Ultrahigh throughput screening to train generative protein models for engineering specificity into unspecific peroxygenases

Pradeep M. Nair,[1,3,*] Daniel M. Steinberg,[2,3] Tiago Resende,[1] Angelica Faith Suarez,[1] Helen Power,[2] Cheng Soon Ong,[2] Vijay Sairam,[1] Shelly Hsiao-Ying Cheng,[1] Lorivie Fragata,[1] Alessandro Serafini,[1] Victor Oh,[1] Say Hwa Tan,[1] Robert E. Speight,[2] Akbar K. Vahidi [1]

[1]Allozymes Pte Ltd, Singapore, 118264

[2]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

[3]These authors contributed equally

*Correspondence: pradeep@allozymes.com (P.M.N.)

Abstract

Enzyme engineering is central to developing biocatalysts with improved activity and specificity, yet traditional approaches are often limited by the scale of experimental screening. Here we combine ultrahigh-throughput microfluidic droplet screening with generative machine learning to engineer specificity into an unspecific peroxygenase (UPO) from *Aspergillus brasiliensis* (AbrUPO). A library of 5e6 variants was expressed in *Komagataella phaffii* (*Pichia pastoris*) and screened for increased production of styrene oxide and reduced production of phenylacetaldehyde and phenylpropanol. From an ultrahigh throughput screening dataset of >30,000 unique sequences paired with function data. These data were used both for direct selection of enriched variants and to train a task-specific generative long short-term memory (LSTM) model using the Variational Search Distributions (VSD) framework. We compared variants selected by rank aggregation from the screening data (R series) with novel sequences generated by the refined generative model (G series). Experimental assays revealed that multiple G variants outperformed both wildtype AbrUPO and R variants, in terms of styrene oxide production and specificity. Correlation analyses further showed that the task-specific generative model outperformed protein models pretrained on large publicly available datasets in predicting experimental outcomes. Our results demonstrate that coupling ultrahigh throughput screening with generative protein models enables efficient discovery of improved enzyme variants beyond those identified by direct screening alone, providing a scalable strategy for task-specific enzyme engineering.

Introduction

Enzymes are uniquely adaptable tools that can catalyze a variety of biochemical reactions with regio- and stereospecificity. They can increase reaction rates by several orders of magnitude, and in contrast to heterogeneous catalysts, the activity of an enzyme can be modulated through the introduction of genetic mutations into the enzyme coding sequence, making them more controllable and potentially versatile tools for biocatalysis.[1–3] With this versatility comes the need for a method to identify the optimal enzyme to produce a target compound.

There is an ever-expanding body of work surrounding the development of tools to predict enzyme function and activity based on sequence, and more recently, structure.[1–6] These tools include maximum entropy models,[7] variational autoencoders,[8] language models,[9] and generative adversarial networks.[10] Modeling methods generally require training using very large datasets (e.g. annotated sequences available on Uniprot[11]) to learn sequence patterns that define enzyme function and may lead to higher activity. The data used for this training are generally collected independently of the protein prediction tasks to be performed, so these datasets are rarely focused on specific biocatalytic reactions of interest. Modeling tools trained in this manner have shown the ability to identify metagenomic sequences that encode enzymes with desired catalytic activity. However, optimization of a wildtype enzyme through amino acid substitutions is often too subtle a change in enzyme primary structure for broadly trained predictive tools to discern significant differences in phenotype.[4,12]

Ultrahigh throughput screening (uHTS) using microfluidic droplet sorting is a method designed to generate large volumes of sequence-function data.[13,14] Unlike publicly available databases that span many different enzyme classes, the data generated from uHTS will be narrowly and deeply focused on the enzymatic reaction that leads to the specific screened phenotype.[4] Prior reports have shown the utility in augmenting pretrained models with a focused experimental dataset.[15] Here, we reasoned that an empirical dataset using an assay predictive of desired phenotype, and exploring sufficient diversity in enzyme sequence, could be used to train a completely task specific model to generate mutation combinations that most improve enzyme performance. To test this hypothesis, a high-volume dataset generated from uHTS was used to train a task specific generative long short-term memory (LSTM) recurrent neural network (RNN) model. We also make use of the variational search distributions (VSD)[16] training framework for learning an LSTM, which imparts some ability for the LSTM to generate sequences that extrapolate away from its training data (Figure 1). This workflow is simpler than those suggested previously,[15] as our generative model requires no pretraining beyond what is in the low-fidelity uHTS dataset. Furthermore, the method is backed by theoretical guarantees[16] for improving sequence performance.

For this study, our goal was to engineer higher activity and specificity into an unspecific peroxygenase (UPO) enzyme. UPOs can perform a wide array of oxyfunctionalizations of non-activated carbons.[17,18] Unlike cytochrome P450s, they require only hydrogen peroxide as the terminal oxidant without the need for cofactors and appropriate cofactor recycling systems. These attributes make them attractive candidates for use in an industrial manufacturing setting. AbrUPO, natively expressed in *Aspergillus brasiliensis*, has been shown to react with at least 16 different substrates to produce more than 50 different products.[19] One of these products, styrene oxide, is a fine chemical intermediate in the production of epoxy resins, cosmetics, and drugs. Chemical synthesis of the epoxide from styrene can be accomplished using organic peracids or hydrogen peroxide and heterogeneous catalysts, but these methods have potential environmental considerations leading to growing interest in more benign biocatalytic synthesis. Here, we explored the enzymatic production of styrene oxide from styrene as well as two alternative substrate-product conversions that can be catalyzed by AbrUPO: styrene to phenylacetaldehyde and propylbenzene to phenylpropanol (Figure 2A). Our goal was to increase production of styrene oxide, while reducing production of phenylacetaldehyde and phenylpropanol, thereby improving the absolute specificity of the UPO biocatalyst. We used the promiscuous AbrUPO wildtype enzyme as a starting point and designed an enzyme library based on these engineering goals (Figure 1A).

We applied combinatorial saturation mutagenesis to generate a library with a theoretical diversity of 5 x 10^6 AbrUPO variants (Figure 1B), and expressed these variants in *K. phaffii*, a yeast host selected for its ability to secrete high titers of active enzymes (Figure 1C). This library was subjected to two uHTS assays (Figure 1D); first to identify variants that displayed higher reactivity with a desired substrate, styrene, and second to identify variants that displayed lower reactivity with an undesired substrate, propylbenzene. Data from these assays were used to train and condition an LSTM[16,20] (Figure 1E) to generate AbrUPO sequences with higher activity and absolute specificity for styrene conversion to styrene oxide. Several of these generated sequences were expressed in *K. phaffii*

clones (Figure 1F). Finally, we performed biotransformation reactions with either styrene or propylbenzene substrates to assess the biocatalytic potential of engineered UPO enzymes, in terms of productivity and specificity for the production of styrene oxide (Figure 1G).

