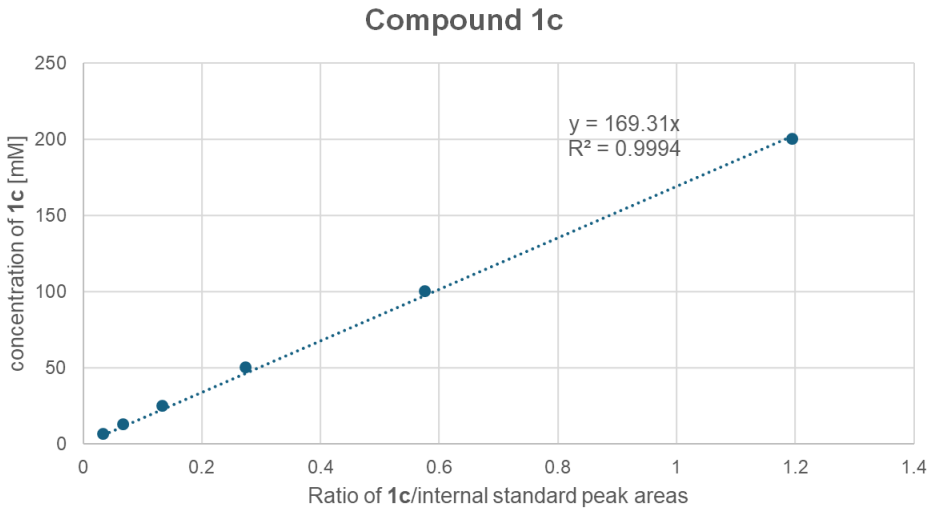


Figure A17. Calibration curve of compound 1c

A.11.5 Calibration Curve of 1d

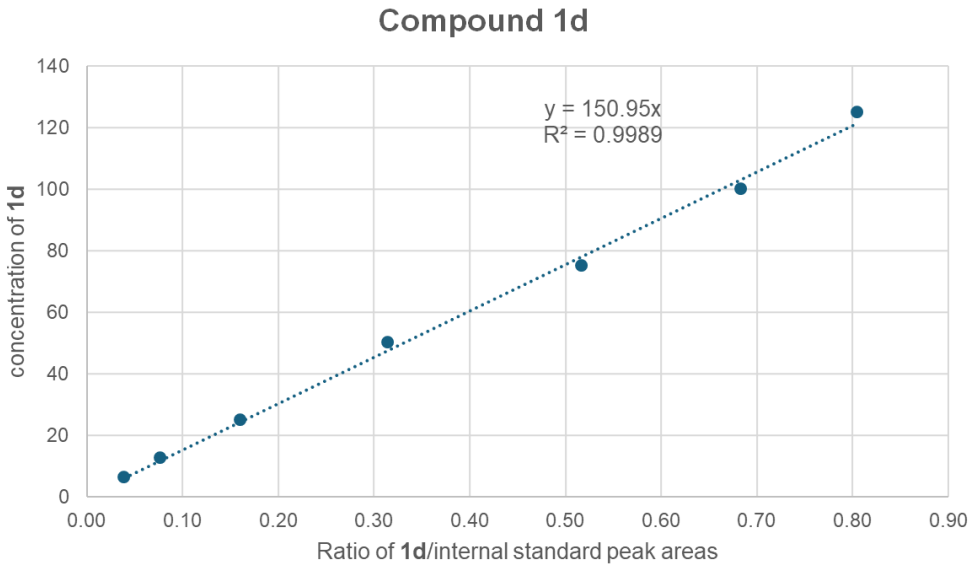


Figure A18. Calibration curve of compound 1d

A.11.6 Calibration Curve of 2a

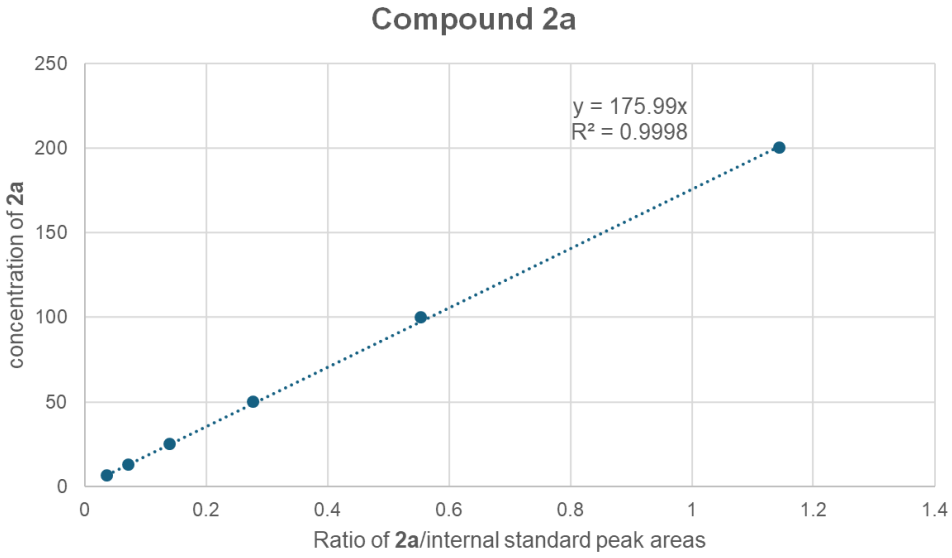


Figure A19. Calibration curve of compound 2a

A.11.7 Calibration Curve of 2b

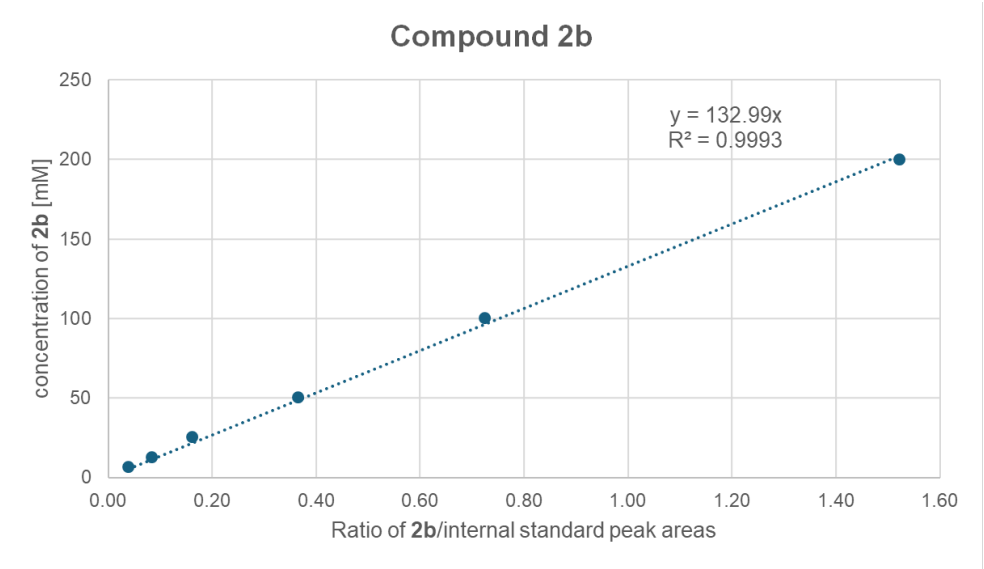


Figure A20. Calibration curve of compound 2b

A.11.8 Calibration Curve of 2c

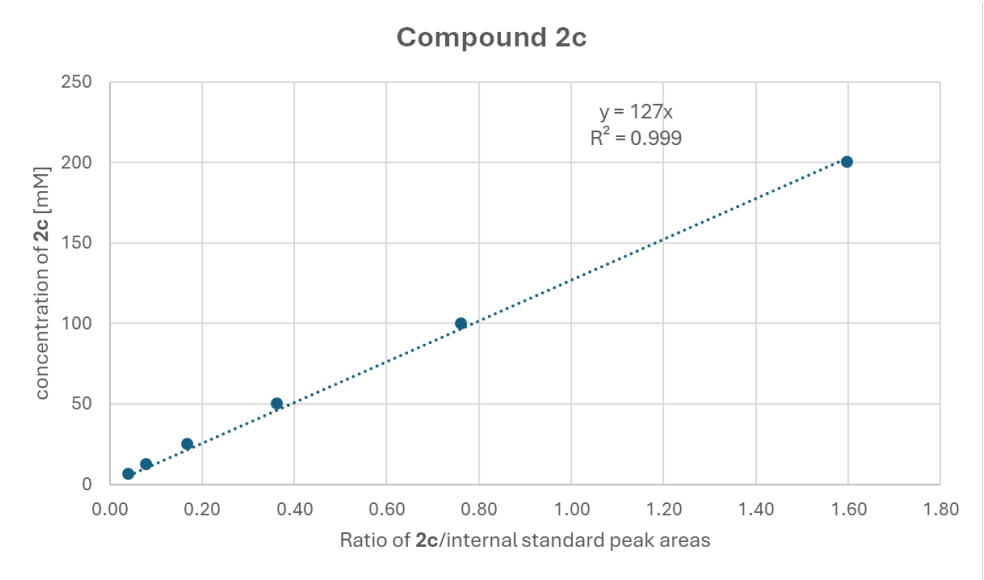


Figure 21. Calibration curve of compound 2c

A.11.9 Calibration Curve of 2d

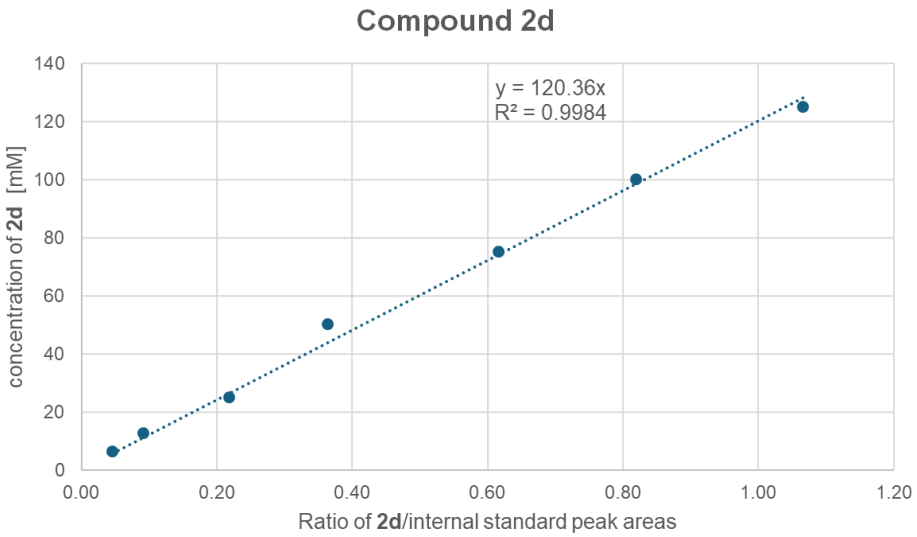


Figure A22. Calibration curve for compound 2d

A.12 GC-FID Traces for Yield Determination of Enzymatic Reactions

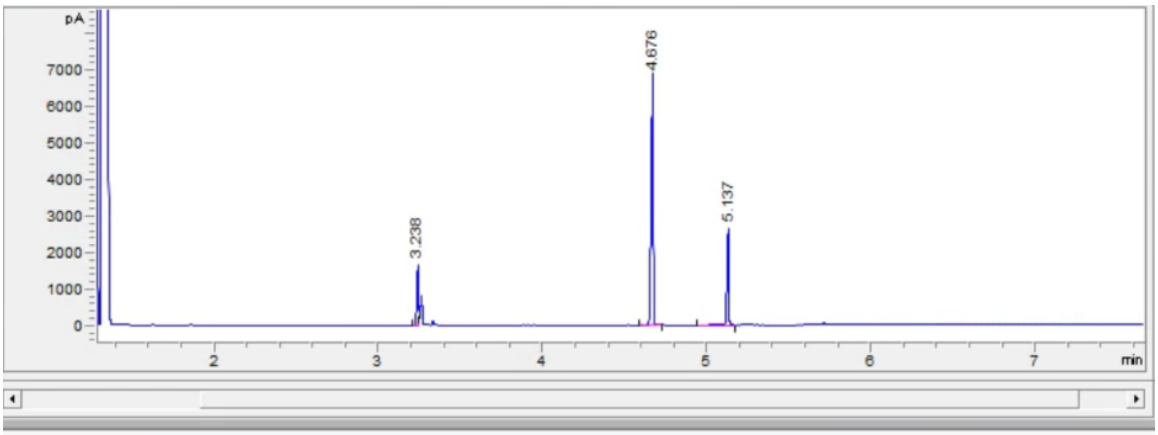
Enzymatic reactions were performed using whole-cell catalysts following the general procedure described in Section A.1.5. Reactions using lyophilized lysate were performed following the general procedure described in Section A.1.9. In all cases, the yield (%) of products 2a-d, as well as remaining starting material (%) of substrates 1a-d were quantified by comparison to an internal standard (1,2-diphenylethane) using GC-FID and the corresponding calibration curve (Section A.11). Analyses of all reactions were performed with an Agilent HP-5 GC column (1 μL injection; 2 min at 100 °C, ramp to 300 °C at 50 °C/min, hold at 300 °C for 2 min). The GC-FID traces shown in this section were used to quantify the evolution lineages for compounds 2a-c shown in Section A.4, the yields of compound 2d in Section A.6, and the yields of compounds 2a,b in Section A.7.

A.12.1 GC-FID traces for enzymatic reactions of 1a to 2a

Table A14. Retention times of 1a, 2a, and 1,2-diphenylethane

Compound	GC-FID Retention time (min)
1a	3.24
1,2-diphenylethane (standard)	4.68
2a	5.13

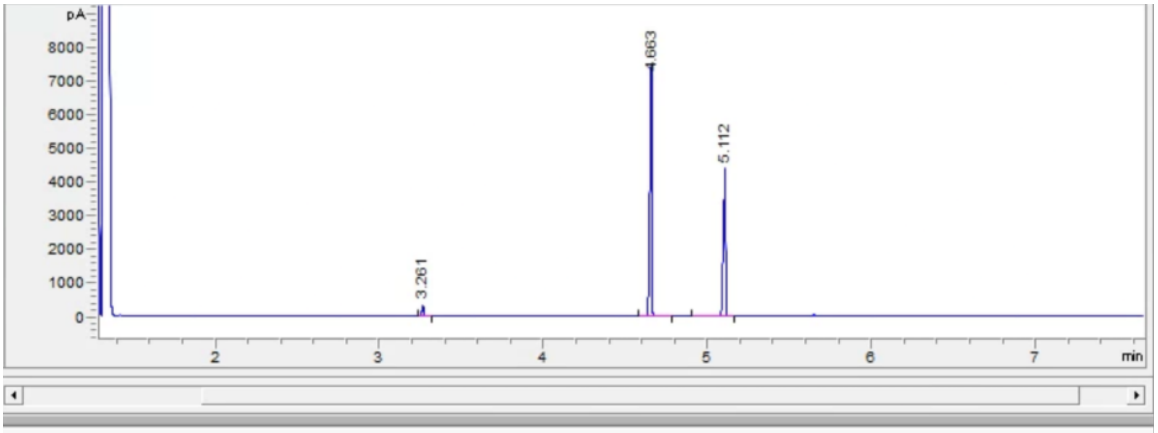
Reaction of 1a to 2a using ApePgb-xHC-5311 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.238	BV	1317.1	1640.1	0.0126	13.899	1.251
2	4.676	VB R	6070.5	6784.6	0.0144	64.057	1.765
3	5.137	VB R	2089.1	2631	0.0124	22.044	1.65

Figure A23. GC-FID trace for the enzymatic reaction of 1a to 2a using ApePgb-xHC-5311 in whole cell

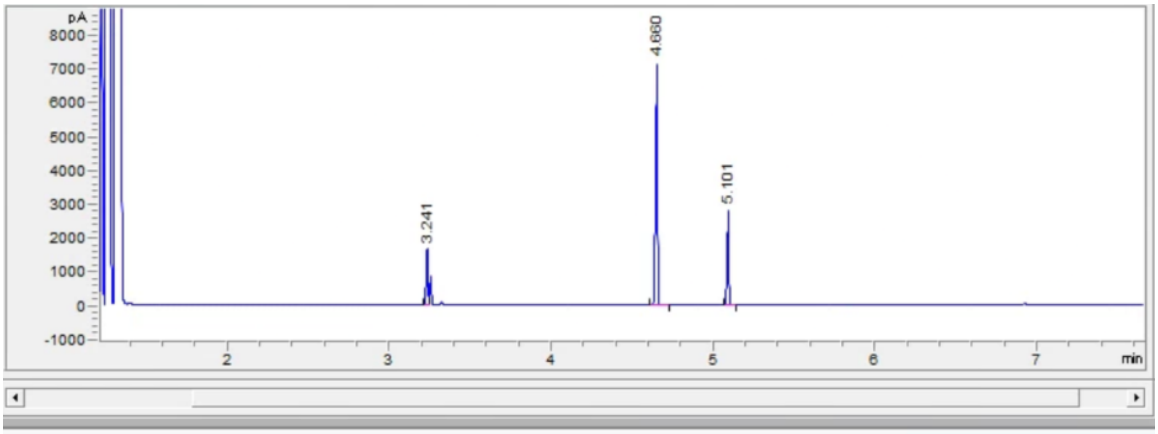
Reaction of 1a to 2a using ApePgb-xHC-5312 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.261	BV R	262.2	328.4	0.0117	2.399	0.922
2	4.663	VB R	6406.3	7090.3	0.0145	58.629	1.665
3	5.112	BV R	4258.5	4346.7	0.0146	38.972	2.132

Figure A24. GC-FID trace for the enzymatic reaction of 1a to 2a using ApePgb-xHC-5312 in whole cell

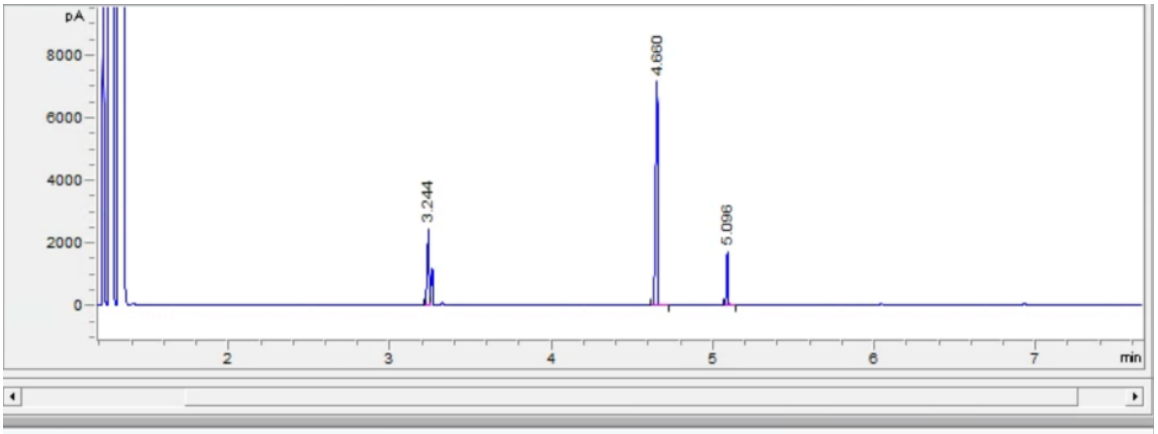
Reaction of 1a to 2a using TamPgb-xHC-5316 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.241	BV	1339.3	1662.1	0.0124	13.987	1.449
2	4.66	VB S	6166.7	7164.6	0.0124	64.401	2.211
3	5.101	VB S	2069.3	2790.3	0.0113	21.611	1.966

Figure A25. GC-FID trace for the enzymatic reaction of 1a to 2a using TamPgb-xHC-5316 in whole cell

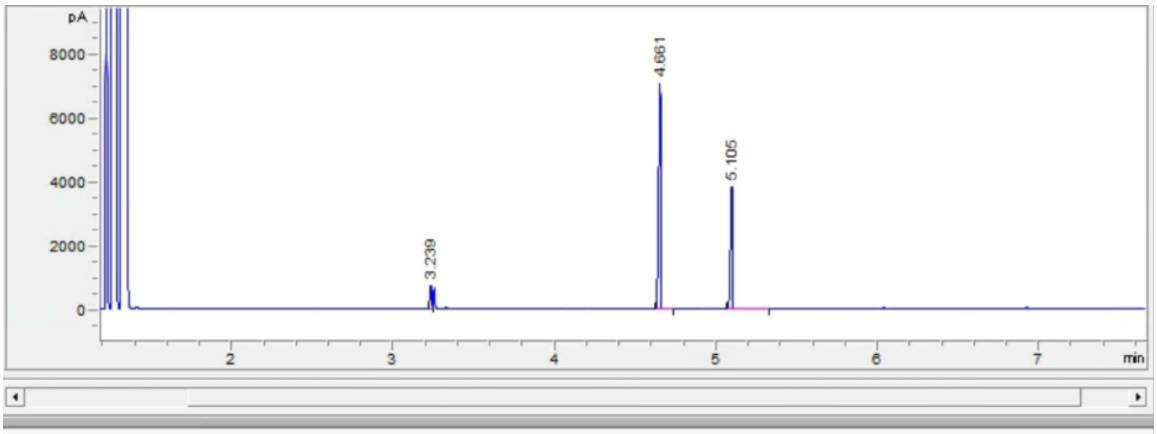
Reaction of 1a to 2a using TamPgb-xHC-5317 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.244	BV S	2094.9	2460.1	0.0131	22.472	1.765
2	4.66	VB S	6135.8	7131.3	0.0134	65.819	2.215
3	5.096	VB S	1091.5	1711.4	9.841E-3	11.709	1.481

Figure A26. GC-FID trace for the enzymatic reaction of 1a to 2a using TamPgb-xHC-5317 in whole cell

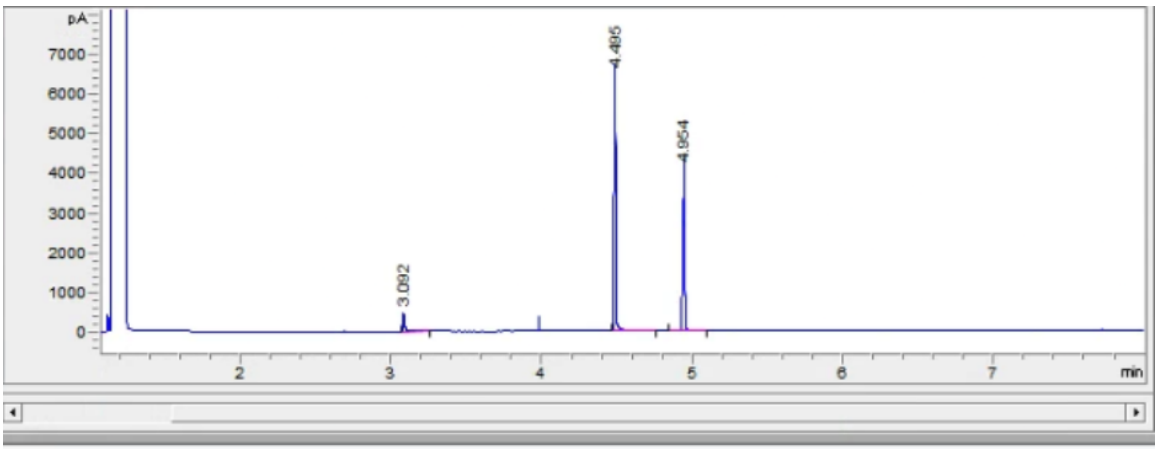
Reaction of 1a to 2a using TamPgb-xHC-5318 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.239	BV	571.9	757.2	0.012	5.617	1.14
2	4.661	VB S	6289.4	7060.6	0.0128	61.771	2.331
3	5.105	VB S	3320.6	3842.4	0.0127	32.613	2.332

Figure A27. GC-FID trace for the enzymatic reaction of 1a to 2a using TamPgb-xHC-5318 in whole cell

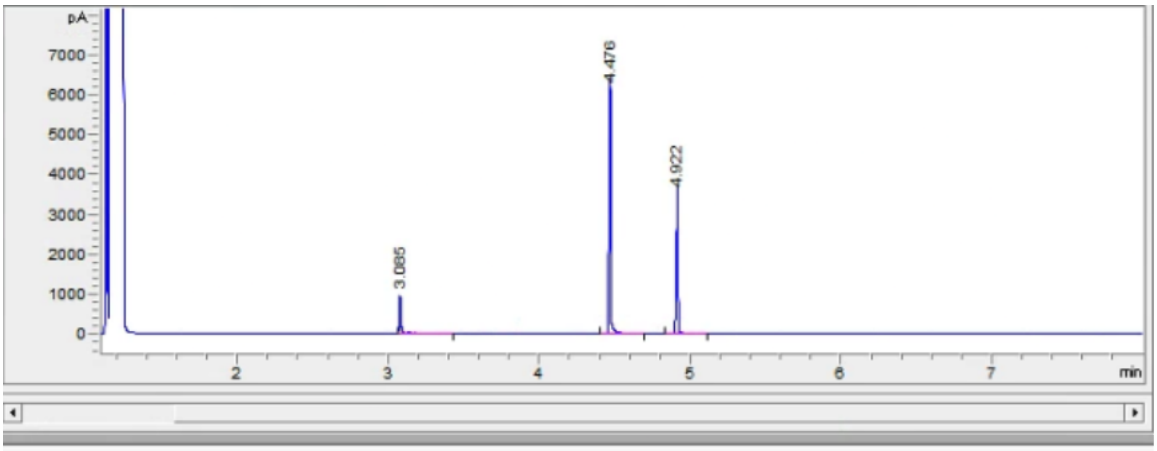
Reaction of 1a to 2a using ApePgb-xHC-5312 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.092	BV R	486	450.2	0.0133	5.081	0.566
2	4.495	VB	5214	6446.9	0.0134	54.504	1.039
3	4.954	BV R	3866.3	4059.4	0.0151	40.415	1.61

Figure A28. GC-FID trace for the enzymatic reaction of 1a to 2a using ApePgb-xHC-5312 in lyophilized lysate

Reaction of 1a to 2a using TamPgb-xHC-5318 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.085	BV R	971	938.6	0.013	11.208	0.632
2	4.476	VB R	4725.7	6116.6	0.0114	54.545	1.065
3	4.922	BV R	2967.1	3470.3	0.0139	34.247	1.315

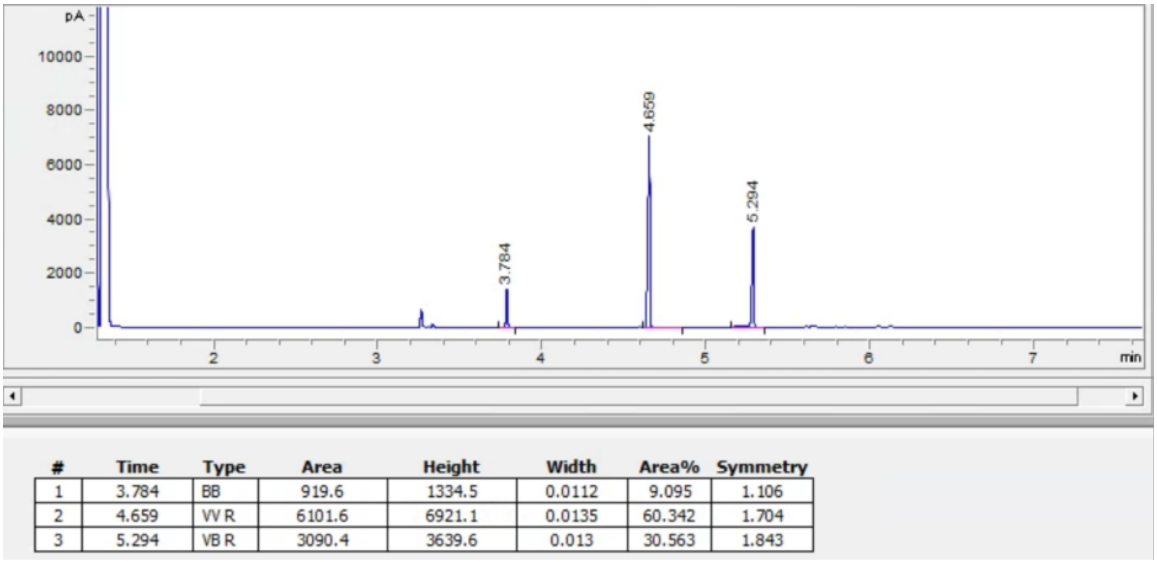
Figure A29. GC-FID trace for the enzymatic reaction of 1a to 2a using TamPgb-xHC-5318 in lyophilized lysate

A.12.2 GC-FID traces for enzymatic reactions of 1b to 2b

Table A15. Retention times of 1b, 2b, and 1,2-diphenylethane

Compound	GC-FID Retention time (min)
1b	3.78
1,2-diphenylethane (standard)	4.68
2b	5.29

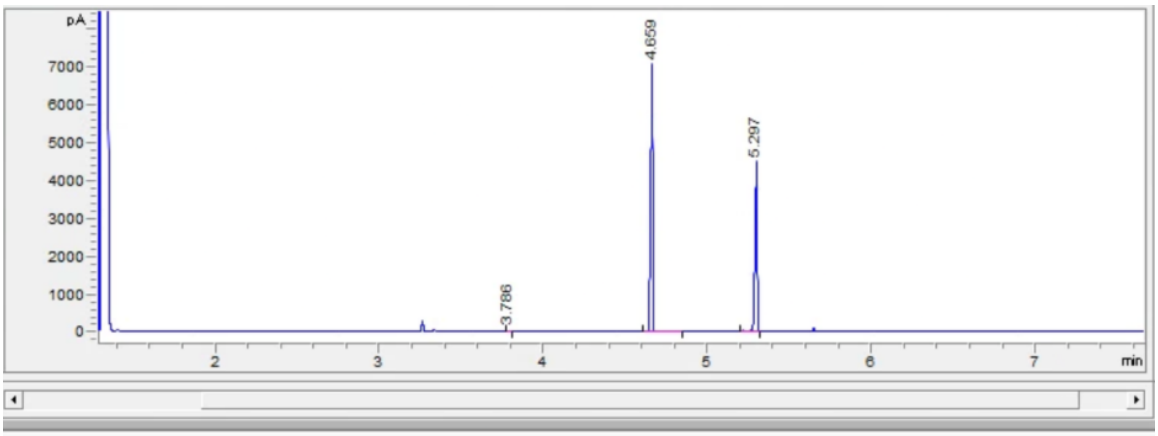
Reaction of 1b to 2b using ApePgb-xHC-5311 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.784	BB	919.6	1334.5	0.0112	9.095	1.106
2	4.659	VV R	6101.6	6921.1	0.0135	60.342	1.704
3	5.294	VB R	3090.4	3639.6	0.013	30.563	1.843

Figure A30. GC-FID trace for the enzymatic reaction of 1b to 2b using ApePgb-xHC-5311 in whole cell

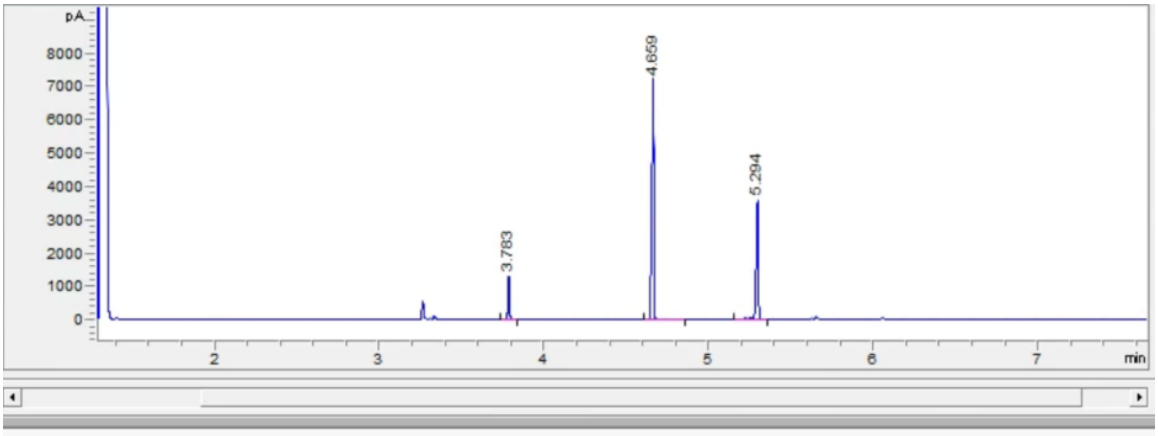
Reaction of 1b to 2b using ApePgb-xHC-5312 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.786	BB	1.9	2.6	0.012	0.019	0.737
2	4.659	VV R	6021.3	7006.6	0.0132	58.705	1.716
3	5.297	VB R	4233.7	4437.9	0.0143	41.276	1.899

Figure A31. GC-FID trace for the enzymatic reaction of 1b to 2b using ApePgb-xHC-5312 in whole cell

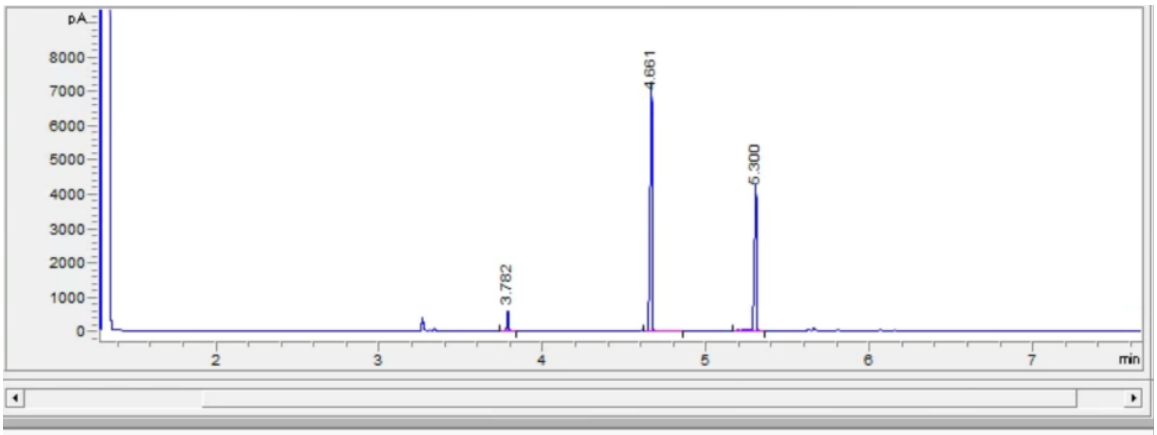
Reaction of 1b to 2b using ApePgb-xHC-5321 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.783	BB	820.4	1230.9	0.011	8.174	1.032
2	4.659	VV R	6035.8	7084.5	0.0123	60.136	1.714
3	5.294	VB R	3180.6	3533.4	0.0136	31.689	1.687

Figure A32. GC-FID trace for the enzymatic reaction of 1b to 2b using ApePgb-xHC-5321 in whole cell

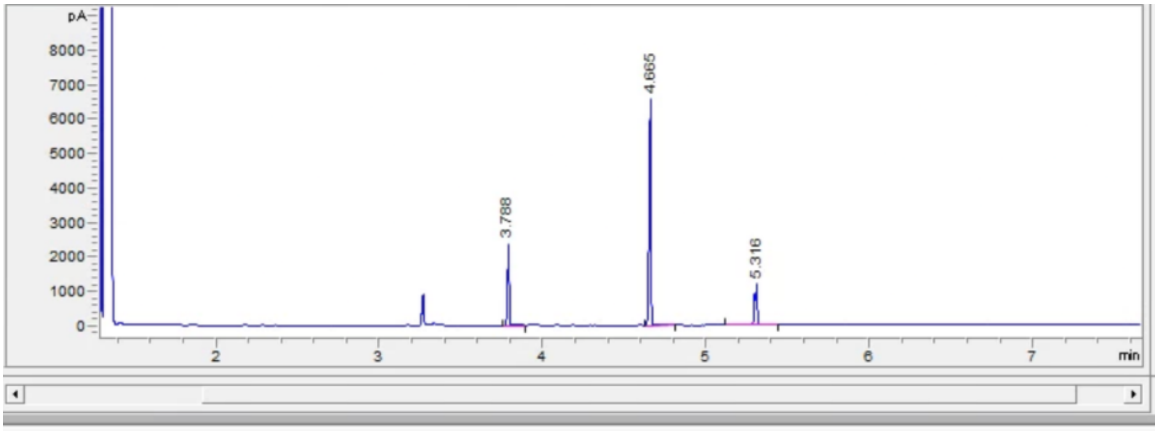
Reaction of 1b to 2b using ApePgb-xHC-5322 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.782	BB	364.7	551.5	0.0109	3.551	0.992
2	4.661	VV R	6063	6832.5	0.0135	59.039	1.613
3	5.3	VB R	3841.9	4089.8	0.014	37.410	1.973

Figure A33. GC-FID trace for the enzymatic reaction of 1b to 2b using ApePgb-xHC-5322 in whole cell

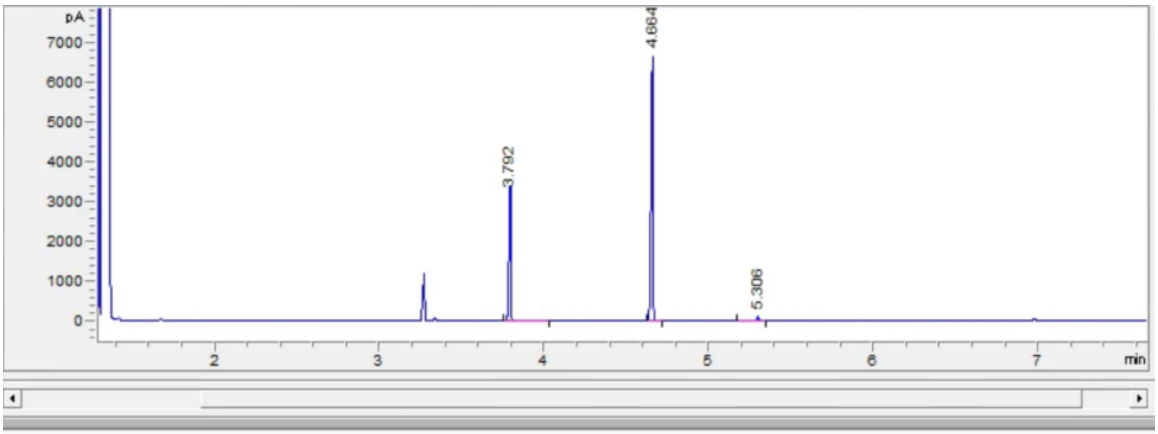
Reaction of 1b to 2b using TamPgb-xHC-5316 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.788	BB	1760.6	2336	0.012	20.647	1.351
2	4.665	BV R	5394.6	6514.4	0.0129	63.266	1.641
3	5.316	BV R	1371.8	1181.1	0.0159	16.087	2.825

Figure A34. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5316 in whole cell

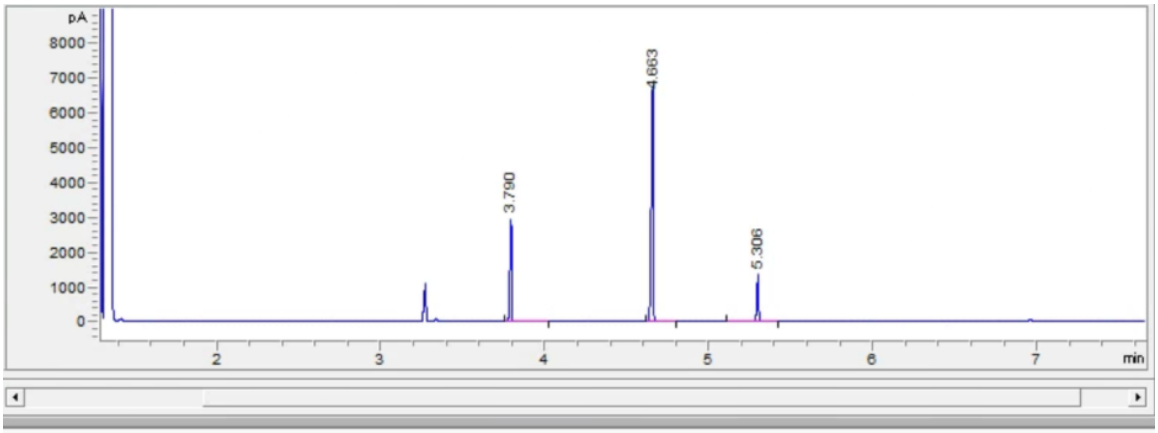
Reaction of 1b to 2b using TamPgb-xHC-5317 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.792	BV R	2782.1	3189	0.0134	32.728	1.546
2	4.664	BB	5625.7	6669.6	0.0138	66.180	1.754
3	5.306	VB R	92.8	110.7	0.0111	1.092	2.201

