Benchmarking and Experimental Validation of Machine Learning Strategies for Enzyme Engineering

Zishuo Zeng[1], Jiao Jin[2], Rufang Xu[1], Xiaozhou Luo[1]*

1. State Key Laboratory of Quantitative Synthetic Biology, Center for Synthetic Biochemistry, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

2. Synceres Biosciences Co. Ltd. 1100 Huanli Road, Guangming District, Shenzhen, Guangdong, China 518083

* Corresponding authors

Emails: xz.luo@siat.ac.cn

Abstract

Enzyme-directed evolution increasingly relies on computational tools to prioritize mutations, yet their practical value is difficult to assess because kinetic data are often aggregated across heterogeneous assay conditions, inflating apparent generalization. Here we introduce EnzyArena, a curated benchmark that groups kinetic parameters (*kcat*, *Km*, *kcat/Km*) into condition-matched experimental subsets to enable realistic evaluation. Using this resource, we benchmark 10 representative models from two arising strategy families—zero-shot fitness prediction and supervised kinetic-parameter prediction—across BRENDA- and SABIO-RK–derived subsets and 25 independent mutagenesis datasets. Kinetic-parameter predictors perform strongly on database-derived subsets but lose their advantage on independent datasets, whereas zero-shot predictors show more consistent generalization. A simple consensus of multiple zero-shot models further improves the precision of identifying beneficial mutants. We prospectively validated these findings in a wet-lab campaign (150 mutants) comparing random mutants, UniKP-prioritized mutants and ESM-1v-prioritized mutants (representing supervised kinetic-parameter prediction and zero-shot fitness prediction, respectively), where ESM-1v achieved the highest utility and UniKP underperformed the random baseline. Together, this study establishes realistic baselines for computational mutant prioritization and highlights consensus zero-shot strategies as a practical starting point for enzyme engineering.

Keywords

Enzyme directed evolution, benchmark dataset, zero-shot prediction, enzyme kinetic parameter prediction

Introduction

Enzyme-directed evolution underpins modern biocatalysis, green chemistry [1]and industrial biotechnology [2] by enabling the rapid optimization of catalysts for non-native substrates, process conditions and reaction networks. Over the past few years, advances in machine learning have rapidly transformed enzyme mutant prioritization [3]. In particular, two emerging classes of computational strategies are now increasingly used to guide mutagenesis campaigns: protein language model–based zero-shot fitness prediction [4-7] and supervised kinetic-parameter (KP) predictors [8-13]. These approaches have been successfully applied to prioritize mutations, accelerate screening, and guide enzyme engineering campaigns across diverse systems [14, 15-17]. In principle, such tools promise to compress the search space, reduce experimental burden, and enable more systematic exploration of sequence–function relationships than traditional intuition-driven approaches.

Despite this rapid adoption, it remains unclear which of these emerging strategies provides the most practical value for enzyme engineering workflows. In real directed-evolution campaigns, experimental screening capacity is limited, and researchers must decide which computational signals to trust when prioritizing variants [18]. Although zero-shot fitness predictors and KP predictors are based on fundamentally different sources of information—evolutionary sequence statistics [4] versus supervised learning from kinetic measurements[10], they are often used for enzyme directed evolution in practice [16, 19]. A systematic and fair comparison of these approaches under realistic evaluation settings is therefore essential to determine which strategy most effectively enriches beneficial mutations.

To meaningfully compare these emerging strategies, both rigorous in silico benchmarking and prospective wet-lab validation are required. Fair computational benchmarking depends on reliable datasets that allow realistic and comparable evaluation across enzyme engineering scenarios. However, current kinetic datasets are difficult to use for this purpose. Many studies [9-13] combine experimental measurements obtained under different assay conditions (e.g.,

temperature, pH, substrate identity, cofactors, host background, and assay format), and then treat the resulting *kcat*, *Km*, or *kcat/Km* values as directly comparable. Endpoints with different meanings, such as catalytic turnover, evolutionary fitness, and expression level, are also frequently used as interchangeable proxies. Even when nominal conditions appear similar, measurements can differ across laboratories and instruments because of calibration and protocol differences [20]. In addition, models trained on public resources such as BRENDA [21] and SABIO-RK [22] are often evaluated on datasets that share homologous sequences or experimental campaigns, which can lead to overly optimistic estimates of generalization [23 24 25]. Together, these issues make it difficult to determine which computational strategies truly improve mutant prioritization in practice.