Results

Combinatorial Library Design

We initially applied two library design concepts to generate diversity in the wildtype that were predicted to improve enzyme performance. First, we used a zero-shot approach to identify residues that are evolutionarily unconstrained and thus more likely to be tolerant to variation. For this purpose, we used a protein language model (PLM)[21] to the target sequence. PLMs can be used to provide a probability distribution over all 20 amino acids for each sequence position, representing the likelihood that a given amino acid is compatible with the protein family context learned by the model. Using this approach, we extracted the maximum probability value across all amino acids for every position. Positions where this maximum probability was low indicated that the model does not strongly favor any particular amino acid, and that the site may tolerate genetic drift without negative impacts on enzyme function. Based on this criterion, the five positions with the lowest maximum probabilities, D129, L28 D177, F8, and H65, were selected as model-guided mutable sites (Figure 1A, red residues). We targeted these sites for combinatorial mutagenesis, to build a mutability sub-library with a theoretical diversity of 2.5 x 10^6 (STAR Methods).

As our engineering goals were to both increase reactivity toward styrene and decrease reactivity toward propylbenzene, we targeted a second set of residues predicted to be directly involved in substrate recognition. Here, we focused on identifying active site residues directly shaping the styrene binding pocket. A three-dimensional structural model of AbrUPO was generated using protein structure prediction tools[22], followed by cavity detection to locate the putative active site[23]. Docking simulations were then performed with styrene as the target ligand (Figure 1A, yellow structure).[24] Residues were ranked according to the spatial distance of their side-chain heavy atoms to the docked substrate. The five closest residues, L84, F179, G182, A186, and L231 (Figure 1A, blue residues), were selected as substrate-proximal positions, under the assumption that mutations at these sites are most likely to alter substrate orientation, binding affinity, or reactivity at the heme center. The substrate binding sub-library was built by saturating these five sites combinatorially, creating a second sub-library also with theoretical diversity of 2.5 x 10^6.

Together, these ten positions, five chosen by predicted high evolutionary likelihood and five chosen by structural proximity to styrene, were used as the basis for combinatorial library design. The two sub-libraries were pooled together during expression host transformation to build a single library in *K. phaffii* with theoretical diversity of 5.0 x 10^6, and screening of this library identified 3.1 x 10^4 unique AbrUPO variant sequences (Table S1 and SI Data1). This dual strategy ensured that the library captured residues predicted to be permissive to sequence variability, as well as positions with direct influence on substrate binding, thereby increasing the likelihood of sampling functional and diverse variants for machine learning–guided enzyme engineering.

Ultrahigh throughput Library Screening

We designed a microfluidic droplet sorting assay (Figure 2A) in which UPO-expressing cells were grown in droplets and then substrate, either styrene or propylbenzene, and hydrogen peroxide cofactor were picoinjected into the droplets and allowed to react. A fluorescent sensor was then picoinjected to convert residual hydrogen peroxide into a fluorescent species. Droplets were then subjected to microfluidic droplet sorting, where droplets were sorted into “Dark” and “Bright” pools based on relative fluorescence intensity. More active UPO variants consume the hydrogen peroxide that the fluorescent sensor detects, so these variants are enriched in the dark droplet populations, and inactive variants are enriched in the bright droplet populations. For styrene, out of a sorted library of

144,398 total droplets, those with the lowest 0.6% fluorescence intensity, corresponding to highest UPO activity, were collected in the dark pool (Figure 2B). For propylbenzene, multiple sorting thresholds were set to create matched pairs of dark and bright pools with varying stringency (Figure 2C). Propylbenzene sort 1 separated the brightest (most inactive) 95% of 9441 total droplets, sort 2 separated the brightest 99% of 5691 total droplets, and sort 3 separated the brightest 93% of 9724 total droplets. Populations of cells in different droplet pools were sequenced to determine the reactivity of different UPO variants to the two different substrates.

Direct Lead Selection

We first used sequencing information directly from uHTS screening to select a subset of variants expected to be more reactive toward styrene, and less reactive toward propylbenzene, compared to wildtype AbrUPO (Table S1). We designed lead identification criteria based on these data (see STAR Methods for how leads were selected). Twenty variants that had been screened in uHTS were clonally built into the *K. phaffii* host and induced to express AbrUPO variants. Analytical assays were used to measure the production of the desired compound, styrene oxide, as well as undesired side products, phenylacetaldehyde and phenylpropanol. Of these twenty initial variants, ten were selected for re-testing with higher concentrations of either styrene or propylbenzene substrate to generate a high fidelity validation dataset. From these validation experiments, we identified a single AbrUPO variant, UPO 5, that produced 2-4x more styrene oxide and phenylacetaldehyde, and more than 5x more phenylpropanol (Figure 3A-B). Another variant, UPO 20, displayed about 3x more specificity for production of styrene oxide compared to phenylpropanol, and about 3.5x more specificity for production of styrene oxide compared to phenylacetaldehyde, while still producing > 90% yield of styrene oxide as the wildtype enzyme.

Protein Modeling to Generate Novel Diversity

We next asked whether the high volume of screening data focused on only the initial biocatalytic reactions of interest could be used to train a machine learning model that could then generate novel diversity with desired reactivity profiles. For this task, we built multiple generative models trained with the uHTS dataset (summarized in Table S2, and visualized in Figure S1). We then tested these models for their correlation to the high fidelity, low volume validation dataset with direct uHTS leads (SI Data2). These test results were used to select the highest performing generative model for subsequent variant generation in the 10 pre-selected sites. We found that an LSTM generative model yielded the best performance on the initial low volume, high fidelity validation dataset (Table S3, Figure S2). This LSTM model was trained using a classifier scorer, as opposed to directly learning from the uHTS data. This classifier was itself trained to predict highly active and substrate specific sequences from the uHTS dataset. We detail how we created classifier targets in the methodology. This classifier scorer allows the LSTM to extrapolate beyond the uHTS dataset using the variational search distributions (VSD) algorithm[16]. For a baseline to compare the generative model to, we used the uHTS data directly as a "model free" variant selection method. We summarize the two compared selection methods as:

- ***R*** – direct selection from 30,801 variants appearing in the microfluidics sequencing data that had more than one styrene "dark" read, using an aggregated ranking scores[25] (R-scores) which preference styrene substrate specificity. See Figure 4A, and STAR Methods for details.
- ***G*** – selection from 4,374 *entirely novel* variants generated by a task specific LSTM, trained using the VSD algorithm and the R-scores. The generated variants were scored using the log-likelihood of the generative model (G-scores), and generated variants that also appeared in the uHTS data were removed. See Figure 4B and STAR Methods for details.