Figure A35. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5317 in whole cell

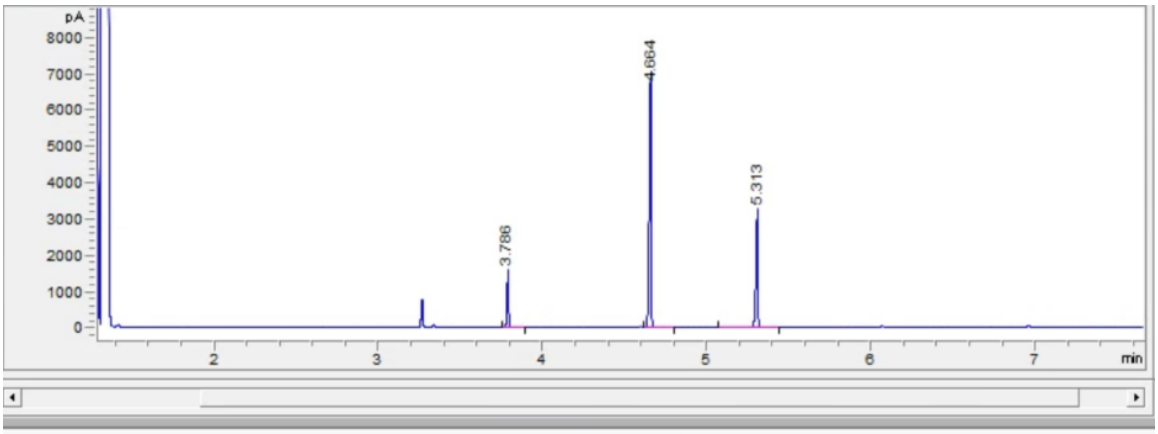
Reaction of 1b to 2b using TamPgb-xHC-5323 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.79	BV R	2276.7	2916.3	0.0123	25.231	1.441
2	4.663	BV R	5776.6	6475.5	0.0136	64.020	1.649
3	5.306	VB R	969.9	1316.5	0.0109	10.749	1.498

Figure A36. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5323 in whole cell

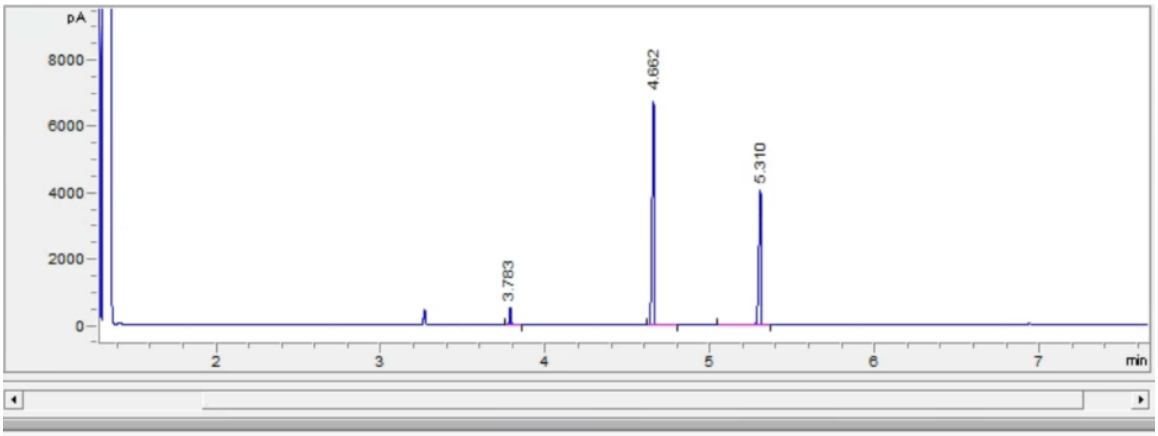
Reaction of 1b to 2b using TamPgb-xHC-5324 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.786	BB	1045.8	1527.1	0.0112	10.811	1.093
2	4.664	BV R	5902.4	6595.6	0.0136	61.012	1.729
3	5.313	BV R	2725.9	3179.5	0.0124	28.177	1.743

Figure A37. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5324 in whole cell

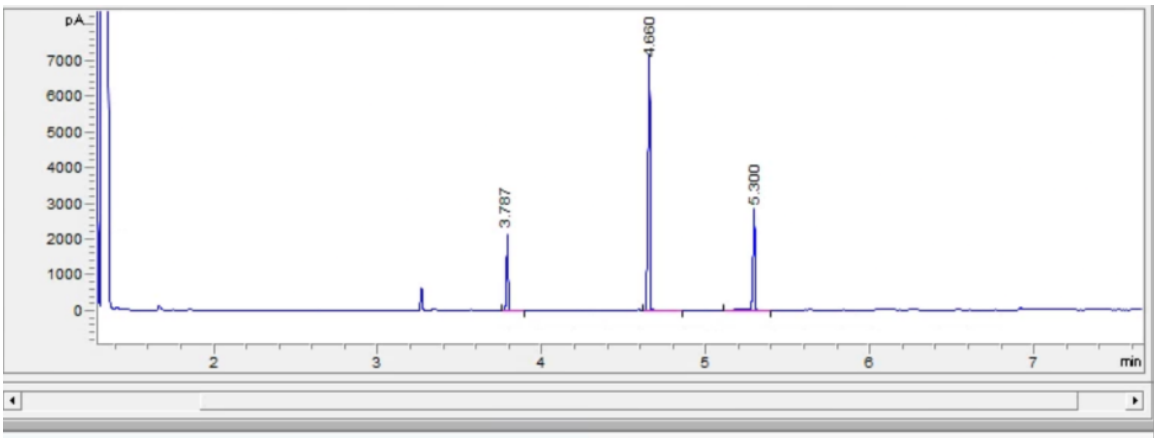
Reaction of 1b to 2b using TamPgb-xHC-5325 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.783	BB	335.4	497.9	0.0111	3.456	0.892
2	4.662	BV R	5760.4	6875.2	0.0138	59.365	1.716
3	5.31	BB	3607.6	4051.1	0.0136	37.179	1.957

Figure A38. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5325 in whole cell

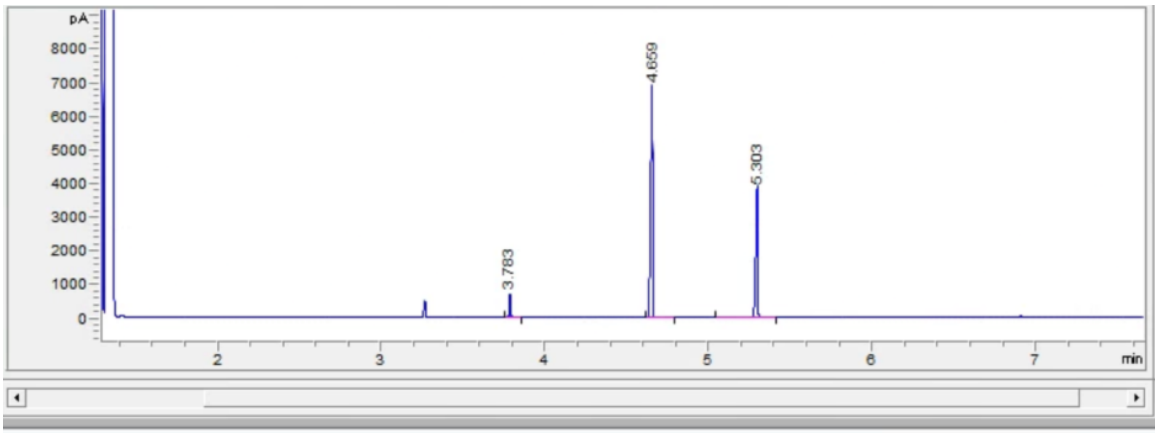
Reaction of 1b to 2b using TamPgb-xHC-5326 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.787	BV	1485.1	2119.1	0.0114	15.323	1.272
2	4.66	BV R	6002.5	6884	0.0142	61.932	1.586
3	5.3	BV R	2204.4	2732.5	0.0118	22.744	1.638

Figure A39. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5326 in whole cell

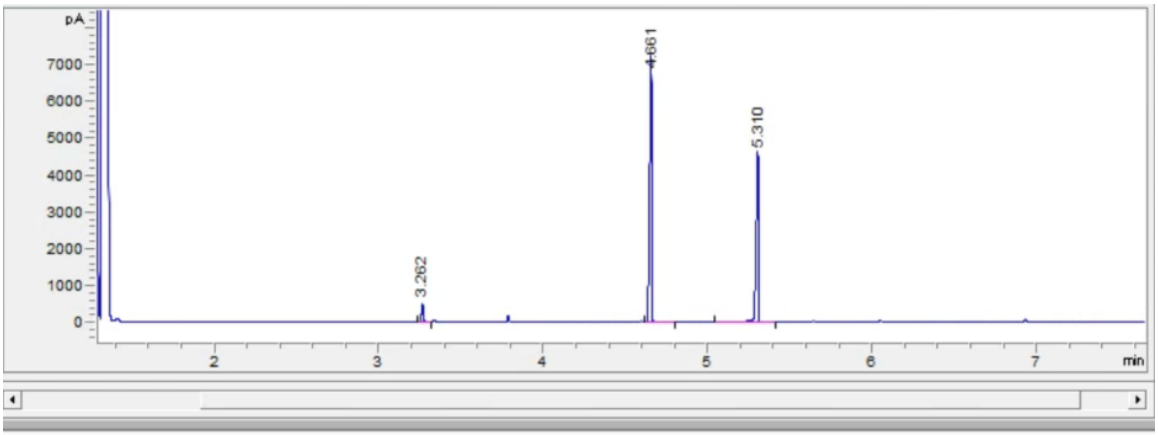
Reaction of 1b to 2b using TamPgb-xHC-5327 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.783	BB	443.8	664.4	0.011	4.529	0.924
2	4.659	BV R	5797.4	6796.6	0.0131	59.168	1.684
3	5.303	BV R	3557.1	3745.8	0.0151	36.303	1.828

Figure A40. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5327 in whole cell

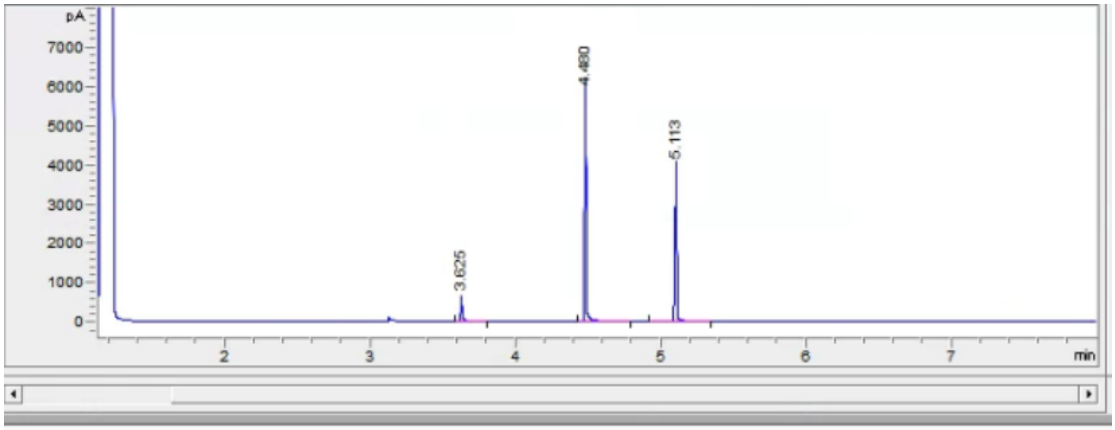
Reaction of 1b to 2b using TamPgb-xHC-5328 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.262	BV R	365.7	503.2	0.0116	3.447	1.027
2	4.661	BV R	6044.8	6708.3	0.0145	56.977	1.662
3	5.31	BV R	4198.7	4567.9	0.0131	39.576	2.073

Figure A41. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5328 in whole cell

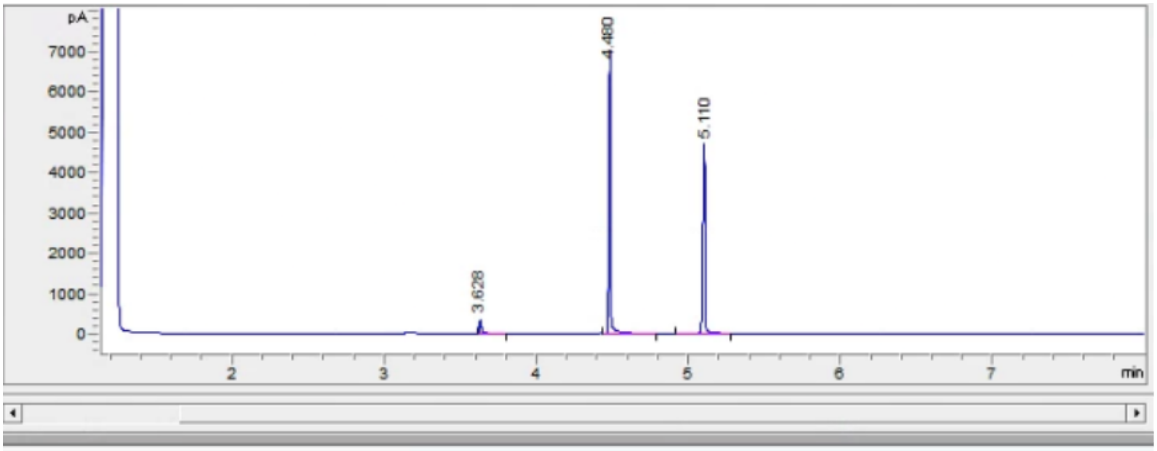
Reaction of 1b to 2b using TamPgb-xHC-5312 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.625	BV R	567	641	0.0133	6.304	0.501
2	4.48	VB	4749.2	5841.2	0.0127	52.801	0.875
3	5.113	VB R	3678.3	3970.7	0.014	40.895	1.326

Figure A42. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5312 in whole cell

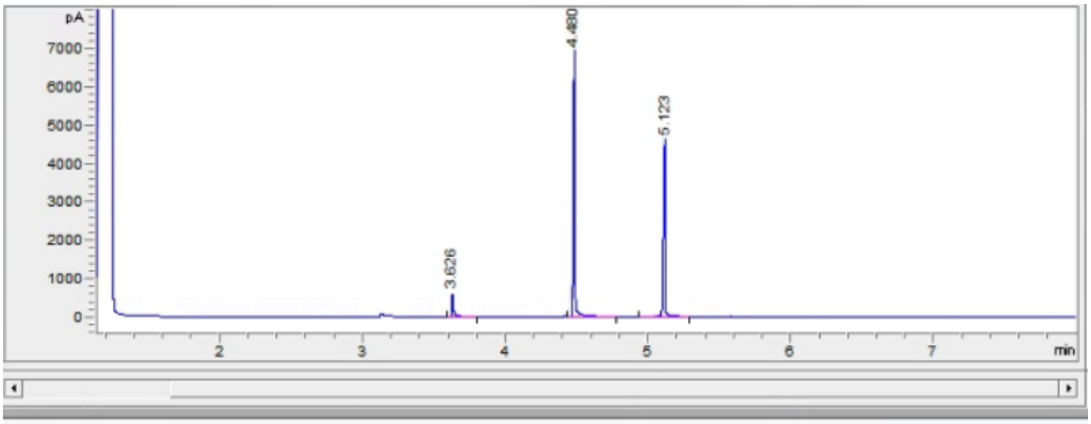
Reaction of 1b to 2b using TamPgb-xHC-5322 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.628	BV R	334.7	335.4	0.0136	3.284	0.416
2	4.48	VB	5515.5	6622.1	0.0129	54.108	0.961
3	5.11	VB R	4343.3	4661.2	0.0149	42.608	1.49

Figure A43. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5322 in whole cell

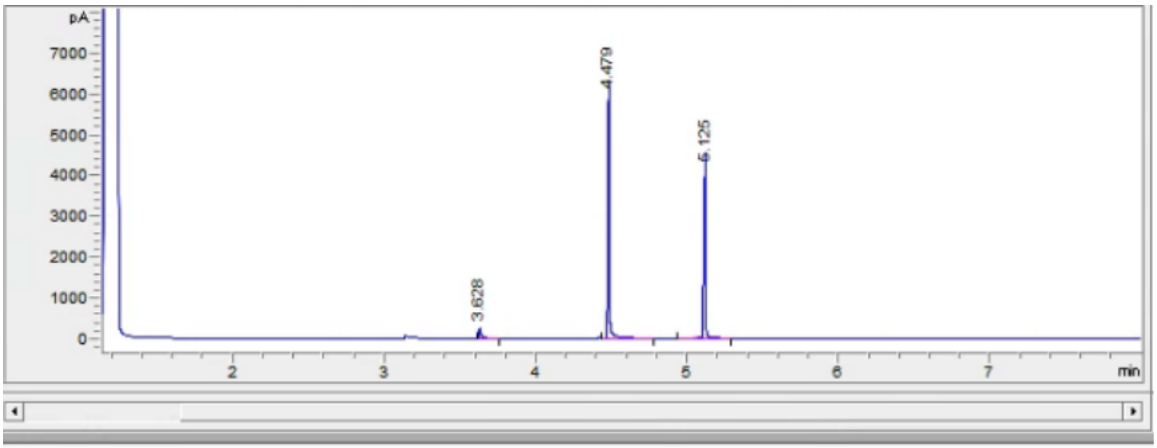
Reaction of 1b to 2b using TamPgb-xHC-5325 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.626	BV R	508.7	562.2	0.0127	5.035	0.468
2	4.48	VB	5364.2	6962	0.0122	53.095	1.021
3	5.123	VB R	4230.1	4560.4	0.0156	41.870	1.722

Figure A44. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5325 in lyophilized lysate

Reaction of 1b to 2b using TamPgb-xHC-5328 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.628	BV R	245.7	242.6	0.014	2.779	0.402
2	4.479	VB	4708.4	5947.5	0.0116	53.263	0.974
3	5.125	VB R	3885.9	4176.3	0.014	43.958	1.532

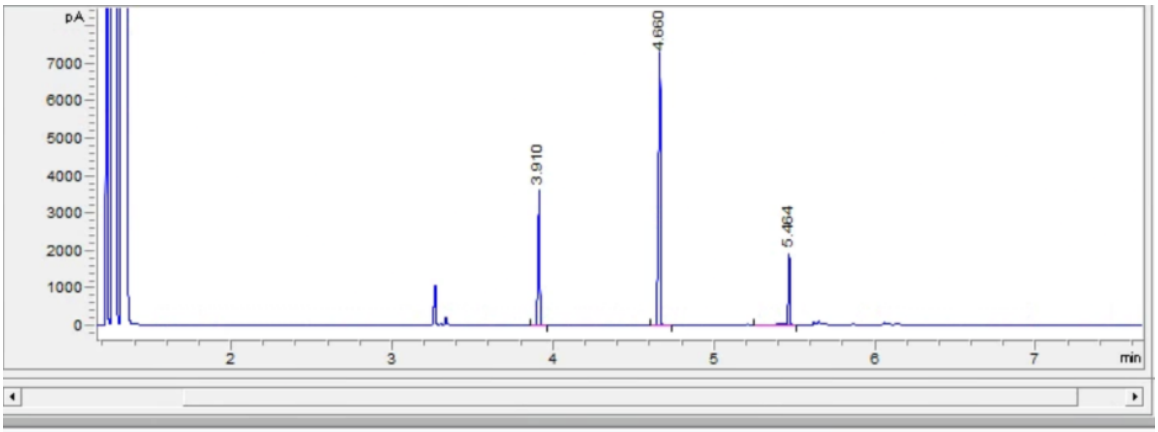
Figure A45. GC-FID trace for the enzymatic reaction of 1b to 2b using TamPgb-xHC-5328 in lyophilized lysate

A.12.3 GC-FID traces for enzymatic reactions of 1c to 2c

Table A16. Retention times of 1c, 2c, and 1,2-diphenylethane

Compound	GC-FID Retention time (min)
1c	3.91
1,2-diphenylethane (standard)	4.68
2c	5.46

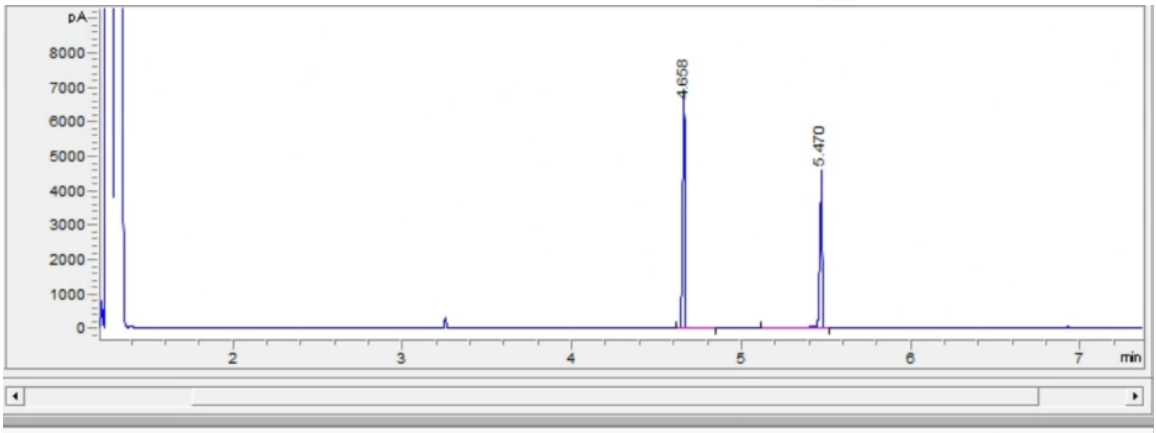
Reaction of 1c to 2c using ApePgb-xHC-5311 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.91	BV R	2803.7	3552.6	0.0124	26.808	1.523
2	4.66	BB	6315.7	7148	0.0135	60.388	1.772
3	5.464	VB R	1339.1	1895.3	0.0114	12.804	1.472

Figure A46. GC-FID trace for the enzymatic reaction of 1c to 2c using ApePgb-xHC-5311 in whole cell

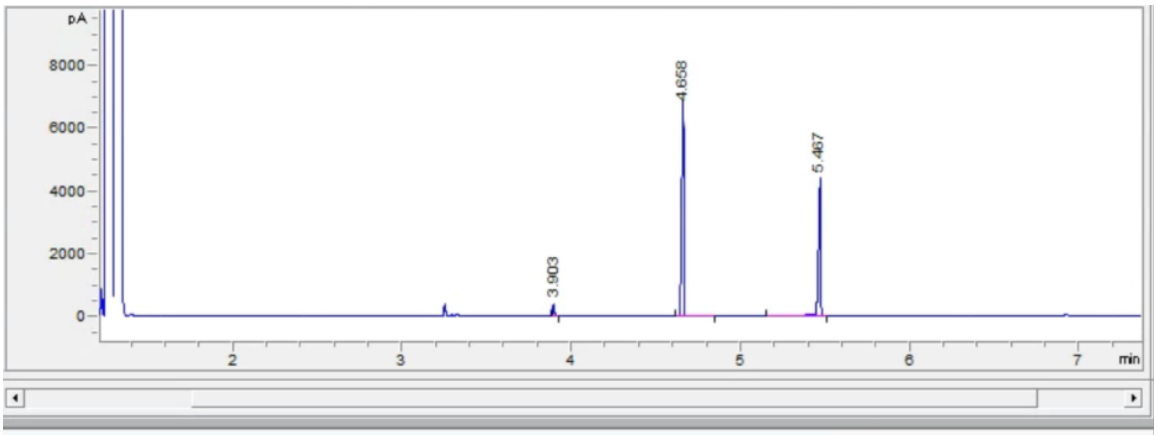
Reaction of 1c to 2c using ApePgb-xHC-5313 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.658	VV R	5690.9	6463.9	0.0135	55.437	1.617
2	5.47	VB R	4574.6	4520	0.0157	44.563	2.325

Figure A47. GC-FID trace for the enzymatic reaction of 1c to 2c using ApePgb-xHC-5313 in whole cell

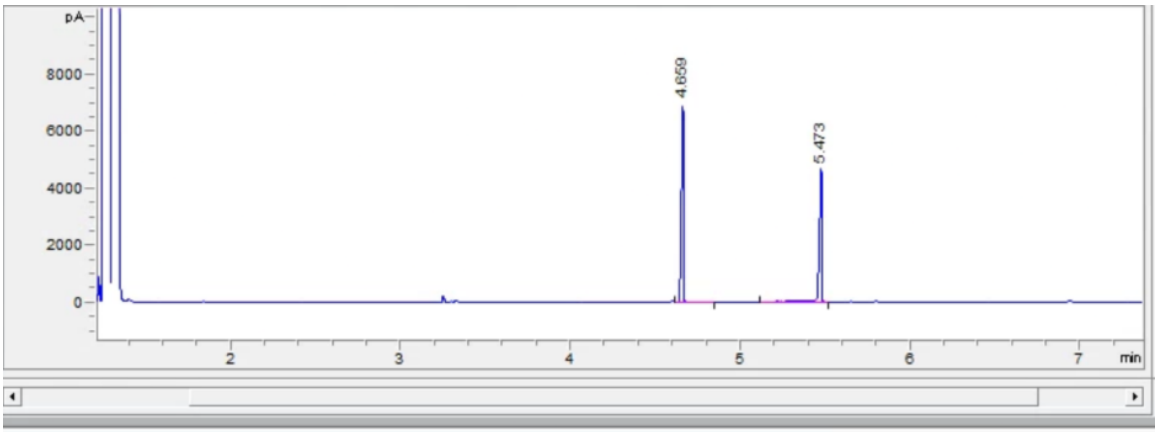
Reaction of 1c to 2c using ApePgb-xHC-5314 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.903	BB	243.6	380.5	9.8691E-3	2.349	0.97
2	4.658	VV R	5844.7	6629.1	0.0143	56.361	1.565
3	5.467	VB R	4281.9	4375.8	0.0153	41.290	2.295

Figure A48. GC-FID trace for the enzymatic reaction of 1c to 2c using ApePgb-xHC-5314 in whole cell

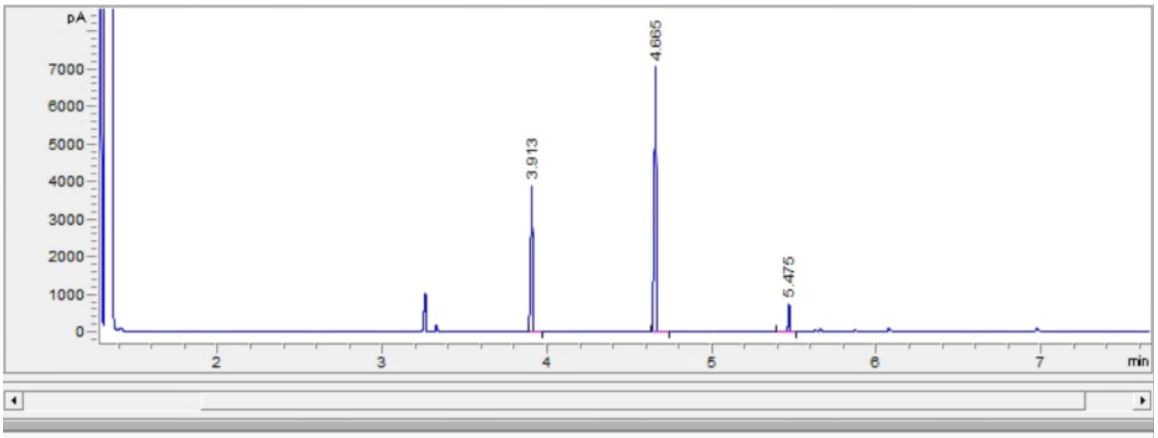
Reaction of 1c to 2c using ApePgb-xHC-5315 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.659	VV R	5811.8	6928.4	0.013	55.685	1.729
2	5.473	VB R	4625	4594.9	0.014	44.315	2.298

Figure 49. GC-FID trace for the enzymatic reaction of 1c to 2c using ApePgb-xHC-5315 in whole cell

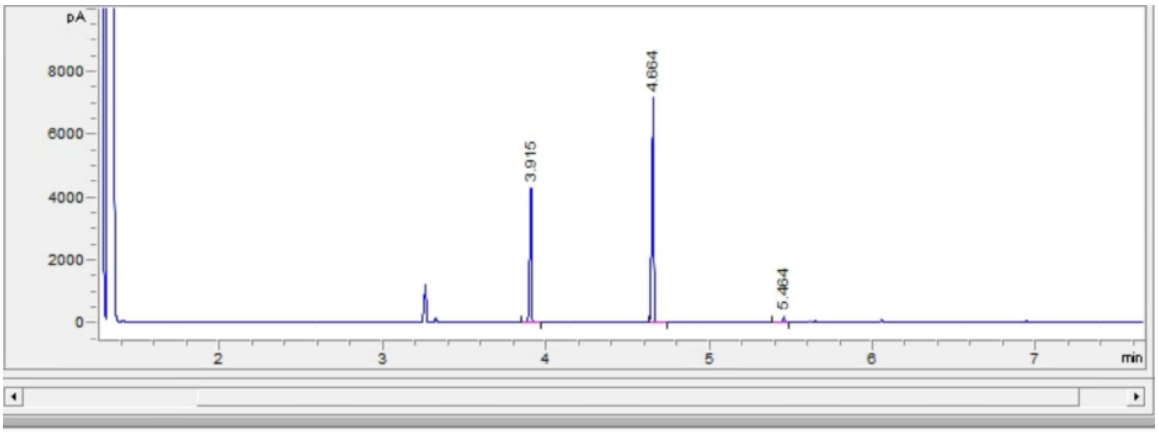
Reaction of 1c to 2c using TamPgb-xHC-5316 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.913	VB S	3082.7	3873.4	0.0121	32.673	1.9
2	4.665	VB S	5902.4	7028.6	0.0124	62.559	2.211
3	5.475	BB	449.9	728.5	9.6047E-3	4.768	1.216

Figure A50. GC-FID trace for the enzymatic reaction of 1c to 2c using TamPgb-xHC-5316 in whole cell

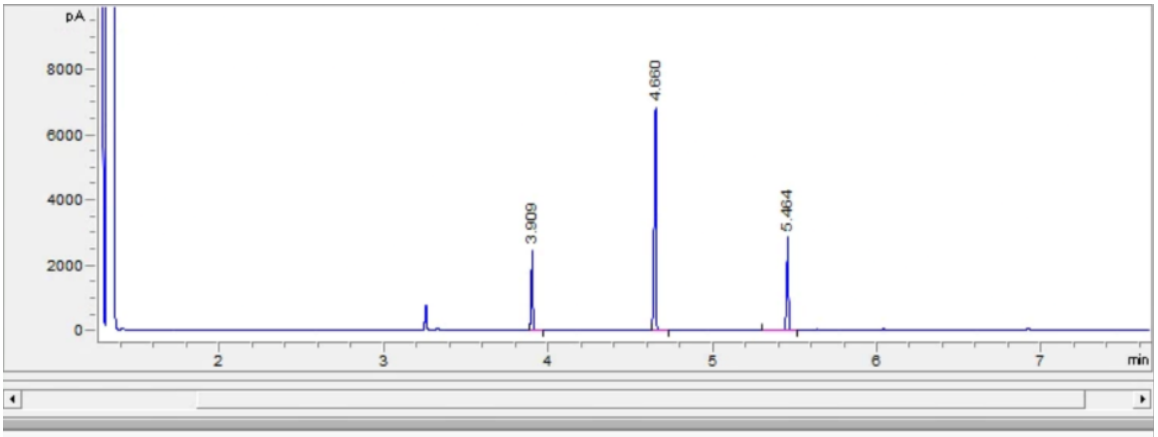
Reaction of 1c to 2c using TamPgb-xHC-5317 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.915	BB S	3560.4	4281.1	0.0125	36.491	2.06
2	4.664	VB S	6071.7	7147.7	0.0125	62.230	2.237
3	5.464	BB	124.8	206.5	9.4509E-3	1.279	1.057

Figure A51. GC-FID trace for the enzymatic reaction of 1c to 2c using TamPgb-xHC-5317 in whole cell

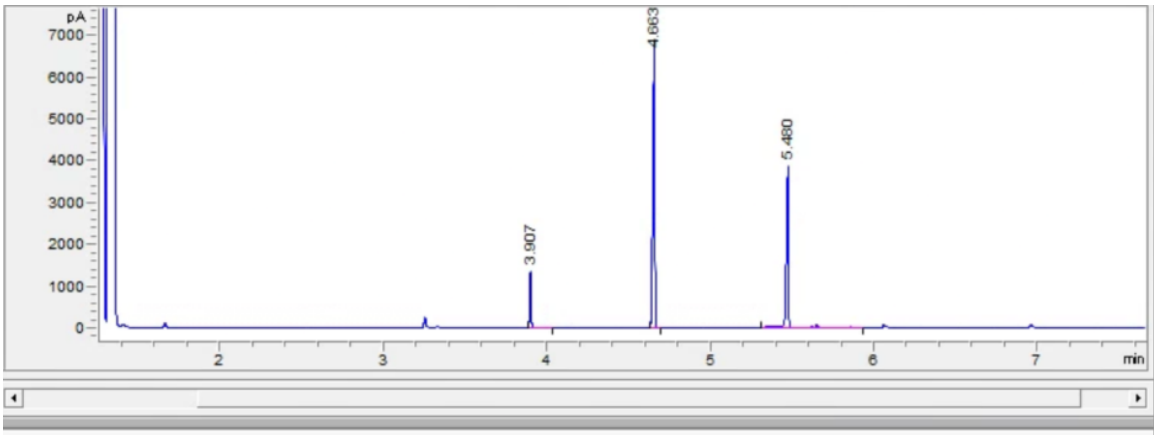
Reaction of 1c to 2c using TamPgb-xHC-5318 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.909	VB S	1688	2440.3	0.0109	17.201	1.489
2	4.66	VB S	5828.4	6808.1	0.013	59.392	2.183
3	5.464	VB S	2297	2847.3	0.012	23.407	2.091

Figure A52. GC-FID trace for the enzymatic reaction of 1c to 2c using TamPgb-xHC-5318 in whole cell

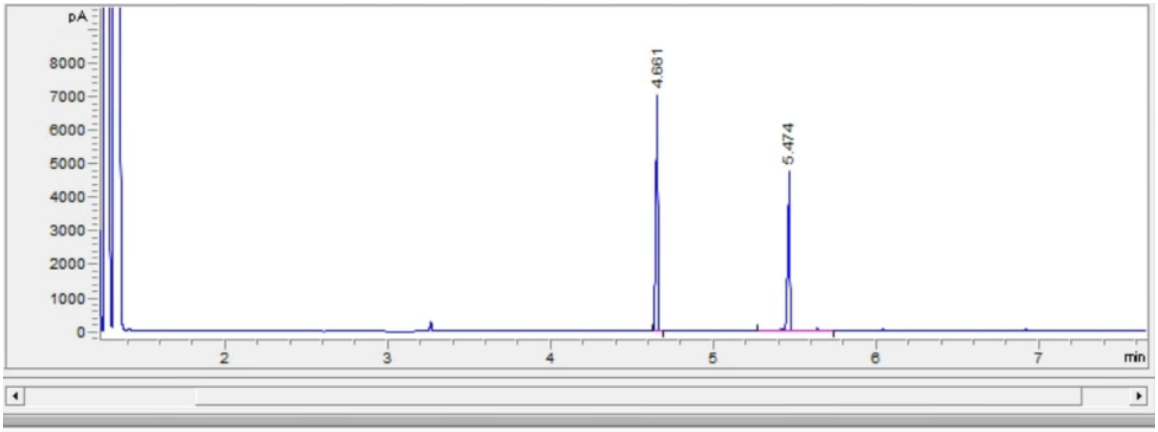
Reaction of 1c to 2c using TamPgb-xHC-5319 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.907	VB S	868.9	1351.9	0.0103	8.397	1.162
2	4.663	VB S	5919.8	6806.9	0.0143	57.206	2.283
3	5.48	VB S	3559.4	3849.8	0.0144	34.396	2.283

Figure A53. GC-FID trace for the enzymatic reaction of 1c to 2c using TamPgb-xHC-5319 in whole cell

Reaction of 1c to 2c using TamPgb-xHC-5320 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	4.661	VB S	5895.8	7023.5	0.0136	55.982	2.22
2	5.474	VB S	4635.9	4763.8	0.0138	44.018	2.784

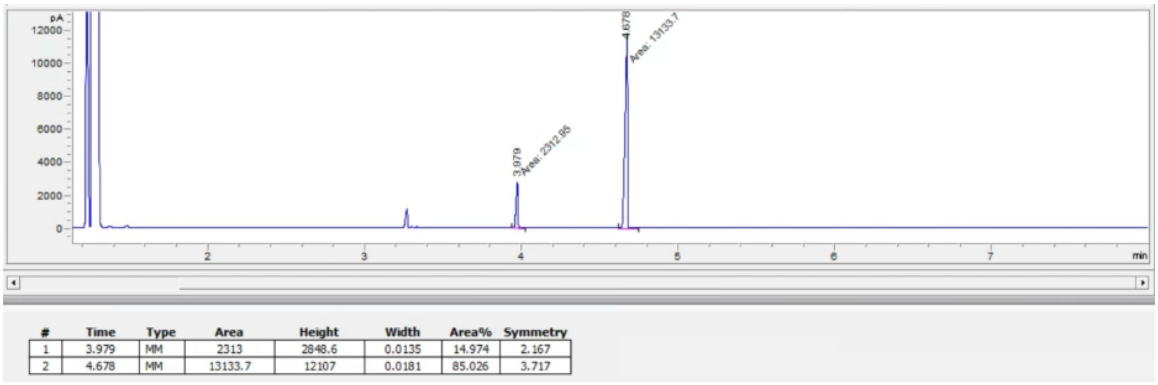
Figure A54. GC-FID trace for the enzymatic reaction of 1c to 2c using TamPgb-xHC-5320 in whole cell

A.12.4 Reaction of 1d to 2d using ApePgb-xHC-5315 in whole cell

Table A17. Retention times of 1d, 2d, and 1,2-diphenylethane

Compound	GC-FID Retention time (min)
1d	1.47
1,2-diphenylethane (standard)	4.68
2d	3.97

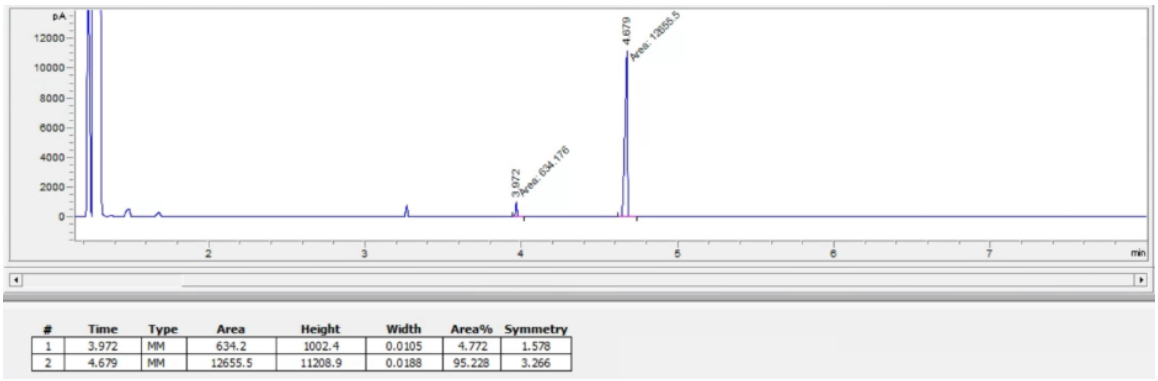
Reaction of 1d to 2d using ApePgb-xHC-5315 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.979	MM	2313	2848.6	0.0135	14.974	2.167
2	4.678	MM	13133.7	12107	0.0181	85.026	3.717

Figure A55. GC-FID trace for the enzymatic reaction of 1d to 2d using ApePgb-xHC-5315 in whole cell

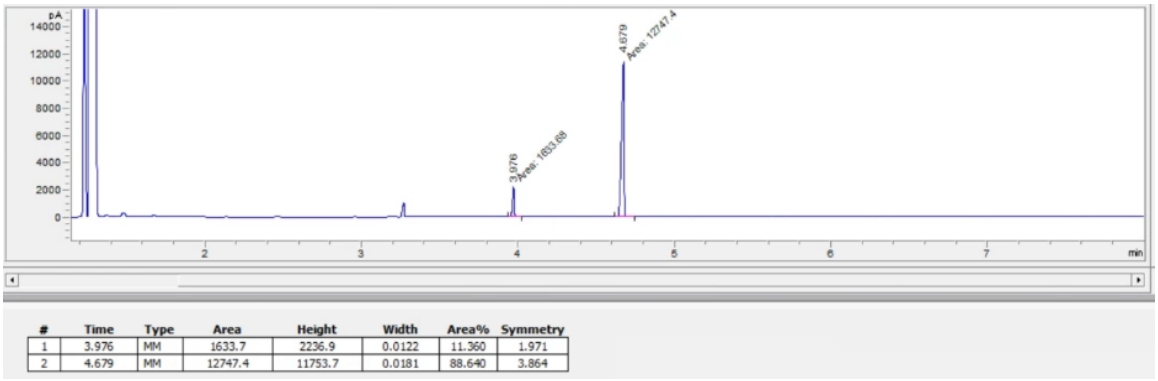
Reaction of 1d to 2d using TamPgb-xHC-5319 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.972	MM	634.2	1002.4	0.0105	4.772	1.578
2	4.679	MM	12655.5	11208.9	0.0188	95.228	3.266

Figure A56. GC-FID trace for the enzymatic reaction of 1d to 2d using TamPgb-xHC-5319 in whole cell

Reaction of 1d to 2d using TamPgb-xHC-5320 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	3.976	MM	1633.7	2236.9	0.0122	11.360	1.971
2	4.679	MM	12747.4	11753.7	0.0181	88.640	3.864

Figure A57. GC-FID trace for the enzymatic reaction of 1d to 2d using TamPgb-xHC-5320 in whole cell

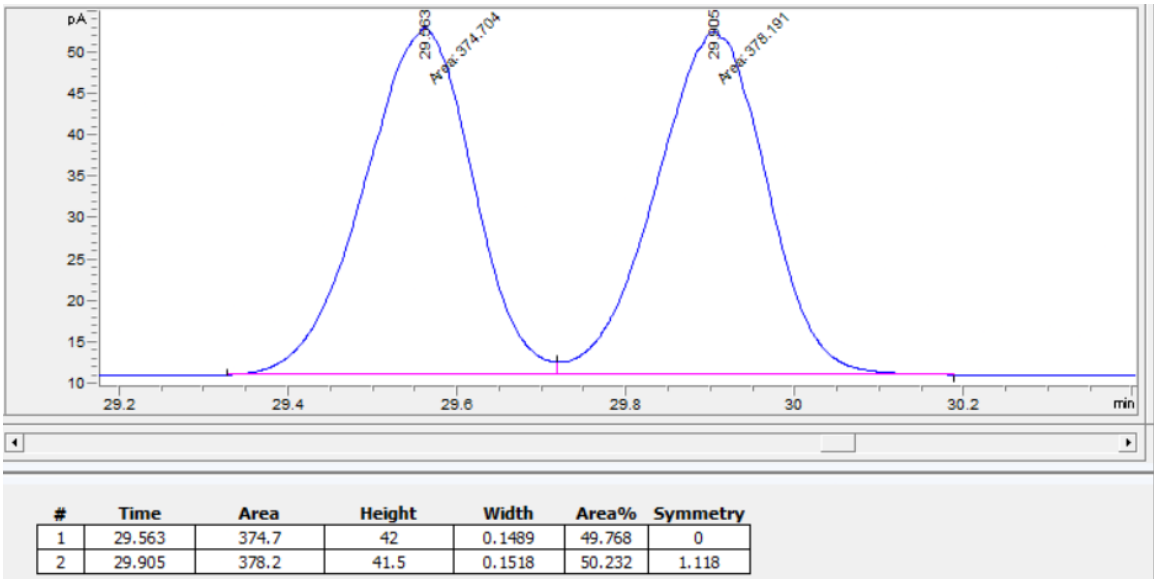
A.13 GC-FID Traces for Enantioselectivity Determination

Enzymatic reactions on analytic scale were performed following the general procedure described in Section A.1.5. The enantiomeric ratio (er) of products 2a-d was determined using GC-FID equipped with a CYCLOSIL-B column (Agilent) as the chiral stationary phase. The GC-FID traces shown in this section were used to determine the er's of the final variants shown in the evolution lineages in Section A.4, the er's of compound 2d in Section A.6, and the er's of compounds 2a-b in Section A.7. Specific separation conditions for each product are detailed in their respective sections below.

