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

Yueming Long[a,†] , Ariane Mora[a,†] , Emre Gürsoy[a,¶], Kadina E. Johnston[b,§], Francesca Zhoufan-Li[b], Frances H. Arnold[*,a,b]

[a] Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125

[¶]current address: Department of Biosystems Science and Engineering, ETH Zurich, Schanzenstrasse 44, 4056 Basel

[b] Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA 91125

[§] current address: Discovery Biologics, Merck & Co., Inc., South San Francisco, CA 94080

[*] Corresponding Author: Frances H. Arnold frances@cheme.caltech.edu

[†] These authors contributed equally.

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.

Introduction

Directed evolution (DE) has been key to the discovery and engineering of biocatalysts for new-to-nature chemistry[1], development of sustainable bioprocesses for pharmaceutical synthesis[2,3], and for engineering proteins for environmental sensing[4] and bioremediation[5], 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[6,7]. 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[8]. To streamline directed evolution, machine learning (ML) can be employed[9–11] to guide sequence-function exploration to variants with high fitness[12–15].

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[3]. 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[16–18]. While valuable, these approaches are costly and capture only a fraction of the sequence diversity most useful for protein evolution[19]. To advance machine learning in protein engineering, we need a method for collecting, analyzing, and pairing sequence-function data from diverse mutagenesis approaches[20,21]. 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 genes[22] and the lack of a standardized format to create and distribute the data. 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[23]. 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[24], but nanopore sequencing is characterized by a

high error rate[25–27]. Previously published high-throughput nanopore sequencing methods, parSEQ[28] and SequenceGenie[27], 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[29]. 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.[27] 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 ten 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[30]. 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.

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. (Supplementary Information 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).

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 basecalled bases per 96 variants and super accurate basecalling model are recommended as sequencing parameters. One Flongle flow cell is recommended for sequencing up to 1600 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).

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, π_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 Supporting Information, Table S13. Different concentrations (100 mM, 200 mM, 300 mM, 400 mM) of $MnCl_2$ 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

L24040-5000.0) 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 Supporting Information, LevSeq Library Preparation and Sequencing; the LevSeq software was run using all default parameters. Barcode-linked primer plates used are in Supporting information, Tables S5–S12; the barcode plates were paired to libraries as given in Supporting Information, Table S14.

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 5182-0715) and analyzed using GCMS (Agilent 7820A(G4350A)).

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[31]. The primer sequences are designed for pET-22b(+) backbones and can be redesigned for other cloning vectors following standard design techniques (Supplementary Information 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)[23]. 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 1A). 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[29–32].

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.