For experimental testing, we chose the top 5 scored variants from each method, as well as 5 logarithmically spaced scored variants from each method, i.e. 10 selections per method. We refer to the variant scoring methodologies as R-scores and G-scores respectively, and the final selected variants as the R-series and G-series respectively – see Tables S4 and S5 for the selected variants. To differentiate between the initial selection pools, we refer to the 30,801 sequences from uHTS

screening with styrene substrate as R, and the 4,374 LSTM generated variants as G. This approach yielded a total of 20 variants, in addition to the wildtype AbrUPO. The encoding genes for each were individually integrated into the *K. phaffii* host genome. Resulting strains were induced to express UPO variant enzymes. Supernatant was collected from each sample, then used in biotransformation reactions to convert either styrene or propylbenzene substrates into styrene oxide, phenylacetaldehyde, or phenylpropanol products. All biotransformation reactions were carried out in duplicate and assayed by gas chromatography to measure product formation (Fig. 5A). Several of the supernatant samples showed no reactivity against either substrate. Western blotting confirmed protein expression upon induction for all transformants (Figure S3)[26]. We cannot assess whether unreactive samples were unreactive because of inactive enzyme or poor enzyme expression or folding. Among those samples that did show reactivity, we observed that two of the G-series, G3 and G167, produced ~2x more of the target styrene oxide product, and produced ~2x more styrene oxide per mole of phenylacetaldehyde, compared to the wildtype AbrUPO (Fig. 5B-C). A third member of the G-series, G929, produced roughly half as much styrene oxide as wildtype AbrUPO, but produced ~3x more styrene oxide per mole of phenylacetaldehyde and nearly 2x more styrene oxide per mole of phenylpropanol, compared to wildtype AbrUPO. In contrast, only R10K from the R-series produced significantly more styrene oxide per mole of phenylacetaldehyde, ~2x higher than wildtype AbrUPO.

To quantitatively assess the ensemble differences between the R-series and the G-series, we compared the top-5 selected variants from both strategies, and all selected variants from both strategies in Table 1. We found that in most cases, the generated variants, G, were significantly (α = 0.01) better performing than the aggregated rank variants, R, using a Brunner Munzel[27] test with an alternative hypothesis of $P(y_G > y_R) > \frac{1}{2}$. The top-5 selected variants demonstrate the difference in the predicted best performing variants under either strategy. This subset also emulates how one would typically use the R- and G-scores to select variants for testing, e.g. one would select the top-k scored variants. Whereas the comparison between all selected variants compares the ensemble distributions of variants selected or generated by the R and G strategies respectively, providing an estimate of how the pools of variants under each strategy compare. To measure absolute specificity (aspec), we compute the fractions of normalized yield, e.g.

$$aspec(A, B) = \frac{max(0,A)}{max(0,A)+max(0,B)+\epsilon}$$

Where we substitute e.g. A = styrene oxide and B = phenylacetaldehyde or phenylpropanol in μM, max guards against lower-than-detection limit negative readings, and $\epsilon = 10^{-5}$ is a numerical guard to avoid dividing by zero if A and B are both zero or negative (Figure 5D). This will yield a value in [0, 1] (modulo the effect of ϵ), where 0 means preference for B *or* no products detected, 1 means complete preference for A and 0.5 means no preference for either substrate.

Table 1. Average experimental results for the **top 5** and **all** selected variants from the G and R methods, and Brunner Munzel test p-values under the alternative hypothesis that the distribution of G is stochastically greater than that of R.

Strategy	Styrene oxide (μM)	Phenylpropanol (μM)	Phenylacetaldehyde (μM)	aspec(SO, PP)	aspec(SO, PA)	Samples (N)
Top 5 selected variants						
G mean	51.48	24.91	22.44	0.8294	0.5768	10
R mean	10.07	2.37	6.00	0.2396	0.2054	10
Brunner Munzel p-value	0.0011	0.0234	0.0091	0.0004	4.275×10^{-5}	20

All selected variants						
G mean	29.97	20.46	10.32	0.6587	0.4984	20
R mean	8.83	4.64	2.87	0.1975	0.1788	20
Brunner Munzel p-value	0.0010	0.0068	0.0007	6.186×10^{-5}	0.0009	40

In the VSD framework, the LSTM was trained in two stages; firstly on the microfluidics sequences only (no experimental outcomes), and secondly it is refined using a classifier trained to discriminate the top 10% of R-scored sequences to score sequences generated by the LSTM. We refer to these as the "prior LSTM" and "VSD LSTM", respectively. The function of the prior LSTM is to constrain the VSD LSTM generated sequences to not be too different from those seen in the microfluidics sequencing data, and the function of the classifier is to push the VSD LSTM to generate predicted high-performing sequences, possibly outside the set of seen sequences.

We also tested the Spearman correlation of all experimental outcomes and the likelihood of the sequences under the refined generative model (VSD LSTM), its unrefined prior model (Prior LSTM) that had not seen the microfluidics measurements, and a host of pretrained models in Table 2 and STAR Methods. We only show the highest correlated pretrained models. See SI Data3 for the complete tested set. In all cases the VSD LSTM model had the highest correlation. The prior LSTM also performs quite well even though it has only seen sequences and microfluidics measurements. This relatively high performance is because most sequences detected by the uHTS process were taken from the styrene-active pool of variants. We must caution however, that these results may suffer from some selection bias, as VSD LSTM was used to select almost half of the samples.

Table 2. Spearman correlations and p-values (in brackets) between the generative and pretrained models on the 42 experimental samples (duplicate samples of 10 R-series, 10 G-series, and AbrUPO wildtype).

Assay	VSD LSTM	Prior LSTM	Netsolp solubility	Boltz unconstrained classification (relevant substrate)	Boltz unconstrained affinity (relevant substrate)
Phenylpropanol (μM)	**0.435 (0.004)**	0.330 (0.033)	0.343 (0.026)	0.395 (0.010)	-0.345 (0.025)
Phenylacetaldehyde (μM)	**0.447 (0.003)**	0.360 (0.019)	0.392 (0.010)	0.163 (0.303)	-0.154 (0.331)
Styrene oxide (μM)	**0.443 (0.003)**	0.363 (0.018)	0.416 (0.006)	0.146 (0.357)	-0.308 (0.047)

We also sought to more thoroughly understand the VSD LSTM's selection performance by testing the model's ability to select "useful" variants from the 42 experimental outcomes on four binarized target criteria, with the proportion of positive examples:

1. All product (SO, PP, PA) AUC > 1, Proportion positive labels: 0.429
2. Styrene Oxide > WT (92.87), Proportion positive labels: 0.143
3. Specificity aspec(SO, PA) > WT (0.5657), Proportion positive labels: 0.429
4. Specificity aspec(SO, PP) > WT (0.7868), Proportion positive labels: 0.238

The first target is a permissive criterion meant to determine how successful the model is at predicting enzymes that will express and fold correctly. The second target captures the engineering goal of

increased styrene oxide production. The third and fourth targets are stringent criteria that identify variants more specific than the wildtype enzyme to produce styrene oxide compared to phenylacetaldehyde or phenylpropanol, respectively.