A.13.1 Chiral GC-FID traces for enzymatic reactions of 1a to 2a

Analysis was performed using an Agilent CYCLOSIL-B column (3 µL injection; 0 min at 50 °C, ramp to 170 °C at 10 °C/min and hold for 19 minutes; ramp to 200 °C at 30 °C/minute and hold for 8 minutes.

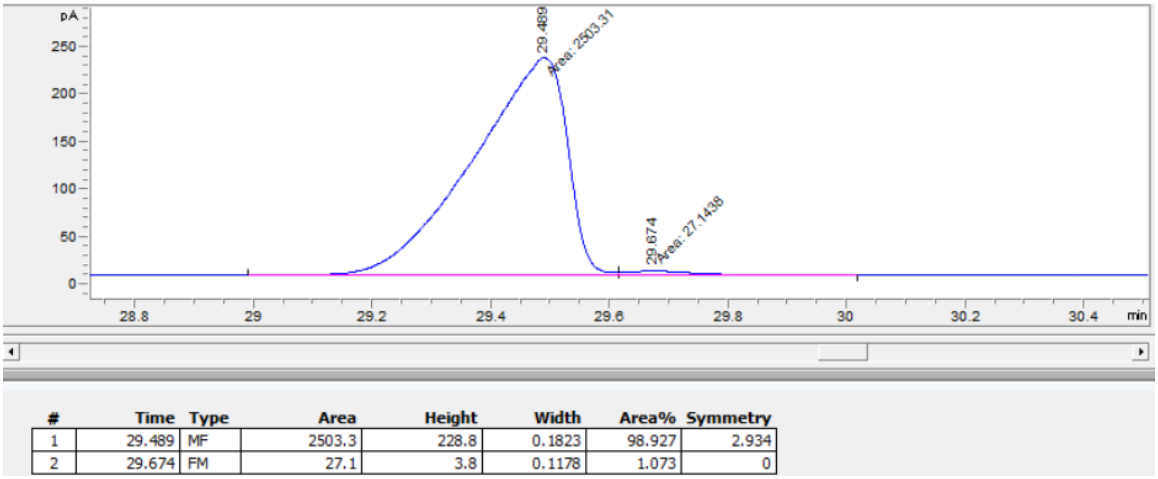
Racemic standard of 2a



#	Time	Area	Height	Width	Area%	Symmetry
1	29.563	374.7	42	0.1489	49.768	0
2	29.905	378.2	41.5	0.1518	50.232	1.118

Figure A58. Chiral GC-FID trace for racemic standard of 2a

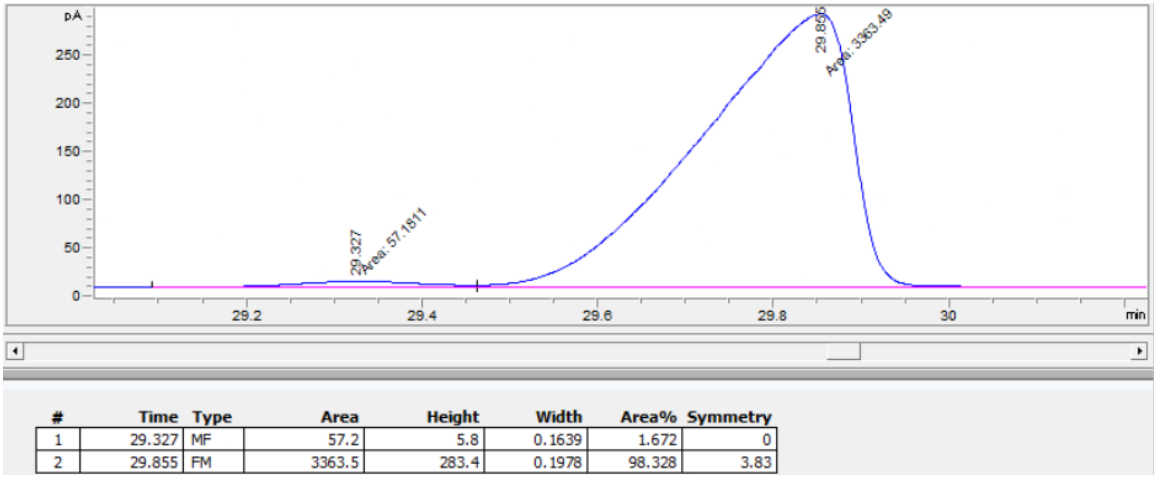
Chiral GC-FID trace of 2a from the enzymatic reaction using TamPgb-xHC-5318 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	29.489	MF	2503.3	228.8	0.1823	98.927	2.934
2	29.674	FM	27.1	3.8	0.1178	1.073	0

Figure A59. Chiral GC-FID trace of 2a from the enzymatic reaction using TamPgb-xHC-5318 in whole cell

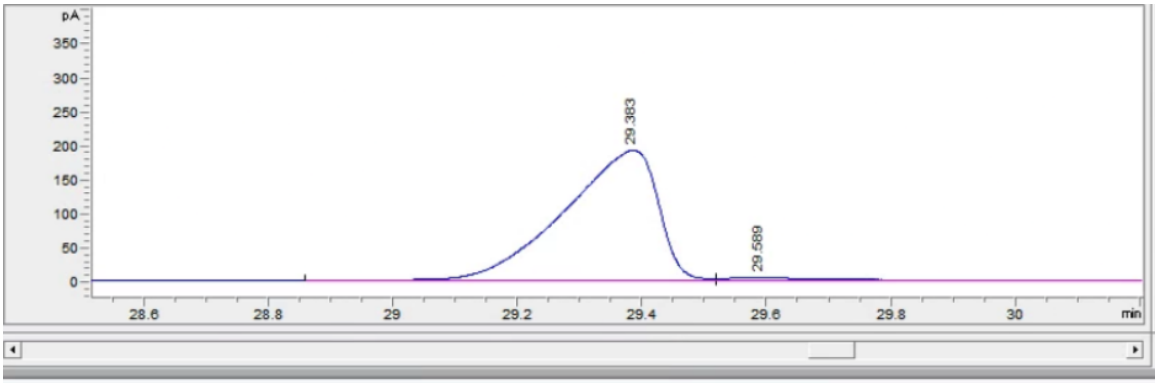
Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5312 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	29.327	MF	57.2	5.8	0.1639	1.672	0
2	29.855	FM	3363.5	283.4	0.1978	98.328	3.83

Figure A60. Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5312 in whole cell

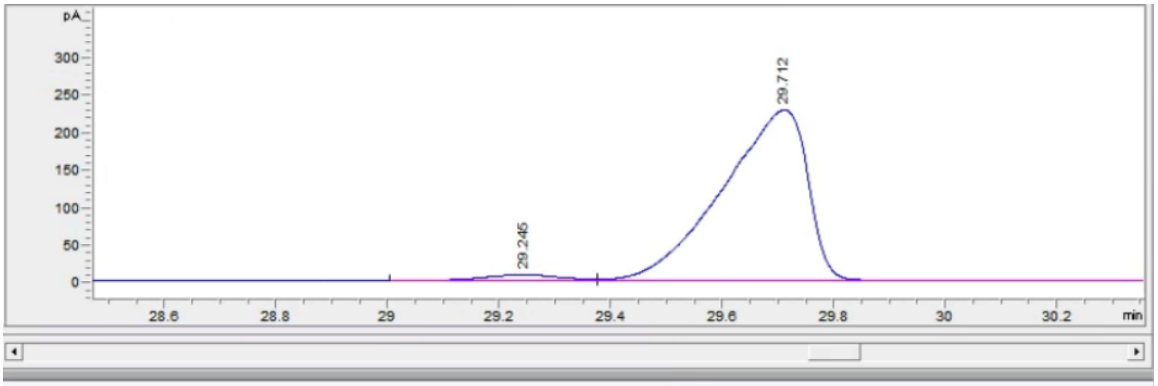
Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5318 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	29.383	BV	1973.2	190.1	0.1571	98.588	2.498
2	29.589	VV	28.3	3.2	0.1312	1.412	0.55

Figure A61. Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5318 in lyophilized lysate

Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5312 in lyophilized lysate



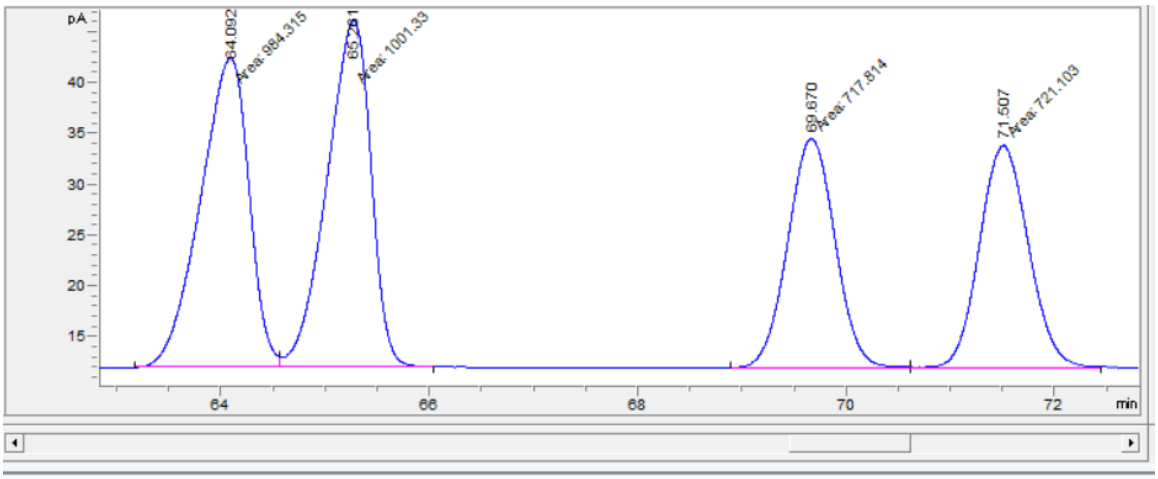
#	Time	Type	Area	Height	Width	Area%	Symmetry
1	29.245	BV	71.9	7.6	0.1368	2.896	1.035
2	29.712	VB	2409.1	227	0.1488	97.104	2.592

Figure A62. Chiral GC-FID trace of 2a from the enzymatic reaction using ApePgb-xHC-5312 in lyophilized lysate

A.13.2 Chiral GC-FID traces for enzymatic reactions of 1b to 2b

Analysis was performed using an Agilent CYCLOSIL-B column (3 μL injection; 0 min at 120 °C, ramp to 150 °C at 10 °C/min and hold for 75 minutes; ramp to 170 °C at 10 °C/minute and hold for 10 minutes.

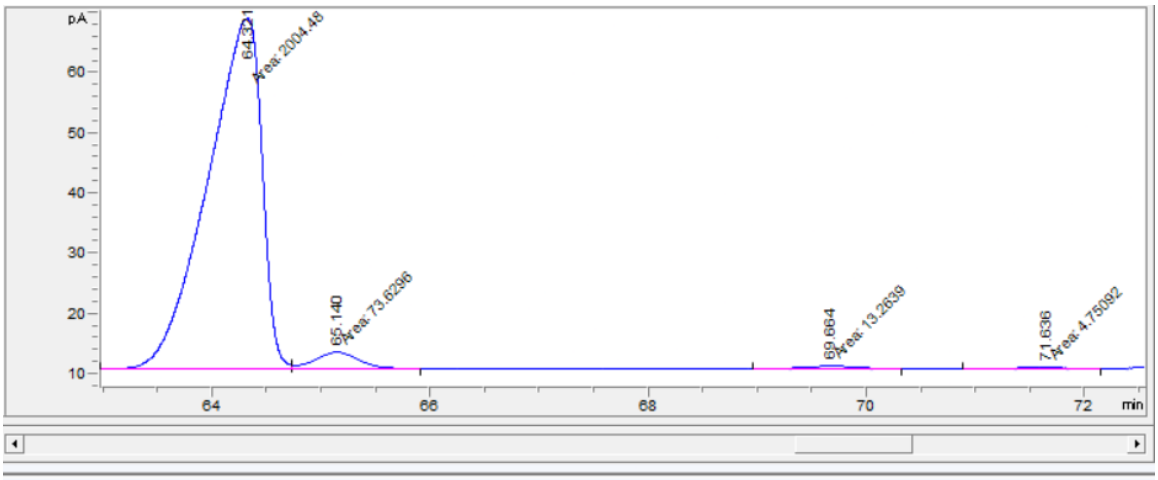
Racemic standard of 2b



#	Time	Area	Height	Width	Area%	Symmetry
1	64.092	984.3	30.4	0.5396	28.743	0
2	65.261	1001.3	34.2	0.4881	29.240	1.372
3	69.67	717.8	22.6	0.5292	20.961	0
4	71.507	721.1	22	0.5471	21.057	0.936

Figure A63. Chiral GC-FID trace for racemic standard of 2b

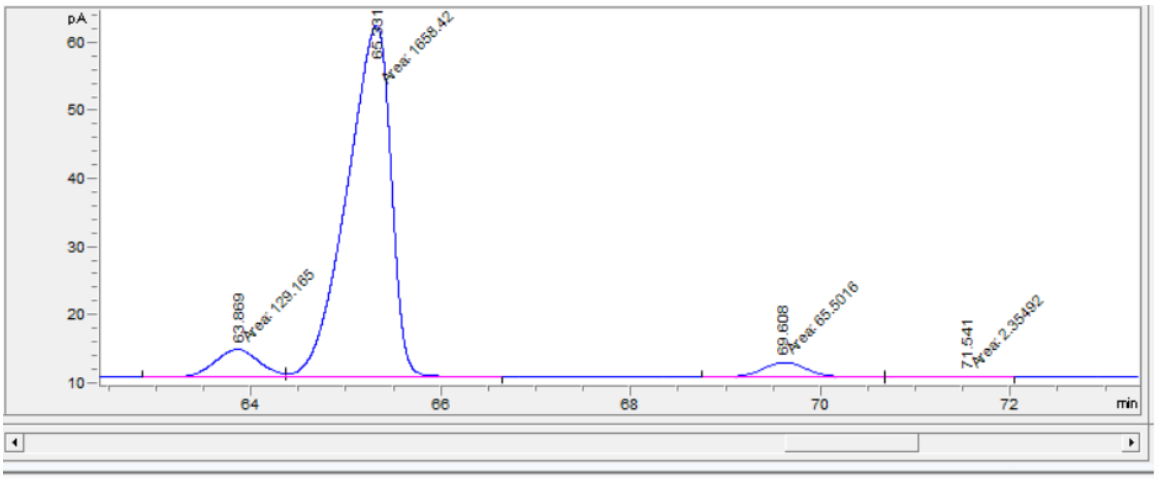
Chiral GC-FID trace of 2b from the enzymatic reaction using ApePgb-xHC-5322 in whole cell



#	Time	Area	Height	Width	Area%	Symmetry
1	64.321	2004.5	58.3	0.5732	95.628	2.609
2	65.14	73.6	2.7	0.4512	3.513	1.043
3	69.664	13.3	4.1E-1	0.5381	0.633	0.983
4	71.636	4.8	1.4E-1	0.5489	0.227	1.674

Figure A64. Chiral GC-FID trace of 2b from the enzymatic reaction using ApePgb-xHC-5322 in whole cell

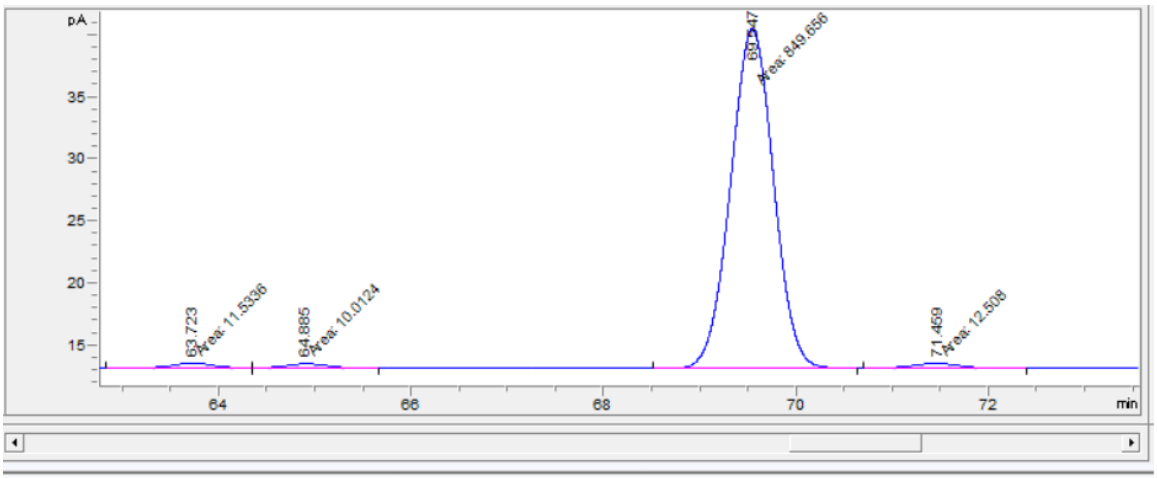
Chiral GC-FID trace of 2b from the enzymatic reaction using ApePgb-xHC-5312 in whole cell



#	Time	Area	Height	Width	Area%	Symmetry
1	63.869	129.2	4	0.5324	6.961	0
2	65.331	1658.4	51.6	0.5359	89.381	1.968
3	69.608	65.5	2.1	0.5161	3.530	0
4	71.541	2.4	1E-1	0.383	0.127	1.099

Figure A65. Chiral GC-FID trace of 2b from the enzymatic reaction using ApePgb-xHC-5312 in whole cell

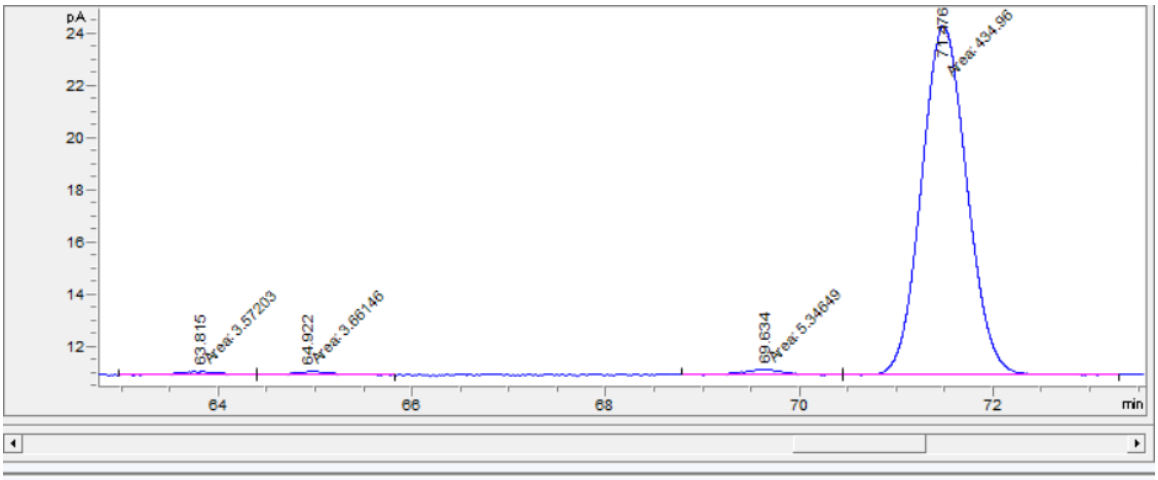
Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5328 in whole cell



#	Time	Area	Height	Width	Area%	Symmetry
1	63.723	11.5	4E-1	0.4827	1.305	1.138
2	64.885	10	3.4E-1	0.4907	1.133	0.777
3	69.547	849.7	27.4	0.5169	96.146	0.941
4	71.459	12.5	3.8E-1	0.5547	1.415	1.012

Figure A66. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5328 in whole cell

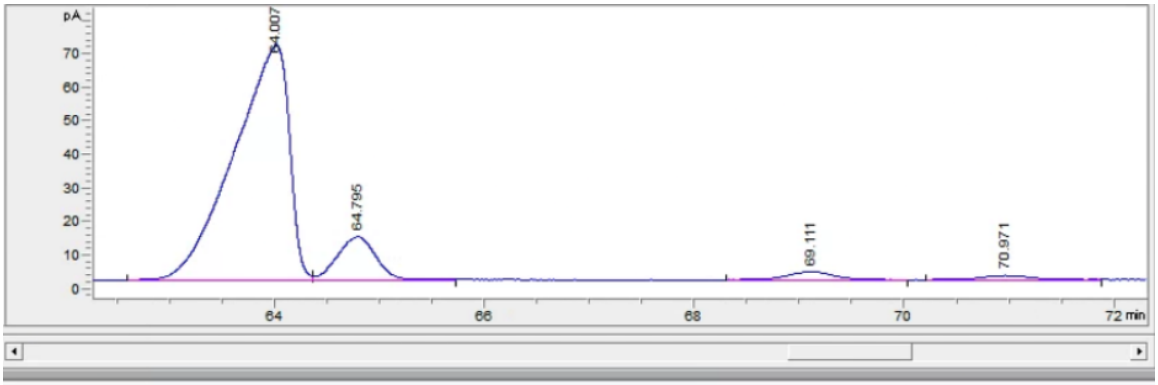
Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5325 in whole cell



#	Time	Area	Height	Width	Area%	Symmetry
1	63.815	3.6	1.3E-1	0.448	0.798	0
2	64.922	3.7	1.4E-1	0.4472	0.818	0.654
3	69.634	5.3	1.9E-1	0.4607	1.195	0
4	71.476	435	13.4	0.5415	97.189	0.78

Figure A67. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5325 in whole cell

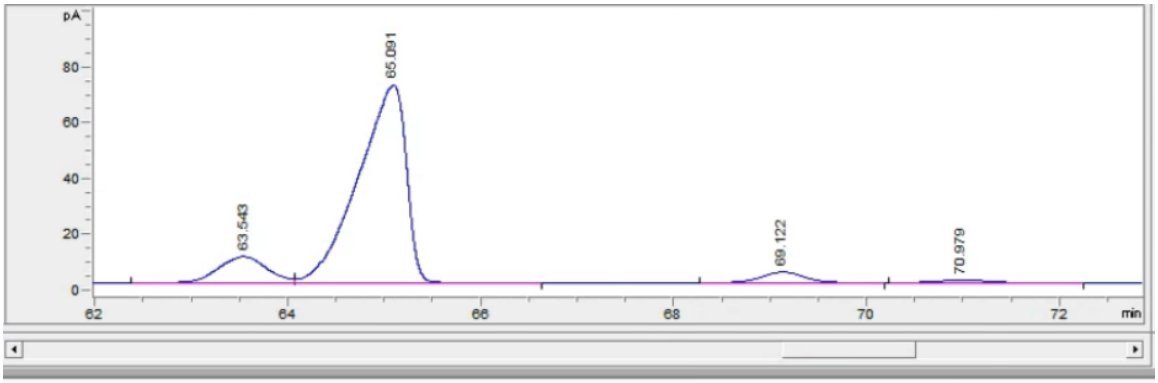
Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5322 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	64.007	BV	2555.7	70.1	0.4896	85.454	3.153
2	64.795	VB	322.6	12.8	0.3818	10.787	1.087
3	69.111	BB	71.8	2.3	0.4692	2.400	0.983
4	70.971	BB	40.7	1.3	0.4861	1.360	0.958

Figure A68. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5322 in lyophilized lysate

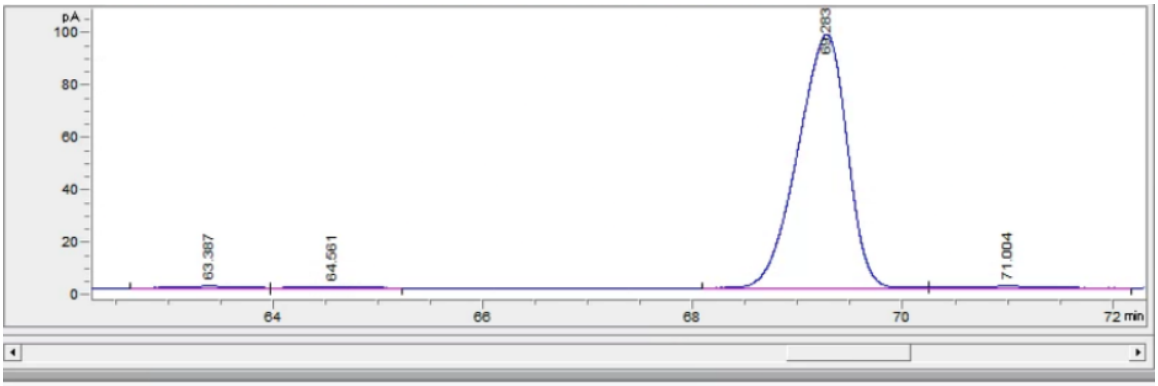
Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5312 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	63.543	BV	321.7	9.3	0.5386	11.077	1.021
2	65.091	VB	2435.4	71	0.476	83.865	2.651
3	69.122	BB	114.7	3.8	0.4669	3.951	0.982
4	70.979	BB	32.2	1	0.4881	1.108	0.935

Figure A69. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5312 in lyophilized lysate

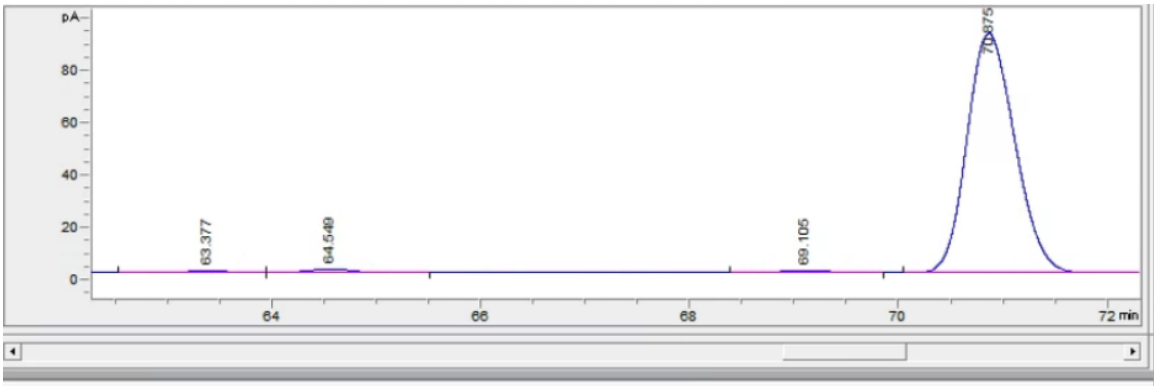
Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5328 in lyophilized lysate



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	63.387	BV	18.4	6.2E-1	0.4261	0.576	1.063
2	64.561	VB	14.2	4.7E-1	0.4332	0.444	0.994
3	69.283	BV	3144.4	96.6	0.4919	98.337	1.367
4	71.004	VB	20.6	6.2E-1	0.5082	0.643	1.001

Figure A70. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5328 in lyophilized lysate

Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5325 in lyophilized lysate



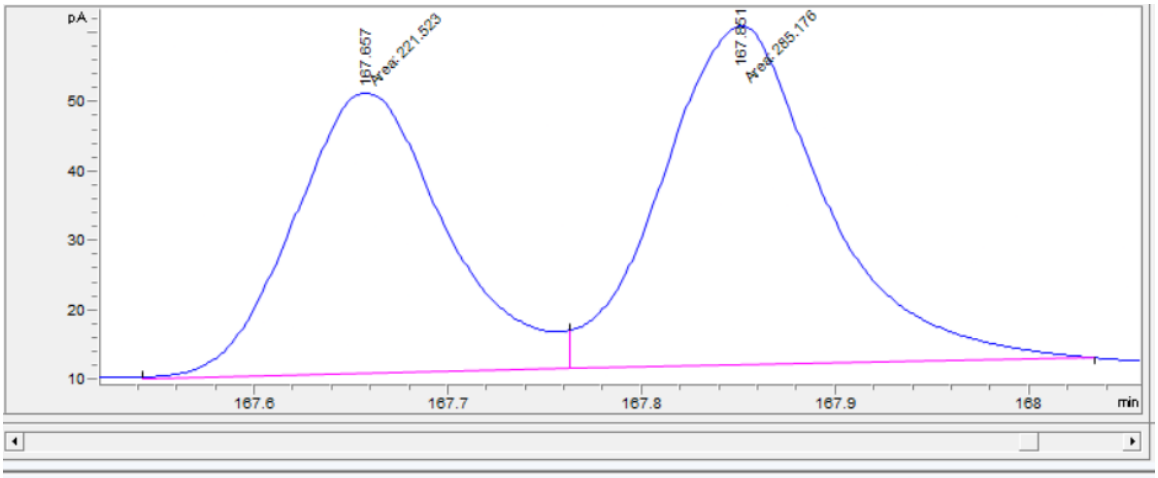
#	Time	Type	Area	Height	Width	Area%	Symmetry
1	63.377	BV	21.7	7.4E-1	0.4125	0.722	1.058
2	64.549	VB	43.2	1.5	0.4192	1.435	1.001
3	69.105	BB	27.6	8.7E-1	0.4788	0.919	0.946
4	70.875	BB	2916.9	91.4	0.5003	96.925	0.809

Figure A71. Chiral GC-FID trace of 2b from the enzymatic reaction using TamPgb-xHC-5325 in lyophilized lysate

A.13.3 Chiral GC-FID traces for enzymatic reactions of 1c to 2c

Analysis was performed using an Agilent CYCLOSIL-B column (3 μL injection; 165 min at 135 °C, ramp to 200 °C at 30 °C/min and hold for 15 minutes)

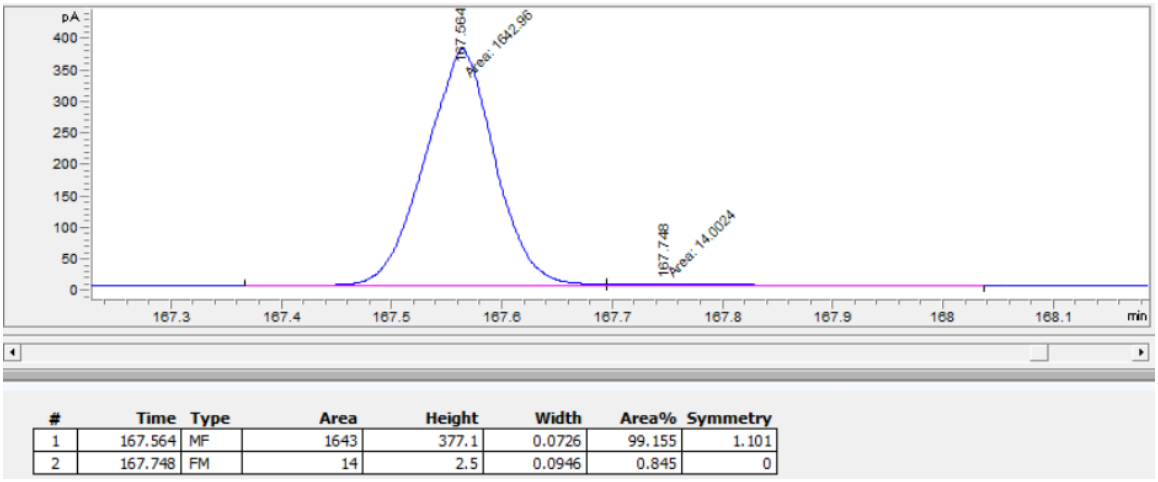
Racemic standard of 2c



#	Time	Area	Height	Width	Area%	Symmetry
1	167.657	221.5	40.3	0.0915	43.719	0
2	167.851	285.2	48.8	0.0975	56.281	0.942

Figure A72. Chiral GC-FID trace for racemic standard of 2c

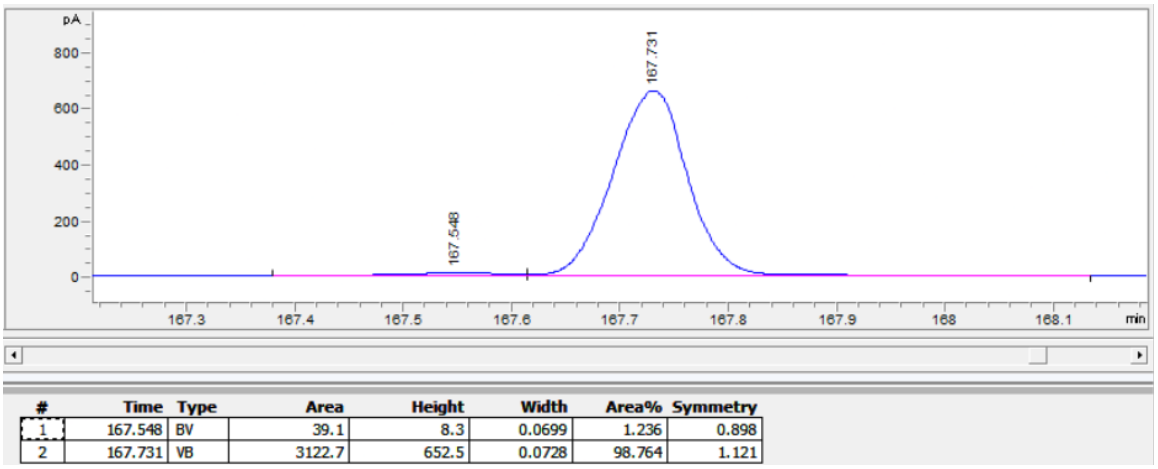
Chiral GC-FID trace of 2c from the enzymatic reaction using TamPgb-xHC-5320 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	167.564	MF	1643	377.1	0.0726	99.155	1.101
2	167.748	FM	14	2.5	0.0946	0.845	0

Figure A73. Chiral GC-FID trace of 2c from the enzymatic reaction using TamPgb-xHC-5320 in whole cell

Chiral GC-FID trace of 2c from the enzymatic reaction using ApePgb-xHC-5315 in whole cell



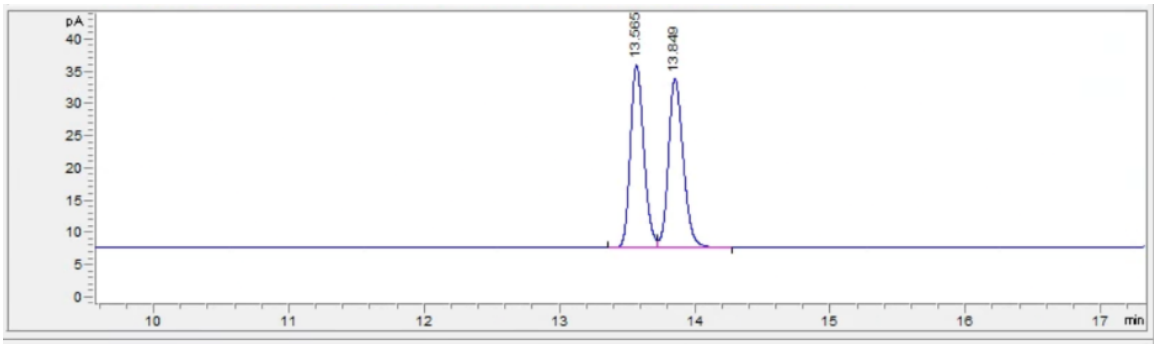
#	Time	Type	Area	Height	Width	Area%	Symmetry
1	167.548	BV	39.1	8.3	0.0699	1.236	0.898
2	167.731	VB	3122.7	652.5	0.0728	98.764	1.121

Figure A74. Chiral GC-FID trace of 2c from the enzymatic reaction using ApePgb-xHC-5315 in whole cell

A.13.4 Chiral GC-FID traces for enzymatic reactions of 1d to 2d

Analysis was performed using an Agilent CYCLOSIL-B column (3 μL injection; 16 min at 135 °C,; ramp to 200 °C at 30 °C/minute and hold for 3 minutes.