In this work, we take two complementary approaches to enable a comprehensive evaluation of computational strategies for enzyme engineering: rigorous benchmarking and prospective wet-lab validation. First, we construct a curated benchmark, EnzyArena, for computationally guided enzyme-directed evolution built around kinetic parameters measured under strictly matched experimental conditions and grouped at the level of individual experimental campaigns. We assemble datasets in which temperature, pH, substrate, host, and assay format are controlled within each subset. We treat these experimental conditions as fundamental sources of variation rather than negligible noise (Figure 1A). Since the KP predictors are commonly trained on public resources such as BRENDA and SABIO-RK, resulting substantial overlap between these database-derived data and the KP models' training data. To further examine model robustness and generalizability, we therefore complement EnzyArena with independent mutagenesis datasets that are not present in these resources. Using this framework, we compare two major model families used for mutation prioritization, zero-shot fitness predictors and KP predictors (Figure 1B), under evaluation protocols designed to reflect realistic directed-evolution scenarios. We then complement this retrospective benchmarking with a prospective wet-lab enzyme engineering campaign to directly test whether model-guided recommendations improve mutant prioritization under practical experimental constraints.

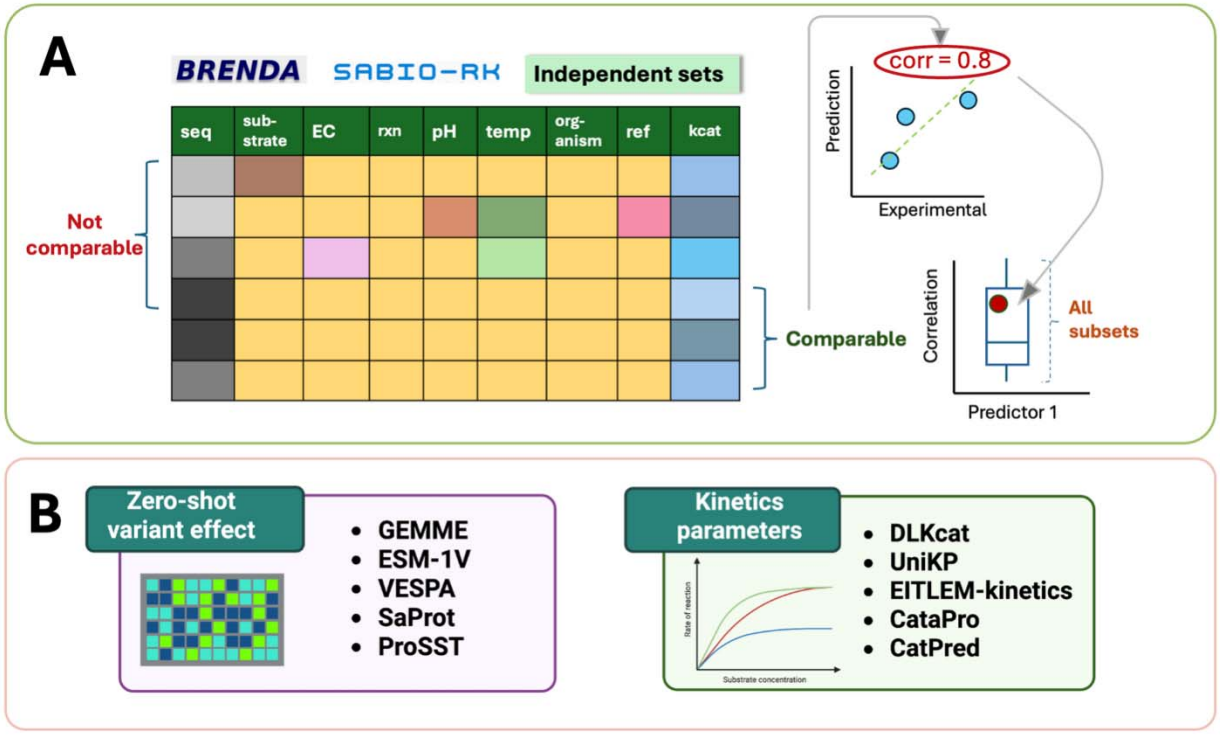

Figure 1. Overview of the methodology in this study. (A) Development of EnzyArena. Data of enzyme KP were obtained from public databases and literature and organized into multiple subsets where all experimental conditions are consistent across samples within subset. Predictions from various computational tools were made for each of the subsets and the Spearman correlations for all the subsets were computed and summarized for each individual predictor. (B) computational models included in this study, including zero-shot fitness prediction and enzyme KP prediction. Additional analyses of docking- and stability-based heuristics are provided in the Supplementary Information.

Results and Discussion

Data curation

A fair evaluation of computational enzyme predictors requires a dataset that minimizes experimental confounders while maintaining broad coverage across enzyme classes. We argue that an ideal benchmarking dataset should meet two key criteria: (1) it should encompass a wide range of enzyme types to reduce functional bias, and (2) experimental conditions—such as pH, temperature, and expression system—should remain consistent within subsets to avoid confounding factors. However, many existing kinetic datasets lack detailed metadata describing assay conditions, which limits reproducibility and cross-study comparability [26]. In particular, the training of KP predictors often violates the second criterion, as they aggregate data from heterogeneous sources during training. Likewise, while repositories such as ProteinGym provide