We used the ([0, 1]-normalized) log-likelihoods of the generative model (G-scores) as the quantity to select variants by thresholding, then measured how many positive and negative variants according to the preceding 4 criteria were selected for different threshold levels. We summarize results as precision-recall curves in Figure S4, where we use bootstrap sampling (with 1000 bootstrap samples) for creating intervals. Precision can also be understood as the true discovery rate (true positives / true and false positive) and recall the sensitivity (true positives / true positives and false negatives). When choosing a threshold on the log-likelihood for selection, precision-recall curves can help us understand the trade-offs we face, as choice of threshold will directly affect precision and recall outcomes. For example, if we chose a low threshold so that all variants are selected, then we would capture all the positive variants, maximizing recall, but we would also select all the negative variants (as false positives), lowering precision. Conversely, if we set the threshold so that only the highest log-likelihood variant is selected, then we will achieve perfect precision if this variant is positive, but low recall because we miss all the other positive variants in the set (false negatives). We can also summarize these curves by measuring the average precision (AP) for all recall values in the curve. An AP of 1 means perfect precision and recall, and an AP that is the same as the chance level of precision means the predictor is no better than random. We summarize the APs below (with 95% confidence interval in brackets),

1. All product (SO, PP, PA) AUC > 1: AP = 0.759 (0.626, 0.887)
2. Styrene Oxide > WT: AP = 0.244 (0.183, 0.458)
3. Specificity aspec(SO, PA) > WT: AP = 0.602 (0.483, 0.813)
4. Specificity aspec(SO, PP) > WT: AP = 0.291 (0.229, 0.448)

All APs are better than chance (including the lower confidence limit). We found that the VSD LSTM performed significantly better than chance (randomly selecting variants according to the base positive label rate) for all recall levels, in all cases. The VSD LSTM maintained high precision for all recall levels for the AUC > 1 and aspec(SO, PA) > WT criteria, but aspec(SO, PP) > WT and Styrene Oxide > WT have a lower precision, most likely because the wildtype sequence already exhibits high activity and specificity for styrene compared to propylbenzene substrate, resulting in fewer positive examples. Note that since some of these variants were chosen using the VSD LSTM (particularly the top-5 G series), there may be some slight selection bias present in these results, however we believe that these plots are still useful in indicating how useful the VSD LSTM would be if it was the only method used to select variants for experimentation, and a threshold selection criterion was used, as opposed to our top 5 and logarithmically spaced methodology.

Conclusion

Engineering enzymes to achieve higher activity and specificity remains a central challenge for biocatalysis. In this work, we demonstrated that ultrahigh throughput microfluidic screening (uHTS) can provide high volume, low fidelity measurements sufficient for effectively training task-specific generative models. Using the Variational Search Distributions (VSD) framework to refine a long short-term memory (LSTM) model on these uHTS data, we were able to generate novel *Aspergillus brasiliensis* UPO variants with improved performance compared not only to the wildtype enzyme, but also to the best variants directly identified from uHTS selection. Several generated sequences achieved substantially higher styrene oxide yields and specificity ratios, validating that low-fidelity measurements can be effectively leveraged to guide design toward higher-fidelity experimental outcomes.

Our findings underscore three key insights. First, low fidelity screening data, whilst noisy and indirect, are nonetheless valuable for informing generative models that can extrapolate beyond the training set

to propose superior variants. Second, a relatively simple generative architecture, when refined via VSD conditioning, consistently outperformed direct selection from the same screening data—illustrating how generative modeling can transcend the limits of the assay used for training. Third, while some pretrained protein language models achieved moderate correlations with assay outcomes, their predictive value was inconsistent or hard to arrive at *ex-ante*, emphasizing the advantage of models explicitly trained on task-specific, experimentally grounded data.

Overall, this study illustrates a generalizable strategy: breadth of coverage through ultrahigh throughput, low fidelity screening, coupled with generative modeling to translate that data into high fidelity design gains. Armed with a model trained on an enzyme class of interest, it is now possible to proceed to intelligently designed specific variants without the onerous task of deep mutational scanning. By exploiting the complementary strengths of scale and extrapolation, this approach accelerates the discovery of enzyme variants with improved production of specific compounds, and offers a blueprint for applying machine learning to protein engineering problems where conventional data collection is limited.

Acknowledgements

DMS, HP, CSO and RS were funded by the CSIRO Advanced Engineering Biology Future Science Platform and Science Digital Program, and were supported by resources and expertise provided by CSIRO IMT Scientific Computing. For helpful discussions and support of this work, we thank Michael Burns, Audrey Robic, Jayer Tan, and Muhummed Khabir. We thank Charmaine Lee for graphical depictions of the project workflow.

Contributions

Conceptualization, PMN, DMS; Library design, TR, DS, and PMN; Cloning SHC, LF, AFS, AS, and PMN; Microfluidic screening, SWT, VS, VO; Analytical chemistry, AFS; Sequencing, SHC, AS, AFS, TR; Project supervision, AKV, and RS; Protein modeling, DS, HP, TR, and CSO

Bibliography

(1) 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.

(2) Xie, W. J.; Warshel, A. Harnessing Generative AI to Decode Enzyme Catalysis and Evolution for Enhanced Engineering. *Natl. Sci. Rev.* **2023**, *10* (12), nwad331. https://doi.org/10.1093/nsr/nwad331.

(3) Weissenow, K.; Rost, B. Are Protein Language Models the New Universal Key? *Curr. Opin. Struct. Biol.* **2025**, *91*, 102997. https://doi.org/10.1016/j.sbi.2025.102997.

(4) Fenoy, E.; Edera, A. A.; Stegmayer, G. Transfer Learning in Proteins: Evaluating Novel Protein Learned Representations for Bioinformatics Tasks. *Brief. Bioinform.* **2022**, *23* (4), bbac232. https://doi.org/10.1093/bib/bbac232.

(5) 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.

(6) Wen, S.; Zheng, W.; Bornscheuer, U. T.; Wu, S. Generative Artificial Intelligence for Enzyme Design: Recent Advances in Models and Applications. *Curr. Opin. Green Sustain. Chem.* **2025**, *52*, 101010. https://doi.org/10.1016/j.cogsc.2025.101010.

(7) Figliuzzi, M.; Barrat-Charlaix, P.; Weigt, M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Mol. Biol. Evol.* **2018**, *35* (4), 1018–1027. https://doi.org/10.1093/molbev/msy007.

(8) Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep Generative Models of Genetic Variation Capture the Effects of Mutations. *Nat. Methods* **2018**, *15* (10), 816–822. https://doi.org/10.1038/s41592-018-0138-4.

(9) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. Synthetic Biology July 10, 2021. https://doi.org/10.1101/2021.07.09.450648.

(10) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks. *Nat. Mach. Intell.* **2021**, *3* (4), 324–333. https://doi.org/10.1038/s42256-021-00310-5.

(11) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci.* **2021**, *118* (15), e2016239118. https://doi.org/10.1073/pnas.2016239118.

(12) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-Tuning Protein Language Models Boosts Predictions across Diverse Tasks. *Nat. Commun.* **2024**, *15* (1), 7407. https://doi.org/10.1038/s41467-024-51844-2.

(13) Kintses, B.; Hein, C.; Mohamed, M. F.; Fischlechner, M.; Courtois, F.; Lainé, C.; Hollfelder, F. Picoliter Cell Lysate Assays in Microfluidic Droplet Compartments for Directed Enzyme Evolution. *Chem. Biol.* **2012**, *19* (8), 1001–1009. https://doi.org/10.1016/j.chembiol.2012.06.009.

(14) Gantz, M.; Neun, S.; Medcalf, E. J.; Van Vliet, L. D.; Hollfelder, F. Ultrahigh-Throughput Enzyme Engineering and Discovery in *In Vitro* Compartments. *Chem. Rev.* **2023**, *123* (9), 5571–5611. https://doi.org/10.1021/acs.chemrev.2c00910.