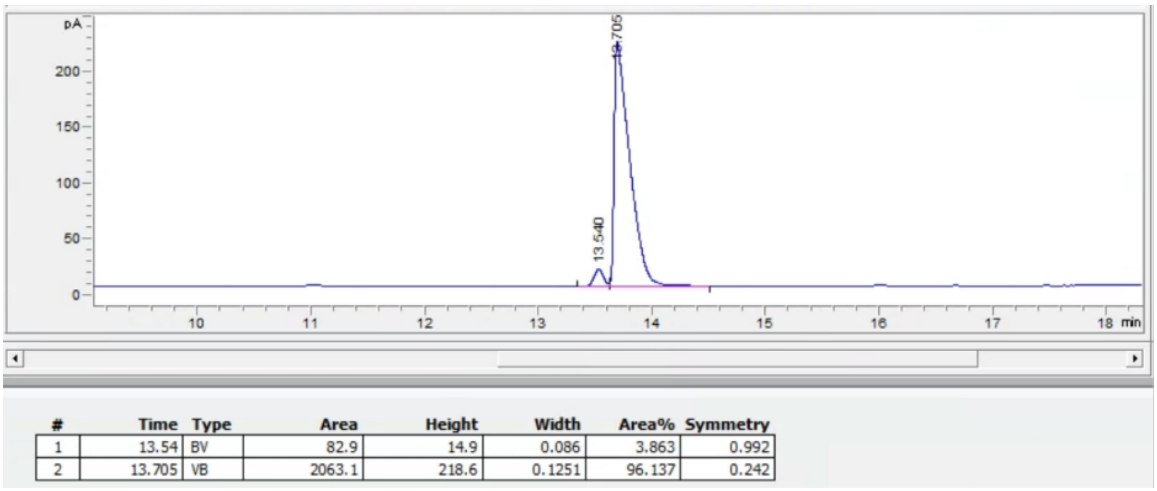
Racemic standard of 2d



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	13.565	BV	197.4	28.3	0.1074	49.653	0.788
2	13.849	VB	200.2	26.3	0.1148	50.347	0.742

Figure A75. Chiral GC-FID trace for racemic standard of 2d

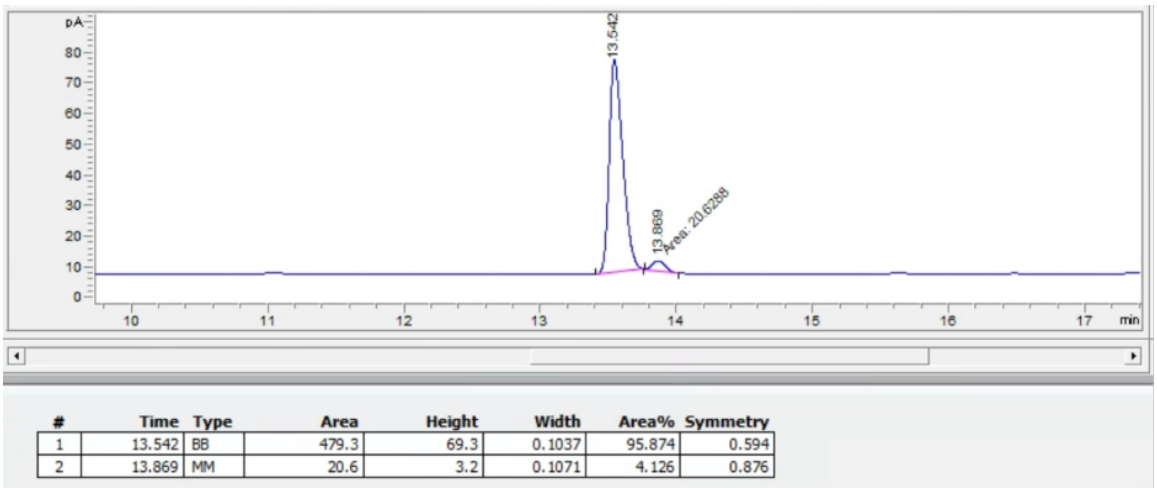
Chiral GC-FID trace of 2d from the enzymatic reaction using ApePgb-xHC-5314 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	13.54	BV	82.9	14.9	0.086	3.863	0.992
2	13.705	VB	2063.1	218.6	0.1251	96.137	0.242

Figure A76. Chiral GC-FID trace of 2d from the enzymatic reaction using ApePgb-xHC-5314 in whole cell

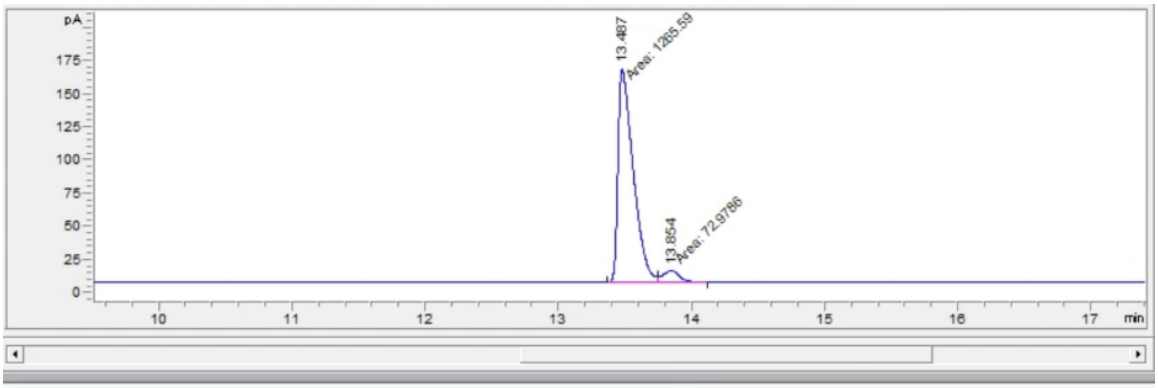
Chiral GC-FID trace of 2d from the enzymatic reaction using TamPgb-xHC-5319 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	13.542	BB	479.3	69.3	0.1037	95.874	0.594
2	13.869	MM	20.6	3.2	0.1071	4.126	0.876

Figure A77. Chiral GC-FID trace of 2d from the enzymatic reaction using TamPgb-xHC-5319 in whole cell

Chiral GC-FID trace of 2d from the enzymatic reaction using TamPgb-xHC-5320 in whole cell



#	Time	Type	Area	Height	Width	Area%	Symmetry
1	13.487	MF	1265.6	160.5	0.1314	94.548	0.376
2	13.854	FM	73	8.5	0.1433	5.452	0

Figure A78. Chiral GC-FID trace of 2d from the enzymatic reaction using TamPgb-xHC-5320 in whole cell

A.14 Spectroscopic Data

A.14.1 ^{1}H NMR of authentic standard of (±)-2a

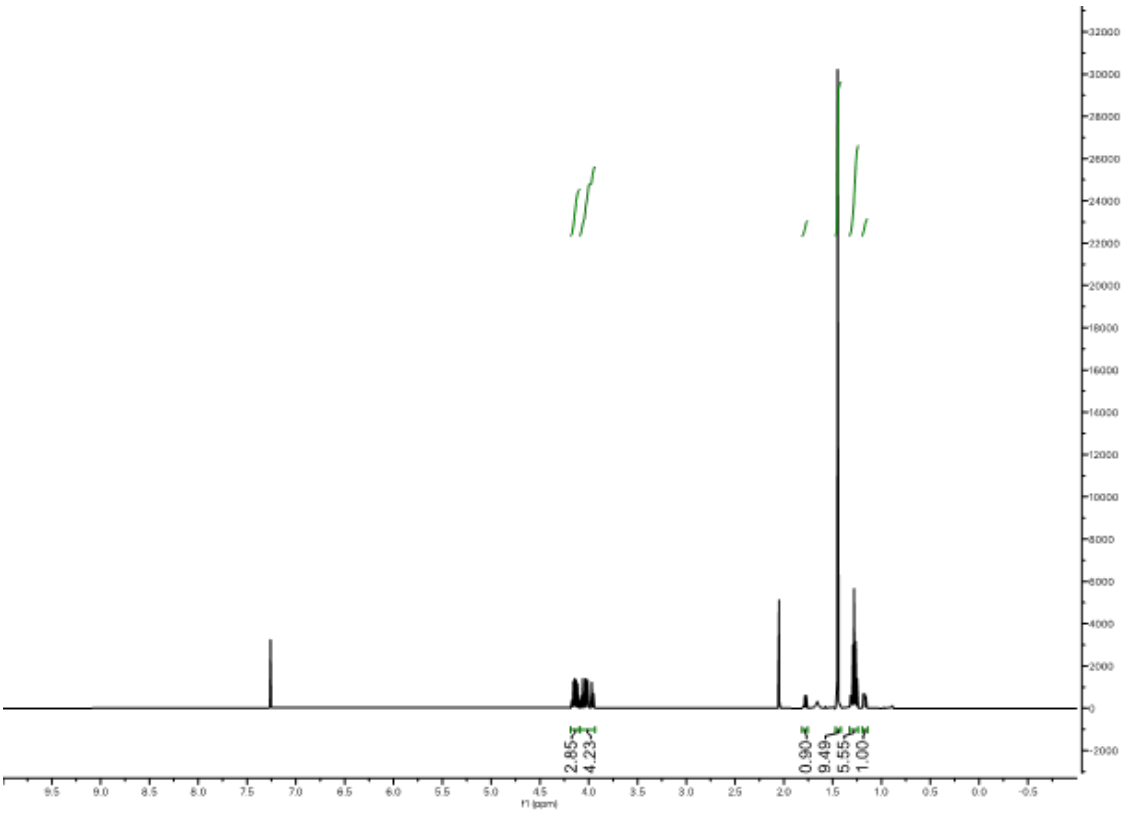


Figure A79. ^{1}H NMR of authentic standard of (±)-2a synthesized according to the procedures outlined in Section 10. Note: this NMR contains EtOAc.

A.14.2 ^{1}H NMR of 2a synthesized via an enzymatic reaction with 1a

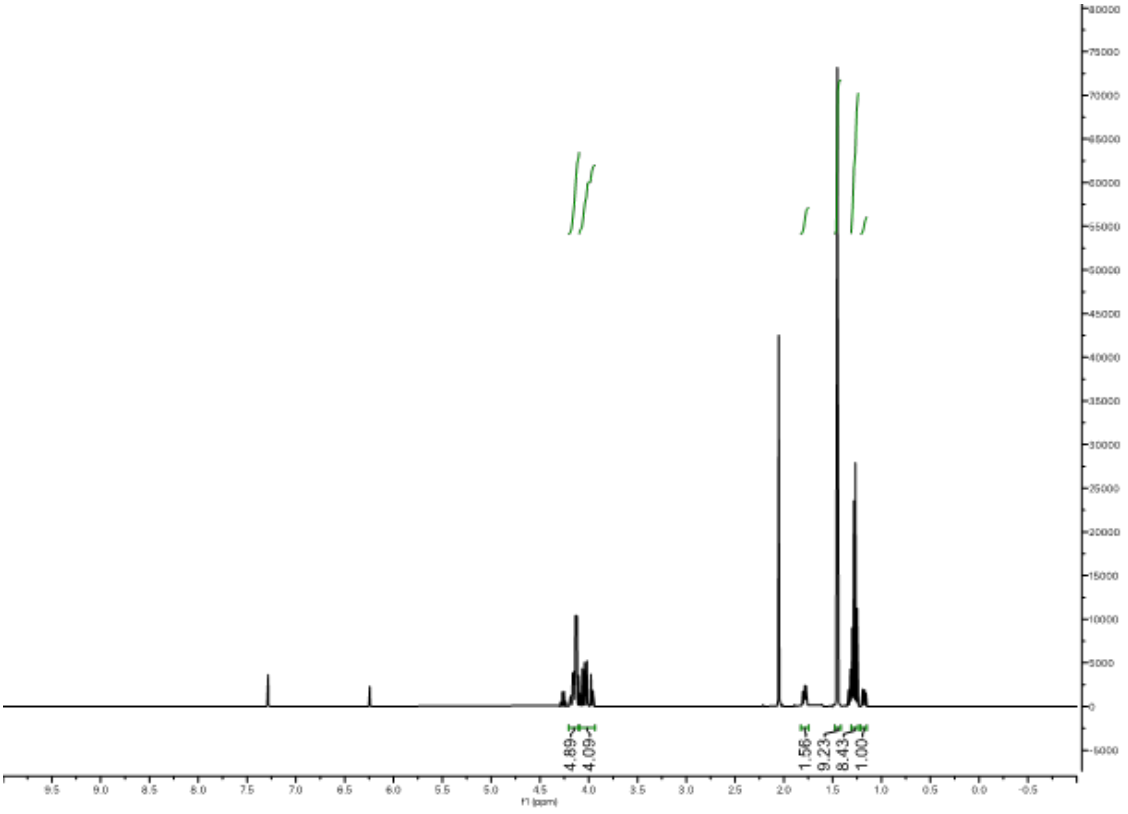


Figure A80. ^{1}H NMR of 2a synthesized enzymatically according to the procedures outlined in Section 1.5. Note: this NMR contains EtOAc and diethyl maleate.

Appendix B

SUPPORTING MATERIAL FOR CHAPTER 3

B.1 Oligonucleotide Design

B.1.1 Barcode Design

LevSeq uses 96 unique 24-nucleotide forward barcodes and 96 unique 24-nucleotide reverse barcodes. All barcodes have at least a Hamming distance of three between each other (different by three substitutions). These barcodes were directly taken from the nanopore native barcoding kit and used with primers designed in the "Primer Design" section below.

B.1.2 Primer Design

The primers of LevSeq are specific to the cloning vectors that host the protein of interest (pET22b(+) in this protocol). Each full-length DNA sequence is captured by the combination of forward and reverse primers. These primers have the below general layout:

F: 5' - XXXXXXXXXXXXXXXXXXXXXXXXATCTCGATCCCGCGAAATTAATACGACTCAC - 3'

R: 5' - XXXXXXXXXXXXXXXXXXXXXXXXGCCCCAAGGGGTTATGCTAGTTATTGCTC - 3'

The 3' region is a backbone-specific primer that binds to the cloning vector near the T7 and T7' promotor sites. The 5' region is the unique 24-nucleotide primer used for differentiating between plates (R) and wells (F).

B.2 Supplementary Protocols

B.2.1 Ordering Barcode Linked primers

The forward and reverse 24-nucleotide barcode sequences are given in Tables S1 and S2; the full-length barcode linked forward and reverse primers for pET22b(+) are given in Tables S3 and S4. If using pET22B(+) as a cloning vector, the sequences listed in Tables S3 and S4 can be directly used for ordering primers at 100 μM concentrations from commercial suppliers. If using a different cloning vector, the user must verify that the primers in the "Primer Design" section would still amplify the desired region.

B.2.2 Preparation of LevSeq Barcode Primer Mixes

There are 96 unique forward barcode-linked primers, corresponding to each well of a 96-well plate; there are 96 reverse barcode-linked primers, corresponding to the ability to sequence 96 different plates as a unique reverse barcode is used for each plate. The barcode-linked primers are listed in the tables below (Tables S3 and S4); the 96 F primers and 96 R primers are both ordered in plate format at 100 μM concentration.

Each well sequenced using LevSeq is encoded by a unique combination of a forward (F) and a reverse (R) barcode. The reverse barcode would identify the sequence to a specific plate, and the forward barcode would identify the sequence to a specific well. Since these barcode-linked primers are compatible with any gene cloned in a specific cloning vector, it is convenient to keep a set of barcode-linked primer plates on hand.

We used the same eight barcode plates throughout this work, and they are named LevSeq01–LevSeq08. The exact barcodes used are given in Tables S5–12. The LevSeq software assumes the forward barcodes used for library preparation are laid out in the order given in Tables S5 and S12. To build the barcode plates depicted in Tables S5–12, we followed the below procedure:

1. A 1fold dilution of the 100 μM forward barcode-linked primer plate from IDT was prepared by adding 10 μL of the forward primer stock to 90 μL ddH2O to a final concentration of 10 μM, keeping the well layout constant. Dilutions were performed in half skirted PCR plates (BIO-RAD HSP9601).

2. A 1fold dilution of the reverse barcode-linked primer from IDT was prepared by adding 150 µL of the 100 µM reverse primer stock to 1350 µL ddH2O to a final concentration of 10 µM. Dilutions were performed in Eppendorf tubes.
3. To create plates for sequencing of forward/reverse primer mixes, 80 µL ddH2O were added to each well of each of the eight plates. Then, 10 µL diluted (10 µM) forward barcode plate was aliquoted to each well keeping the layout constant.
4. A unique reverse barcode was added to every well of each of the eight plates. To do so, 10 µL of diluted (10 µM) reverse barcode from step 2 was used. For instance, LevSeq01 has forward barcode 01 – 96 and reverse barcode 01, LevSeq02 has forward barcode 01 – 96 and reverse barcode 02, and so on. The final concentration of each barcode plate is 1 µM each for the forward and reverse primers (10X stock).
5. Eight of the LevSeq barcode plates (LevSeq 01 – 08) at 10X stock concentration were made in step 4. In order to make 1X reaction-ready barcode plates, the following dilution was performed: To another set of eight fully skirted PCR plates, 90 µL of ddH2O were added, followed by 10 µL of diluted LevSeq barcode stock plates (1 µM). The final concentration of each ready-to-use barcode plate is 0.1 µM for both forward and reverse barcode linked primers.
6. When not in use, the 10X stocks prepared in step 4 were stored at -20 °C, while the barcode plates were stored at 4 °C. The 100 µM, 10 µM, and 1 µM barcode plates can be stored for long periods of time.

B.2.3 LevSeq Library Preparation and Sequencing

The steps below can be followed to complete a LevSeq sample preparation ready for loading onto an Oxford Nanopore flow cell (FLO-MIN114). Note that when designing a new set of barcode-linked backbone-specific primers, it is recommended to test the primers and PCR protocol using a few wells before deploying them for plate-scale reactions. The library protocol below provides part numbers for the materials and reagents used for developing this protocol; reagents from other providers should work as well.

1. Prepare a PCR master mix for the number of plates to be sequenced according to the table below.

Component	Amount per plate (µL)
ThermopoI Buffer (NEB B9004L)	144
10 mM dNTPs (NEB N0447)	28.8
Taq Polymerase (NEB M0267)	7.2
Mol-Bio Grade DMSO (mp 194819)	57.6
ddH2O	770

2. Add 7 µL of master mix to each well of as many half-skirted PCR plates (USA scientific 1402-9700) as will be sequenced. These are referred to as "PCR plates".
3. Stamp 2 µL of the 1-µM barcode-linked primer mix from the barcode plates into the PCR plates.
    a. "Stamp" means "apply to all wells, keeping the plate layout consistent".
4. Stamp 1 µL of overnight culture from each plate to be sequenced into the PCR plates. Record which barcode plate was used with which PCR plate.
5. Complete a PCR using the below thermal cycler program. This colony PCR amplifies the entire gene of interest from the template DNA contained in the cell culture.

Step	Temperature (°C)	Time
1	95	5 min
2	95	20 s
3	TD 68 -> 63.5	20 s
4	68	1 min per kb
5	Return to 2, 9x	
6	95	20 s
7	68	1 min per kb
8	Return to 6, 24x	
9	68	5 min
10	4	Hold

    a. "TD" in step 3 stands for "touchdown", a touchdown step decreases the temperature each cycle. The TD in the above PCR starts at 68 °C and drops to 63.5 °C by the end, decreasing by 0.5 °C per cycle.
    b. Extension time in steps 4 and 7 are recommended to be minimally 1 minute per kb of gene, a 2-min extension time was used for genes below 1 kb during development of this protocol.
6. While the PCR is running, prepare a 1% agarose gel with SYBR gold added (Thermo Fisher Scientific, S11494).
7. Once the PCR is completed, for each plate, pool 5 µL of each reaction into an Eppendorf tube, for a total of 480 µL of pooled PCR products per plate. Pooling will leave you with as many Eppendorf tubes as you have plates.
    a. Note: Pooling can be performed using a 12-channel multichannel pipette. 1) First, transfer 5 µL of reactions from each row in the PCR plate (to be sequenced) into a new row of a 96-well PCR plate. Since there are eight rows in the original plate, each well of the row in the new 96-well PCR plate will have 40 µL of pooled PCR products. 2) Transfer 30 µL from each well in the single row of pooled reactions using a single-channel pipette to a microcentrifuge tube. Since there are 12 columns in the new 96-well PCR plate, the microcentrifuge tube should have 360 µL of pooled PCR products that contain PCR reactions of every well of the PCR plate (to be sequenced).
    b. Alternatively, a liquid handling robot can be used for this task. It is important that 5 µL from each well of each PCR plate is combined into one microcentrifuge tube to improve the evenness of sequencing coverage. There

will now be an equal number of microcentrifuge tubes to as plates being sequenced.

8. For each tube made in step 7, take 100 µL of pooled PCR reactions and add it to 20 µL 6x loading dye (NEB B7025S) in a microcentrifuge tube. The remaining pooled reaction can be stored at -20 °C for future use.
9. Load the contents of each tube made in step 9 into the agarose gel prepared in step 7. Each tube should have separated lanes. Load a 10 µL of 1 kb ladder in the flanking lanes.
10. Run the agarose gel at 120 V until the bands have sufficiently migrated. Reference the ladder to identify the PCR products which should be the size of the gene plus about 100 bp. The extra bases come from the barcodes and the region between the primer attachment sites and the open reading frame.
11. Gel extract the desired bands. Samples from separated PCR reaction plates should have their own microcentrifuge tube to hold the extracted gels. Commercial kits such as the Zymoclean Gel DNA Recovery Kit (Zymo Research, D4001) are typically used for this step. Elution should be performed using 10 µL of ddH2O rather than elution buffer.
12. After gel extraction, measure the DNA concentration of each gel-extracted PCR product from each plate. We use a GE NanoVue Plus to measure DNA concentration in ng/µL unit.
13. Combine the gel-extracted PCR products from each plate in equimolar concentrations. In general, 200 ng of pooled DNA sample per 1-kb gene are sufficient for any sample preparation. For experiments that use the same length template, the weight is directly proportional to molarity. Hence the same amount of DNA in nanograms from each plate can be combined into one microcentrifuge tube. For example, if plate one after gel extraction has a measured concentration of 20 ng/µL and plate two has a concentration of 10 ng/µL, then to make 20 µL of 10 ng/µL final sample, 5 µL of the first plate and 10 µL of the second plate were added to 5 µL of ddH2O to make the final sample.
14. After step 13, there should be one single tube of cleaned, normalized DNA consisting of all genes from all plates to be sequenced. Depending on the Oxford Nanopore sample preparation protocol, this sample is adjusted to the desired concentration before sample preparation for sequencing.
15. A MinION flow cell with the LSK114 ligation sequencing kit was used while developing this protocol, which recommends 200 fmol of amplicon DNA for sample preparation. This translates to 120 ng for a 1-kb gene.
16. From the concentration-normalized sample prepared in step 13, transfer the equivalent volume of 120 ng of pooled DNA into a new microcentrifuge tube. This pooled DNA sample (120 ng) will be carried forward for Oxford Nanopore sequencing sample preparation. Protocols are linked here (https://community.nanoporetech.com/docs/prepare/library_prep_protocols/genomic-dna-by-ligation-sqk-lsk114/v/gde_9161_v114_revu_29jun2022/dna-repair-and-end-prep?devices=minion).
17. Once the sample is loaded onto the flow cell for sequencing, we recommend choosing the super accurate basecalling option to ensure retrieval of the highest quality

sequences while setting up the sequencing run. It is recommended to stop the flow cell after the basecalled bases reach a data volume of 0.02 Gb per plate to be sequenced, so an experiment with 20 plates would be stopped after collecting 0.4 Gb of basecalled bases.

B.2.4 LevSeq Data Analysis

If basecalling is enabled, a "fastq_pass" folder with the basecalled sequencing results will be returned at the designated folder location. This folder location is given at the end of the sequencing run set up and is usually where the Minknow software is installed in the system. (In our case, data can be accessed from "/var/lib/minknow/data/name_of_experiment"). Once sequencing is complete, LevSeq software can be used to process results (in the format of .fastq.gz files) and assign variants to their original wells. Detailed instructions on how to use the LevSeq software are provided on the LevSeq Github page (https://github.com/fhalab/LevSeq).

B.2.5 LevSeq Post Data Analysis Workflow and Interpretation

This detailed workflow outlines steps to ensure data quality and rigor in variant calling and downstream analyses. The process includes alignment quality checks, variant visualization, and decision-making for further experimental steps.

Initial Data Quality Assessment

1. Check Plate Map:
    a. Verify the processed data against the plate map.
    b. Ensure proper labeling and data integrity for all wells.
2. Evaluate alignment counts and probability values for all wells:
    a. Alignment Count: Must be >20 for each well.
    b. Probability Value: Mean error <10% for each well.
3. Proceed with data analysis only if all wells meet these criteria.

B.2.6 Handling Variants with Suboptimal Metrics

If a variant is called in any well where: Alignment Count is ≤20, or mean error > 10%, or the has been classified as a"mixed well".
Still proceed with data collection and follow these steps for sequencing validation:

1. Variant Visualization using IGV Web App:
    a. Open the IGV Web App (https://igv.org/app/) and prepare the necessary files:
        i. Template FASTA File: Found in the respective plate folder.
        ii. Generate the corresponding FAI index using: <samtools faidx temp.fasta>
        iii. BAM and BAM.BAI Files: Located in the NB folder for each well.
    b. Upload files into IGV as follows:
        i. Genome Track: Use the FASTA and FAI files.
        ii. Alignment Track: Upload the BAM and BAM.BAI files.
2. Analyze Population Distribution:
    a. Visually evaluate the distribution of reads and alignment quality to confirm variants presence
3. Decision-Making for Additional Experiments
    a. If the variant' s authenticity remains unclear or if population distribution suggests potential artifacts, decide if resequencing or alternative method is needed
    b. If the variant's authenticity seems valid, only one species is shown in the alignment, use the consensus reads on top as correct sequence

A.2.6 Integrating Experimental Data with Sequencing Data

1. Document experimental values and methodologies for inclusion in subsequent reporting or publication.
2. Consider uploading to centralized location for data handling and visualization.

B.3 Alternative Methods and Recommendations

For protein engineering and mutant library sequencing, workflows like Sequence Genie[97], parSEQ[104], UMIC-Seq[105], DuBA.flow[98], evSeq[100], and LevSeq can generate high-throughput sequence data for synthetic biology applications. Earlier methods designed to sequence large gene libraries used Illumina sequencing, while later methods exclusively used nanopore long read sequencing. Sequence Genie, parSEQ, LevSeq and DuBA.flow all utilize similar barcoding strategies, and report similar pros (speed of turnaround, cost), and cons (PCR bias). While the experimental approach to the pipeline is similar, differences remain in the level of automation each method enables and throughput (see details below). parSEQ excels in large-scale, cloud-based sequencing with automated analysis, making it ideal for labs with advanced infrastructure and cloud computing skills, in particular for sequencing

gene libraries (e.g. oligoPools). UMIC-Seq, integrates microfluidic screening with high-accuracy nanopore sequencing and provides insights into evolutionary trajectories. This workflow is ideal for studies that focus on positive mutations, omitting negative or neutral data, and have access to a droplet sorter. evSeq offers a cost-effective solution for sequencing predefined target regions of a gene, is accessible to smaller labs but has some limitations in turnaround time. LevSeq provides an accessible, real-time sequencing solution for full-length protein-coding genes, similar to Sequence Genie and DuBA.flow; it is optimized for iterative directed evolution and machine learning-guided protein engineering. LevSeq is designed for labs without centralized sequencing facilities; however, the throughput is limited to 9,216 samples. Meanwhile, other workflows specialize in scalable, cost-effective construct validation, focusing on plasmid library verification, such as the verification of oligopools, and synthetic construct quality control.

In the section below, we summarize several methods which utilize next generation sequencing to gather sequence information for pooled DNA samples. This section aims to be informative and enable researchers to select the most suitable workflow for their sequencing needs.

Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process[95] (2015, Illumina MiSeq)

The protocol describes an approach that uses acoustic liquid handling to reduce reagent costs combined with a MiSeq sequencer to sequence 4,000–5,000 plasmids per run. This method supports multiplexing up to tens of thousands of samples and uses rolling circle amplification for consistent DNA preparation. This method requires several specific equipment (e.g., Illumina MiSeq and LabCyte Echo systems) making it best suited for sequencing facilities or large-scale industrial or academic labs with high-throughput demands and access to automation infrastructure. The setup costs and technical requirements may not be feasible for smaller academic labs looking to validate sequencing libraries or to perform directed evolution/machine learning-guided protein engineering (MLPE) experiments.

Miniaturization of High-Throughput Plasmid DNA Library Preparation for Next-Generation Sequencing Using Multifactorial Optimization[96] (2019, Illumina MiSeq)

This study presents a miniaturized, high-throughput workflow for preparing plasmid DNA libraries for next-generation sequencing (NGS) using the Nextera XT kit and Labcyte Echo acoustic liquid dispensing system, providing advancements compared to the Shapland method (2015). Key benefits include significantly reduced reagent costs relative to the previously reported method (approximately £10.90 per plasmid); however, it should be noted that this is a higher cost than any of the nanopore-based methods reported below. The authors report high-quality sequencing with minimal input DNA (1 ng) by avoiding contamination issues through quality control for genomic DNA. Challenges include reliance on specialized acoustic dispensing equipment and variability in fragment sizes across samples. Like the previous method, this approach is best suited for sequencing facilities and academic labs specializing in molecular and synthetic biology that handle large-scale DNA assemblies.

evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library (2022, Illumina)

The evSeq (every variant sequencing) method offers a cost-effective and accessible solution for sequencing protein variant libraries, enabling high-throughput sequencing of targeted regions at costs as low as $0.01 per variant under optimal conditions. The method seamlessly integrates with existing protein engineering workflows to produce sequence-fitness datasets for machine learning-assisted protein engineering (MLPE). While powerful for targeted sequencing, evSeq is limited to predefined target regions, making it less suited for whole gene sequencing. This workflow is ideal for small to medium-sized labs focused on protein engineering within regions of up to 300 base pairs, particularly for those without immediate sequencing needs or in-house sequencing capabilities.

Highly Multiplexed, Fast and Accurate Nanopore Sequencing for Verification of Synthetic DNA Constructs and Sequence Libraries (2019, Nanopore)

This study presents a similar approach to LevSeq and provides a cost-effective, rapid nanopore sequencing workflow for verifying synthetic DNA constructs. While the paper presents barcodes specifically for BglBrick libraries, the method can be generalized by ordering different primer libraries. Using PCR-based dual barcoding, the authors sequenced 576 samples directly from bacterial cultures without plasmid extraction, with higher theoretical sample numbers achievable. The software corrects nanopore errors through strand bias analysis, enabling accurate single-nucleotide variant detection. The paper reports processing 576 samples in 72 hours at £2.20 per sample, comparable to LevSeq, providing a cheaper alternative to Sanger sequencing. This workflow is ideal for small to medium synthetic biology labs seeking cost-effective in-house sequencing for high-throughput DNA assembly validation and DNA construct verification. The software enables demultiplexing reads to wells and variant calling, running with a docker image that requires specialized docker building knowledge. The workflow is not ideal for DE and MLDE applications because the variant calling is not optimized for mutations from mutagenesis libraries and does not integrate sequence data with fitness measurements. The output of this method can be used as input to LevSeq downstream analysis.

UMI-Linked Consensus Sequencing Enables Phylogenetic Analysis of Directed Evolution (2020, Nanopore)

The UMI-linked consensus sequencing (UMIC-seq) workflow integrates microfluidic droplet-based screening with nanopore sequencing to enhance directed evolution campaigns. It delivers highly accurate long-read sequencing (>99.99% accuracy), enabling the identification of evolutionary trajectories and epistatic interactions across full-length gene sequences. The use of ultrahigh-throughput microfluidic screening ensures that only functionally relevant variants are sequenced, streamlining data processing. Key strengths include scalability, low per-variant sequencing costs, and the ability to reveal complex mutation interactions. However, the workflow does not directly establish sequence-function relationships, as each beneficial mutation still requires individual experimental validation. UMI-seq only captures positive functional data, leaving potential gaps in understanding

negative or neutral mutations, which are critical for mapping the full fitness landscape. This method is best suited for labs employing microfluidic droplet screening and aiming to obtain sequence information before validating variants.

parSEQ: Probe and Rescue Sequencing for Advanced Variant Retrieval from DNA Pool (2023, Nanopore)

parSEQ presents a similar experimental pipeline to LevSeq and provides a generalized high-throughput, cost-effective approach to sequence and retrieve DNA variants. Using a sequence-first, screen-later approach with 384-well plates and two-tier barcoding method, it processes up to 36,864 variants per run. By integrating automation and NGS platforms like ONT MinION or Illumina MiSeq, it reduces sequencing costs to under $3 per variant. ParSeq uses cloud-based analysis making it ideal for labs with scalable, generalizable workflows that benefit from automation. ParSEQ is well-suited for large academic labs, biotech startups, and sequencing facilities, while smaller labs can adapt the methodology using external sequencing providers.

DuBA.flow─A Low-Cost, Long-Read Amplicon Sequencing Workflow for the Validation of Synthetic DNA Constructs (2024, Nanopore)

The DuBA.flow workflow introduces a low-cost, highly scalable long-read sequencing approach using dual barcode amplicon sequencing with Oxford Nanopore technology. This method efficiently validates synthetic DNA constructs directly from bacterial colonies, bypassing plasmid extraction. Key strengths include its cost-effectiveness (under €0.10 per sample), high throughput (up to 1,536 samples per run), and flexibility, supported by an automated computational pipeline that simplifies analysis and interpretation. However, challenges include optimization requirements for primer design, potential issues with unspecific amplification, and reduced reliability for non-standard constructs or unknown sequences. Additionally, while DuBA.flow accelerates validation, fitness measurements of sequences are not directly addressed, requiring further downstream processing and analysis if that information is desired. This workflow is ideal for labs aiming for cost-efficient

construct validation with moderate sequencing needs, it is not specifically designed for protein engineering.

Arrayed in vivo Barcoding for Multiplexed Sequence Verification of Plasmid DNA and Demultiplexing of Pooled Libraries (2024, Nanopore)

The Bacterial Positioning System (BPS) workflow is a high-throughput, cost-effective platform for sequence validation of plasmid libraries using in vivo barcoding and Oxford Nanopore sequencing, utilizing a fundamentally different approach from other methods. It enables multiplexing of up to 73,728 plasmids without requiring individual plasmid extractions, significantly reducing costs (as low as $0.12 per plasmid) and preparation time. Its primary purpose is to efficiently validate and balance complex DNA pools, such as variant libraries ordered as oligo pools, with high accuracy and scalability. The computational setup combines a Nextflow-based pipeline with Singularity for containerized execution. This workflow is ideal for whole plasmid construct validation in small to medium-sized labs that aim to verify the sequence identity of large oligo pools and DNA libraries, though it requires familiarity with complex software environments.

LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning

The LevSeq workflow, an enhancement to the evSeq method, provides a pipeline to generate sequence-function datasets for directed evolution and machine learning-guided protein engineering (MLPE). By integrating dual barcoding with nanopore sequencing, LevSeq enables rapid and cost-effective in-house sequencing of full-length protein-coding genes. This real-time sequencing capability allows researchers to obtain sequence information immediately during the experimental workflow. The method supports scalable sequencing of thousands of variants, offering a cost-efficient alternative to traditional methods like Sanger sequencing. LevSeq's software is open-source and user-friendly, making it accessible to researchers with limited bioinformatics expertise.

LevSeq's power comes with specific technical requirements: robust PCR amplification, precise sample normalization, and consistent read coverage. Meeting these requirements consistently over time can be challenging, potentially affecting the reproducibility of sequencing results across different experimental batches. Despite these considerations, the workflow is ideally suited for small to medium-sized protein engineering laboratories focusing on machine learning-guided protein engineering (MLPE) or directed evolution. It particularly benefits labs seeking rapid, cost-effective in-house sequencing capabilities to generate full-length gene sequences while uniquely integrating this sequence information with functional data to accelerate their iterative optimization processes. The software developed for pairing these data functions as a standalone package that can process sequencing data from the aforementioned methods at two stages, beginning with raw sequencing output, which is compatible with other barcoding-based methods (e.g., parSEQ and evSeq). Additionally, for other methods, variant data can be used as input to the LevSeq software to combine sequence and function data for protein engineering experiments, provided it follows the same output format as LevSeq.

B.4 Supplementary Figures

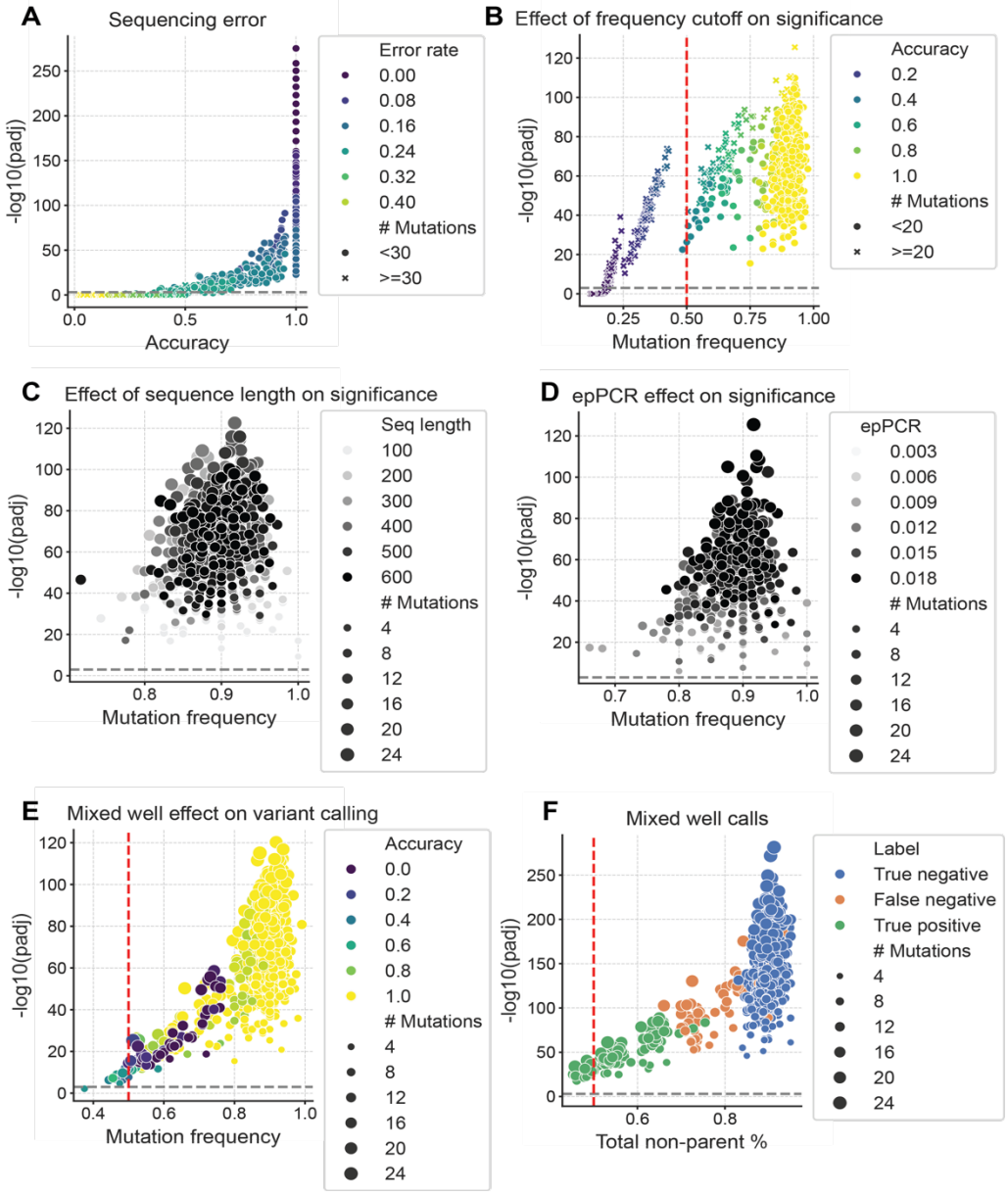
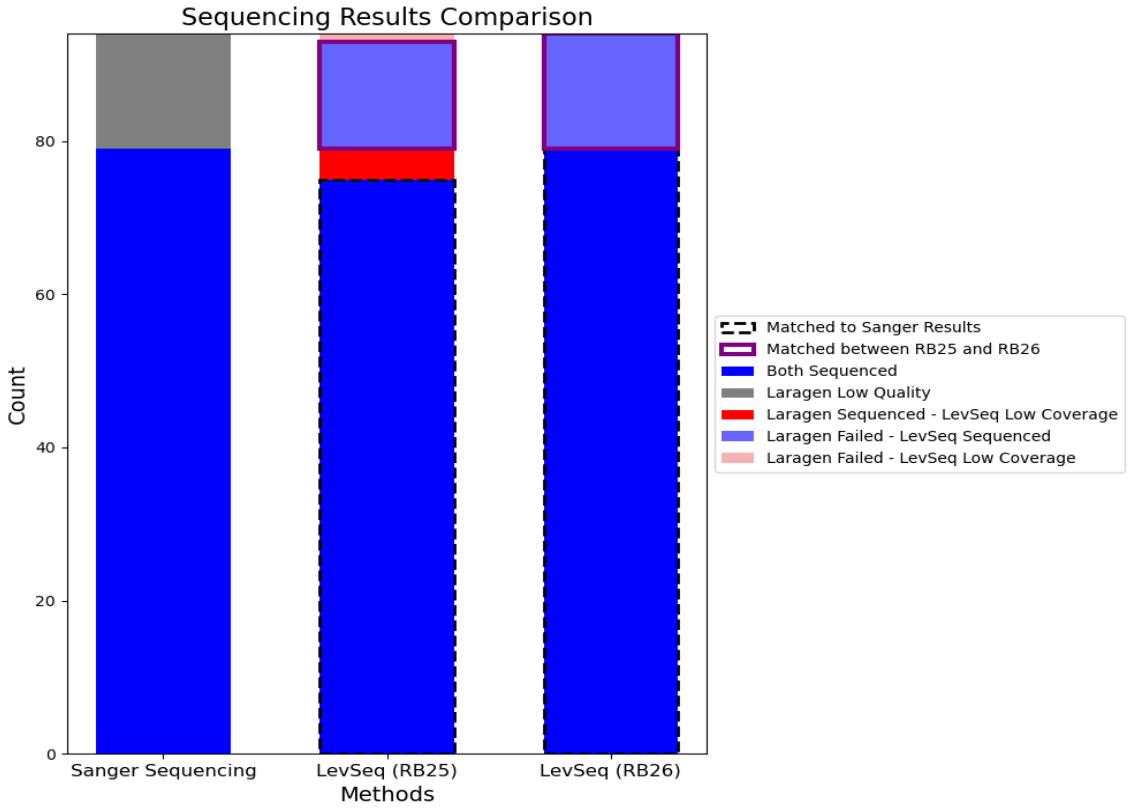


Figure B4-1. epPCR simulation study results. A. Nanopore sequencing error was varied from 0 to 50% at a step size of 5, holding the epPCR mutation rate constant at 2%, read depth of 10, and a frequency cutoff of 50%, which is default in LevSeq. B. Mutation frequency is a parameter that defines when a mutation is tested to be a true mutation from PCR (significant) rather than a sequencing error. The default is 50%, this means there must be 50% non-parent mutations within the position for significance to be tested, this reduces false positive rates and increases the accuracy of variant calling. C. Sequence length was varied while holding the other parameters constant, no trend is observed showing sequence length does not affect the identification of correct sequence, so gene length does not affect our software's variant calling ability. D. epPCR error rate was varied up to 2%, and no

adverse effects were observed on the variant calling accuracy. E. Mixed wells were simulated by combining wells (doping in one sequence at a certain percentage), for these tests it was up to 50%, holding the other parameters constant. For 20 reads there is a high false negative rate, with mixed wells unable to be accurately predicted at the low sequence rate, at less than 10 reads we can't detect true positives. F. Simulated as in E, except colored by the call (e.g. whether the well was called as a mixed-well) rather than the accuracy of calling each mutation, the total accuracy for calling the well was 0.92, precision of 1.0 and recall of 0.61.