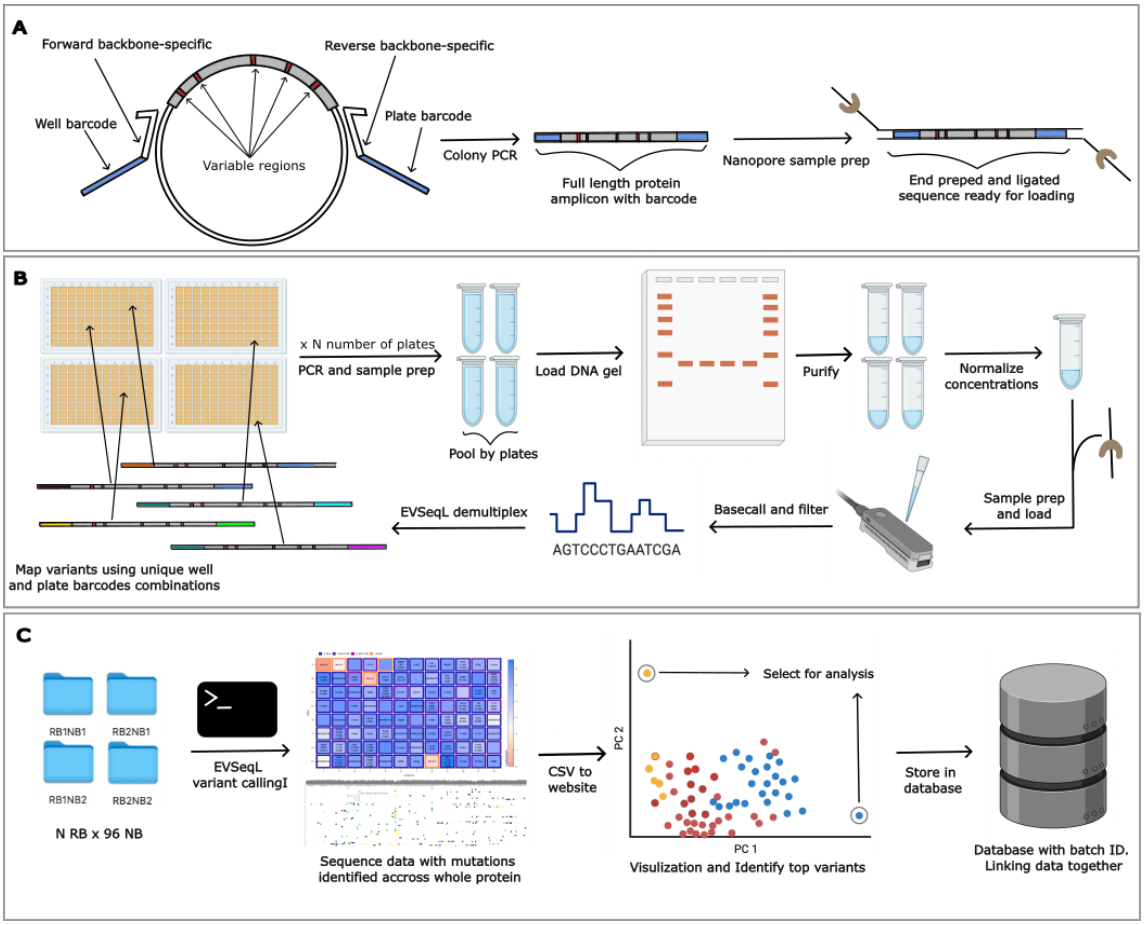

Figure 1: Overview of LevSeq library preparation, variant sequencing, and data visualization. A. The first step of LevSeq involves a one-step PCR using backbone-specific 5' and 3' end barcoded primers to amplify the full-length targeted gene. B. All PCR products from one 96-well plate are pooled for 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 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 1B). 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 1C). 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 2).

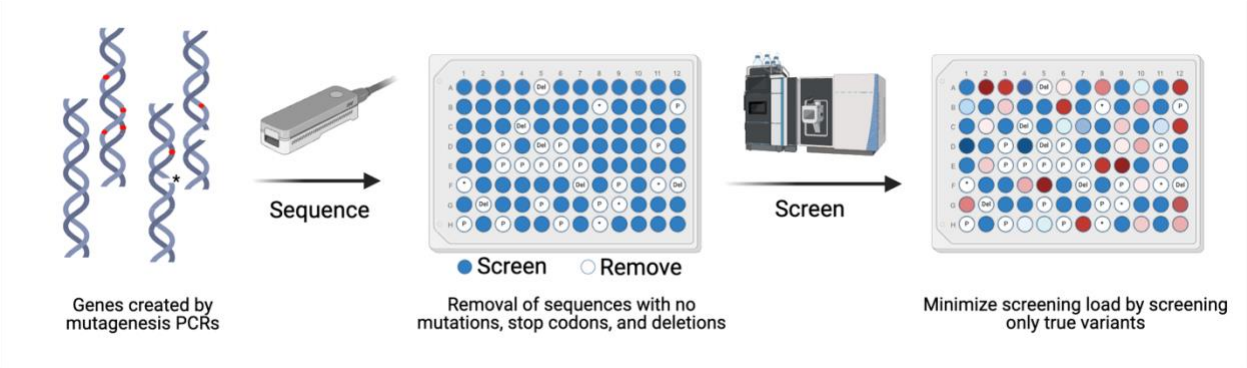

Figure 2: LevSeq reduces screening burden by enabling removal of sequences with no mutations, stop codons, and deletions.

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[36]. 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. Experimental variation, defined here as protein sequence length and ePCR error rate showed no effect on the efficacy of variant calling (Supplementary Figure 1E–F). 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%), (Supplementary Figure 1A–B). We showed that variants are accurately detected (>99%) with 10 reads[37] to generate a consensus sequence[38,39] up to a sequencing error rate of

20%, which exceeds the expected error of nanopore devices[40] (Figure 3A). Previous methods have opted for alternative approaches to overcome the high error rate: SequenceGenie leverages strand bias and Bayesian statistics to identify true variants[27], while parSEQ uses a conservative consensus threshold of 0.9[28].