(15) Thomas, N.; Belanger, D.; Xu, C.; Lee, H.; Hirano, K.; Iwai, K.; Polic, V.; Nyberg, K. D.; Hoff, K. G.; Frenz, L.; Emrich, C. A.; Kim, J. W.; Chavarha, M.; Ramanan, A.; Agresti, J. J.; Colwell, L. J. Engineering Highly Active Nuclease Enzymes with Machine Learning and High-Throughput Screening. *Cell Syst.* **2025**, *16* (3), 101236. https://doi.org/10.1016/j.cels.2025.101236.

(16) Steinberg, D. M.; Oliveira, R.; Ong, C. S.; Bonilla, E. V. Variational Search Distributions. In *The Thirteenth International Conference on Learning Representations (ICLR)*; 2025.

(17) Melling, B.; Mielke, T.; Whitwood, A. C.; O'Riordan, T. J. C.; Mulholland, N.; Cartwright, J.;

Unsworth, W. P.; Grogan, G. Complementary Specificity of Unspecific Peroxygenases Enables Access to Diverse Products from Terpene Oxygenation. *Chem Catal.* **2024**, *4* (2), 100889. https://doi.org/10.1016/j.checat.2023.100889.

(18) Rytkönen, E.; Hakulinen, N.; Jänis, J.; Rouvinen, J. Exploring Unspecific Peroxygenase Selectivity with Diverse Hydrocarbon Substrates. *ChemistryOpen* **2025**, e202400521. https://doi.org/10.1002/open.202400521.

(19) Schmitz, F.; Koschorreck, K.; Hollmann, F.; Urlacher, V. B. Aromatic Hydroxylation of Substituted Benzenes by an Unspecific Peroxygenase from *Aspergillus Brasiliensis*. *React. Chem. Eng.* **2023**, *8* (9), 2177–2186. https://doi.org/10.1039/d3re00209h.

(20) Saka, K.; Kakuzaki, T.; Metsugi, S.; Kashiwagi, D.; Yoshida, K.; Wada, M.; Tsunoda, H.; Teramoto, R. Antibody Design Using LSTM Based Deep Generative Model from Phage Display Library for Affinity Maturation. *Sci. Rep.* **2021**, *11* (1), 5852. https://doi.org/10.1038/s41598-021-85274-7.

(21) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci.* **2021**, *118* (15), e2016239118. https://doi.org/10.1073/pnas.2016239118.

(22) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(23) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv February 11, 2023. https://doi.org/10.48550/arXiv.2210.01776.

(24) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.

(25) Bedö, J.; Ong, C. S. Multivariate Spearman's ρ for Aggregating Ranks Using Copulas. *J. Mach. Learn. Res.* **2016**, *17* (201), 1–30.

(26) Ebner, K.; Pfeifenberger, L. J.; Rinnofner, C.; Schusterbauer, V.; Glieder, A.; Winkler, M. Discovery and Heterologous Expression of Unspecific Peroxygenases. *Catalysts* **2023**, *13* (1), 206. https://doi.org/10.3390/catal13010206.

(27) Brunner, E.; Munzel, U. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biom. J.* **2000**, *42* (1), 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1%253C17::AID-BIMJ17%253E3.0.CO;2-U.

(28) Liu, D.; Xuanyuan, T.; Liu, X.; Fu, W.; Liu, W. Massive and Efficient Encapsulation of Single Cells in Monodisperse Droplets and Collagen–Alginate Microgels Using a Microfluidic Device. *Front. Bioeng. Biotechnol.* **2023**, *11*. https://doi.org/10.3389/fbioe.2023.1281375.

(29) Abate, A. R.; Hung, T.; Mary, P.; Agresti, J. J.; Weitz, D. A. High-Throughput Injection with Microfluidics Using Picoinjectors. *Proc. Natl. Acad. Sci.* **2010**, *107* (45), 19163–19166. https://doi.org/10.1073/pnas.1006888107.

(30) Baret, J.-C.; Miller, O. J.; Taly, V.; Ryckelynck, M.; El-Harrak, A.; Frenz, L.; Rick, C.; Samuels, M. L.; Hutchison, J. B.; Agresti, J. J.; Link, D. R.; Weitz, D. A.; Griffiths, A. D. Fluorescence-Activated Droplet Sorting (FADS): Efficient Microfluidic Cell Sorting Based on Enzymatic Activity. *Lab. Chip* **2009**, *9* (13), 1850–1858. https://doi.org/10.1039/B902504A.

(31) Teo, A. J. T.; Yan, M.; Dong, J.; Xi, H.-D.; Fu, Y.; Tan, S. H.; Nguyen, N.-T. Controllable Droplet Generation at a Microfluidic T-Junction Using AC Electric Field. *Microfluid. Nanofluidics* **2020**, *24* (3), 21. https://doi.org/10.1007/s10404-020-2327-6.

(32) Guo, W. Pressure-Driven Filling of Closed-End Microchannel: Realization of Comb-Shaped

Transducers for Acoustofluidics. *Phys. Rev. Appl.* **2018**, *10* (5). https://doi.org/10.1103/PhysRevApplied.10.054045.

(33) Collins, D. J.; Neild, A.; deMello, A.; Liu, A.-Q.; Ai, Y. The Poisson Distribution and beyond: Methods for Microfluidic Droplet Production and Single Cell Encapsulation. *Lab. Chip* **2015**, *15* (17), 3439–3459. https://doi.org/10.1039/C5LC00614G.

(34) Mika, T.; Kalnins, M.; Spalvins, K. The Use of Droplet-Based Microfluidic Technologies for Accelerated Selection of Yarrowia Lipolytica and Phaffia Rhodozyma Yeast Mutants. *Biol. Methods Protoc.* **2024**, *9* (1), bpae049. https://doi.org/10.1093/biomethods/bpae049.

(35) Phuong, M.; Hutter, M. Formal Algorithms for Transformers. arXiv July 19, 2022. http://arxiv.org/abs/2207.09238 (accessed 2024-07-18).

(36) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

(37) Thumuluri, V.; Martiny, H.-M.; Almagro Armenteros, J. J.; Salomon, J.; Nielsen, H.; Johansen, A. R. NetSolP: Predicting Protein Solubility in *Escherichia Coli* Using Language Models. *Bioinformatics* **2022**, *38* (4), 941–946. https://doi.org/10.1093/bioinformatics/btab801.

(38) Mikhael, P. G.; Chinn, I.; Barzilay, R. CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes.

(39) Kroll, A.; Ranjan, S.; Lercher, M. J. A Multimodal Transformer Network for Protein-Small Molecule Interactions Enhances Predictions of Kinase Inhibition and Enzyme-Substrate Relationships. *PLOS Comput. Biol.* **2024**, *20* (5), e1012100. https://doi.org/10.1371/journal.pcbi.1012100.

(40) Wilson, M.; Coudrat, T.; Warden, A. SELFprot: Effective and Efficient Multitask Finetuning Methods for Protein Parameter Prediction. *J. Chem. Inf. Model.* **2025**, *65* (7), 3226–3238. https://doi.org/10.1021/acs.jcim.4c02230.

(41) Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction.