*For low quality or low coverage sequences see Supplementary Figure S6 – S25 for details

Figure B4-2. Comparison of sequencing results between Sanger (Laragen) and LevSeq pipelines (LevSeq-RB25 and LevSeq-RB26). The same plate of samples was sequenced using Laragen (miniprep sequencing) and LevSeq (with two barcode plates: LevSeq-RB25 and LevSeq-RB26) to assess reproducibility and accuracy. Out of 94 total samples, Laragen successfully sequenced 79 samples, while 15 samples failed to sequence. For the 79 successfully sequenced samples, LevSeq-RB26 matched all 79 sequences, whereas LevSeq-RB25 matched 75 sequences, with 4 wells marked as #N.A.# due to insufficient sequencing coverage. For the 15 samples that failed to sequence with Laragen, LevSeq-RB26 successfully sequenced all 15, while LevSeq-RB25 sequenced 14, with 1 well marked as #N.A.#. Importantly, no mismatches were found among the successful sequences by both pipelines.

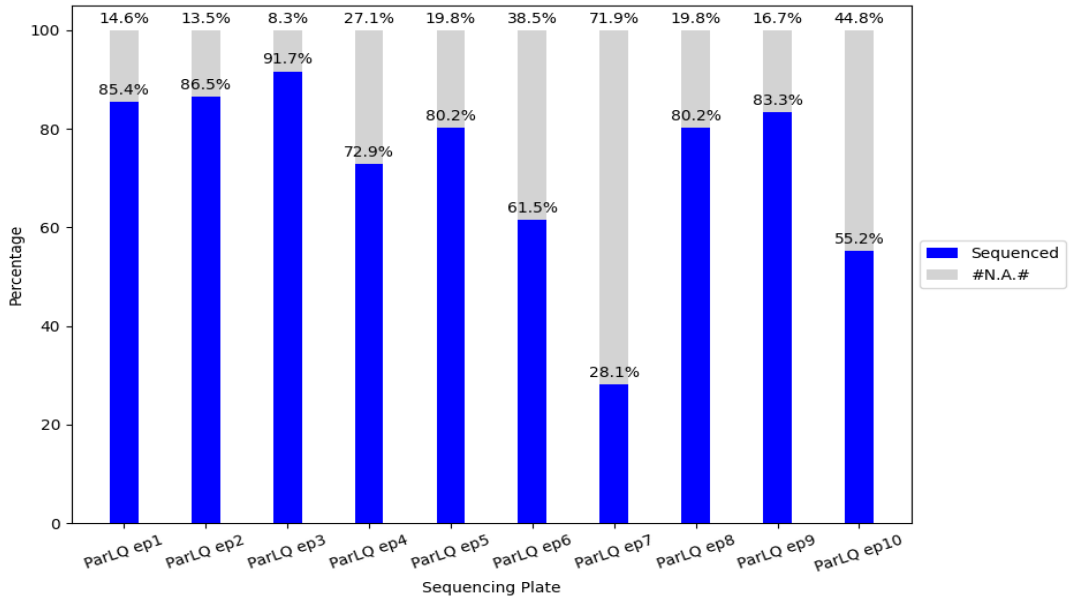


Figure B4-3. ParLQ (~200 amino acids) error prone library sequencing percentage by experiment plate. Plate 6, 7, and 10 had below 60% variants sequenced while the other seven plates have above 70% variants sequenced. This sequencing run was performed without optimizing PCR procedure.

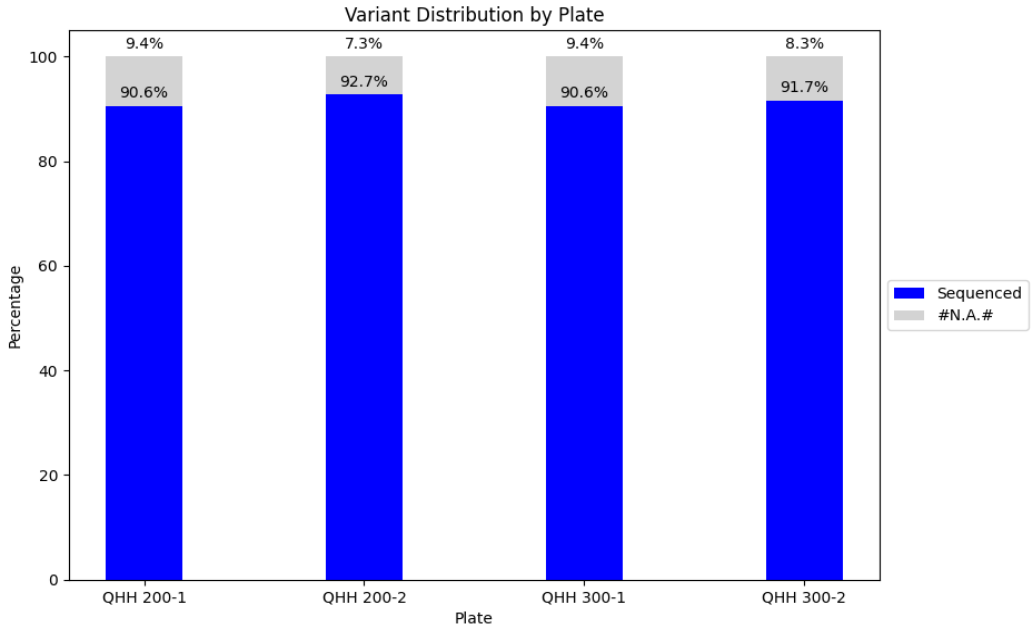


Figure B.4-4. QHH (~500 amino acids) error-prone library sequencing percentage by experiment plate. This sequencing run is performed with updated PCR procedure.

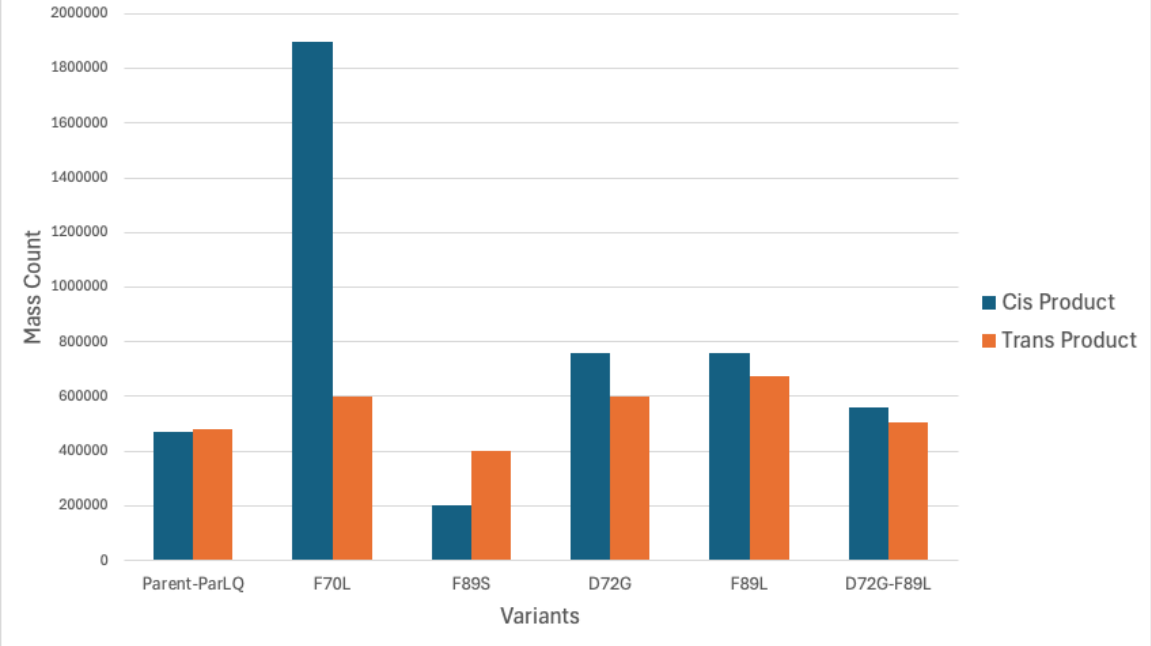


Figure B4-5. Mass spec ion count of cis and trans products for individually tested parents and selected variants. The figure is generated using ion counts and are not validated with standard curve.

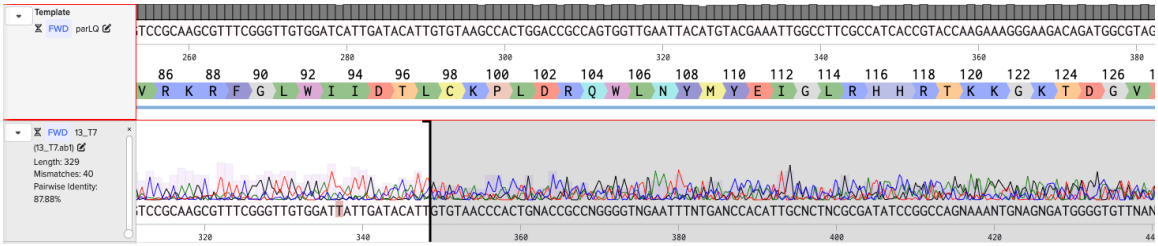


Figure B4-6. MAFFT alignment of Laragen low quality sequence B1.

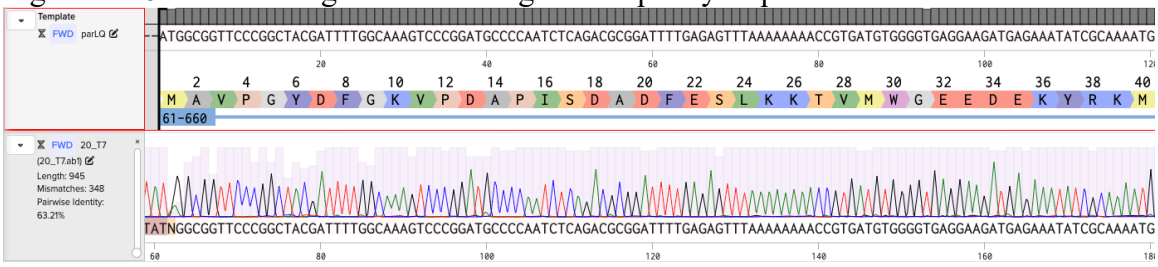


Figure B4-7. MAFFT alignment of Laragen low quality sequence B8.

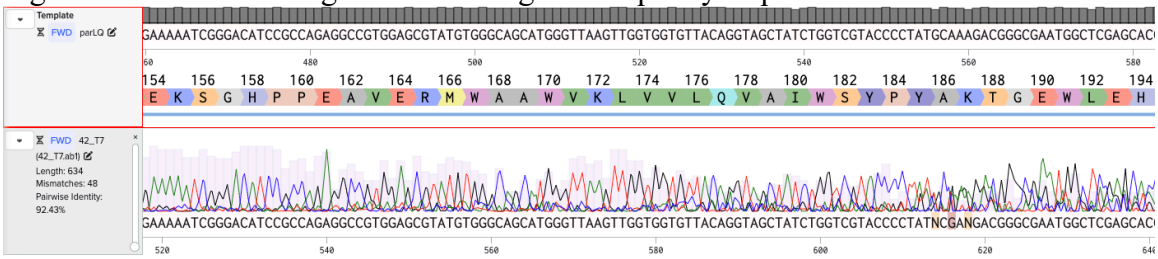


Figure B4-8. MAFFT alignment of Laragen low quality sequence D6.

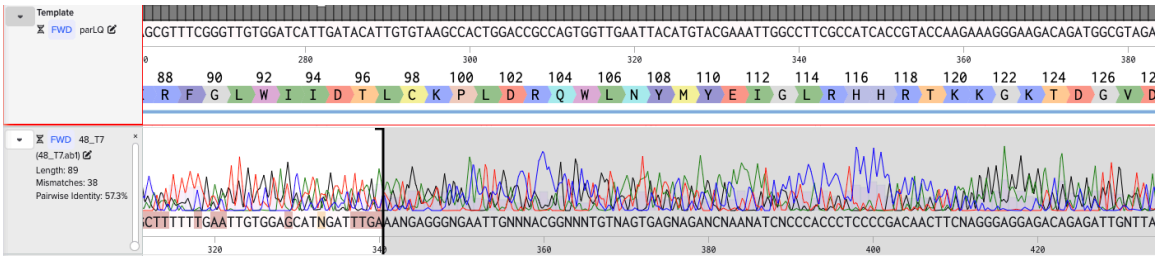


Figure B4-9. MAFFT alignment of Laragen low quality sequence D12.

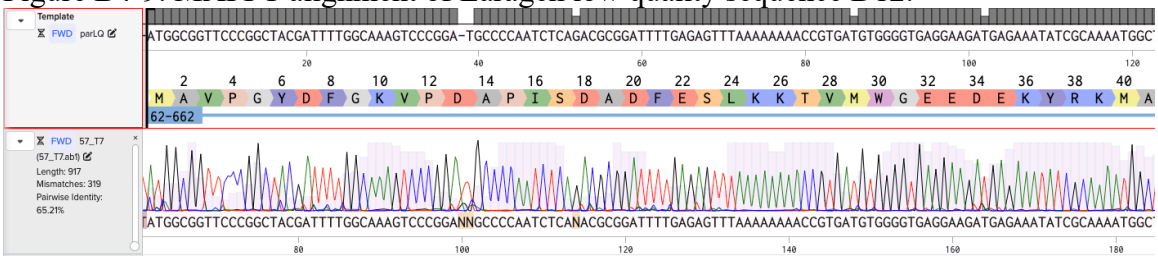


Figure B4-10. MAFFT alignment of Laragen low quality sequence E9.

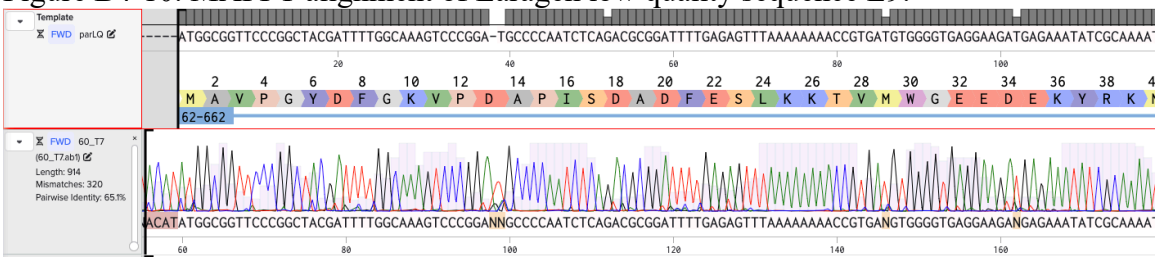


Figure B4-11. MAFFT alignment of Laragen low quality sequence E12.

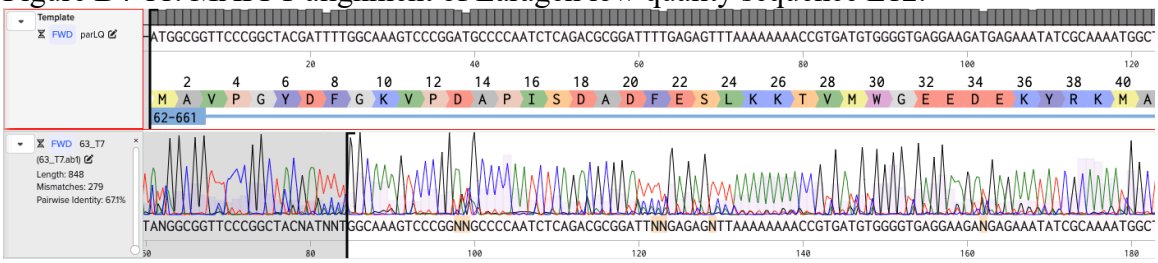


Figure B4-12. MAFFT alignment of Laragen low quality sequence F3.

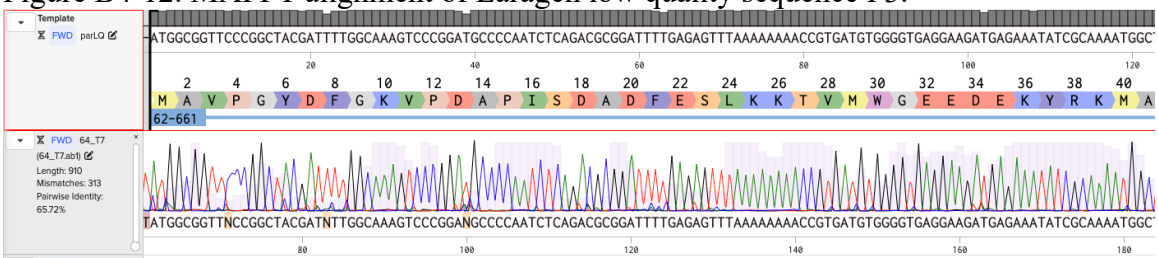


Figure B4-13. MAFFT alignment of Laragen low quality sequence F4.

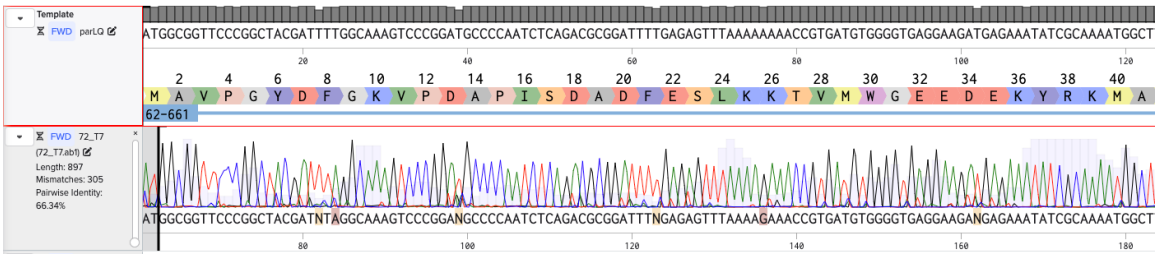


Figure B4-14. MAFFT alignment of Laragen low quality sequence F12.

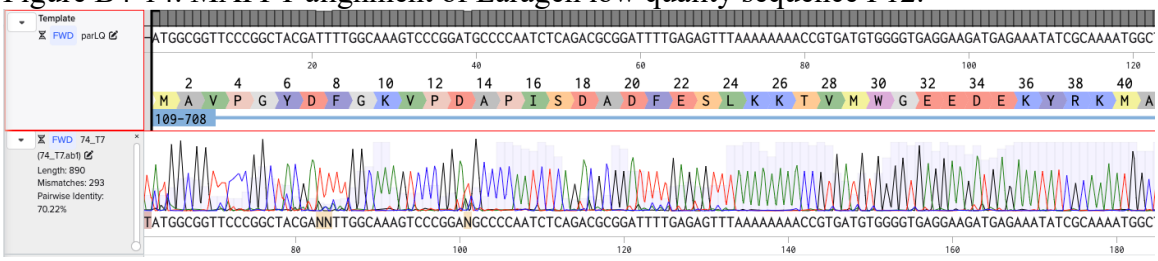


Figure B4-15. MAFFT alignment of Laragen low quality sequence G2.

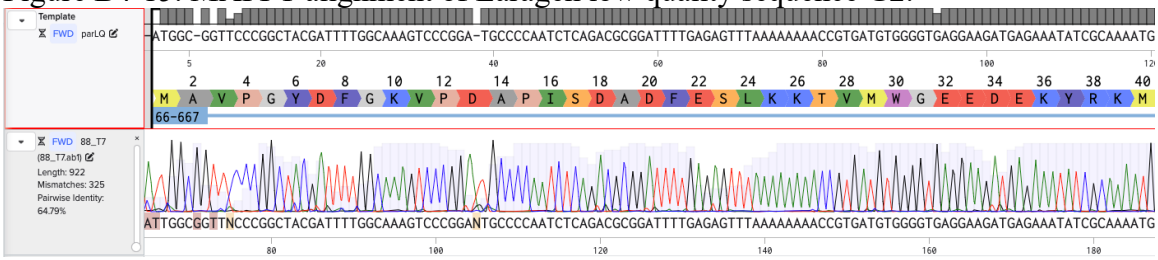


Figure B4-16. MAFFT alignment of Laragen low quality sequence H4.

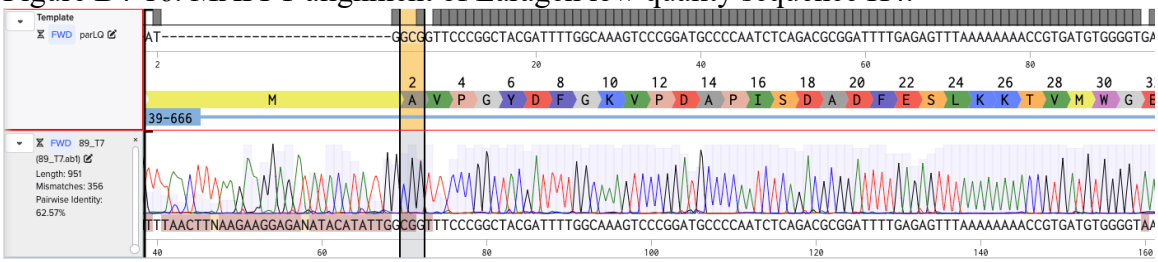


Figure B4-17. MAFFT alignment of Laragen low quality sequence H5.

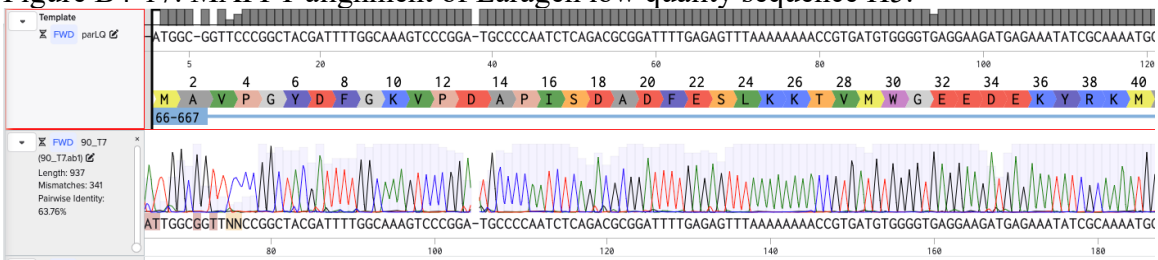


Figure B4-18. MAFFT alignment of Laragen low quality sequence H6.

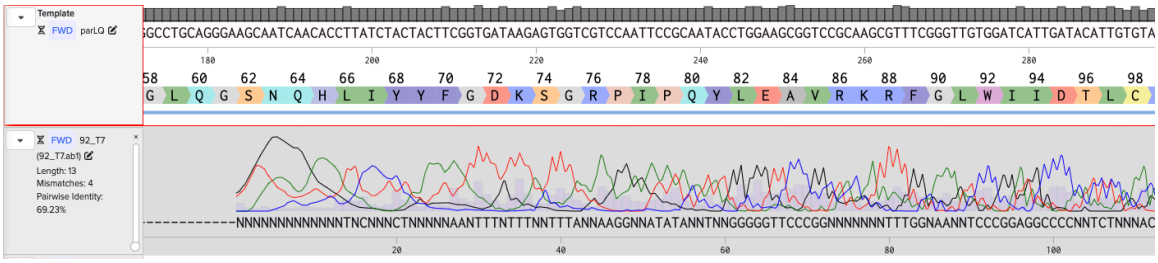


Figure B4-19. MAFFT alignment of Laragen low quality sequence H11.

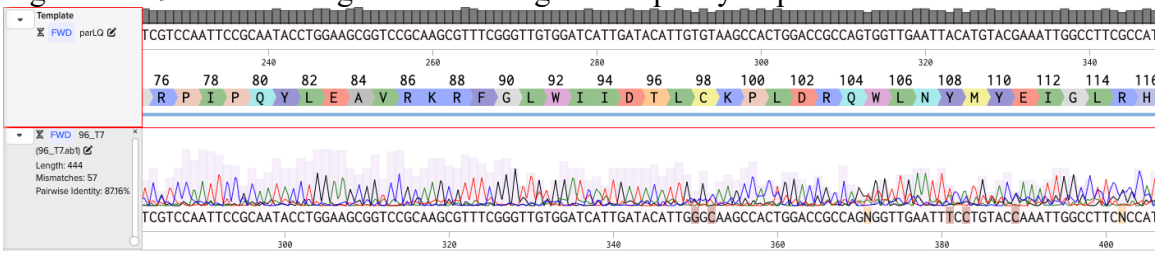


Figure B4-20. MAFFT alignment of Laragen low quality sequence H12.

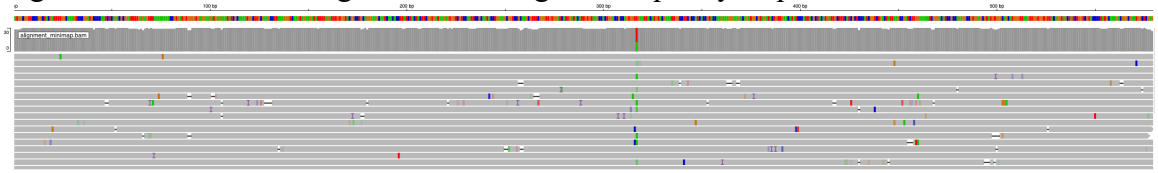


Figure B4-21. Bam alignment of 12 counts from LevSeq (RB25) low coverage sequence D1 (12 counts by LevSeq), mixed well with T317A majority.

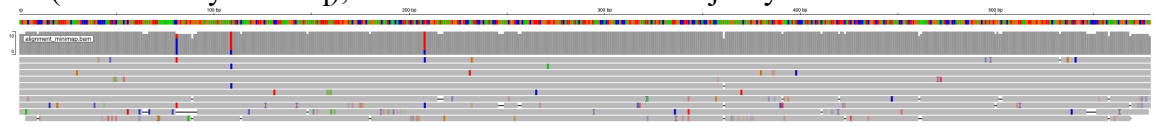


Figure B4-22. Bam alignment of LevSeq (RB25) low coverage sequence E4 (4 counts by LevSeq), C81T_T208C_A232G majority.

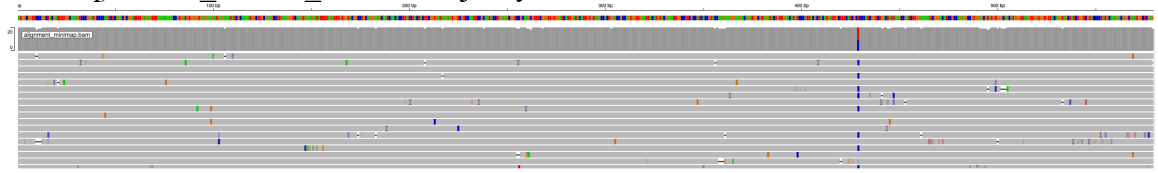


Figure B4-23. Bam alignment of LevSeq (RB25) low coverage sequence E7 (9 counts by LevSeq), mixed well with T213C majority.

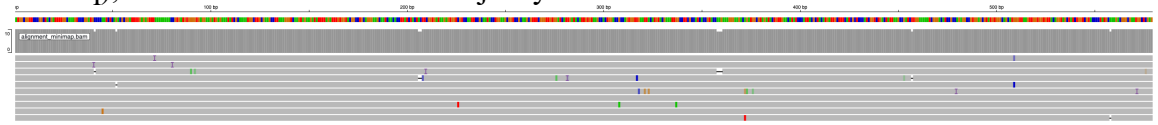


Figure B4-24. Bam alignment of LevSeq (RB25) low coverage sequence E8 (5 counts by LevSeq), it is calling a #PARENT# variant.

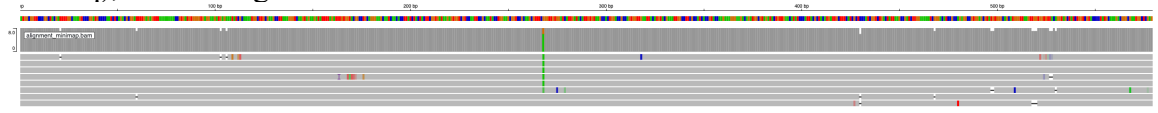


Figure B4-25. Bam alignment of LevSeq (RB25) low coverage sequence F3 (7 counts by LevSeq), G268A majority.

B.5 Supplementary Tables

B.5.1 Barcode and Primer Sequences

Table B1. LevSeq forward barcode sequences used in this work. NB indicates forward barcodes.

Well Position	Name	Barcode
A1	NB01	CACAAAGACACCGACAACTTTCTT
A2	NB02	ACAGACGACTACAAACGGAATCGA
A3	NB03	CCTGGTAACTGGGACACAAGACTC
A4	NB04	TAGGGAAACACGATAGAATCCGAA
A5	NB05	AAGGTTACACAAACCCTGGACAAG
A6	NB06	GACTACTTTCTGCCTTTGCGAGAA
A7	NB07	AAGGATTCATTCCCACGGTAACAC
A8	NB08	ACGTAACTTGGTTTGTTCCCTGAA
A9	NB09	AACCAAGACTCGCTGTGCCTAGTT
A10	NB10	GAGAGGACAAAGGTTTCAACGCTT
A11	NB11	TCCATTCCCTCCGATAGATGAAAC
A12	NB12	TCCGATTCTGCTTCTTTCTACCTG
B1	NB13	AGAACGACTTCCATACTCGTGTGA
B2	NB14	AACGAGTCTCTTGGGACCCATAGA
B3	NB15	AGGTCTACCTCGCTAACACCACTG
B4	NB16	CGTCAACTGACAGTGGTTCGTACT
B5	NB17	ACCCTCCAGGAAAGTACCTCTGAT
B6	NB18	CCAAACCCAACAACCTAGATAGGC
B7	NB19	GTTCCTCGTGCAGTGTCAAGAGAT
B8	NB20	TTGCGTCCTGTTACGAGAACTCAT
B9	NB21	GAGCCTCTCATTGTCCGTTCTCTA
B10	NB22	ACCACTGCCATGTATCAAAGTACG
B11	NB23	CTTACTACCCAGTGAACCTCCTCG
B12	NB24	GCATAGTTCTGCATGATGGGTTAG
C1	NB25	GTAAGTTGGGTATGCAACGCAATG
C2	NB26	CATACAGCGACTACGCATTCTCAT
C3	NB27	CGACGGTTAGATTCACCTCTTACA
C4	NB28	TGAAACCTAAGAAGGCACCGTATC
C5	NB29	CTAGACACCTTGGGTTGACAGACC
C6	NB30	TCAGTGAGGATCTACTTCGACCCA

C7	NB31	TGCGTACAGCAATCAGTTACATTG
C8	NB32	CCAGTAGAAGTCCGACAACGTCAT
C9	NB33	CAGACTTGGTACGGTTGGGTAACT
C10	NB34	GGACGAAGAACTCAAGTCAAAGGC
C11	NB35	CTACTTACGAAGCTGAGGGACTGC
C12	NB36	ATGTCCCAGTTAGAGGAGGAAACA
D1	NB37	GCTTGCGATTGATGCTTAGTATCA
D2	NB38	ACCACAGGAGGACGATACAGAGAA
D3	NB39	CCACAGTGTCAACTAGAGCCTCTC
D4	NB40	TAGTTTGGATGACCAAGGATAGCC
D5	NB41	GGAGTTCGTCCAGAGAAGTACACG
D6	NB42	CTACGTGTAAGGCATACCTGCCAG
D7	NB43	CTTTCGTTGTTGACTCGACGGTAG
D8	NB44	AGTAGAAAGGGTTCCTTCCCACTC
D9	NB45	GATCCAACAGAGATGCCTTCAGTG
D10	NB46	GCTGTGTTCCACTTCATTCTCCTG
D11	NB47	GTGCAACTTTCCCACAGGTAGTTC
D12	NB48	CATCTGGAACGTGGTACACCTGTA
E1	NB49	ACTGGTGCAGCTTTGAACATCTAG
E2	NB50	ATGGACTTTGGTAACTTCCTGCGT
E3	NB51	GTTGAATGAGCCTACTGGGTCCTC
E4	NB52	TGAGAGACAAGATTGTTCGTGGAC
E5	NB53	AGATTCAGACCGTCTCATGCAAAG
E6	NB54	CAAGAGCTTTGACTAAGGAGCATG
E7	NB55	TGGAAGATGAGACCCTGATCTACG
E8	NB56	TCACTACTCAACAGGTGGCATGAA
E9	NB57	GCTAGGTCAATCTCCTTCGGAAGT
E10	NB58	CAGGTTACTCCTCCGTGAGTCTGA
E11	NB59	TCAATCAAGAAGGGAAAGCAAGGT
E12	NB60	CATGTTCAACCAAGGCTTCTATGG
F1	NB61	AGAGGGTACTATGTGCCTCAGCAC
F2	NB62	CACCCACACTTACTTCAGGACGTA
F3	NB63	TTCTGAAGTTCCTGGGTCTTGAAC
F4	NB64	GACAGACACCGTTCATCGACTTTC
F5	NB65	TTCTCAGTCTTCCTCCAGACAAGG
F6	NB66	CCGATCCTTGTGGCTTCTAACTTC
F7	NB67	GTTTGTCATACTCGTGTGCTCACC

F8	NB68	GAATCTAAGCAAACACGAAGGTGG
F9	NB69	TACAGTCCGAGCCTCATGTGATCT
F10	NB70	ACCGAGATCCTACGAATGGAGTGT
F11	NB71	CCTGGGAGCATCAGGTAGTAACAG
F12	NB72	TAGCTGACTGTCTTCCATACCGAC
G1	NB73	AAGAAACAGGATGACAGAACCCTC
G2	NB74	TACAAGCATCCCAACACTTCCACT
G3	NB75	GACCATTGTGATGAACCCTGTTGT
G4	NB76	ATGCTTGTTACATCAACCCTGGAC
G5	NB77	CGACCTGTTTCTCAGGGATACAAC
G6	NB78	AACAACCGAACCTTTGAATCAGAA
G7	NB79	TCTCGGAGATAGTTCTCACTGCTG
G8	NB80	CGGATGAACATAGGATAGCGATTC
G9	NB81	CCTCATCTTGTGAAGTTGTTTCGG
G10	NB82	ACGGTATGTCGAGTTCCAGGACTA
G11	NB83	TGGCTTGATCTAGGTAAGGTCGAA
G12	NB84	GTAGTGGACCTAGAACCTGTGCCA
H1	NB85	AACGGAGGAGTTAGTTGGATGATC
H2	NB86	AGGTGATCCCAACAAGCGTAAGTA
H3	NB87	TACATGCTCCTGTTGTTAGGGAGG
H4	NB88	TCTTCTACTACCGATCCGAAGCAG
H5	NB89	ACAGCATCAATGTTTGGCTAGTTG
H6	NB90	GATGTAGAGGGTACGGTTTGAGGC
H7	NB91	GGCTCCATAGGAACTCACGCTACT
H8	NB92	TTGTGAGTGGAAAGATACAGGACC
H9	NB93	AGTTTCCATCACTTCAGACTTGGG
H10	NB94	GATTGTCCTCAAACTGCCACCTAC
H11	NB95	CCTGTCTGGAAGAAGAATGGACTT
H12	NB96	CTGAACGGTCATAGAGTCCACCAT

Table B2. LevSeq reverse barcode sequences used in this work. RB stands for reverse barcodes.

Well Position	Name	Barcode
A1	RB01	AAGAAAGTTGTCGGTGTCTTTGTG
A2	RB02	TCGATTCCGTTTGTAGTCGTCTGT
A3	RB03	GAGTCTTGTGTCCCAGTTACCAGG

A4	RB04	TTCGGATTCTATCGTGTTTCCCTA
A5	RB05	CTTGTCCAGGGTTTGTGTAACCTT
A6	RB06	TTCTCGCAAAGGCAGAAAGTAGTC
A7	RB07	GTGTTACCGTGGGAATGAATCCTT
A8	RB08	TTCAGGGAACAAACCAAGTTACGT
A9	RB09	AACTAGGCACAGCGAGTCTTGGTT
A10	RB10	AAGCGTTGAAACCTTTGTCCTCTC
A11	RB11	GTTTCATCTATCGGAGGGAATGGA
A12	RB12	CAGGTAGAAAGAAGCAGAATCGGA
B1	RB13	AGAACGACTTCCATACTCGTGTGA
B2	RB14	AACGAGTCTCTTGGGACCCATAGA
B3	RB15	AGGTCTACCTCGCTAACACCACTG
B4	RB16	CGTCAACTGACAGTGGTTCGTACT
B5	RB17	ACCCTCCAGGAAAGTACCTCTGAT
B6	RB18	CCAAACCCAACAACCTAGATAGGC
B7	RB19	GTTCCTCGTGCAGTGTCAAGAGAT
B8	RB20	TTGCGTCCTGTTACGAGAACTCAT
B9	RB21	GAGCCTCTCATTGTCCGTTCTCTA
B10	RB22	ACCACTGCCATGTATCAAAGTACG
B11	RB23	CTTACTACCCAGTGAACCTCCTCG
B12	RB24	GCATAGTTCTGCATGATGGGTTAG
C1	RB25	GTAAGTTGGGTATGCAACGCAATG
C2	RB26	CATACAGCGACTACGCATTCTCAT
C3	RB27	CGACGGTTAGATTCACCTCTTACA
C4	RB28	TGAAACCTAAGAAGGCACCGTATC
C5	RB29	CTAGACACCTTGGGTTGACAGACC
C6	RB30	TCAGTGAGGATCTACTTCGACCCA
C7	RB31	TGCGTACAGCAATCAGTTACATTG
C8	RB32	CCAGTAGAAGTCCGACAACGTCAT
C9	RB33	CAGACTTGGTACGGTTGGGTAACT
C10	RB34	GGACGAAGAACTCAAGTCAAAGGC
C11	RB35	CTACTTACGAAGCTGAGGGACTGC
C12	RB36	ATGTCCCAGTTAGAGGAGGAAACA
D1	RB37	GCTTGCGATTGATGCTTAGTATCA
D2	RB38	ACCACAGGAGGACGATACAGAGAA
D3	RB39	CCACAGTGTCAACTAGAGCCTCTC
D4	RB40	TAGTTTGGATGACCAAGGATAGCC

D5	RB41	GGAGTTCGTCCAGAGAAGTACACG
D6	RB42	CTACGTGTAAGGCATACCTGCCAG
D7	RB43	CTTTCGTTGTTGACTCGACGGTAG
D8	RB44	AGTAGAAAGGGTTCCTTCCCACTC
D9	RB45	GATCCAACAGAGATGCCTTCAGTG
D10	RB46	GCTGTGTTCCACTTCATTCTCCTG
D11	RB47	GTGCAACTTTCCCACAGGTAGTTC
D12	RB48	CATCTGGAACGTGGTACACCTGTA
E1	RB49	ACTGGTGCAGCTTTGAACATCTAG
E2	RB50	ATGGACTTTGGTAACTTCCTGCGT
E3	RB51	GTTGAATGAGCCTACTGGGTCCTC
E4	RB52	TGAGAGACAAGATTGTTCGTGGAC
E5	RB53	AGATTCAGACCGTCTCATGCAAAG
E6	RB54	CAAGAGCTTTGACTAAGGAGCATG
E7	RB55	TGGAAGATGAGACCCTGATCTACG
E8	RB56	TCACTACTCAACAGGTGGCATGAA
E9	RB57	GCTAGGTCAATCTCCTTCGGAAGT
E10	RB58	CAGGTTACTCCTCCGTGAGTCTGA
E11	RB59	TCAATCAAGAAGGGAAAGCAAGGT
E12	RB60	CATGTTCAACCAAGGCTTCTATGG
F1	RB61	AGAGGGTACTATGTGCCTCAGCAC
F2	RB62	CACCCACACTTACTTCAGGACGTA
F3	RB63	TTCTGAAGTTCCTGGGTCTTGAAC
F4	RB64	GACAGACACCGTTCATCGACTTTC
F5	RB65	TTCTCAGTCTTCCTCCAGACAAGG
F6	RB66	CCGATCCTTGTGGCTTCTAACTTC
F7	RB67	GTTTGTCATACTCGTGTGCTCACC
F8	RB68	GAATCTAAGCAAACACGAAGGTGG
F9	RB69	TACAGTCCGAGCCTCATGTGATCT
F10	RB70	ACCGAGATCCTACGAATGGAGTGT
F11	RB71	CCTGGGAGCATCAGGTAGTAACAG
F12	RB72	TAGCTGACTGTCTTCCATACCGAC
G1	RB73	AAGAAACAGGATGACAGAACCCTC
G2	RB74	TACAAGCATCCCAACACTTCCACT
G3	RB75	GACCATTGTGATGAACCCTGTTGT
G4	RB76	ATGCTTGTTACATCAACCCTGGAC
G5	RB77	CGACCTGTTTCTCAGGGATACAAC

G6	RB78	AACAACCGAACCTTTGAATCAGAA
G7	RB79	TCTCGGAGATAGTTCTCACTGCTG
G8	RB80	CGGATGAACATAGGATAGCGATTC
G9	RB81	CCTCATCTTGTGAAGTTGTTTCGG
G10	RB82	ACGGTATGTCGAGTTCCAGGACTA
G11	RB83	TGGCTTGATCTAGGTAAGGTCGAA
G12	RB84	GTAGTGGACCTAGAACCTGTGCCA
H1	RB85	AACGGAGGAGTTAGTTGGATGATC
H2	RB86	AGGTGATCCCAACAAGCGTAAGTA
H3	RB87	TACATGCTCCTGTTGTTAGGGAGG
H4	RB88	TCTTCTACTACCGATCCGAAGCAG
H5	RB89	ACAGCATCAATGTTTGGCTAGTTG
H6	RB90	GATGTAGAGGGTACGGTTTGAGGC
H7	RB91	GGCTCCATAGGAACTCACGCTACT
H8	RB92	TTGTGAGTGGAAAGATACAGGACC
H9	RB93	AGTTTCCATCACTTCAGACTTGGG
H10	RB94	GATTGTCCTCAAACTGCCACCTAC
H11	RB95	CCTGTCTGGAAGAAGAATGGACTT
H12	RB96	CTGAACGGTCATAGAGTCCACCAT

Table B3. Full-length LevSeq forward barcode-linked primer sequences used in this work. The primer-linked barcodes below are directly ordered from IDT.