The second component of the software suite is a website that processes LevSeq 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[27]. ParSEQ is an extensive software suite and utilizes comprehensive cloud computing, making it ideal for larger scale operations. However, it requires knowledge of cloud infrastructure[28], an uncommon skill set in a typical protein engineering lab. 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. We envision these datasets will be compatible with existing databases[41–44] and 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 (*Par*Pgb) LQ variants[45]. *Par*Pgb 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[46–49].

Three of the ten plates exhibited unusually low sequencing coverage, contributing to a large number of samples with insufficient coverage (Supplementary Figure 2). (The insufficient coverage in this case resulted from improper PCR amplification. This can be mitigated by adjusting the PCR setup method (Supplementary Protocols) and was not observed in subsequent sequencing experiments (Supplementary Figure 3). The final dataset for *Par*Pgb 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 3B and 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.

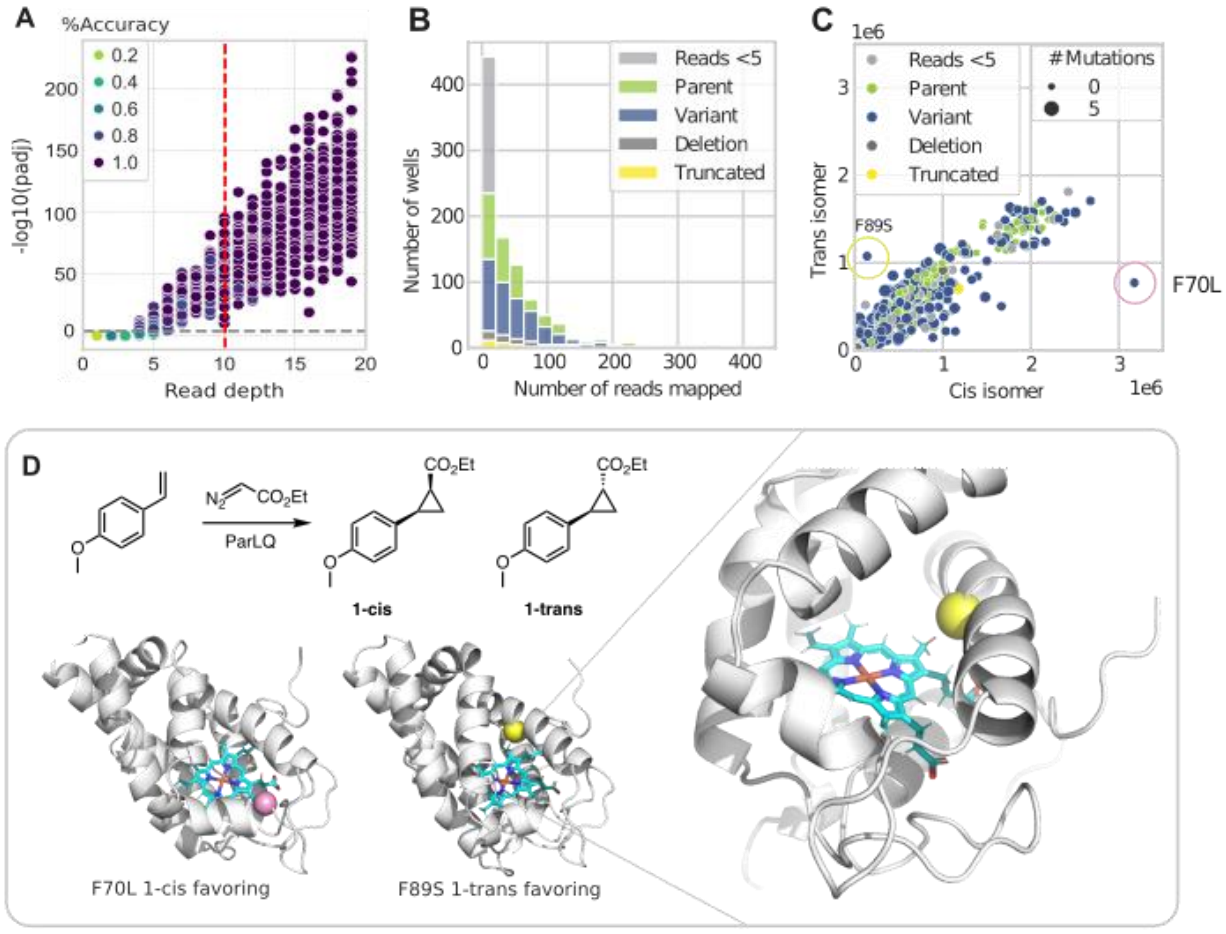

Figure 3: A. Accuracy of detecting variants using simulation studies on ParPgb LQ, varying read depth from 1 to 20 using an epPCR mutation rate of 2% and a nanopore sequencing error rate of 10%. B. Beneficial epPCR mutations favoring the *cis* and *trans* products are shown on

the ParPgb structure. C. Clustered reads across all ten plates from the epPCR experiment. 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 selected variants are highlighted on an AlphaFold3 structure of ParPgb.

All ten plates were screened for activity and specificity for catalyzing the formation of the *cis* and *trans* cyclopropanation products of 4-methoxystyrene (Figure 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 F89L 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[50,51] between sites 72 and 89, which could be further investigated with a double-site saturation mutagenesis experiment.

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[27]. 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[27,28]. 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[28].

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)[30] by Yang et al. to sequence and analyze four 96-well plates of *Par*Pgb 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[30]. 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)[13], cluster learning-assisted directed evolution (CLADE)[52], and degenerate codon optimization for informed libraries (DeCOIL)[53].

Conclusion

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.

Acknowledgments

The authors thank Daniel Graves for initial primer design and Ravi Lai, Jason Yang, Kathleen Sicinski, Ziyang Qin, Julia Reisenbauer, Casey Ritts, and Jae Kennemeur for providing samples for testing and developing protocols. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award