Graphical abstract

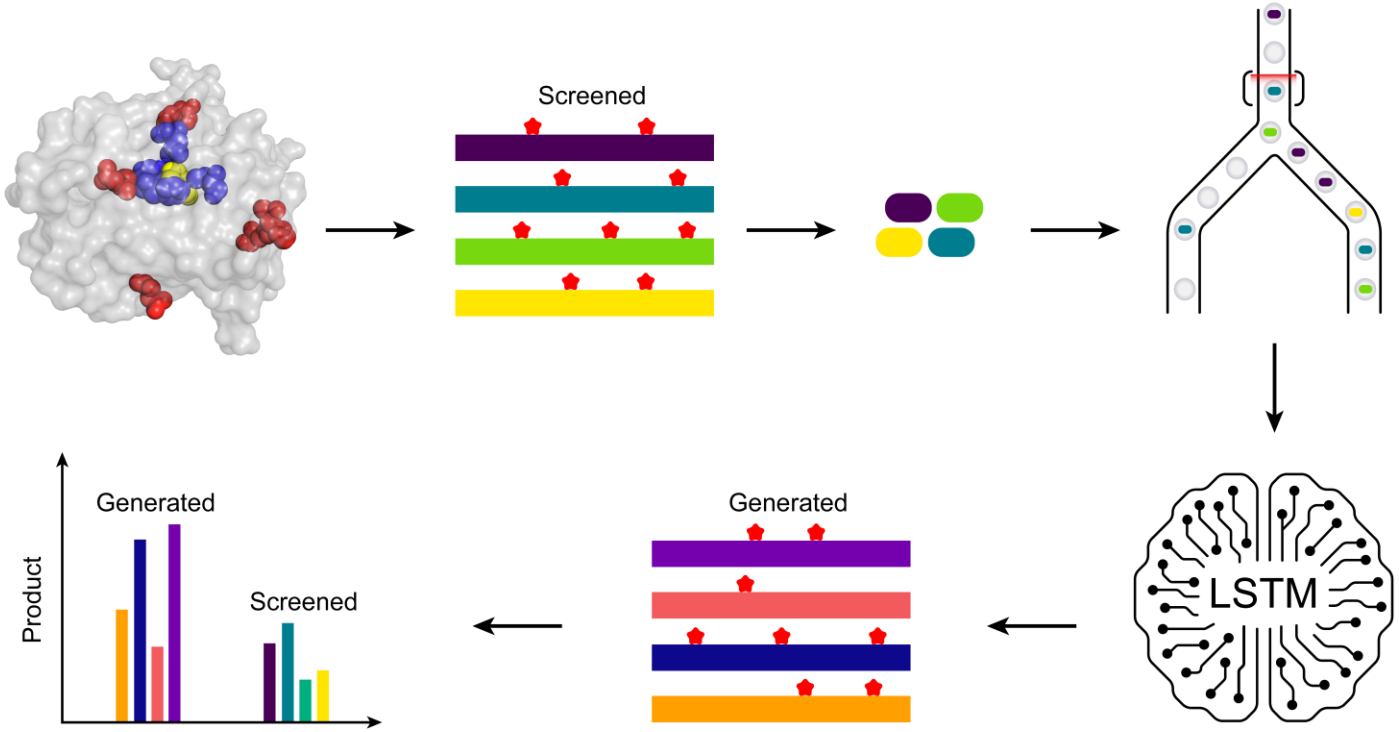
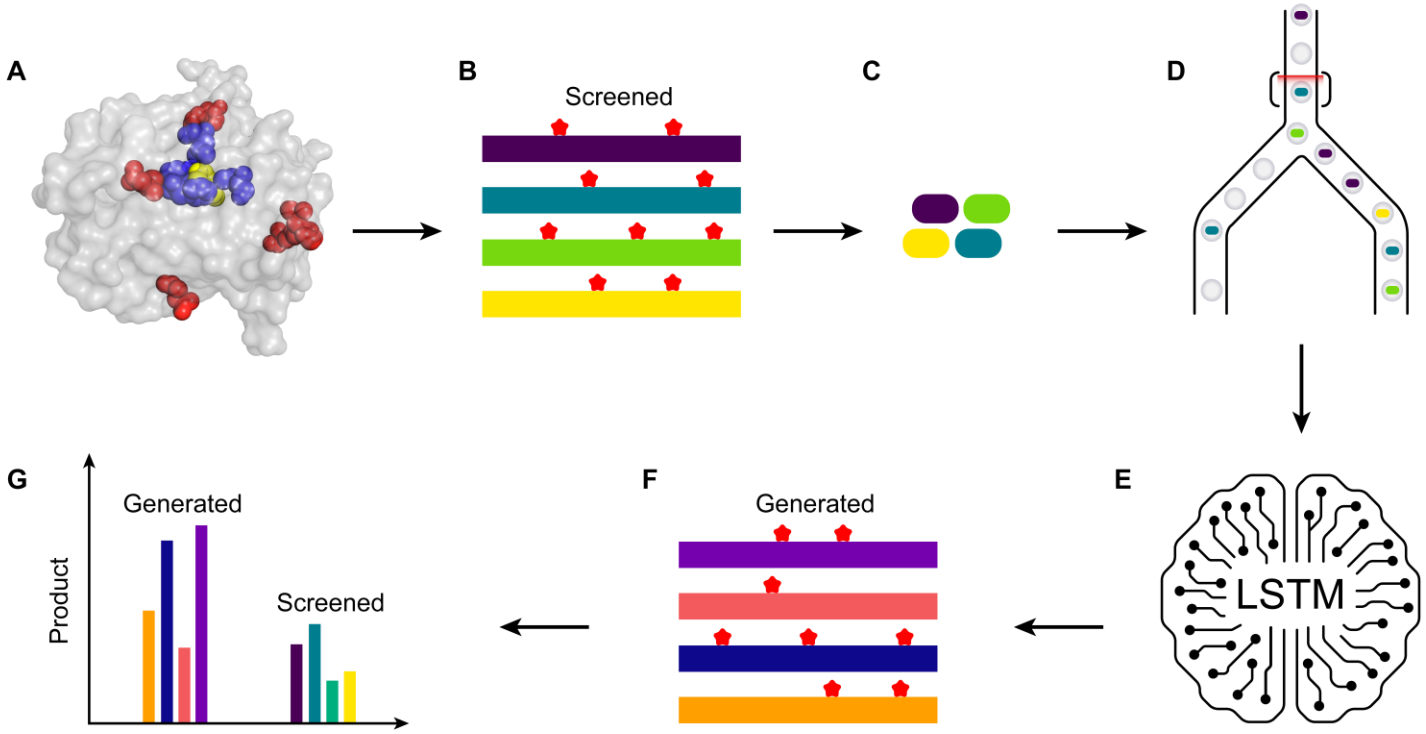