Well Position	**Name**	**Barcode Linked Primer Sequence**
A1	NB01	CACAAAGACACCGACAACTTTCTTATCTCGATCCCGCGAAATTAATACGACTCAC
A2	NB02	ACAGACGACTACAAACGGAATCGAATCTCGATCCCGCGAAATTAATACGACTCAC
A3	NB03	CCTGGTAACTGGGACACAAGACTCATCTCGATCCCGCGAAATTAATACGACTCAC
A4	NB04	TAGGGAAACACGATAGAATCCGAAATCTCGATCCCGCGAAATTAATACGACTCAC
A5	NB05	AAGGTTACACAAACCCTGGACAAGATCTCGATCCCGCGAAATTAATACGACTCAC
A6	NB06	GACTACTTTCTGCCTTTGCGAGAAATCTCGATCCCGCGAAATTAATACGACTCAC
A7	NB07	AAGGATTCATTCCCACGGTAACACATCTCGATCCCGCGAAATTAATACGACTCAC
A8	NB08	ACGTAACTTGGTTTGTTCCCTGAAATCTCGATCCCGCGAAATTAATACGACTCAC

A9	NB09	AACCAAGACTCGCTGTGCCTAGTTATCTCGATCCCGCGAAATTAATACGACTCAC
A10	NB10	GAGAGGACAAAGGTTTCAACGCTTATCTCGATCCCGCGAAATTAATACGACTCAC
A11	NB11	TCCATTCCCTCCGATAGATGAAACATCTCGATCCCGCGAAATTAATACGACTCAC
A12	NB12	TCCGATTCTGCTTCTTTCTACCTGATCTCGATCCCGCGAAATTAATACGACTCAC
B1	NB13	AGAACGACTTCCATACTCGTGTGAATCTCGATCCCGCGAAATTAATACGACTCAC
B2	NB14	AACGAGTCTCTTGGGACCCATAGAATCTCGATCCCGCGAAATTAATACGACTCAC
B3	NB15	AGGTCTACCTCGCTAACACCACTGATCTCGATCCCGCGAAATTAATACGACTCAC
B4	NB16	CGTCAACTGACAGTGGTTCGTACTATCTCGATCCCGCGAAATTAATACGACTCAC
B5	NB17	ACCCTCCAGGAAAGTACCTCTGATATCTCGATCCCGCGAAATTAATACGACTCAC
B6	NB18	CCAAACCCAACAACCTAGATAGGCATCTCGATCCCGCGAAATTAATACGACTCAC
B7	NB19	GTTCCTCGTGCAGTGTCAAGAGATATCTCGATCCCGCGAAATTAATACGACTCAC
B8	NB20	TTGCGTCCTGTTACGAGAACTCATATCTCGATCCCGCGAAATTAATACGACTCAC
B9	NB21	GAGCCTCTCATTGTCCGTTCTCTAATCTCGATCCCGCGAAATTAATACGACTCAC
B10	NB22	ACCACTGCCATGTATCAAAGTACGATCTCGATCCCGCGAAATTAATACGACTCAC
B11	NB23	CTTACTACCCAGTGAACCTCCTCGATCTCGATCCCGCGAAATTAATACGACTCAC
B12	NB24	GCATAGTTCTGCATGATGGGTTAGATCTCGATCCCGCGAAATTAATACGACTCAC
C1	NB25	GTAAGTTGGGTATGCAACGCAATGATCTCGATCCCGCGAAATTAATACGACTCAC
C2	NB26	CATACAGCGACTACGCATTCTCATATCTCGATCCCGCGAAATTAATACGACTCAC
C3	NB27	CGACGGTTAGATTCACCTCTTACAATCTCGATCCCGCGAAATTAATACGACTCAC
C4	NB28	TGAAACCTAAGAAGGCACCGTATCATCTCGATCCCGCGAAATTAATACGACTCAC
C5	NB29	CTAGACACCTTGGGTTGACAGACCATCTCGATCCCGCGAAATTAATACGACTCAC
C6	NB30	TCAGTGAGGATCTACTTCGACCCAATCTCGATCCCGCGAAATTAATACGACTCAC
C7	NB31	TGCGTACAGCAATCAGTTACATTGATCTCGATCCCGCGAAATTAATACGACTCAC
C8	NB32	CCAGTAGAAGTCCGACAACGTCATATCTCGATCCCGCGAAATTAATACGACTCAC
C9	NB33	CAGACTTGGTACGGTTGGGTAACTATCTCGATCCCGCGAAATTAATACGACTCAC

C10	NB34	GGACGAAGAACTCAAGTCAAAGGCATCTCGATCCCGCGAAATTAATACG ACTCAC
C11	NB35	CTACTTACGAAGCTGAGGGACTGCATCTCGATCCCGCGAAATTAATACGA CTCAC
C12	NB36	ATGTCCCAGTTAGAGGAGGAAACAATCTCGATCCCGCGAAATTAATACG ACTCAC
D1	NB37	GCTTGCGATTGATGCTTAGTATCAATCTCGATCCCGCGAAATTAATACGA CTCAC
D2	NB38	ACCACAGGAGGACGATACAGAGAAATCTCGATCCCGCGAAATTAATACG ACTCAC
D3	NB39	CCACAGTGTCAACTAGAGCCTCTCATCTCGATCCCGCGAAATTAATACGA CTCAC
D4	NB40	TAGTTTGGATGACCAAGGATAGCCATCTCGATCCCGCGAAATTAATACGA CTCAC
D5	NB41	GGAGTTCGTCCAGAGAAGTACACGATCTCGATCCCGCGAAATTAATACG ACTCAC
D6	NB42	CTACGTGTAAGGCATACCTGCCAGATCTCGATCCCGCGAAATTAATACGA CTCAC
D7	NB43	CTTTCGTTGTTGACTCGACGGTAGATCTCGATCCCGCGAAATTAATACGA CTCAC
D8	NB44	AGTAGAAAGGGTTCCTTCCCACTCATCTCGATCCCGCGAAATTAATACGA CTCAC
D9	NB45	GATCCAACAGAGATGCCTTCAGTGATCTCGATCCCGCGAAATTAATACGA CTCAC
D10	NB46	GCTGTGTTCCACTTCATTCTCCTGATCTCGATCCCGCGAAATTAATACGAC TCAC
D11	NB47	GTGCAACTTTCCCACAGGTAGTTCATCTCGATCCCGCGAAATTAATACGA CTCAC
D12	NB48	CATCTGGAACGTGGTACACCTGTAATCTCGATCCCGCGAAATTAATACGA CTCAC
E1	NB49	ACTGGTGCAGCTTTGAACATCTAGATCTCGATCCCGCGAAATTAATACGA CTCAC
E2	NB50	ATGGACTTTGGTAACTTCCTGCGTATCTCGATCCCGCGAAATTAATACGA CTCAC
E3	NB51	GTTGAATGAGCCTACTGGGTCCTCATCTCGATCCCGCGAAATTAATACGA CTCAC
E4	NB52	TGAGAGACAAGATTGTTCGTGGACATCTCGATCCCGCGAAATTAATACG ACTCAC
E5	NB53	AGATTCAGACCGTCTCATGCAAAGATCTCGATCCCGCGAAATTAATACGA CTCAC
E6	NB54	CAAGAGCTTTGACTAAGGAGCATGATCTCGATCCCGCGAAATTAATACG ACTCAC
E7	NB55	TGGAAGATGAGACCCTGATCTACGATCTCGATCCCGCGAAATTAATACG ACTCAC
E8	NB56	TCACTACTCAACAGGTGGCATGAAATCTCGATCCCGCGAAATTAATACGA CTCAC
E9	NB57	GCTAGGTCAATCTCCTTCGGAAGTATCTCGATCCCGCGAAATTAATACGA CTCAC
E10	NB58	CAGGTTACTCCTCCGTGAGTCTGAATCTCGATCCCGCGAAATTAATACGA CTCAC

E11	NB59	TCAATCAAGAAGGGAAAGCAAGGTATCTCGATCCCGCGAAATTAATACGACTCAC
E12	NB60	CATGTTCAACCAAGGCTTCTATGGATCTCGATCCCGCGAAATTAATACGACTCAC
F1	NB61	AGAGGGTACTATGTGCCTCAGCACATCTCGATCCCGCGAAATTAATACGACTCAC
F2	NB62	CACCCACACTTACTTCAGGACGTAATCTCGATCCCGCGAAATTAATACGACTCAC
F3	NB63	TTCTGAAGTTCCTGGGTCTTGAACATCTCGATCCCGCGAAATTAATACGACTCAC
F4	NB64	GACAGACACCGTTCATCGACTTTCATCTCGATCCCGCGAAATTAATACGACTCAC
F5	NB65	TTCTCAGTCTTCCTCCAGACAAGGATCTCGATCCCGCGAAATTAATACGACTCAC
F6	NB66	CCGATCCTTGTGGCTTCTAACTTCATCTCGATCCCGCGAAATTAATACGACTCAC
F7	NB67	GTTTGTCATACTCGTGTGCTCACCATCTCGATCCCGCGAAATTAATACGACTCAC
F8	NB68	GAATCTAAGCAAACACGAAGGTGGATCTCGATCCCGCGAAATTAATACGACTCAC
F9	NB69	TACAGTCCGAGCCTCATGTGATCTATCTCGATCCCGCGAAATTAATACGACTCAC
F10	NB70	ACCGAGATCCTACGAATGGAGTGTATCTCGATCCCGCGAAATTAATACGACTCAC
F11	NB71	CCTGGGAGCATCAGGTAGTAACAGATCTCGATCCCGCGAAATTAATACGACTCAC
F12	NB72	TAGCTGACTGTCTTCCATACCGACATCTCGATCCCGCGAAATTAATACGACTCAC
G1	NB73	AAGAAACAGGATGACAGAACCCTCATCTCGATCCCGCGAAATTAATACGACTCAC
G2	NB74	TACAAGCATCCCAACACTTCCACTATCTCGATCCCGCGAAATTAATACGACTCAC
G3	NB75	GACCATTGTGATGAACCCTGTTGTATCTCGATCCCGCGAAATTAATACGACTCAC
G4	NB76	ATGCTTGTTACATCAACCCTGGACATCTCGATCCCGCGAAATTAATACGACTCAC
G5	NB77	CGACCTGTTTCTCAGGGATACAACATCTCGATCCCGCGAAATTAATACGACTCAC
G6	NB78	AACAACCGAACCTTTGAATCAGAAATCTCGATCCCGCGAAATTAATACGACTCAC
G7	NB79	TCTCGGAGATAGTTCTCACTGCTGATCTCGATCCCGCGAAATTAATACGACTCAC
G8	NB80	CGGATGAACATAGGATAGCGATTCATCTCGATCCCGCGAAATTAATACGACTCAC
G9	NB81	CCTCATCTTGTGAAGTTGTTTCGGATCTCGATCCCGCGAAATTAATACGACTCAC
G10	NB82	ACGGTATGTCGAGTTCCAGGACTAATCTCGATCCCGCGAAATTAATACGACTCAC
G11	NB83	TGGCTTGATCTAGGTAAGGTCGAAATCTCGATCCCGCGAAATTAATACGACTCAC

G12	NB84	GTAGTGGACCTAGAACCTGTGCCAATCTCGATCCCGCGAAATTAATACGACTCAC
H1	NB85	AACGGAGGAGTTAGTTGGATGATCATCTCGATCCCGCGAAATTAATACGACTCAC
H2	NB86	AGGTGATCCCAACAAGCGTAAGTAATCTCGATCCCGCGAAATTAATACGACTCAC
H3	NB87	TACATGCTCCTGTTGTTAGGGAGGATCTCGATCCCGCGAAATTAATACGACTCAC
H4	NB88	TCTTCTACTACCGATCCGAAGCAGATCTCGATCCCGCGAAATTAATACGACTCAC
H5	NB89	ACAGCATCAATGTTTGGCTAGTTGATCTCGATCCCGCGAAATTAATACGACTCAC
H6	NB90	GATGTAGAGGGTACGGTTTGAGGCATCTCGATCCCGCGAAATTAATACGACTCAC
H7	NB91	GGCTCCATAGGAACTCACGCTACTATCTCGATCCCGCGAAATTAATACGACTCAC
H8	NB92	TTGTGAGTGGAAAGATACAGGACCATCTCGATCCCGCGAAATTAATACGACTCAC
H9	NB93	AGTTTCCATCACTTCAGACTTGGGATCTCGATCCCGCGAAATTAATACGACTCAC
H10	NB94	GATTGTCCTCAAACTGCCACCTACATCTCGATCCCGCGAAATTAATACGACTCAC
H11	NB95	CCTGTCTGGAAGAAGAATGGACTTATCTCGATCCCGCGAAATTAATACGACTCAC
H12	NB96	CTGAACGGTCATAGAGTCCACCATATCTCGATCCCGCGAAATTAATACGACTCAC

Table B4. Full-length LevSeq reverse barcode-linked primer sequences used in this work. The primer-linked barcodes below are directly ordered from IDT.

Well Position	**Name**	**Full Length Order**
A1	RB01	AAGAAAGTTGTCGGTGTCTTTGTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
A2	RB02	TCGATTCCGTTTGTAGTCGTCTGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
A3	RB03	GAGTCTTGTGTCCCAGTTACCAGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
A4	RB04	TTCGGATTCTATCGTGTTTCCCTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
A5	RB05	CTTGTCCAGGGTTTGTGTAACCTTGCCCCAAGGGGTTATGCTAGTTATTGCTC
A6	RB06	TTCTCGCAAAGGCAGAAAGTAGTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
A7	RB07	GTGTTACCGTGGGAATGAATCCTTGCCCCAAGGGGTTATGCTAGTTATTGCTC

A8	RB08	TTCAGGGAACAAACCAAGTTACGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
A9	RB09	AACTAGGCACAGCGAGTCTTGGTTGCCCCAAGGGGTTATGCTAGTTATTGCTC
A10	RB10	AAGCGTTGAAACCTTTGTCCTCTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
A11	RB11	GTTTCATCTATCGGAGGGAATGGAGCCCCAAGGGGTTATGCTAGTTATTGCTC
A12	RB12	CAGGTAGAAAGAAGCAGAATCGGAGCCCCAAGGGGTTATGCTAGTTATTGCTC
B1	RB13	AGAACGACTTCCATACTCGTGTGAGCCCCAAGGGGTTATGCTAGTTATTGCTC
B2	RB14	AACGAGTCTCTTGGGACCCATAGAGCCCCAAGGGGTTATGCTAGTTATTGCTC
B3	RB15	AGGTCTACCTCGCTAACACCACTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
B4	RB16	CGTCAACTGACAGTGGTTCGTACTGCCCCAAGGGGTTATGCTAGTTATTGCTC
B5	RB17	ACCCTCCAGGAAAGTACCTCTGATGCCCCAAGGGGTTATGCTAGTTATTGCTC
B6	RB18	CCAAACCCAACAACCTAGATAGGCGCCCCAAGGGGTTATGCTAGTTATTGCTC
B7	RB19	GTTCCTCGTGCAGTGTCAAGAGATGCCCCAAGGGGTTATGCTAGTTATTGCTC
B8	RB20	TTGCGTCCTGTTACGAGAACTCATGCCCCAAGGGGTTATGCTAGTTATTGCTC
B9	RB21	GAGCCTCTCATTGTCCGTTCTCTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
B10	RB22	ACCACTGCCATGTATCAAAGTACGGCCCCAAGGGGTTATGCTAGTTATTGCTC
B11	RB23	CTTACTACCCAGTGAACCTCCTCGGCCCCAAGGGGTTATGCTAGTTATTGCTC
B12	RB24	GCATAGTTCTGCATGATGGGTTAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
C1	RB25	GTAAGTTGGGTATGCAACGCAATGGCCCCAAGGGGTTATGCTAGTTATTGCTC
C2	RB26	CATACAGCGACTACGCATTCTCATGCCCCAAGGGGTTATGCTAGTTATTGCTC
C3	RB27	CGACGGTTAGATTCACCTCTTACAGCCCCAAGGGGTTATGCTAGTTATTGCTC
C4	RB28	TGAAACCTAAGAAGGCACCGTATCGCCCCAAGGGGTTATGCTAGTTATTGCTC
C5	RB29	CTAGACACCTTGGGTTGACAGACCGCCCCAAGGGGTTATGCTAGTTATTGCTC
C6	RB30	TCAGTGAGGATCTACTTCGACCCAGCCCCAAGGGGTTATGCTAGTTATTGCTC

C7	RB31	TGCGTACAGCAATCAGTTACATTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
C8	RB32	CCAGTAGAAGTCCGACAACGTCATGCCCCAAGGGGTTATGCTAGTTATTGCTC
C9	RB33	CAGACTTGGTACGGTTGGGTAACTGCCCCAAGGGGTTATGCTAGTTATTGCTC
C10	RB34	GGACGAAGAACTCAAGTCAAAGGCGCCCCAAGGGGTTATGCTAGTTATTGCTC
C11	RB35	CTACTTACGAAGCTGAGGGACTGCGCCCCAAGGGGTTATGCTAGTTATTGCTC
C12	RB36	ATGTCCCAGTTAGAGGAGGAAACAGCCCCAAGGGGTTATGCTAGTTATTGCTC
D1	RB37	GCTTGCGATTGATGCTTAGTATCAGCCCCAAGGGGTTATGCTAGTTATTGCTC
D2	RB38	ACCACAGGAGGACGATACAGAGAAGCCCCAAGGGGTTATGCTAGTTATTGCTC
D3	RB39	CCACAGTGTCAACTAGAGCCTCTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
D4	RB40	TAGTTTGGATGACCAAGGATAGCCGCCCCAAGGGGTTATGCTAGTTATTGCTC
D5	RB41	GGAGTTCGTCCAGAGAAGTACACGGCCCCAAGGGGTTATGCTAGTTATTGCTC
D6	RB42	CTACGTGTAAGGCATACCTGCCAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
D7	RB43	CTTTCGTTGTTGACTCGACGGTAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
D8	RB44	AGTAGAAAGGGTTCCTTCCCACTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
D9	RB45	GATCCAACAGAGATGCCTTCAGTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
D10	RB46	GCTGTGTTCCACTTCATTCTCCTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
D11	RB47	GTGCAACTTTCCCACAGGTAGTTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
D12	RB48	CATCTGGAACGTGGTACACCTGTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
E1	RB49	ACTGGTGCAGCTTTGAACATCTAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
E2	RB50	ATGGACTTTGGTAACTTCCTGCGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
E3	RB51	GTTGAATGAGCCTACTGGGTCCTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
E4	RB52	TGAGAGACAAGATTGTTCGTGGACGCCCCAAGGGGTTATGCTAGTTATTGCTC
E5	RB53	AGATTCAGACCGTCTCATGCAAAGGCCCCAAGGGGTTATGCTAGTTATTGCTC

E6	RB54	CAAGAGCTTTGACTAAGGAGCATGGCCCCAAGGGGTTATGCTAGTTATTGCTC
E7	RB55	TGGAAGATGAGACCCTGATCTACGGCCCCAAGGGGTTATGCTAGTTATTGCTC
E8	RB56	TCACTACTCAACAGGTGGCATGAAGCCCCAAGGGGTTATGCTAGTTATTGCTC
E9	RB57	GCTAGGTCAATCTCCTTCGGAAGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
E10	RB58	CAGGTTACTCCTCCGTGAGTCTGAGCCCCAAGGGGTTATGCTAGTTATTGCTC
E11	RB59	TCAATCAAGAAGGGAAAGCAAGGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
E12	RB60	CATGTTCAACCAAGGCTTCTATGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
F1	RB61	AGAGGGTACTATGTGCCTCAGCACGCCCCAAGGGGTTATGCTAGTTATTGCTC
F2	RB62	CACCCACACTTACTTCAGGACGTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
F3	RB63	TTCTGAAGTTCCTGGGTCTTGAACGCCCCAAGGGGTTATGCTAGTTATTGCTC
F4	RB64	GACAGACACCGTTCATCGACTTTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
F5	RB65	TTCTCAGTCTTCCTCCAGACAAGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
F6	RB66	CCGATCCTTGTGGCTTCTAACTTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
F7	RB67	GTTTGTCATACTCGTGTGCTCACCGCCCCAAGGGGTTATGCTAGTTATTGCTC
F8	RB68	GAATCTAAGCAAACACGAAGGTGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
F9	RB69	TACAGTCCGAGCCTCATGTGATCTGCCCCAAGGGGTTATGCTAGTTATTGCTC
F10	RB70	ACCGAGATCCTACGAATGGAGTGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
F11	RB71	CCTGGGAGCATCAGGTAGTAACAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
F12	RB72	TAGCTGACTGTCTTCCATACCGACGCCCCAAGGGGTTATGCTAGTTATTGCTC
G1	RB73	AAGAAACAGGATGACAGAACCCTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
G2	RB74	TACAAGCATCCCAACACTTCCACTGCCCCAAGGGGTTATGCTAGTTATTGCTC
G3	RB75	GACCATTGTGATGAACCCTGTTGTGCCCCAAGGGGTTATGCTAGTTATTGCTC
G4	RB76	ATGCTTGTTACATCAACCCTGGACGCCCCAAGGGGTTATGCTAGTTATTGCTC

G5	RB77	CGACCTGTTTCTCAGGGATACAACGCCCCAAGGGGTTATGCTAGTTATTGCTC
G6	RB78	AACAACCGAACCTTTGAATCAGAAGCCCCAAGGGGTTATGCTAGTTATTGCTC
G7	RB79	TCTCGGAGATAGTTCTCACTGCTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
G8	RB80	CGGATGAACATAGGATAGCGATTCGCCCCAAGGGGTTATGCTAGTTATTGCTC
G9	RB81	CCTCATCTTGTGAAGTTGTTTCGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
G10	RB82	ACGGTATGTCGAGTTCCAGGACTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
G11	RB83	TGGCTTGATCTAGGTAAGGTCGAAGCCCCAAGGGGTTATGCTAGTTATTGCTC
G12	RB84	GTAGTGGACCTAGAACCTGTGCCAGCCCCAAGGGGTTATGCTAGTTATTGCTC
H1	RB85	AACGGAGGAGTTAGTTGGATGATCGCCCCAAGGGGTTATGCTAGTTATTGCTC
H2	RB86	AGGTGATCCCAACAAGCGTAAGTAGCCCCAAGGGGTTATGCTAGTTATTGCTC
H3	RB87	TACATGCTCCTGTTGTTAGGGAGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
H4	RB88	TCTTCTACTACCGATCCGAAGCAGGCCCCAAGGGGTTATGCTAGTTATTGCTC
H5	RB89	ACAGCATCAATGTTTGGCTAGTTGGCCCCAAGGGGTTATGCTAGTTATTGCTC
H6	RB90	GATGTAGAGGGTACGGTTTGAGGCGCCCCAAGGGGTTATGCTAGTTATTGCTC
H7	RB91	GGCTCCATAGGAACTCACGCTACTGCCCCAAGGGGTTATGCTAGTTATTGCTC
H8	RB92	TTGTGAGTGGAAAGATACAGGACCGCCCCAAGGGGTTATGCTAGTTATTGCTC
H9	RB93	AGTTTCCATCACTTCAGACTTGGGGCCCCAAGGGGTTATGCTAGTTATTGCTC
H10	RB94	GATTGTCCTCAAACTGCCACCTACGCCCCAAGGGGTTATGCTAGTTATTGCTC
H11	RB95	CCTGTCTGGAAGAAGAATGGACTTGCCCCAAGGGGTTATGCTAGTTATTGCTC
H12	RB96	CTGAACGGTCATAGAGTCCACCATGCCCCAAGGGGTTATGCTAGTTATTGCTC

B.5.2 Plate Maps

This section contains all plate maps for the barcode plates (LevSeq plates) used in this study. The tables that follow show how the primers from the forward and reverse barcode plates (Tables S3 and S4) were arrayed to produce the barcode plates. Each entry in the plate maps below lists the forward primer on top and the reverse primer on the bottom. An "F" after the delimiter indicates that the barcode preceding the delimiter was from the forward barcode plate ("NB" in Table S3) and an "R" indicates that the well was from the reverse barcode plate ("RB" in Table S4). A detailed protocol for how the barcode plates were produced is given in Preparation of LevSeq Barcode Primer Mixes, above.

Table B5. Plate map for LevSeq01 used in this study.

LevSeq01	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB01-R	NB02-F RB01-R	NB03-F RB01-R	NB04-F RB01-R	NB05-F RB01-R	NB06-F RB01-R	NB07-F RB01-R	NB08-F RB01-R	NB09-F RB01-R	NB1F RB01-R	NB11-F RB01-R	NB12-F RB01-R
B	NB13-F RB01-R	NB14-F RB01-R	NB15-F RB01-R	NB16-F RB01-R	NB17-F RB01-R	NB18-F RB01-R	NB19-F RB01-R	NB2F RB01-R	NB21-F RB01-R	NB22-F RB01-R	NB23-F RB01-R	NB24-F RB01-R
C	NB25-F RB01-R	NB26-F RB01-R	NB27-F RB01-R	NB28-F RB01-R	NB29-F RB01-R	NB3F RB01-R	NB31-F RB01-R	NB32-F RB01-R	NB33-F RB01-R	NB34-F RB01-R	NB35-F RB01-R	NB36-F RB01-R
D	NB37-F RB01-R	NB38-F RB01-R	NB39-F RB01-R	NB4F RB01-R	NB41-F RB01-R	NB42-F RB01-R	NB43-F RB01-R	NB44-F RB01-R	NB45-F RB01-R	NB46-F RB01-R	NB47-F RB01-R	NB48-F RB01-R
E	NB49-F RB01-R	NB5F RB01-R	NB51-F RB01-R	NB52-F RB01-R	NB53-F RB01-R	NB54-F RB01-R	NB55-F RB01-R	NB56-F RB01-R	NB57-F RB01-R	NB58-F RB01-R	NB59-F RB01-R	NB6F RB01-R
F	NB61-F RB01-R	NB62-F RB01-R	NB63-F RB01-R	NB64-F RB01-R	NB65-F RB01-R	NB66-F RB01-R	NB67-F RB01-R	NB68-F RB01-R	NB69-F RB01-R	NB7F RB01-R	NB71-F RB01-R	NB72-F RB01-R
G	NB73-F RB01-R	NB74-F RB01-R	NB75-F RB01-R	NB76-F RB01-R	NB77-F RB01-R	NB78-F RB01-R	NB79-F RB01-R	NB8F RB01-R	NB81-F RB01-R	NB82-F RB01-R	NB83-F RB01-R	NB84-F RB01-R
H	NB85-F RB01-R	NB86-F RB01-R	NB87-F RB01-R	NB88-F RB01-R	NB89-F RB01-R	NB9F RB01-R	NB91-F RB01-R	NB92-F RB01-R	NB93-F RB01-R	NB94-F RB01-R	NB95-F RB01-R	NB96-F RB01-R

Table B6. Plate map for LevSeq02 used in this study.

LevSeq02	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB02-R	NB02-F RB02-R	NB03-F RB02-R	NB04-F RB02-R	NB05-F RB02-R	NB06-F RB02-R	NB07-F RB02-R	NB08-F RB02-R	NB09-F RB02-R	NB1F RB02-R	NB11-F RB02-R	NB12-F RB02-R
B	NB13-F RB02-R	NB14-F RB02-R	NB15-F RB02-R	NB16-F RB02-R	NB17-F RB02-R	NB18-F RB02-R	NB19-F RB02-R	NB2F RB02-R	NB21-F RB02-R	NB22-F RB02-R	NB23-F RB02-R	NB24-F RB02-R
C	NB25-F RB02-R	NB26-F RB02-R	NB27-F RB02-R	NB28-F RB02-R	NB29-F RB02-R	NB3F RB02-R	NB31-F RB02-R	NB32-F RB02-R	NB33-F RB02-R	NB34-F RB02-R	NB35-F RB02-R	NB36-F RB02-R
D	NB37-F RB02-R	NB38-F RB02-R	NB39-F RB02-R	NB4F RB02-R	NB41-F RB02-R	NB42-F RB02-R	NB43-F RB02-R	NB44-F RB02-R	NB45-F RB02-R	NB46-F RB02-R	NB47-F RB02-R	NB48-F RB02-R
E	NB49-F RB02-R	NB5F RB02-R	NB51-F RB02-R	NB52-F RB02-R	NB53-F RB02-R	NB54-F RB02-R	NB55-F RB02-R	NB56-F RB02-R	NB57-F RB02-R	NB58-F RB02-R	NB59-F RB02-R	NB6F RB02-R
F	NB61-F RB02-R	NB62-F RB02-R	NB63-F RB02-R	NB64-F RB02-R	NB65-F RB02-R	NB66-F RB02-R	NB67-F RB02-R	NB68-F RB02-R	NB69-F RB02-R	NB7F RB02-R	NB71-F RB02-R	NB72-F RB02-R
G	NB73-F RB02-R	NB74-F RB02-R	NB75-F RB02-R	NB76-F RB02-R	NB77-F RB02-R	NB78-F RB02-R	NB79-F RB02-R	NB8F RB02-R	NB81-F RB02-R	NB82-F RB02-R	NB83-F RB02-R	NB84-F RB02-R
H	NB85-F RB02-R	NB86-F RB02-R	NB87-F RB02-R	NB88-F RB02-R	NB89-F RB02-R	NB9F RB02-R	NB91-F RB02-R	NB92-F RB02-R	NB93-F RB02-R	NB94-F RB02-R	NB95-F RB02-R	NB96-F RB02-R

Table B7. Plate map for LevSeq03 used in this study.

LevSeq03	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB03-R	NB02-F RB03-R	NB03-F RB03-R	NB04-F RB03-R	NB05-F RB03-R	NB06-F RB03-R	NB07-F RB03-R	NB08-F RB03-R	NB09-F RB03-R	NB1F RB03-R	NB11-F RB03-R	NB12-F RB03-R
B	NB13-F RB03-R	NB14-F RB03-R	NB15-F RB03-R	NB16-F RB03-R	NB17-F RB03-R	NB18-F RB03-R	NB19-F RB03-R	NB2F RB03-R	NB21-F RB03-R	NB22-F RB03-R	NB23-F RB03-R	NB24-F RB03-R
C	NB25-F RB03-R	NB26-F RB03-R	NB27-F RB03-R	NB28-F RB03-R	NB29-F RB03-R	NB3F RB03-R	NB31-F RB03-R	NB32-F RB03-R	NB33-F RB03-R	NB34-F RB03-R	NB35-F RB03-R	NB36-F RB03-R
D	NB37-F RB03-R	NB38-F RB03-R	NB39-F RB03-R	NB4F RB03-R	NB41-F RB03-R	NB42-F RB03-R	NB43-F RB03-R	NB44-F RB03-R	NB45-F RB03-R	NB46-F RB03-R	NB47-F RB03-R	NB48-F RB03-R
E	NB49-F RB03-R	NB5F RB03-R	NB51-F RB03-R	NB52-F RB03-R	NB53-F RB03-R	NB54-F RB03-R	NB55-F RB03-R	NB56-F RB03-R	NB57-F RB03-R	NB58-F RB03-R	NB59-F RB03-R	NB6F RB03-R
F	NB61-F RB03-R	NB62-F RB03-R	NB63-F RB03-R	NB64-F RB03-R	NB65-F RB03-R	NB66-F RB03-R	NB67-F RB03-R	NB68-F RB03-R	NB69-F RB03-R	NB7F RB03-R	NB71-F RB03-R	NB72-F RB03-R
G	NB73-F RB03-R	NB74-F RB03-R	NB75-F RB03-R	NB76-F RB03-R	NB77-F RB03-R	NB78-F RB03-R	NB79-F RB03-R	NB8F RB03-R	NB81-F RB03-R	NB82-F RB03-R	NB83-F RB03-R	NB84-F RB03-R
H	NB85-F RB03-R	NB86-F RB03-R	NB87-F RB03-R	NB88-F RB03-R	NB89-F RB03-R	NB9F RB03-R	NB91-F RB03-R	NB92-F RB03-R	NB93-F RB03-R	NB94-F RB03-R	NB95-F RB03-R	NB96-F RB03-R

Table B8. Plate map for LevSeq04 used in this study.

LevSeq04	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB04-R	NB02-F RB04-R	NB03-F RB04-R	NB04-F RB04-R	NB05-F RB04-R	NB06-F RB04-R	NB07-F RB04-R	NB08-F RB04-R	NB09-F RB04-R	NB1F RB04-R	NB11-F RB04-R	NB12-F RB04-R
B	NB13-F RB04-R	NB14-F RB04-R	NB15-F RB04-R	NB16-F RB04-R	NB17-F RB04-R	NB18-F RB04-R	NB19-F RB04-R	NB2F RB04-R	NB21-F RB04-R	NB22-F RB04-R	NB23-F RB04-R	NB24-F RB04-R
C	NB25-F RB04-R	NB26-F RB04-R	NB27-F RB04-R	NB28-F RB04-R	NB29-F RB04-R	NB3F RB04-R	NB31-F RB04-R	NB32-F RB04-R	NB33-F RB04-R	NB34-F RB04-R	NB35-F RB04-R	NB36-F RB04-R
D	NB37-F RB04-R	NB38-F RB04-R	NB39-F RB04-R	NB4F RB04-R	NB41-F RB04-R	NB42-F RB04-R	NB43-F RB04-R	NB44-F RB04-R	NB45-F RB04-R	NB46-F RB04-R	NB47-F RB04-R	NB48-F RB04-R
E	NB49-F RB04-R	NB5F RB04-R	NB51-F RB04-R	NB52-F RB04-R	NB53-F RB04-R	NB54-F RB04-R	NB55-F RB04-R	NB56-F RB04-R	NB57-F RB04-R	NB58-F RB04-R	NB59-F RB04-R	NB6F RB04-R
F	NB61-F RB04-R	NB62-F RB04-R	NB63-F RB04-R	NB64-F RB04-R	NB65-F RB04-R	NB66-F RB04-R	NB67-F RB04-R	NB68-F RB04-R	NB69-F RB04-R	NB7F RB04-R	NB71-F RB04-R	NB72-F RB04-R
G	NB73-F RB04-R	NB74-F RB04-R	NB75-F RB04-R	NB76-F RB04-R	NB77-F RB04-R	NB78-F RB04-R	NB79-F RB04-R	NB8F RB04-R	NB81-F RB04-R	NB82-F RB04-R	NB83-F RB04-R	NB84-F RB04-R
H	NB85-F RB04-R	NB86-F RB04-R	NB87-F RB04-R	NB88-F RB04-R	NB89-F RB04-R	NB9F RB04-R	NB91-F RB04-R	NB92-F RB04-R	NB93-F RB04-R	NB94-F RB04-R	NB95-F RB04-R	NB96-F RB04-R

Table B9. Plate map for LevSeq05 used in this study.

LevSeq05	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB05-R	NB02-F RB05-R	NB03-F RB05-R	NB04-F RB05-R	NB05-F RB05-R	NB06-F RB05-R	NB07-F RB05-R	NB08-F RB05-R	NB09-F RB05-R	NB1F RB05-R	NB11-F RB05-R	NB12-F RB05-R
B	NB13-F RB05-R	NB14-F RB05-R	NB15-F RB05-R	NB16-F RB05-R	NB17-F RB05-R	NB18-F RB05-R	NB19-F RB05-R	NB2F RB05-R	NB21-F RB05-R	NB22-F RB05-R	NB23-F RB05-R	NB24-F RB05-R
C	NB25-F RB05-R	NB26-F RB05-R	NB27-F RB05-R	NB28-F RB05-R	NB29-F RB05-R	NB3F RB05-R	NB31-F RB05-R	NB32-F RB05-R	NB33-F RB05-R	NB34-F RB05-R	NB35-F RB05-R	NB36-F RB05-R
D	NB37-F RB05-R	NB38-F RB05-R	NB39-F RB05-R	NB4F RB05-R	NB41-F RB05-R	NB42-F RB05-R	NB43-F RB05-R	NB44-F RB05-R	NB45-F RB05-R	NB46-F RB05-R	NB47-F RB05-R	NB48-F RB05-R
E	NB49-F RB05-R	NB5F RB05-R	NB51-F RB05-R	NB52-F RB05-R	NB53-F RB05-R	NB54-F RB05-R	NB55-F RB05-R	NB56-F RB05-R	NB57-F RB05-R	NB58-F RB05-R	NB59-F RB05-R	NB6F RB05-R
F	NB61-F RB05-R	NB62-F RB05-R	NB63-F RB05-R	NB64-F RB05-R	NB65-F RB05-R	NB66-F RB05-R	NB67-F RB05-R	NB68-F RB05-R	NB69-F RB05-R	NB7F RB05-R	NB71-F RB05-R	NB72-F RB05-R
G	NB73-F RB05-R	NB74-F RB05-R	NB75-F RB05-R	NB76-F RB05-R	NB77-F RB05-R	NB78-F RB05-R	NB79-F RB05-R	NB8F RB05-R	NB81-F RB05-R	NB82-F RB05-R	NB83-F RB05-R	NB84-F RB05-R
H	NB85-F RB05-R	NB86-F RB05-R	NB87-F RB05-R	NB88-F RB05-R	NB89-F RB05-R	NB9F RB05-R	NB91-F RB05-R	NB92-F RB05-R	NB93-F RB05-R	NB94-F RB05-R	NB95-F RB05-R	NB96-F RB05-R

Table B10. Plate map for LevSeq06 used in this study.