Number DE-SC0022218. A.M. is Supported by the Schmidt Science Fellows, in partnership with the Rhodes Trust.

Supporting Information Available

The following files are available free of charge.

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

References

(1) 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.

(2) *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).

(3) 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.

(4) Merkx, M.; Smith, B.; Jewett, M. Engineering Sensor Proteins. *ACS Sens.* **2019**, *4* (12), 3089–3091. https://doi.org/10.1021/acssensors.9b02459.

(5) Sarai, N. S.; Fulton, T. J.; O'Meara, R. L.; Johnston, K. E.; Brinkmann-Chen, S.; Maar, R. R.; Tecklenburg, R. E.; Roberts, J. M.; Reddel, J. C. T.; Katsoulis, D. E.; Arnold, F. H. Directed Evolution of Enzymatic Silicon-Carbon Bond Cleavage in Siloxanes. *Science* **2024**, *383* (6681), 438–443. https://doi.org/10.1126/science.adi5554.

(6) Smith, J. M. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225* (5232), 563–564. https://doi.org/10.1038/225563a0.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) Wittmann, B. J.; Yue, Y.; Arnold, F. H. Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution. *Cell Syst.* **2021**, *12* (11), 1026-1045.e7. https://doi.org/10.1016/j.cels.2021.07.008.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

(30) Yang, J.; Lal, R. G.; Bowden, J. C.; Astudillo, R.; Hameedi, M. A.; Kaur, S.; Hill, M.; Yue, Y.; Arnold, F. H. Active Learning-Assisted Directed Evolution. bioRxiv July 31, 2024, p 2024.07.27.605457. https://doi.org/10.1101/2024.07.27.605457.

(31) 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.

(32) 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.

(33) 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.

(34) 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.

(35) 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.

(36) 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.

(37) 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.

(38) 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.

(39) 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.

(40) Ni, Y.; Liu, X.; Simeneh, Z. M.; Yang, M.; Li, R. Benchmarking of Nanopore R10.4 and R9.4.1 Flow Cells in Single-Cell Whole-Genome Amplification and Whole-Genome Shotgun Sequencing. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 2352–2364. https://doi.org/10.1016/j.csbj.2023.03.038.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) 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.

(46) 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.

(47) 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.

(48) 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.

(49) 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.

(50) 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.

(51) Starr, T. N.; Thornton, J. W. Epistasis in Protein Evolution. *Protein Sci. Publ. Protein Soc.* **2016**, *25* (7), 1204–1218. https://doi.org/10.1002/pro.2897.

(52) 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.

(53) 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.