Figure 1

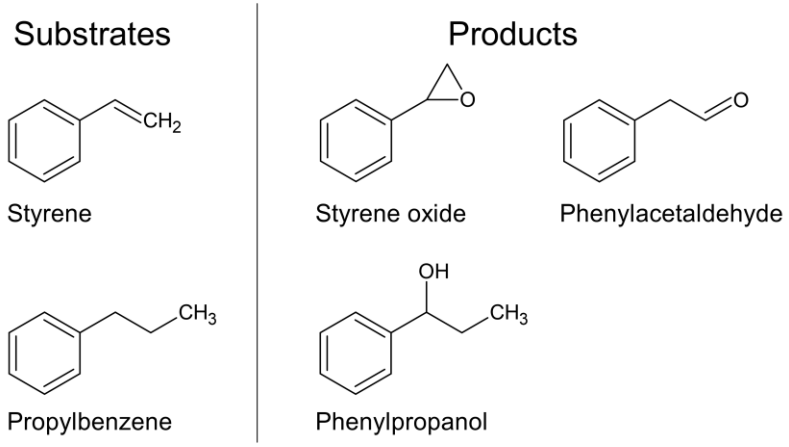


Ultrahigh throughput-guided enzyme modeling. (A) Diverse libraries are designed for the target enzyme based on engineering goals that may include increased activity and specificity. (B). Gene variants are constructed, targeting the sites and mutations identified during design. (C) The expression host is transformed to express the enzyme library. (D) The enzyme library is screened using microfluidic droplet sorting to generate a high volume of data. (E) Microfluidic droplet sorting data is used to train a task-specific long short-term memory (LSTM) model. (F) The LSTM generates novel enzyme variants that are predicted to have improved performance. (G) The generated variants are assayed against a subset of screened variants in low volume, high fidelity testing.

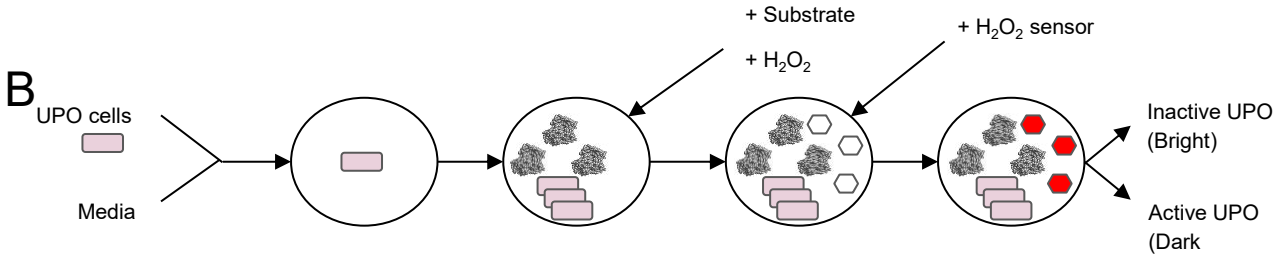
Figure 2

Ultrahigh throughput screening (uHTS). (A) Either styrene or propylbenzene are substrates added during screen. UPO catalyzes their conversion to either styrene oxide or phenylacetaldehyde (for styrene) or phenylpropanol (for propylbenzene). (B) Cells are singly encapsulated in microfluidic droplets in growth media containing methanol for induction of protein expression. Cells are allowed to grow, replicate, and express UPO protein. Substrate and hydrogen peroxide are picoinjected into droplets and UPO is allowed to react. A hydrogen peroxide sensor is picoinjected into droplets to convert any residual hydrogen peroxide into a fluorescent species. Droplets are then sorted by fluorescence. Active UPO will consume hydrogen peroxide, leading to darker droplets. (C) Droplet fluorescence histogram from droplet screening for styrene selection. The red line indicates sorting threshold to separate bright and dark droplets. (D) Droplet fluorescence histograms from droplet screening for propylbenzene (PB) selections. The red lines indicate weighted mean sorting thresholds to separate bright and dark droplets.