LevSeq06	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB06-R	NB02-F RB06-R	NB03-F RB06-R	NB04-F RB06-R	NB05-F RB06-R	NB06-F RB06-R	NB07-F RB06-R	NB08-F RB06-R	NB09-F RB06-R	NB1F RB06-R	NB11-F RB06-R	NB12-F RB06-R
B	NB13-F RB06-R	NB14-F RB06-R	NB15-F RB06-R	NB16-F RB06-R	NB17-F RB06-R	NB18-F RB06-R	NB19-F RB06-R	NB2F RB06-R	NB21-F RB06-R	NB22-F RB06-R	NB23-F RB06-R	NB24-F RB06-R
C	NB25-F RB06-R	NB26-F RB06-R	NB27-F RB06-R	NB28-F RB06-R	NB29-F RB06-R	NB3F RB06-R	NB31-F RB06-R	NB32-F RB06-R	NB33-F RB06-R	NB34-F RB06-R	NB35-F RB06-R	NB36-F RB06-R
D	NB37-F RB06-R	NB38-F RB06-R	NB39-F RB06-R	NB4F RB06-R	NB41-F RB06-R	NB42-F RB06-R	NB43-F RB06-R	NB44-F RB06-R	NB45-F RB06-R	NB46-F RB06-R	NB47-F RB06-R	NB48-F RB06-R
E	NB49-F RB06-R	NB5F RB06-R	NB51-F RB06-R	NB52-F RB06-R	NB53-F RB06-R	NB54-F RB06-R	NB55-F RB06-R	NB56-F RB06-R	NB57-F RB06-R	NB58-F RB06-R	NB59-F RB06-R	NB6F RB06-R
F	NB61-F RB06-R	NB62-F RB06-R	NB63-F RB06-R	NB64-F RB06-R	NB65-F RB06-R	NB66-F RB06-R	NB67-F RB06-R	NB68-F RB06-R	NB69-F RB06-R	NB7F RB06-R	NB71-F RB06-R	NB72-F RB06-R
G	NB73-F RB06-R	NB74-F RB06-R	NB75-F RB06-R	NB76-F RB06-R	NB77-F RB06-R	NB78-F RB06-R	NB79-F RB06-R	NB8F RB06-R	NB81-F RB06-R	NB82-F RB06-R	NB83-F RB06-R	NB84-F RB06-R
H	NB85-F RB06-R	NB86-F RB06-R	NB87-F RB06-R	NB88-F RB06-R	NB89-F RB06-R	NB9F RB06-R	NB91-F RB06-R	NB92-F RB06-R	NB93-F RB06-R	NB94-F RB06-R	NB95-F RB06-R	NB96-F RB06-R

Table B11. Plate map for LevSeq07 used in this study.

LevSeq07	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB07-R	NB02-F RB07-R	NB03-F RB07-R	NB04-F RB07-R	NB05-F RB07-R	NB06-F RB07-R	NB07-F RB07-R	NB08-F RB07-R	NB09-F RB07-R	NB1F RB07-R	NB11-F RB07-R	NB12-F RB07-R
B	NB13-F RB07-R	NB14-F RB07-R	NB15-F RB07-R	NB16-F RB07-R	NB17-F RB07-R	NB18-F RB07-R	NB19-F RB07-R	NB2F RB07-R	NB21-F RB07-R	NB22-F RB07-R	NB23-F RB07-R	NB24-F RB07-R
C	NB25-F RB07-R	NB26-F RB07-R	NB27-F RB07-R	NB28-F RB07-R	NB29-F RB07-R	NB3F RB07-R	NB31-F RB07-R	NB32-F RB07-R	NB33-F RB07-R	NB34-F RB07-R	NB35-F RB07-R	NB36-F RB07-R
D	NB37-F RB07-R	NB38-F RB07-R	NB39-F RB07-R	NB4F RB07-R	NB41-F RB07-R	NB42-F RB07-R	NB43-F RB07-R	NB44-F RB07-R	NB45-F RB07-R	NB46-F RB07-R	NB47-F RB07-R	NB48-F RB07-R
E	NB49-F RB07-R	NB5F RB07-R	NB51-F RB07-R	NB52-F RB07-R	NB53-F RB07-R	NB54-F RB07-R	NB55-F RB07-R	NB56-F RB07-R	NB57-F RB07-R	NB58-F RB07-R	NB59-F RB07-R	NB6F RB07-R
F	NB61-F RB07-R	NB62-F RB07-R	NB63-F RB07-R	NB64-F RB07-R	NB65-F RB07-R	NB66-F RB07-R	NB67-F RB07-R	NB68-F RB07-R	NB69-F RB07-R	NB7F RB07-R	NB71-F RB07-R	NB72-F RB07-R
G	NB73-F RB07-R	NB74-F RB07-R	NB75-F RB07-R	NB76-F RB07-R	NB77-F RB07-R	NB78-F RB07-R	NB79-F RB07-R	NB8F RB07-R	NB81-F RB07-R	NB82-F RB07-R	NB83-F RB07-R	NB84-F RB07-R
H	NB85-F RB07-R	NB86-F RB07-R	NB87-F RB07-R	NB88-F RB07-R	NB89-F RB07-R	NB9F RB07-R	NB91-F RB07-R	NB92-F RB07-R	NB93-F RB07-R	NB94-F RB07-R	NB95-F RB07-R	NB96-F RB07-R

Table B12. Plate map for LevSeq08 used in this study.

LevSeq08	01	02	03	04	05	06	07	08	09	10	11	12
A	NB01-F RB08-R	NB02-F RB08-R	NB03-F RB08-R	NB04-F RB08-R	NB05-F RB08-R	NB06-F RB08-R	NB07-F RB08-R	NB08-F RB08-R	NB09-F RB08-R	NB1F RB08-R	NB11-F RB08-R	NB12-F RB08-R
B	NB13-F RB08-R	NB14-F RB08-R	NB15-F RB08-R	NB16-F RB08-R	NB17-F RB08-R	NB18-F RB08-R	NB19-F RB08-R	NB2F RB08-R	NB21-F RB08-R	NB22-F RB08-R	NB23-F RB08-R	NB24-F RB08-R
C	NB25-F RB08-R	NB26-F RB08-R	NB27-F RB08-R	NB28-F RB08-R	NB29-F RB08-R	NB3F RB08-R	NB31-F RB08-R	NB32-F RB08-R	NB33-F RB08-R	NB34-F RB08-R	NB35-F RB08-R	NB36-F RB08-R
D	NB37-F RB08-R	NB38-F RB08-R	NB39-F RB08-R	NB4F RB08-R	NB41-F RB08-R	NB42-F RB08-R	NB43-F RB08-R	NB44-F RB08-R	NB45-F RB08-R	NB46-F RB08-R	NB47-F RB08-R	NB48-F RB08-R
E	NB49-F RB08-R	NB5F RB08-R	NB51-F RB08-R	NB52-F RB08-R	NB53-F RB08-R	NB54-F RB08-R	NB55-F RB08-R	NB56-F RB08-R	NB57-F RB08-R	NB58-F RB08-R	NB59-F RB08-R	NB6F RB08-R
F	NB61-F RB08-R	NB62-F RB08-R	NB63-F RB08-R	NB64-F RB08-R	NB65-F RB08-R	NB66-F RB08-R	NB67-F RB08-R	NB68-F RB08-R	NB69-F RB08-R	NB7F RB08-R	NB71-F RB08-R	NB72-F RB08-R
G	NB73-F RB08-R	NB74-F RB08-R	NB75-F RB08-R	NB76-F RB08-R	NB77-F RB08-R	NB78-F RB08-R	NB79-F RB08-R	NB8F RB08-R	NB81-F RB08-R	NB82-F RB08-R	NB83-F RB08-R	NB84-F RB08-R
H	NB85-F RB08-R	NB86-F RB08-R	NB87-F RB08-R	NB88-F RB08-R	NB89-F RB08-R	NB9F RB08-R	NB91-F RB08-R	NB92-F RB08-R	NB93-F RB08-R	NB94-F RB08-R	NB95-F RB08-R	NB96-F RB08-R

Table B13. Primers specific to the backbone of pET22b(+) used in ParLQ error prone library construction.

Name	Direction	Sequence
005	Forward	GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATG
006	Reverse	CAGTGCTAGGTGAAGGAATACCGCCAAGCGGAA

Table B14. The LevSeq barcode plates used for sequencing ParLQ error prone mutagenesis libraries.

Library ID	Barcode plate
*Par*LQ-ep1	LevSeq01
*Par*LQ-ep2	LevSeq02
*Par*LQ-ep3	LevSeq03
*Par*LQ-ep4	LevSeq04
*Par*LQ-ep5	LevSeq01
*Par*LQ-ep6	LevSeq02
*Par*LQ-ep7	LevSeq03
*Par*LQ-ep8	LevSeq05
*Par*LQ-ep9	LevSeq06
*Par*LQ-ep10	LevSeq07

Table B15. The LevSeq barcode sequences used for sequencing for Supplementary Figure S2. *Note, RB26 is different from the Supplementary Table S2 and S4 due to order mistake.

Plate ID	RB Barcode Sequences
LevSeq RB25	GTAAGTTGGGTATGCAACGCAATG
LevSeq RB26*	CATACAGCGACTACGCATTCTCAT

Appendix C

SUPPORTING MATERIAL FOR CHAPTER 4

C.1 Data Format

The final CSV contains one row per variant per reaction with columns in Figure S1. There are 15 columns in total.

Four columns are essential for retaining all information about an enzymatic reaction: aa_sequence, reaction_smiles, fitness_value and additional_information.

- The aa_sequence column stores the cleaned full-length amino-acid string
- The reaction_smiles provides a machine-readable representation of the model reaction
- The fitness_value reports the activity metrics
- The additional_information preserves all other assay parameters in dictionary format.

Additional columns required for upload: id, plate, well, amino_acid_substitutions. These are essential for visualization and direct ingestion by DEDB while retaining full experimental context for downstream analysis.

Extra columns are information retained during the extraction pipeline and will be stored in addition to the essential columns and the user can download the full CSV with all information.

id	barcode_plate	plate	well	smiles_string	smiles_reaction	alignment_count	alignment_probability	nucleotide_mutation	amino_acid_substitutions	nt_sequence	aa_sequence	fitness_value	fitness_type	additional_information
sL407C-1	1	Plate_1	A01	CCCCCCCCCO	O=C(O)CCCCCC	1	1				TIKEMPQPKTFC	230	ttn	{"x_axis_label":"sL407C-1",
dF393W-1	1	Plate_1	A02	CCCCCCCCCO	O=C(O)CCCCCC	1	1		F393W		TIKEMPQPKTFC	192	ttn	{"x_axis_label":"dF393W-1"
dQ403W-	1	Plate_1	A03	CCCCCCCCCO	O=C(O)CCCCCC	1	1		Q403W		TIKEMPQPKTFC	420	ttn	{"x_axis_label":"dQ403W-1"

Figure C1. Final output CSV format compatible with the Enzyme Engineering Database upload.

C.2 PubMed Search Query

The search query was performed on July 1, 2025. A set of keywords for directed evolution and a list of relevant authors were manually identified, and the specific query is shown below.

```
"""( ( "enzyme engineering"[tiab] OR "engineered enzyme"[tiab] OR "protein engineering"[tiab] OR "directed evolution"[tiab] OR "site-directed mutagenesis"[tiab] OR "enzyme design"[tiab] OR "designer enzyme"[tiab] OR "enzyme reprogramming"[tiab] OR "artificial enzym*"[tiab] OR "artificial metalloenzyme"[tiab] OR "de novo enzyme"[tiab] OR "synthetic enzyme"[tiab] OR "Hemeproteins"[mh] OR "Protein Engineering"[mh] OR "Biocatalysis"[mh] OR ( ("novel"[tiab] OR "new"[tiab]) AND ( "halogenase*"[tiab] OR "protoglobin*"[tiab] OR "cytochrome P450"[tiab] OR "P450"[tiab] OR "CYP"[tiab] OR "peroxidase*"[tiab] OR "monooxygenase*"[tiab] OR "transaminase*"[tiab] OR "aminotransferase*"[tiab] OR "hydrolase*"[tiab] OR "esterase*"[tiab] OR "lipase*"[tiab] OR "lyase*"[tiab] OR "aldolase*"[tiab] OR "oxidoreductase*"[tiab] OR "dehydrogenase*"[tiab] OR "metalloenzyme*"[tiab] OR "carbene"[tiab] OR "nitrene"[tiab] ) ) ) AND ( "non-natural"[tiab] OR "unnatural"[tiab] OR "abiotic"[tiab] OR "new-to-nature"[tiab] OR "non-native"[tiab] OR "xenobiotic"[tiab] OR "abiological"[tiab] OR "noncanonical"[tiab] ) AND ( "reaction"[tiab] OR "catalysis"[tiab] OR "chemistry"[tiab] OR "transformation"[tiab] ) ) OR ( ( "Arnold FH"[lastau] OR "Reetz MT"[lastau] OR "Fasan R"[lastau] OR "Zhao H"[lastau] OR "Baker D"[lastau] OR "Roelfes G"[lastau] OR "Ward TR"[lastau] OR "Lu Y"[lastau] OR "Hyster TA"[lastau] OR "Liu Z"[lastau] OR "Lewis JC"[lastau] OR "Coelho PS"[lastau] OR "Buller AR"[lastau] OR "Hilvert D"[lastau] OR "Kast P"[lastau] OR "Garcia-Borràs M"[lastau] ) AND ( ("engineering"[tiab] AND "halogenase"[tiab]) OR ("designer"[tiab] AND "enzyme"[tiab]) OR "enzyme engineering"[tiab] OR ("artificial"[tiab] AND "enzym*"[tiab]) OR ("artificial"[tiab] AND "metalloenzym*"[tiab]) OR "directed evolution"[tiab] OR ("non-native"[tiab] AND "reaction"[tiab] AND "enzym*"[tiab]) OR ("abiological"[tiab] AND "catalys*"[tiab] AND "enzym*"[tiab]) OR ("non-natural"[tiab] AND "reacti*"[tiab] AND "enzym*"[tiab]) OR ("noncanonical"[tiab] AND "activity"[tiab] AND "enzym*"[tiab]) OR ("novel"[tiab] AND "chemistr*"[tiab] AND "enzym*"[tiab]) OR "artificial enzym*"[tiab] OR "artificial metalloenzym*"[tiab] OR "engineered enzyme*"[tiab] OR "site-directed mutagenesis"[tiab] OR "enzyme design"[tiab] OR "designer enzyme*"[tiab] OR "enzyme reprogramming"[tiab] OR "de novo enzyme*"[tiab] OR "synthetic enzyme*"[tiab] OR "hemeprotein*"[mh] OR "protein engineering"[mh] OR "biocatalysis"[mh] OR "enzymatic platform"[tiab] OR "biocatal*"[tiab] OR "biocatalyst"[tiab] OR "biocatalytic platform"[tiab] OR "enzymatic synthesis"[tiab] OR "enzymatic reaction*"[tiab] OR "enzymatic transformation*"[tiab] OR "enzyme cataly*"[tiab] OR "enzymatic assembly"[tiab] OR ("enzymatic"[tiab] AND "insertion"[tiab]) OR "cytochrome P450*"[tiab] OR "P450*"[tiab] OR "CYP"[tiab] OR "P411*"[tiab] OR "peroxidase*"[tiab] OR "monooxygenase*"[tiab] OR "transaminase*"[tiab] OR "aminotransferase*"[tiab] OR "hydrolase*"[tiab] OR "esterase*"[tiab] OR "lipase*"[tiab] ) ) AND ( "biocatal*"[tiab] OR "catalytic activity"[tiab] OR "enzym*"[tiab] ) AND english[Language] AND ("2015/01/01"[PDAT] : "2025/12/31"[PDAT]) NOT (review[Publication Type] OR preprint[Publication Type] OR editorial[Publication Type]) """
```

C.3 Literature Retrieval and LLM Data Extraction

Using the search query above, we compiled a set of relevant publications and then asked Gemini-2.5-Flash to classify each manuscript as either relevant or irrelevant to our database, returning a confidence score based on the title and abstract. The abstract classification prompt is shown below.

Gemini-2.5-Flash prompt for abstract classification:

```
"""
You are a scientific literature reviewer. Your task is to determine if a research paper is
relevant to specific requirements.
REQUIREMENTS:
I want to extract the following information from the scientific literature:
Sequences and functions of enzymes that perform new-to-nature chemistry.
Therefore, I am searching for papers that report enzyme engineering research aimed at
enabling new-to-nature transformations.
The papers must include the amino acid sequence (or at least sequence IDs) and functional
data of enzymes (e.g., substrate, product, yield, enantioselectivity [ee], total turnover
number [TTN], etc.).
The following are considered irrelevant:
- Purely computational studies (unless the predictions are experimentally confirmed)
- Binding proteins or non-catalytic proteins — only enzymes with catalytic activity are
relevant
- Method papers without experimentally validated enzyme examples
- Review articles (as they do not contain novel information)
- Strain engineering or catalytic cascade studies unless data on individual enzymes is
provided

PAPER TITLE: {title}
ABSTRACT: {abstract}
Analyze if this paper is directly relevant to the requirements. Consider:
- Does the paper address the main concepts of the requirements?
- Are the research methods, findings, or conclusions related to the requirements?
- Would this paper be useful for someone researching the requirements?
Respond in the following JSON format:

{{
       "is_relevant": boolean,
       "confidence": 0.1.0,
       "reasoning": "Brief explanation of your decision"
}}
Be strict in your evaluation - only mark as relevant if there's a clear connection to the
requirements.
"""
```

For any paper with a confidence score above 0.9, we manually downloaded the manuscript and APPENDIX Cin PDF format. The two PDFs were then processed through the API-based automatic sequence–fitness CSV generation pipeline. (https://github.com/YuemingLong/DEBase/)

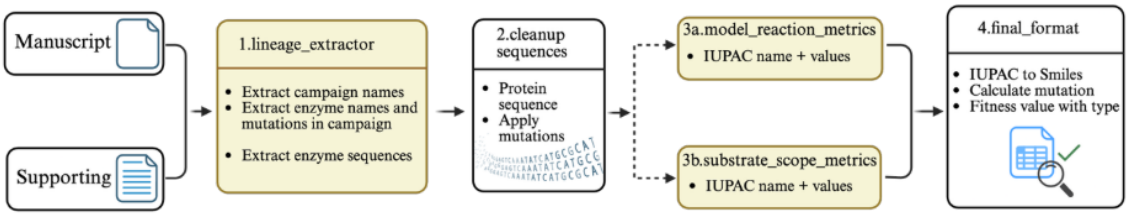


Figure C2. Pipeline for LLM based extraction.

We divided the generation of the sequence-fitness CSV into five steps, each guided by specific prompts to reduce context drift in the language model.

First, the code instructs Gemini-2.5-Flash to count the engineering campaigns reported in the paper; each campaign ID then serves as a primary key for retrieving every enzyme name and its mutations from the single table containing the complete mutation list.

Second, these enzyme names serve as a secondary key for extracting their DNA sequences. The prompt directs the model to copy each sequence exactly as printed. Because DNA strings are repetitive and error-prone, all spaces are removed and any sequence containing a stop codon before the terminus is discarded.

Third, the remaining DNA is cleaned and translated into protein. The listed mutations are applied only when the wild-type residue at the target position matches; otherwise, the sequence cell is left blank and flagged for manual review. This yields a complete sequence CSV, which is then passed to the model reaction extractor to gather substrate and product identifiers, IUPAC names, and performance metrics, prioritizing the table containing the full dataset.

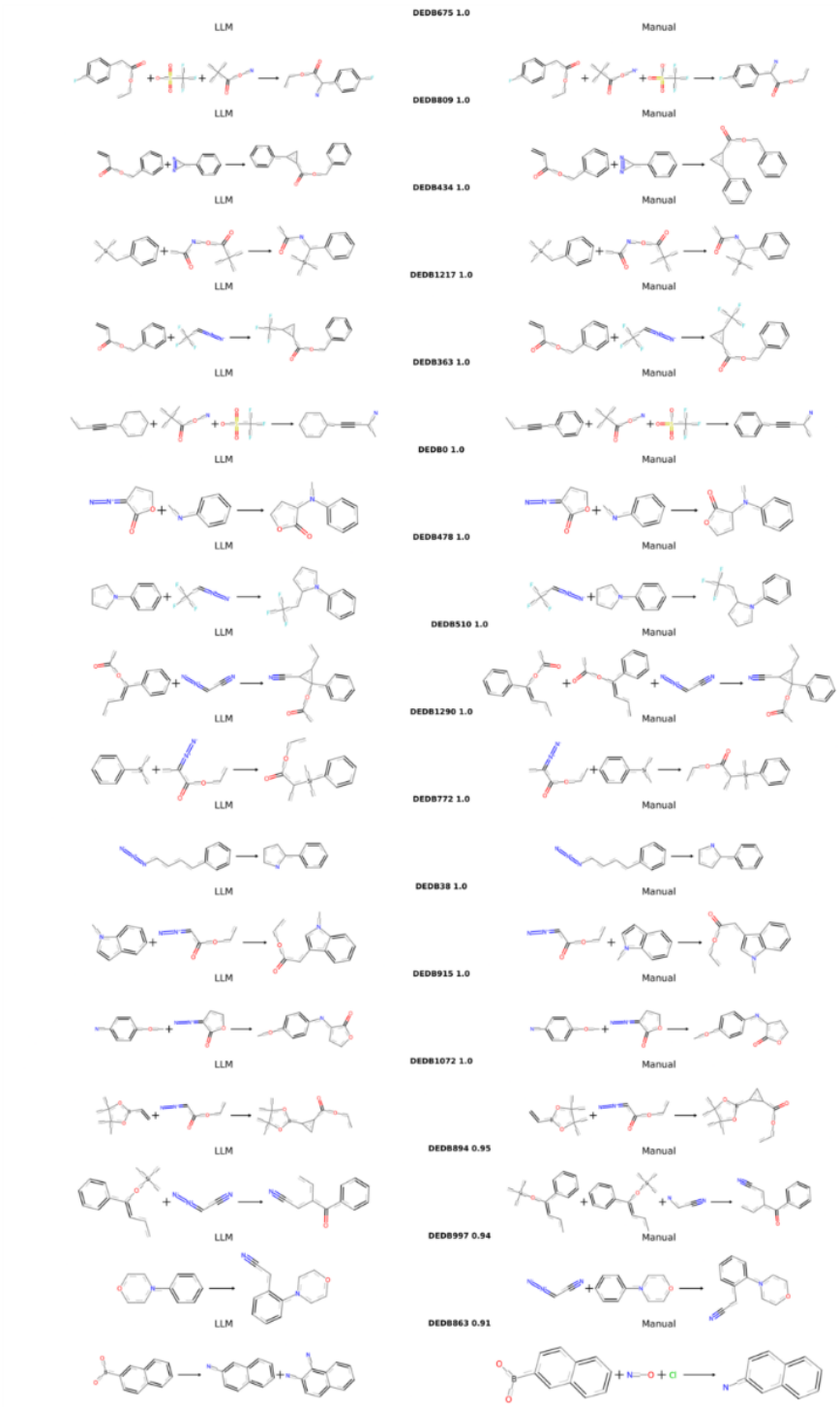
Fourth, the substrate scope extractor runs with similar logic, but its prompt excludes the model reaction and asks Gemini to state why each entry lies outside that scope.

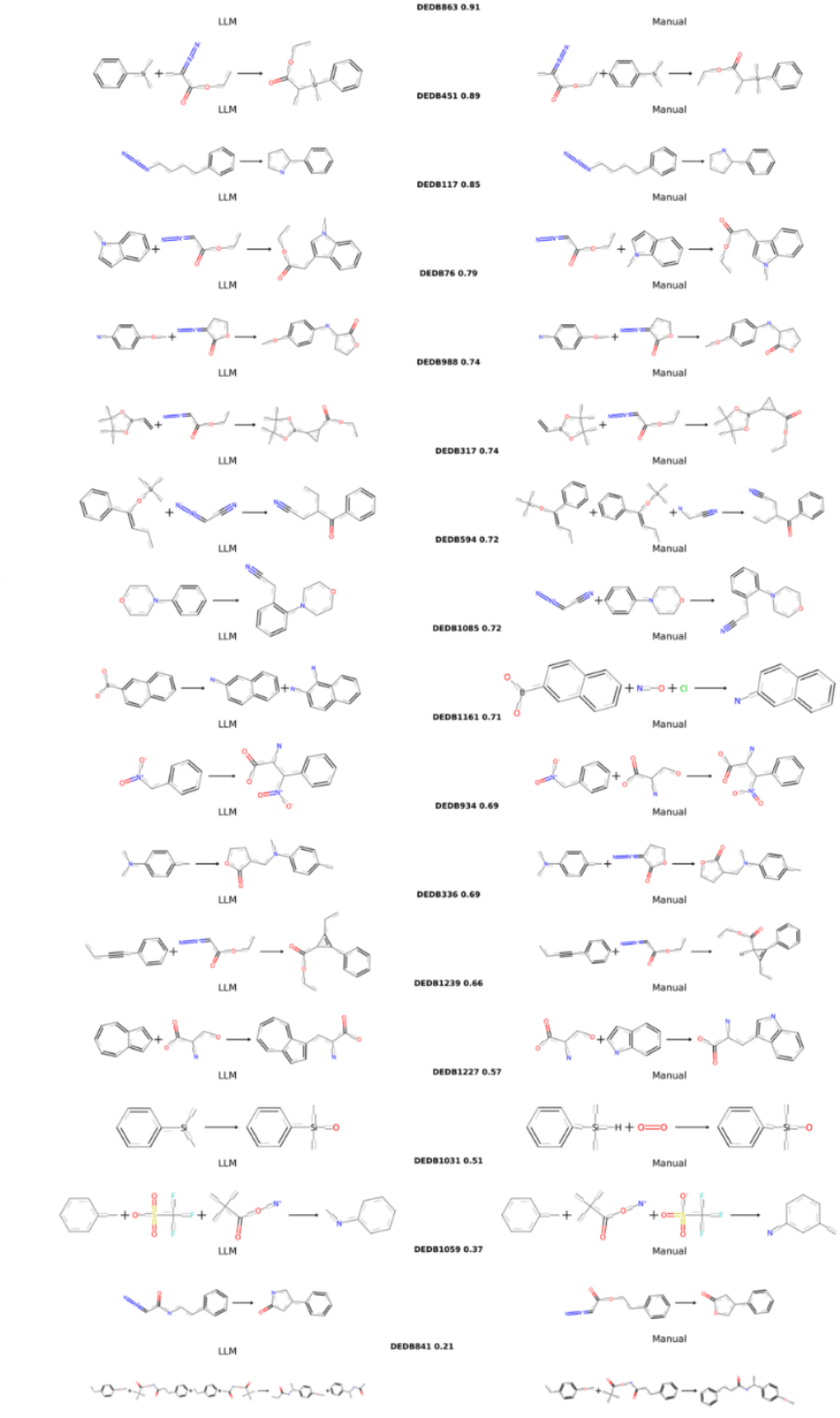
Finally, a cleanup script converts all IUPAC names to canonical SMILES via NCI, OPSIN, and PubMed, recalculates amino acid changes against the first variant of each campaign, reverse-translates DNA substitutions, and appends both the fitness value and its type to the finished output.

C.4 Additional Details About the Dashboard

Rows in the table include a fold-change column, color-coded red for gain-of-function and blue for loss-of-function, and bidirectionally linked to the structural model. Selecting a variant highlights the mutated residues in the model and automatically zooms to the affected region. For campaigns with plate-scale data, the dashboard displays a 96-well grid heat map, where each well is shaded by activity and annotated with its corresponding mutation string. A retention-of-function curve is plotted alongside the heat map, enabling rapid assessment of library quality and mutation effects. Users can sort or filter the variant table, choose which activity metric populates the heat map, switch between plates when multiple plates are available, and hover over any cell to reveal variant details. All actions dynamically update both the colored heat map and the 3-D viewer, highlighting the corresponding residues and zooming to their local environment.

C.5 Reactions from LLM Extraction and the Manually Curated Gold-Standard Dataset




LLM
DEDB863 0.91
Manual
LLM
DEDB451 0.89
Manual
LLM
DEDB117 0.85
Manual
LLM
DED876 0.79
Manual
LLM
DEDB988 0.74
Manual
LLM
DEDB317 0.74
Manual
LLM
DEDB594 0.72
Manual
LLM
DEDB1085 0.72
Manual
LLM
DEDB1161 0.71
Manual
LLM
DEDB934 0.69
Manual
LLM
DEDB336 0.69
Manual
LLM
DEDB1239 0.66
Manual
LLM
DEDB1227 0.57
Manual
LLM
DEDB1031 0.51
Manual
LLM
DEDB1059 0.37
Manual
LLM
DEDB841 0.21
Manual

C.6 Evaluation of the LLM Extraction Pipeline Against the Gold-Standard Dataset

To evaluate the LLM extraction pipeline, we computationally compared the sequences and reactions extracted by the LLM with those from the gold standard dataset. First, we assessed how many extracted papers contained both the original parent sequence and the corresponding reaction. For this, we used the first sequence and reaction from the manual extraction as the reference features. The corresponding LLM-extracted data was searched for reactions matching the manual reaction, selecting only the most similar row (i.e., the sequence corresponding to the most similar reaction). These rows were then filtered for missing amino_acid_substitutions (i.e., parent sequences), which were subsequently checked for sequence similarity using Hamming distance and Levenshtein distance, allowing for truncation and variable start positions (as it is common for researchers to start at the first index and omit the methionine). Any disagreements in residues or truncations were recorded. The most common truncation difference was omission of the His-tag, which was filtered out during postprocessing of the manually curated dataset.

This process resulted in 10 correct sequences, five single mutations (two of which were identified as errors in the manually curated data, and three attributable to the LLM), seven sequences deemed completely incorrect (Hamming distance > 100), and one missing sequence. We next checked whether the substitutions were part of the correct lineage or randomly introduced. We found that some positions were from the correct lineage, suggesting these errors were likely due to incorrect additions; however, this was inconsistent beyond the single substitutions. Therefore, we considered a case successful only in the first 15 of 32 instances (four cases were omitted as they could not be extracted by the LLM). Of these, eight had correct reactions (similarity score > 0.8), resulting in a 9 out of 27 paper success rate (33%) for the LLM extraction, which included two cases where the parent sequence was correctly identified by the LLM and improved the manual curation.

C.7 Automated and Manual Validation Steps for LLM-Extracted Papers

Extraction Pipeline Sequence Validation

The sequence extractor checks whether a DNA sequence contains internal stop codons and rejects it if any are found. Additionally, up to five sequence queries are performed to generate consensus reads for the extracted sequences. The cleanup step also includes a sanity check, ensuring that a sequence is only populated if the identified mutations match those of the parent sequence. For example, if mutations V56Y and A90S are identified, the parent sequence must contain V56 and A90 for the child sequence to be populated. The same logic applies in reverse for child-to-parent sequence generation.

Extraction Pipeline Reaction Validation

The LLM extractor is specifically prompted to extract only the ID and map it to an existing IUPAC name in the file. If no IUPAC name is found, the extraction pipeline leaves the IUPAC section empty or marks it as "not mentioned in file."

Manual Validation

For each final CSV, the extractor first reviews the amino_acid_substitution column and verifies that each variant's substitution matches the paper's description. The extractor then manually copies and pastes the sequence from the file (if available) or retrieves it from the PDB if the notes field in enzyme_sequence_data.CSV indicates a PDB source. If the sequence matches and the mutation matches, the sequence extraction for the paper is deemed successful. This is because mutations are not propagated through the pipeline but are calculated only at the final formatting step. Therefore, if one sequence and its mutation are correct, all related sequences are also correct. In some cases, sequences are manually filled in, and the final formatting step is run manually to generate the final CSV.

For reaction validation, the extractor checks the manuscript and supplementary information (SI) for an IUPAC name section. If it exists, each substrate and product ID is manually verified to ensure correct mapping. If no IUPAC section exists, the extractor verifies that the name

extracted in 3b_substrate_scope.CSV appears anywhere in the text; if not, the entire entry is removed. Since IUPAC names are programmatically converted to SMILES using NCI, PubChem, and OPSIN, accurate IUPAC extraction ensures correct SMILES strings. If all IUPAC names map correctly to their corresponding IDs, the reaction extraction is deemed successful. In rare cases, IUPAC names are manually added, and the final formatting step is run manually to produce the final CSV.

BIBLIOGRAPHY

(1) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53* (D1), D609–D617. https://doi.org/10.1093/nar/gkae1010.

(2) *AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences | Nucleic Acids Research | Oxford Academic.* https://academic.oup.com/nar/article/52/D1/D368/7337620?utm_source=chatgpt.com&login=false (accessed 2026-02-02).

(3) Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N.; Davis, L.; Dogan, T.; Atalay, V.; Rifaioglu, A. S.; Dalkıran, A.; Cetin Atalay, R.; Zhang, C.; Hurto, R. L.; Freddolino, P. L.; Zhang, Y.; Bhat, P.; Supek, F.; Fernández, J. M.; Gemovic, B.; Perovic, V. R.; Davidović, R. S.; Sumonja, N.; Veljkovic, N.; Asgari, E.; Mofrad, M. R. K.; Profiti, G.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Boecker, F.; Schoof, H.; Kahanda, I.; Thurlby, N.; McHardy, A. C.; Renaux, A.; Saidi, R.; Gough, J.; Freitas, A. A.; Antczak, M.; Fabris, F.; Wass, M. N.; Hou, J.; Cheng, J.; Wang, Z.; Romero, A. E.; Paccanaro, A.; Yang, H.; Goldberg, T.; Zhao, C.; Holm, L.; Törönen, P.; Medlar, A. J.; Zosa, E.; Borukhov, I.; Novikov, I.; Wilkins, A.; Lichtarge, O.; Chi, P.-H.; Tseng, W.-C.; Linial, M.; Rose, P. W.; Dessimoz, C.; Vidulin, V.; Dzeroski, S.; Sillitoe, I.; Das, S.; Lees, J. G.; Jones, D. T.; Wan, C.; Cozzetto, D.; Fa, R.; Torres, M.; Warwick Vesztrocy, A.; Rodriguez, J. M.; Tress, M. L.; Frasca, M.; Notaro, M.; Grossi, G.; Petrini, A.; Re, M.; Valentini, G.; Mesiti, M.; Roche, D. B.; Reeb, J.; Ritchie, D. W.; Aridhi, S.; Alborzi, S. Z.; Devignes, M.-D.; Koo, D. C. E.; Bonneau, R.; Gligorijević, V.; Barot, M.; Fang, H.; Toppo, S.; Lavezzo, E.; Falda, M.; Berselli, M.; Tosatto, S. C. E.; Carraro, M.; Piovesan, D.; Ur Rehman, H.; Mao, Q.; Zhang, S.; Vucetic, S.; Black, G. S.; Jo, D.; Suh, E.; Dayton, J. B.; Larsen, D. J.; Omdahl, A. R.; McGuffin, L. J.; Brackenridge, D. A.; Babbitt, P. C.; Yunes, J. M.; Fontana, P.; Zhang, F.; Zhu, S.; You, R.; Zhang, Z.; Dai, S.; Yao, S.; Tian, W.; Cao, R.; Chandler, C.; Amezola, M.; Johnson, D.; Chang, J.-M.; Liao, W.-H.; Liu, Y.-W.; Pascarelli, S.; Frank, Y.; Hoehndorf, R.; Kulmanov, M.; Boudellioua, I.; Politano, G.; Di Carlo, S.; Benso, A.; Hakala, K.; Ginter, F.; Mehryary, F.; Kaewphan, S.; Björne, J.; Moen, H.; Tolvanen, M. E. E.; Salakoski, T.; Kihara, D.; Jain, A.; Šmuc, T.; Altenhoff, A.; Ben-Hur, A.; Rost, B.; Brenner, S. E.; Orengo, C. A.; Jeffery, C. J.; Bosco, G.; Hogan, D. A.; Martin, M. J.; O'Donovan, C.; Mooney, S. D.; Greene, C. S.; Radivojac, P.; Friedberg, I. The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens. *Genome Biol.* **2019**, *20*, 244. https://doi.org/10.1186/s13059-019-1835-8.

(4) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.;

Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.
(5) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16* (8), 687–694. https://doi.org/10.1038/s41592-019-0496-6.
(6) Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D. ProtaBank: A Repository for Protein Design and Engineering Data. *Protein Sci. Publ. Protein Soc.* **2018**, *27* (6), 1113–1124. https://doi.org/10.1002/pro.3406.
(7) Hauenstein, J.; Jeske, L.; Jäde, A.; Krull, M.; Dümmer, K.; Koblitz, J.; Tietz, A.; Jahn, D.; Reimer, L. C.; Bunk, B. BRENDA in 2026: A Global Core Biodata Resource for Functional Enzyme and Metabolic Data within the DSMZ Digital Diversity. *Nucleic Acids Res.* **2026**, *54* (D1), D527–D534. https://doi.org/10.1093/nar/gkaf1113.
(8) Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. BRENDA in 2019: A European ELIXIR Core Data Resource. *Nucleic Acids Res.* **2019**, *47* (D1), D542–D549. https://doi.org/10.1093/nar/gky1048.
(9) Kraut, D. A.; Carroll, K. S.; Herschlag, D. Challenges in Enzyme Mechanism and Energetics. *Annu. Rev. Biochem.* **2003**, *72* (1), 517–571. https://doi.org/10.1146/annurev.biochem.72.121801.161617.
(10) Hanoian, P.; Liu, C. T.; Hammes-Schiffer, S.; Benkovic, S. Perspectives on Electrostatics and Conformational Motions in Enzyme Catalysis. *Acc. Chem. Res.* **2015**, *48* (2), 482–489. https://doi.org/10.1021/ar500390e.
(11) Schwartz, S. D. Protein Dynamics and Enzymatic Catalysis. *J. Phys. Chem. B* **2023**, *127* (12), 2649–2660. https://doi.org/10.1021/acs.jpcb.3c00477.
(12) Yabukarski, F.; Doukov, T.; Pinney, M. M.; Biel, J. T.; Fraser, J. S.; Herschlag, D. Ensemble-Function Relationships to Dissect Mechanisms of Enzyme Catalysis. *Sci. Adv.* *8* (41), eabn7738. https://doi.org/10.1126/sciadv.abn7738.
(13) Medina, E.; R. Latham, D.; Sanabria, H. Unraveling Protein's Structural Dynamics: From Configurational Dynamics to Ensemble Switching Guides Functional Mesoscale Assemblies. *Curr. Opin. Struct. Biol.* **2021**, *66*, 129–138. https://doi.org/10.1016/j.sbi.2020.10.016.
(14) Nussinov, R.; Liu, Y.; Zhang, W.; Jang, H. Protein Conformational Ensembles in Function: Roles and Mechanisms. *RSC Chem. Biol.* *4* (11), 850–864. https://doi.org/10.1039/d3cb00114h.
(15) Miton, C. M.; Buda, K.; Tokuriki, N. Epistasis and Intramolecular Networks in Protein Evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 160–168. https://doi.org/10.1016/j.sbi.2021.04.007.
(16) Pandya, C.; Farelli, J. D.; Dunaway-Mariano, D.; Allen, K. N. Enzyme Promiscuity: Engine of Evolutionary Innovation *. *J. Biol. Chem.* **2014**, *289* (44), 30229–30236. https://doi.org/10.1074/jbc.R114.572990.
(17) Guzmán, G. I.; Sandberg, T. E.; LaCroix, R. A.; Nyerges, Á.; Papp, H.; de Raad, M.; King, Z. A.; Hefner, Y.; Northen, T. R.; Notebaart, R. A.; Pál, C.; Palsson, B. O.; Papp, B.; Feist, A. M. Enzyme Promiscuity Shapes Adaptation to Novel Growth Substrates. *Mol. Syst. Biol.* **2019**, *15* (4), MSB188462. https://doi.org/10.15252/msb.20188462.
(18) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed Engl.* **2018**, *57* (16), 4143–4148. https://doi.org/10.1002/anie.201708408.