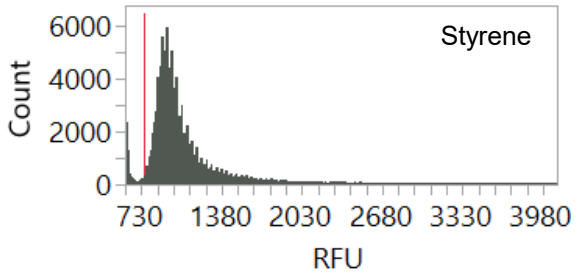
A



B



C



D

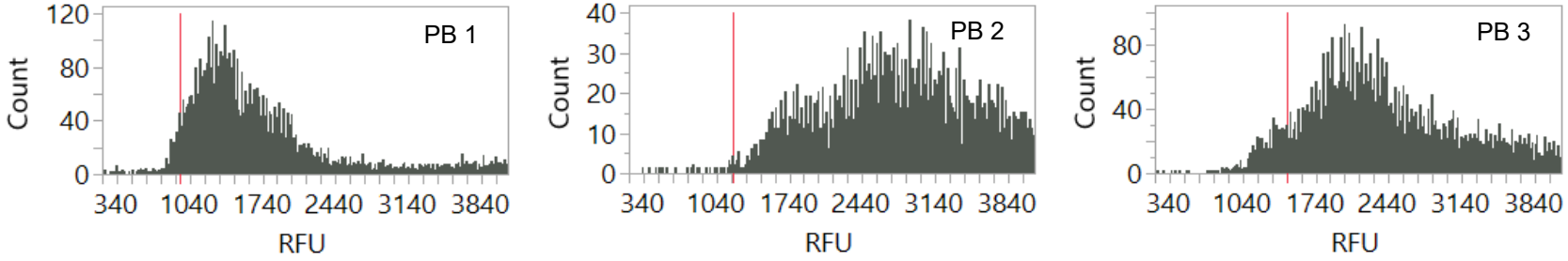
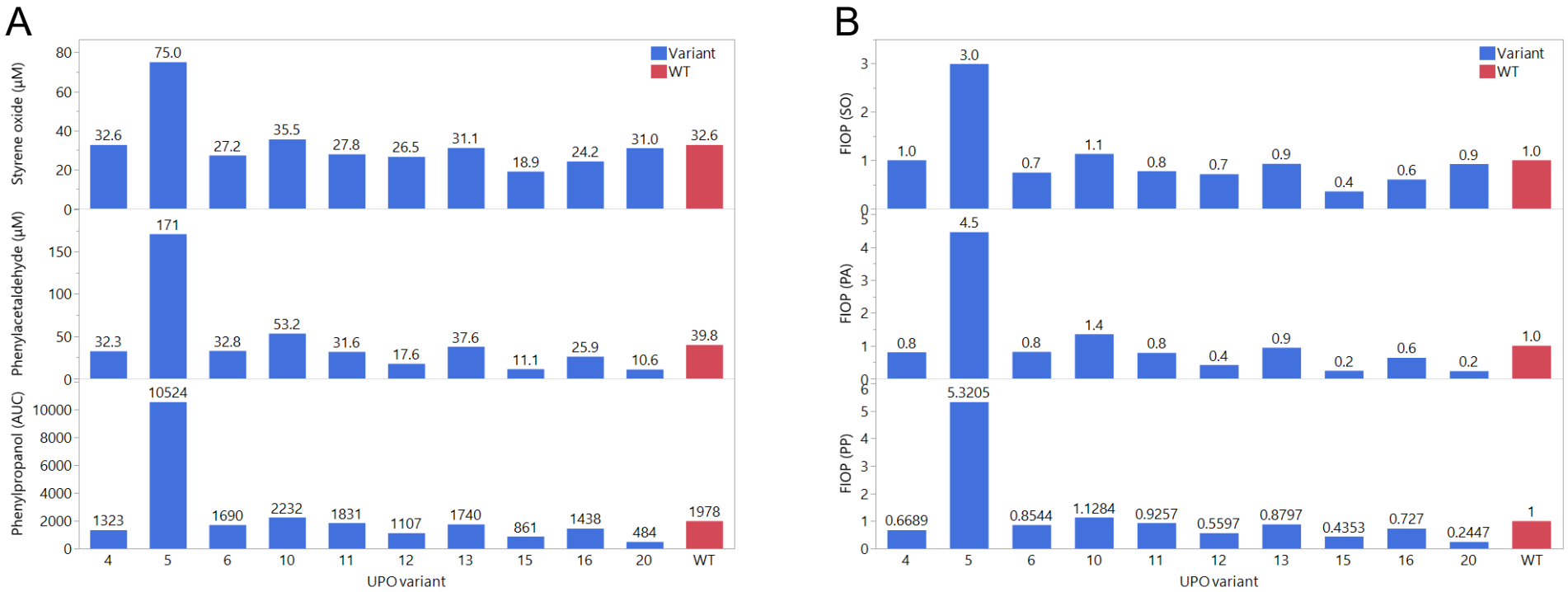
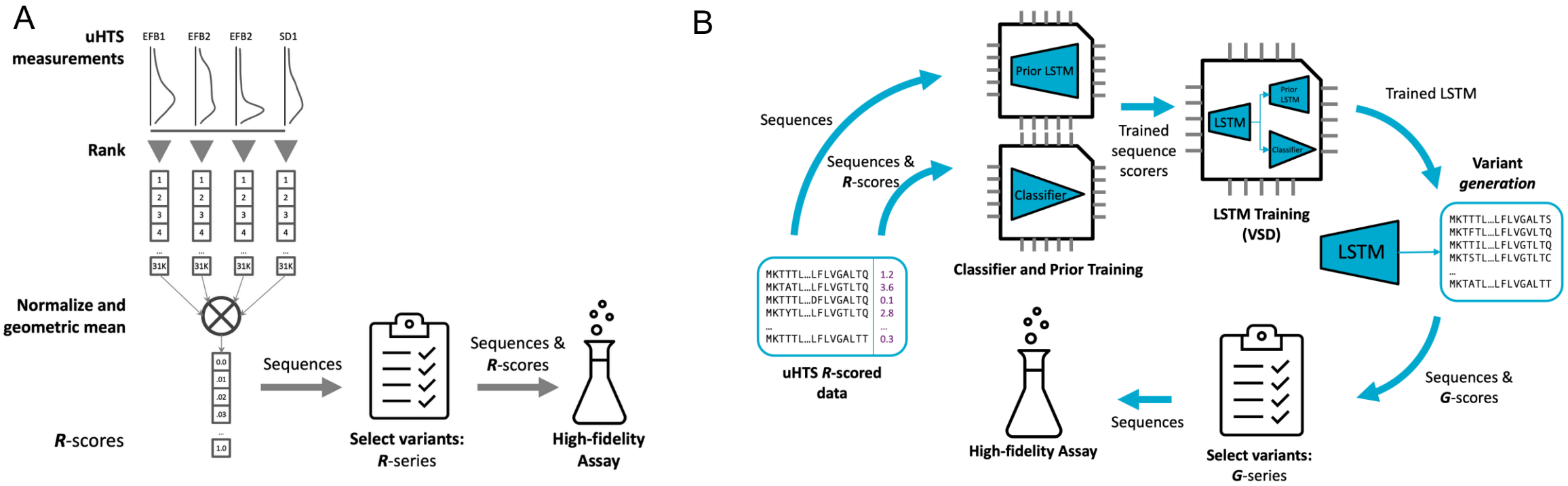


Figure 3



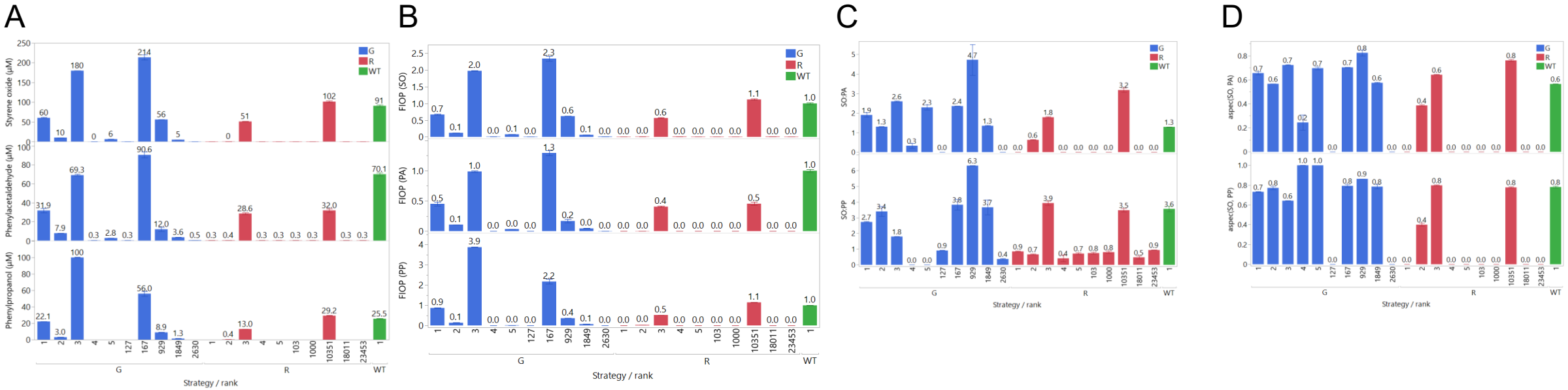
Validation of uHTS. (A) Lead AbrUPO variants from uHTS screening were assessed for productivity of styrene oxide and phenylacetaldehyde by gas chromatography, and for production of phenylpropanol by high performance liquid chromatography. (B) Fold improvement over parent (FIOP) for production by lead UPO variants selected directly from uHTS.

Figure 4



Model training data from uHTS. (A) Visualisation of the rank aggregation process to compute R-scores. (B) Visualisation of the VSD LSTM training process and computation of G-scores.

Figure 5



R- and G-series performance. For R- and G-series UPO variants: (A) Productivity of styrene oxide, phenylacetaldehyde, and phenylpropanol measured by gas chromatography. (B) Fold improvement over parent (FIOP) for production of styrene oxide (SO), phenylacetaldehyde (PA), and phenylpropanol (PP). (C) Ratio of styrene oxide (SO) to phenylacetaldehyde (PA) or phenylpropanol (PP). (D) Absolute specificity of styrene oxide to phenylacetaldehyde (aspec(SO, PA)) and styrene oxide to phenylpropanol (aspec(SO, PP)). Error bars denote standard deviation across duplicate biotransformation reactions.