(19) Miller, D. C.; Athavale, S. V.; Arnold, F. H. Combining Chemistry and Protein Engineering for New-to-Nature Biocatalysis. *Nat. Synth.* **2022**, *1* (1), 18–23. https://doi.org/10.1038/s44160-021-00008-x.
(20) Terwilliger, T. C.; Liebschner, D.; Croll, T. I.; Williams, C. J.; McCoy, A. J.; Poon, B. K.; Afonine, P. V.; Oeffner, R. D.; Richardson, J. S.; Read, R. J.; Adams, P. D. AlphaFold Predictions Are Valuable Hypotheses and Accelerate but Do Not Replace Experimental Structure Determination. *Nat. Methods* **2024**, *21* (1), 110–116. https://doi.org/10.1038/s41592-023-02087-4.
(21) *Strengths and limitations of AlphaFold 2 | AlphaFold*. https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/strengths-and-limitations-of-alphafold/?utm_source=chatgpt.com (accessed 2026-02-02).
(22) Rocha, R. A.; Speight, R. E.; Scott, C. Engineering Enzyme Properties for Improved Biocatalytic Processes in Batch and Continuous Flow. *Org. Process Res. Dev.* **2022**, *26* (7), 1914–1924. https://doi.org/10.1021/acs.oprd.1c00424.
(23) Buller, R.; Lutz, S.; Kazlauskas, R. J.; Snajdrova, R.; Moore, J. C.; Bornscheuer, U. T. From Nature to Industry: Harnessing Enzymes for Biocatalysis. *Science* **2023**, *382* (6673), eadh8615. https://doi.org/10.1126/science.adh8615.
(24) Cobb, R. E.; Chao, R.; Zhao, H. Directed Evolution: Past, Present and Future. *AIChE J. Am. Inst. Chem. Eng.* **2013**, *59* (5), 1432–1440. https://doi.org/10.1002/aic.13995.
(25) Johnston, K. E.; Fannjiang, C.; Wittmann, B. J.; Hie, B. L.; Yang, K. K.; Wu, Z. Machine Learning for Protein Engineering. arXiv May 26, 2023. https://doi.org/10.48550/arXiv.2305.16634.
(26) Lannelongue, L.; Inouye, M. Pitfalls of Machine Learning Models for Protein–Protein Interaction Networks. *Bioinformatics* **2024**, *40* (2), btae012. https://doi.org/10.1093/bioinformatics/btae012.
(27) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (12), 866–876. https://doi.org/10.1038/nrm2805.
(28) Starr, T. N.; Thornton, J. W. Epistasis in Protein Evolution. *Protein Sci.* **2016**, *25* (7), 1204–1218. https://doi.org/10.1002/pro.2897.
(29) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16* (8), 687–694. https://doi.org/10.1038/s41592-019-0496-6.
(30) Caraus, I.; Alsuwailem, A. A.; Nadon, R.; Makarenkov, V. Detecting and Overcoming Systematic Bias in High-Throughput Screening Technologies: A Comprehensive Review of Practical Issues and Methodological Solutions. *Brief. Bioinform.* **2015**, *16* (6), 974–986. https://doi.org/10.1093/bib/bbv004.
(31) The Gene Ontology Consortium. The Gene Ontology Knowledgebase in 2026. *Nucleic Acids Res.* **2026**, *54* (D1), D1779–D1792. https://doi.org/10.1093/nar/gkaf1292.
(32) *Enzyme Nomenclature*. https://iubmb.qmul.ac.uk/enzyme/?utm_source=chatgpt.com (accessed 2026-02-02).
(33) Caraus, I.; Alsuwailem, A. A.; Nadon, R.; Makarenkov, V. Detecting and Overcoming Systematic Bias in High-Throughput Screening Technologies: A Comprehensive Review of Practical Issues and Methodological Solutions. *Brief. Bioinform.* **2015**, *16* (6), 974–986. https://doi.org/10.1093/bib/bbv004.

(34) Reetz, M. T. Directed Evolution of Enantioselective Enzymes: An Unconventional Approach to Asymmetric Catalysis in Organic Chemistry. *J. Org. Chem.* **2009**, *74* (16), 5767–5778. https://doi.org/10.1021/jo901046k.
(35) Fowler, D. M.; Fields, S. Deep Mutational Scanning: A New Style of Protein Science. *Nat. Methods* **2014**, *11* (8), 801–807. https://doi.org/10.1038/nmeth.3027.
(36) Costello, M.; Fleharty, M.; Abreu, J.; Farjoun, Y.; Ferriera, S.; Holmes, L.; Granger, B.; Green, L.; Howd, T.; Mason, T.; Vicente, G.; Dasilva, M.; Brodeur, W.; DeSmet, T.; Dodge, S.; Lennon, N. J.; Gabriel, S. Characterization and Remediation of Sample Index Swaps by Non-Redundant Dual Indexing on Massively Parallel Sequencing Platforms. *BMC Genomics* **2018**, *19*, 332. https://doi.org/10.1186/s12864-018-4703-0.
(37) Costello, M.; Fleharty, M.; Abreu, J.; Farjoun, Y.; Ferriera, S.; Holmes, L.; Granger, B.; Green, L.; Howd, T.; Mason, T.; Vicente, G.; Dasilva, M.; Brodeur, W.; DeSmet, T.; Dodge, S.; Lennon, N. J.; Gabriel, S. Characterization and Remediation of Sample Index Swaps by Non-Redundant Dual Indexing on Massively Parallel Sequencing Platforms. *BMC Genomics* **2018**, *19*, 332. https://doi.org/10.1186/s12864-018-4703-0.
(38) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.
(39) Range, J.; Halupczok, C.; Lohmann, J.; Swainston, N.; Kettner, C.; Bergmann, F. T.; Weidemann, A.; Wittig, U.; Schnell, S.; Pleiss, J. EnzymeML-a Data Exchange Format for Biocatalysis and Enzymology. *FEBS J.* **2022**, *289* (19), 5864–5874. https://doi.org/10.1111/febs.16318.
(40) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.
(41) Esposito, D.; Weile, J.; Shendure, J.; Starita, L. M.; Papenfuss, A. T.; Roth, F. P.; Fowler, D. M.; Rubin, A. F. MaveDB: An Open-Source Platform to Distribute and Interpret Data from Multiplexed Assays of Variant Effect. *Genome Biol.* **2019**, *20*, 223. https://doi.org/10.1186/s13059-019-1845-6.
(42) Greenman, K. P.; Amini, A. P.; Yang, K. K. Benchmarking Uncertainty Quantification for Protein Engineering. *PLOS Comput. Biol.* **2025**, *21* (1), e1012639. https://doi.org/10.1371/journal.pcbi.1012639.
(43) Hiesinger, K.; Dar'in, D.; Proschak, E.; Krasavin, M. Spirocyclic Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2021**, *64* (1), 150–183. https://doi.org/10.1021/acs.jmedchem.0c01473.
(44) Burkhard, J. A.; Wagner, B.; Fischer, H.; Schuler, F.; Müller, K.; Carreira, E. M. Synthesis of Azaspirocycles and Their Evaluation in Drug Discovery. *Angew. Chem. Int. Ed.* **2010**, *49* (20), 3524–3527. https://doi.org/10.1002/anie.200907108.

(45) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52* (21), 6752–6756. https://doi.org/10.1021/jm901241e.
(46) *Four-Membered Ring-Containing Spirocycles: Synthetic Strategies and Opportunities | Chemical Reviews*. https://pubs.acs.org/doi/10.1021/cr500127b (accessed 2026-02-02).
(47) Zak, M.; Yuen, P.; Liu, X.; Patel, S.; Sampath, D.; Oeh, J.; Liederer, B. M.; Wang, W.; O'Brien, T.; Xiao, Y.; Skelton, N.; Hua, R.; Sodhi, J.; Wang, Y.; Zhang, L.; Zhao, G.; Zheng, X.; Ho, Y.-C.; Bair, K. W.; Dragovich, P. S. Minimizing CYP2C9 Inhibition of Exposed-Pyridine NAMPT (Nicotinamide Phosphoribosyltransferase) Inhibitors. *J. Med. Chem.* **2016**, *59* (18), 8345–8368. https://doi.org/10.1021/acs.jmedchem.6b00697.
(48) Brown, D. G.; Bernstein, P. R.; Griffin, A.; Wesolowski, S.; Labrecque, D.; Tremblay, M. C.; Sylvester, M.; Mauger, R.; Edwards, P. D.; Throner, S. R.; Folmer, J. J.; Cacciola, J.; Scott, C.; Lazor, L. A.; Pourashraf, M.; Santhakumar, V.; Potts, W. M.; Sydserff, S.; Giguère, P.; Lévesque, C.; Dasser, M.; Groblewski, T. Discovery of Spirofused Piperazine and Diazepane Amides as Selective Histamine-3 Antagonists with in Vivo Efficacy in a Mouse Model of Cognition. *J. Med. Chem.* **2014**, *57* (3), 733–758. https://doi.org/10.1021/jm4014828.
(49) Berry, A.; Bosanac, T.; Ginn, J. D.; Hopkins, T. D.; Schlyer, S.; Soleymanzadeh, F.; Westbrook, J.; Yu, M.; Zhang, Z. Preparation of Heterocyclic Compounds as Soluble Guanylate Cyclase Activators. WO2013025425, 2013.
(50) Liu, H.; SUN, D.; Wang, Z. Degradation of Irak4 by Conjugation of Irak4 Inhibitors with E3 Ligase Ligand and Methods of Use. WO2023237049A1, December 14, 2023. https://patents.google.com/patent/WO2023237049A1/en?oq=WO+2023%2f237049+A1 (accessed 2026-02-02).
(51) Bentley, J. M.; Brookings, D. C.; Brown, J. A.; Cain, T. P.; Chovatia, P. T.; Foley, A. M.; Gallimore, E. O.; Gleave, L. J.; Heifetz, A.; Horsley, H. T.; Hutchings, M. C.; Jackson, V. E.; Johnson, J. A.; Johnstone, C.; Kroeplien, B.; Lecomte, F. C.; Leigh, D.; Lowe, M. A.; Madden, J.; Porter, J. R.; Quincey, J. R.; Reed, L. C.; Reuberson, J. T.; Richardson, A. J.; Richardson, S. E.; Selby, M. D.; Shaw, M. A.; Zhu, Z. Imidazopyridine Derivatives as Modulators of Tnf Activity. WO2014009295A1, January 16, 2014.
(52) Crawford, J. J.; Lee, W.; Young, W. B. Heteroaryl Pyridone and Aza-Pyridone Amide Compounds. WO2015000949A1, 2015.
(53) Bierer, D.; Flamme, I.; Zubov, D.; Neubauer, T.; Tersteegen, A.; JUHL, C.; GLATZ, M.; DREHER, J.; Holton, S.; TERJUNG, C.; Baumann, L.; POETHKO, T.; Xiong, J.; QIU, Y. Masp Inhibitory Compounds and Uses Thereof. WO2020225095A1, November 12, 2020. https://patents.google.com/patent/WO2020225095A1/en (accessed 2026-02-02).
(54) Ching, K.-C.; Tran, T. N. Q.; Amrun, S. N.; Kam, Y.-W.; Ng, L. F. P.; Chai, C. L. L. Structural Optimizations of Thieno[3,2-b]Pyrrole Derivatives for the Development of Metabolically Stable Inhibitors of Chikungunya Virus. *J. Med. Chem.* **2017**, *60* (7), 3165–3186. https://doi.org/10.1021/acs.jmedchem.7b00180.
(55) Gilbert, E. J.; Zhou, G.; Wong, M. K. C.; Tong, L.; Shankar, B. B.; Huang, C.; Kelly, J.; Lavey, B. J.; McCombie, S. W.; Chen, L.; Rizvi, R.; Dong, Y.; Shu, Y.; Kozlowski, J. A.; Shih, N.-Y.; Hipkin, R. W.; Gonsiorek, W.; Malikzay, A.; Lunn, C. A.; Favreau, L.; Lundell, D. J. Non-Aromatic A-Ring Replacement in the Triaryl Bis-Sulfone CB2 Receptor Inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20* (2), 608–611. https://doi.org/10.1016/j.bmcl.2009.11.084.

(56) Grabowski, K.; Proschak, E.; Baringhaus, K.-H.; Rau, O.; Schubert-Zsilavecz, M.; Schneider, G. Bioisosteric Replacement of Molecular Scaffolds: From Natural Products to Synthetic Compounds. *Nat. Prod. Commun.* **2008**, *3* (8), 1934578X0800300821. https://doi.org/10.1177/1934578X0800300821.
(57) Bauer, M. R.; Fruscia, P. D.; Lucas, S. C. C.; Michaelides, I. N.; Nelson, J. E.; Storer, R. I.; Whitehurst, B. C. Put a Ring on It: Application of Small Aliphatic Rings in Medicinal Chemistry. *RSC Med. Chem.* **2021**, *12* (4), 448–471. https://doi.org/10.1039/D0MD00370K.
(58) Lawson, A. D. G.; MacCoss, M.; Heer, J. P. Importance of Rigidity in Designing Small Molecule Drugs To Tackle Protein–Protein Interactions (PPIs) through Stabilization of Desired Conformers. *J. Med. Chem.* **2018**, *61* (10), 4283–4289. https://doi.org/10.1021/acs.jmedchem.7b01120.
(59) Koul, S.; Bhuniya, D.; Mookhtiar, K.; Bhosale, S.; Kurhade, S.; Naik, K.; Velayutham, R.; Salunkhe, V. Spirocyclic Compounds, Compositions and Medicinal Applications Thereof. WO2014054053A1, 2014.
(60) Ritchie, T. J.; Macdonald, S. J. F. The Impact of Aromatic Ring Count on Compound Developability – Are Too Many Aromatic Rings a Liability in Drug Design? *Drug Discov. Today* **2009**, *14* (21), 1011–1020. https://doi.org/10.1016/j.drudis.2009.07.014.
(61) Reissig, H.-U.; Zimmer, R. Donor−Acceptor-Substituted Cyclopropane Derivatives and Their Application in Organic Synthesis. *Chem. Rev.* **2003**, *103* (4), 1151–1196. https://doi.org/10.1021/cr010016n.
(62) Chawner, S. J.; Cases-Thomas, M. J.; Bull, J. A. Divergent Synthesis of Cyclopropane-Containing Lead-Like Compounds, Fragments and Building Blocks through a Cobalt Catalyzed Cyclopropanation of Phenyl Vinyl Sulfide. *Eur. J. Org. Chem.* **2017**, *2017* (34), 5015–5024. https://doi.org/10.1002/ejoc.201701030.
(63) Corey, E. J.; Chaykovsky, M. Dimethyloxosulfonium Methylide ((CH3)2SOCH2) and Dimethylsulfonium Methylide ((CH3)2SCH2). Formation and Application to Organic Synthesis. *J. Am. Chem. Soc.* **1965**, *87* (6), 1353–1364. https://doi.org/10.1021/ja01084a034.
(64) Sano, S.; Ando, T.; Yokoyama, K.; Nagao, Y. New Reaction Mode of the Horner-Wadsworth-Emmons Reaction for the Preparation of α-Fluoro-α,β-Unsaturated Esters. *Synlett* **2000**, *1998*, 777–779. https://doi.org/10.1055/s-1998-1780.
(65) Ebner, C.; Carreira, E. M. Cyclopropanation Strategies in Recent Total Syntheses. *Chem. Rev.* **2017**, *117* (18), 11651–11679. https://doi.org/10.1021/acs.chemrev.6b00798.
(66) Kim, Y.-H.; Song, W.-S.; Go, H.; Cha, C.-J.; Lee, C.; Yu, M.-H.; Lau, P. C. K.; Lee, K. 2-Nitrobenzoate 2-Nitroreductase (NbaA) Switches Its Substrate Specificity from 2-Nitrobenzoic Acid to 2,4-Dinitrobenzoic Acid under Oxidizing Conditions. *J. Bacteriol.* **2013**, *195* (2), 180–192. https://doi.org/10.1128/JB.02016-12.
(67) Carreira, E. M.; Fessard, T. C. Four-Membered Ring-Containing Spirocycles: Synthetic Strategies and Opportunities. *Chem. Rev.* **2014**, *114* (16), 8257–8322. https://doi.org/10.1021/cr500127b.
(68) Bender, A. M.; Carter, T. R.; Spock, M.; Rodriguez, A. L.; Dickerson, J. W.; Rook, J. M.; Chang, S.; Qi, A.; Presley, C. C.; Engers, D. W.; Harp, J. M.; Bridges, T. M.; Niswender, C. M.; Conn, P. J.; Lindsley, C. W. Synthesis and Characterization of Chiral 6-Azaspiro[2.5]Octanes as Potent and Selective Antagonists of the M4 Muscarinic

Acetylcholine Receptor. *Bioorg. Med. Chem. Lett.* **2022**, *56*, 128479. https://doi.org/10.1016/j.bmcl.2021.128479.
(69) 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* (3), 372–377. https://doi.org/10.1021/acscentsci.7b00548.
(70) Coelho, P. S.; Wang, Z. J.; Ener, M. E.; Baril, S. A.; Kannan, A. 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* (8), 485–487. https://doi.org/10.1038/nchembio.1278.
(71) Liu, Z.; Arnold, F. H. New-to-Nature Chemistry from Old Protein Machinery: Carbene and Nitrene Transferases. *Curr. Opin. Biotechnol.* **2021**, *69*, 43–51. https://doi.org/10.1016/j.copbio.2020.12.005.
(72) Brandenberg, O. F.; Miller, D. C.; Markel, U.; Ould Chaib, A.; Arnold, F. H. Engineering Chemoselectivity in Hemoprotein-Catalyzed Indole Amidation. *ACS Catal.* **2019**, *9* (9), 8271–8275. https://doi.org/10.1021/acscatal.9b02508.
(73) Pesce, A.; Bolognesi, M.; Nardini, M. Chapter Three - Protoglobin: Structure and Ligand-Binding Properties. In *Advances in Microbial Physiology*; Poole, R. K., Ed.; Microbial Globins - Status and Opportunities; Academic Press, 2013; Vol. 63, pp 79–96. https://doi.org/10.1016/B978-0-12-407693-8.00003-0.
(74) Schaus, L.; Das, A.; Knight, A. M.; Jimenez-Osés, G.; Houk, K. N.; Garcia-Borràs, M.; Arnold, F. H.; Huang, X. Protoglobin-Catalyzed Formation of Cis-Trifluoromethyl-Substituted Cyclopropanes by Carbene Transfer. *Angew. Chem. Int. Ed.* **2023**, *62* (4), e202208936. https://doi.org/10.1002/anie.202208936.
(75) Renata, H.; Lewis, R. D.; Sweredoski, M. J.; Moradian, A.; Hess, S.; Wang, Z. J.; Arnold, F. H. Identification of Mechanism-Based Inactivation in P450-Catalyzed Cyclopropanation Facilitates Engineering of Improved Enzymes. *J. Am. Chem. Soc.* **2016**, *138* (38), 12527–12533. https://doi.org/10.1021/jacs.6b06823.
(76) Shen, H.; Tan, X.; Zhou, C.; Zhou, M.; Hu, Y.; Shi, H.; Dey, F.; Liu, Y.; Ding, X. Free Amino Compounds for the Treatment and Prophylaxis of Bacterial Infection. WO2020104436A1, 2020.
(77) Lindsley, C.; Conn, P. J.; Engers, D. W.; Bender, A. M.; Baker, L. A. Antagonists of the Muscarinic Acetylcholine Receptor M4. WO2019126559A1, 2019.
(78) Kattar, S. D.; Gulati, A.; Margrey, K. A.; Keylor, M. H.; Ardolino, M.; Yan, X.; Johnson, R.; Palte, R. L.; McMinn, S. E.; Nogle, L.; Su, J.; Xiao, D.; Piesvaux, J.; Lee, S.; Hegde, L. G.; Woodhouse, J. D.; Faltus, R.; Moy, L. Y.; Xiong, T.; Ciaccio, P. J.; Pearson, K.; Patel, M.; Otte, K. M.; Leyns, C. E. G.; Kennedy, M. E.; Bennett, D. J.; DiMauro, E. F.; Fell, M. J.; Fuller, P. H. Discovery of MK-1468: A Potent, Kinome-Selective, Brain-Penetrant Amidoisoquinoline LRRK2 Inhibitor for the Potential Treatment of Parkinson's Disease. *J. Med. Chem.* **2023**, *66* (21), 14912–14927. https://doi.org/10.1021/acs.jmedchem.3c01486.
(79) *Global Enzymes Market in Industrial Applications*. https://www.bccresearch.com/market-research/biotechnology/global-markets-for-enzymes-in-industrial-applications.html (accessed 2024-06-17).
(80) Wang, Y.; Xue, P.; Cao, M.; Yu, T.; Lane, S. T.; Zhao, H. Directed Evolution: Methodologies and Applications. *Chem. Rev.* **2021**, *121* (20), 12384–12444. https://doi.org/10.1021/acs.chemrev.1c00260.

(81) Merkx, M.; Smith, B.; Jewett, M. Engineering Sensor Proteins. *ACS Sens.* **2019**, *4* (12), 3089–3091. https://doi.org/10.1021/acssensors.9b02459.
(82) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225* (5232), 563–564. https://doi.org/10.1038/225563a0.
(83) Bosley, A. D.; Ostermeier, M. Mathematical Expressions Useful in the Construction, Description and Evaluation of Protein Libraries. *Biomol. Eng.* **2005**, *22* (1), 57–61. https://doi.org/10.1016/j.bioeng.2004.11.002.
(84) Kouba, P.; Kohout, P.; Haddadi, F.; Bushuiev, A.; Samusevich, R.; Sedlar, J.; Damborsky, J.; Pluskal, T.; Sivic, J.; Mazurenko, S. Machine Learning-Guided Protein Engineering. *ACS Catal.* **2023**, *13* (21), 13863–13895. https://doi.org/10.1021/acscatal.3c02743.
(85) Ao, Y.-F.; Dörr, M.; Menke, M. J.; Born, S.; Heuson, E.; Bornscheuer, U. T. Data-Driven Protein Engineering for Improving Catalytic Activity and Selectivity. *ChemBioChem* **2024**, *25* (3), e202300754. https://doi.org/10.1002/cbic.202300754.
(86) Wittmann, B. J.; Yue, Y.; Arnold, F. H. Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution. *Cell Syst.* **2021**, *12* (11), 1026-1045.e7. https://doi.org/10.1016/j.cels.2021.07.008.
(87) Hie, B. L.; Yang, K. K. Adaptive Machine Learning for Protein Engineering. *Curr. Opin. Struct. Biol.* **2022**, *72*, 145–152. https://doi.org/10.1016/j.sbi.2021.11.002.
(88) Mardikoraem, M.; Woldring, D. Machine Learning-Driven Protein Library Design: A Path Toward Smarter Libraries. *Methods Mol. Biol. Clifton NJ* **2022**, *2491*, 87–104. https://doi.org/10.1007/978-1-0716-2285-8_5.
(89) 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. https://doi.org/10.7554/eLife.16965.
(90) 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* (32), e2400439121. https://doi.org/10.1073/pnas.2400439121.
(91) Yang, J.; Li, F.-Z.; Arnold, F. H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **2024**, *10* (2), 226–241. https://doi.org/10.1021/acscentsci.3c01275.
(92) McCullum, E. O.; Williams, B. A. R.; Zhang, J.; Chaput, J. C. Random Mutagenesis by Error-Prone PCR. *Methods Mol. Biol. Clifton NJ* **2010**, *634*, 103–109. https://doi.org/10.1007/978-1-60761-652-8_7.
(93) Zhao, H.; Giver, L.; Shao, Z.; Affholter, J. A.; Arnold, F. H. Molecular Evolution by Staggered Extension Process (StEP) in Vitro Recombination. *Nat. Biotechnol.* **1998**, *16* (3), 258–261. https://doi.org/10.1038/nbt0398-258.
(94) Slatko, B. E.; Gardner, A. F.; Ausubel, F. M. Overview of Next Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **2018**, *122* (1), e59. https://doi.org/10.1002/cpmb.59.
(95) Shapland, E. B.; Holmes, V.; Reeves, C. D.; Sorokin, E.; Durot, M.; Platt, D.; Allen, C.; Dean, J.; Serber, Z.; Newman, J.; Chandran, S. Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process. *ACS Synth. Biol.* **2015**, *4* (7), 860–866. https://doi.org/10.1021/sb500362n.
(96) Suckling, L.; McFarlane, C.; Sawyer, C.; Chambers, S. P.; Kitney, R. I.; McClymont, D. W.; Freemont, P. S. Miniaturisation of High-Throughput Plasmid DNA Library Preparation

for next-Generation Sequencing Using Multifactorial Optimisation. *Synth. Syst. Biotechnol.* **2019**, *4* (1), 57–66. https://doi.org/10.1016/j.synbio.2019.01.002.

(97) Currin, A.; Swainston, N.; Dunstan, M. S.; Jervis, A. J.; Mulherin, P.; Robinson, C. J.; Taylor, S.; Carbonell, P.; Hollywood, K. A.; Yan, C.; Takano, E.; Scrutton, N. S.; Breitling, R. Highly Multiplexed, Fast and Accurate Nanopore Sequencing for Verification of Synthetic DNA Constructs and Sequence Libraries. *Synth. Biol.* **2019**, *4* (1), ysz025. https://doi.org/10.1093/synbio/ysz025.

(98) Ramírez Rojas, A. A.; Brinkmann, C. K.; Köbel, T. S.; Schindler, D. DuBA.flow─A Low-Cost, Long-Read Amplicon Sequencing Workflow for the Validation of Synthetic DNA Constructs. *ACS Synth. Biol.* **2024**, *13* (2), 457–465. https://doi.org/10.1021/acssynbio.3c00522.

(99) Li, W.; Miller, D.; Liu, X.; Tosi, L.; Chkaiban, L.; Mei, H.; Hung, P.-H.; Parekkadan, B.; Sherlock, G.; Levy, S. F. Arrayed in Vivo Barcoding for Multiplexed Sequence Verification of Plasmid DNA and Demultiplexing of Pooled Libraries. *Nucleic Acids Res.* **2024**, *52* (10), e47. https://doi.org/10.1093/nar/gkae332.

(100) 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* (3), 1313–1324. https://doi.org/10.1021/acssynbio.1c00592.

(101) Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore Sequencing Technology, Bioinformatics and Applications. *Nat. Biotechnol.* **2021**, *39* (11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.

(102) Sahlin, K.; Medvedev, P. Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis. *Nat. Commun.* **2021**, *12* (1), 2. https://doi.org/10.1038/s41467-020-20340-8.

(103) Delahaye, C.; Nicolas, J. Sequencing DNA with Nanopores: Troubles and Biases. *PLoS ONE* **2021**, *16* (10), e0257521. https://doi.org/10.1371/journal.pone.0257521.

(104) Houmani, M.; Peterkin, F.; Antoun, G.; Fischer, L.; Hammi, A. parSEQ: Probe and Rescue Sequencing for Advanced Variant Retrieval from DNA Pool. bioRxiv December 12, 2023, p 2023.12.12.571337. https://doi.org/10.1101/2023.12.12.571337.

(105) Zurek, P. J.; Knyphausen, P.; Neufeld, K.; Pushpanath, A.; Hollfelder, F. UMI-Linked Consensus Sequencing Enables Phylogenetic Analysis of Directed Evolution. *Nat. Commun.* **2020**, *11* (1), 6023. https://doi.org/10.1038/s41467-020-19687-9.

(106) Yang, J.; Ravi, L.; James, B.; Raul, A.; Mikhail, H.; Yisong, Y.; Frances, A. Active Learning-Assisted Directed Evolution.

(107) Smith, A. M.; Heisler, L. E.; St Onge, R. P.; Farias-Hesson, E.; Wallace, I. M.; Bodeau, J.; Harris, A. N.; Perry, K. M.; Giaever, G.; Pourmand, N.; Nislow, C. Highly-Multiplexed Barcode Sequencing: An Efficient Method for Parallel Analysis of Pooled Samples. *Nucleic Acids Res.* **2010**, *38* (13), e142. https://doi.org/10.1093/nar/gkq368.

(108) Wierbowski, S. D.; Vo, T. V.; Falter-Braun, P.; Jobe, T. O.; Kruse, L. H.; Wei, X.; Liang, J.; Meyer, M. J.; Akturk, N.; Rivera-Erick, C. A.; Cordero, N. A.; Paramo, M. I.; Shayhidin, E. E.; Bertolotti, M.; Tippens, N. D.; Akther, K.; Sharma, R.; Katayose, Y.; Salehi-Ashtiani, K.; Hao, T.; Ronald, P. C.; Ecker, J. R.; Schweitzer, P. A.; Kikuchi, S.; Mizuno, H.; Hill, D. E.; Vidal, M.; Moghe, G. D.; McCouch, S. R.; Yu, H. A Massively Parallel Barcoded Sequencing Pipeline Enables Generation of the First ORFeome and Interactome Map for Rice. *Proc. Natl. Acad. Sci.* **2020**, *117* (21), 11836–11842. https://doi.org/10.1073/pnas.1918068117.

(109) Srivathsan, A.; Lee, L.; Katoh, K.; Hartop, E.; Kutty, S. N.; Wong, J.; Yeo, D.; Meier, R. ONTbarcoder and MinION Barcodes Aid Biodiversity Discovery and Identification by Everyone, for Everyone. *BMC Biol.* **2021**, *19* (1), 217. https://doi.org/10.1186/s12915-021-01141-x.

(110) Appel, M. J.; Longwell, S. A.; Morri, M.; Neff, N.; Herschlag, D.; Fordyce, P. M. uPIC–M: Efficient and Scalable Preparation of Clonal Single Mutant Libraries for High-Throughput Protein Biochemistry. *ACS Omega* **2021**, *6* (45), 30542–30554. https://doi.org/10.1021/acsomega.1c04180.

(111) Campbell, N. R.; Harmon, S. A.; Narum, S. R. Genotyping-in-Thousands by Sequencing (GT-Seq): A Cost Effective SNP Genotyping Method Based on Custom Amplicon Sequencing. *Mol. Ecol. Resour.* **2015**, *15* (4), 855–867. https://doi.org/10.1111/1755-0998.12357.

(112) Zehra, F.; Javed, M.; Khan, D.; Pasha, M. Comparative Analysis of C++ and Python in Terms of Memory and Time. Preprints December 21, 2020. https://doi.org/10.20944/preprints202012.0516.v1.

(113) Sims, D.; Sudbery, I.; Ilott, N. E.; Heger, A.; Ponting, C. P. Sequencing Depth and Coverage: Key Considerations in Genomic Analyses. *Nat. Rev. Genet.* **2014**, *15* (2), 121–132. https://doi.org/10.1038/nrg3642.

(114) Lang, J. MAECI: A Pipeline for Generating Consensus Sequence with Nanopore Sequencing Long-Read Assembly and Error Correction. *PLOS ONE* **2022**, *17* (5), e0267066. https://doi.org/10.1371/journal.pone.0267066.

(115) Espada, R.; Zarevski, N.; Dramé-Maigné, A.; Rondelez, Y. Accurate Gene Consensus at Low Nanopore Coverage. *GigaScience* **2022**, *11*, giac102. https://doi.org/10.1093/gigascience/giac102.

(116) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

(117) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea, the Reaction Knowledgebase in 2022. *Nucleic Acids Res.* **2022**, *50* (D1), D693–D700. https://doi.org/10.1093/nar/gkab1016.

(118) Knight, A. M. Expanding the Scope of Metalloprotein Families and Substrate Classes in New-to-Nature Reactions. phd, California Institute of Technology, 2021. https://doi.org/10.7907/7qh5-5130.

(119) 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* (37), 20196–20201. https://doi.org/10.1021/jacs.3c08053.

(120) Porter, N. J.; Danelius, E.; Gonen, T.; Arnold, F. H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **2022**, *144* (20), 8892–8896. https://doi.org/10.1021/jacs.2c02723.

(121) Poelwijk, F. J.; Socolich, M.; Ranganathan, R. Learning the Pattern of Epistasis Linking Genotype and Phenotype in a Protein. *Nat. Commun.* **2019**, *10* (1), 4213. https://doi.org/10.1038/s41467-019-12130-8.

(122) Qiu, Y.; Hu, J.; Wei, G.-W. Cluster Learning-Assisted Directed Evolution. *Nat. Comput. Sci.* **2021**, *1* (12), 809–818. https://doi.org/10.1038/s43588-021-00168-y.

(123) 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* (8), 2444–2454. https://doi.org/10.1021/acssynbio.3c00301.
(124) Huffman, M. A.; Fryszkowska, A.; Alvizo, O.; Borra-Garske, M.; Campos, K. R.; Canada, K. A.; Devine, P. N.; Duan, D.; Forstater, J. H.; Grosser, S. T.; Halsey, H. M.; Hughes, G. J.; Jo, J.; Joyce, L. A.; Kolev, J. N.; Liang, J.; Maloney, K. M.; Mann, B. F.; Marshall, N. M.; McLaughlin, M.; Moore, J. C.; Murphy, G. S.; Nawrat, C. C.; Nazor, J.; Novick, S.; Patel, N. R.; Rodriguez-Granillo, A.; Robaire, S. A.; Sherer, E. C.; Truppo, M. D.; Whittaker, A. M.; Verma, D.; Xiao, L.; Xu, Y.; Yang, H. Design of an in Vitro Biocatalytic Cascade for the Manufacture of Islatravir. *Science* **2019**, *366* (6470), 1255–1259. https://doi.org/10.1126/science.aay8484.
(125) Zhao, L.-P.; Mai, B. K.; Cheng, L.; Zhao, Y.; Guo, R.; Liu, P.; Yang, Y. Biocatalytic Enantioselective C(Sp3)-H Fluorination Enabled by Directed Evolution of Non-Haem Iron Enzymes. *Nat. Synth.* **2024**, *3* (8), 967–975. https://doi.org/10.1038/s44160-024-00536-2.
(126) 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. https://doi.org/10.1016/j.cbpa.2024.102536.
(127) Wittmann, B. J.; Johnston, K. E.; Wu, Z.; Arnold, F. H. Advances in Machine Learning for Directed Evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 11–18. https://doi.org/10.1016/j.sbi.2021.01.008.
(128) 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* (1), 230–238. https://doi.org/10.1021/acssynbio.4c00625.
(129) Stafford, R. L.; Zimmerman, E. S.; Hallam, T. J.; Sato, A. K. A General Sequence Processing and Analysis Program for Protein Engineering. *J. Chem. Inf. Model.* **2014**, *54* (10), 3020–3032. https://doi.org/10.1021/ci500362s.
(130) 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* (31), 27165–27171. https://doi.org/10.1021/jacs.5c07015.
(131) Yeh, A. H.-W.; Norn, C.; Kipnis, Y.; Tischer, D.; Pellock, S. J.; Evans, D.; Ma, P.; Lee, G. R.; Zhang, J. Z.; Anishchenko, I.; Coventry, B.; Cao, L.; Dauparas, J.; Halabiya, S.; DeWitt, M.; Carter, L.; Houk, K. N.; Baker, D. De Novo Design of Luciferases Using Deep Learning. *Nature* **2023**, *614* (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.
(132) Völler, J.-S. Enzymatic Zinc Hydride. *Nat. Catal.* **2021**, *4* (3), 181–181. https://doi.org/10.1038/s41929-021-00595-0.
(133) 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* (7899), 79–85. https://doi.org/10.1038/s41586-021-04365-7.
(134) Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured Information Extraction from Scientific Text with Large Language Models. *Nat. Commun.* **2024**, *15* (1), 1418. https://doi.org/10.1038/s41467-024-45563-x.
(135) Jiang, J.; Hu, J.; Xie, S.; Guo, M.; Dong, Y.; Fu, S.; Jiang, X.; Yue, Z.; Shi, J.; Zhang, X.; Song, M.; Chen, G.; Lu, H.; Wu, X.; Guo, P.; Han, D.; Sun, Z.; Qiu, J. Enzyme Co-Scientist: Harnessing Large Language Models for Enzyme Kinetic Data Extraction from Literature.

bioRxiv March 11, 2025, p 2025.03.03.641178. https://doi.org/10.1101/2025.03.03.641178.
(136) Wei, G.; Ran, X.; AI-Abssi, R.; Yang, Z. Finding the Dark Matter: Large Language Model-Based Enzyme Kinetic Data Extractor and Its Validation. *Protein Sci.* **2025**, *34* (9), e70251. https://doi.org/10.1002/pro.70251.
(137) *RDKit*. https://www.rdkit.org/ (accessed 2026-02-03).
(138) Bommarius, A. S. Total Turnover Number – a Key Criterion for Process Evaluation. *Chem. Ing. Tech.* **2023**, *95* (4), 491–497. https://doi.org/10.1002/cite.202200177.
(139) Song, Z.; Zhang, Q.; Wu, W.; Pu, Z.; Yu, H. Rational Design of Enzyme Activity and Enantioselectivity. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1129149. https://doi.org/10.3389/fbioe.2023.1129149.
(140) 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* (18), 8852–8858. https://doi.org/10.1073/pnas.1901979116.
(141) Xie, H.; Liu, K.; Li, Z.; Wang, Z.; Wang, C.; Li, F.; Han, W.; Wang, L. Machine-Learning-Aided Engineering Hemoglobin as Carbene Transferase for Catalyzing Enantioselective Olefin Cyclopropanation. *JACS Au* **2024**, *4* (12), 4957–4967. https://doi.org/10.1021/jacsau.4c01045.
(142) Li, F.-Z.; Radtke, L. A.; Johnston, K. E.; Liu, C.-H.; Yue, Y.; Arnold, F. H. SUBSTRATE-AWARE ZERO-SHOT PREDICTORS FOR NON-NATIVE ENZYME ACTIVITIES. **2025**.
(143) Goldman, S.; Das, R.; Yang, K. K.; Coley, C. W. Machine Learning Modeling of Family Wide Enzyme-Substrate Specificity Screens. *PLOS Comput. Biol.* **2022**, *18* (2), e1009853. https://doi.org/10.1371/journal.pcbi.1009853.
(144) Kiss, G.; Röthlisberger, D.; Baker, D.; Houk, K. Evaluation and Ranking of Enzyme Designs. *Protein Sci. Publ. Protein Soc.* **2010**, *19* (9), 1760–1773. https://doi.org/10.1002/pro.462.
(145) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-Based Enzyme Design Can Recapitulate the Effects of Laboratory Directed Evolution in Silico. *Nat. Commun.* **2020**, *11* (1), 4808. https://doi.org/10.1038/s41467-020-18619-x.
(146) Johnson, S. R.; Fu, X.; Viknander, S.; Goldin, C.; Monaco, S.; Zelezniak, A.; Yang, K. K. Computational Scoring and Experimental Evaluation of Enzymes Generated by Neural Networks. *Nat. Biotechnol.* **2025**, *43* (3), 396–405. https://doi.org/10.1038/s41587-024-02214-2.
(147) Otten, R.; Pádua, R. A. P.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; Kern, D. How Directed Evolution Reshapes the Energy Landscape in an Enzyme to Boost Catalysis. *Science* **2020**, *370* (6523), 1442–1446. https://doi.org/10.1126/science.abd3623.
(148) Bolon, D. N.; Voigt, C. A.; Mayo, S. L. De Novo Design of Biocatalysts. *Curr. Opin. Chem. Biol.* **2002**, *6* (2), 125–129. https://doi.org/10.1016/S1367-5931(02)00303-4.
(149) Kipnis, Y.; Chaib, A. O.; Vorobieva, A. A.; Cai, G.; Reggiano, G.; Basanta, B.; Kumar, E.; Mittl, P. R. E.; Hilvert, D.; Baker, D. Design and Optimization of Enzymatic Activity in a de Novo B-barrel Scaffold. *Protein Sci. Publ. Protein Soc.* **2022**, *31* (11), e4405. https://doi.org/10.1002/pro.4405.

(150) Bunzel, H. A.; Anderson, J. L. R.; Mulholland, A. J. Designing Better Enzymes: Insights from Directed Evolution. *Curr. Opin. Struct. Biol.* **2021**, *67*, 212–218. https://doi.org/10.1016/j.sbi.2020.12.015.
(151) Hossack, E. J.; Hardy, F. J.; Green, A. P. Building Enzymes through Design and Evolution. *ACS Catal.* **2023**, *13* (19), 12436–12444. https://doi.org/10.1021/acscatal.3c02746.
(152) 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. *Science* **2023**, *379* (6637), 1123–1130. https://doi.org/10.1126/science.ade2574.
(153) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.
(154) Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*; ACM Press: Bonn, Germany, 2005; pp 89–96. https://doi.org/10.1145/1102351.1102363.
(155) Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; Li, H. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th international conference on Machine learning*; ACM: Corvalis Oregon USA, 2007; pp 129–136. https://doi.org/10.1145/1273496.1273513.
(156) Schober, P.; Boer, C.; Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126* (5), 1763. https://doi.org/10.1213/ANE.0000000000002864.
(157) Schober, P.; Boer, C.; Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126* (5), 1763. https://doi.org/10.1213/ANE.0000000000002864.
(158) *Dataset Shift in Machine Learning*; Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. D., Eds.; Jordan, M. I., Dietterich, T., Series Eds.; Neural Information Processing series; MIT Press: Cambridge, MA, USA, 2008.
(159) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9689–9701.
(160) 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. U. S. A.* **2021**, *118* (15), e2016239118. https://doi.org/10.1073/pnas.2016239118.
(161) Zhou, Z.; Zhang, L.; Yu, Y.; Wu, B.; Li, M.; Hong, L.; Tan, P. Enhancing Efficiency of Protein Language Models with Minimal Wet-Lab Data through Few-Shot Learning. *Nat. Commun.* **2024**, *15* (1), 5566. https://doi.org/10.1038/s41467-024-49798-6.

(162) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.
(163) Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv December 17, 2019. https://doi.org/10.48550/arXiv.1906.02530.
(164) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*; PMLR, 2016; pp 1050–1059.