large-scale fitness datasets with diverse experimental labels, including binding, organismal fitness, expression, stability, and activity, the majority of activity data involve non-enzymatic functions or enzyme reactions with macromolecular substrates, which contradicts with KP predictors in this study, as they support only small-molecule substrates. These limitations motivated our effort to construct condition-controlled datasets suitable for rigorous benchmarking.

Following these principles, we collected 1,571 *kcat* (turnover number), 2,250 *Km* (Michaelis constant), and 1,351 *kcat*/*Km* (catalytic efficiency) entries from the BRENDA and SABIO-RK databases, applying strict inclusion criteria (i.e., mutagenesis studies with ≥10 variants per assay, including one wild-type and at least nine mutants; see Methods). The curated data were partitioned into 130 *kcat* subsets, 174 *Km* subsets, and 110 *kcat*/*Km* subsets, ensuring internal consistency across experimental parameters such as pH, temperature, source organism, substrate, reaction type, and reference. This structured partitioning ensured that data within each subset were comparable and that observed performance differences among predictors could be attributed to algorithmic rather than experimental variation.

In addition to the KP datasets, we manually curated 25 independent datasets related to enzyme catalytic activity that are not included in BRENDA or SABIO-RK. These comprise four deep mutagenesis scanning (DMS) fitness score datasets, nine *kcat* datasets, nine *Km* datasets, and three *kcat*/*Km* datasets (Table 1). These independent datasets were particularly important for unbiased evaluation of kinetic-parameter predictors, which are frequently trained on data derived from BRENDA and SABIO-RK. Together, these curated and independent datasets form the foundation of **EnzyArena**, enabling fair and reproducible assessment of enzyme activity prediction models under well-controlled experimental conditions.

Selection of Representative Predictors for Each Computational Strategy

To provide a fair and comprehensive evaluation of commonly used computational strategies for enzyme engineering, we selected five representative tools for each of the two major paradigms examined in this study: zero-shot fitness prediction and KP prediction (Table 2). Selection of predictors are based on the considerations of methodological diversity, widespread use, recency

of publication, and performance in independent benchmark when available. Detailed descriptions of the individual models can be found in Supplementary Information.

Zero-shot fitness predictors have gained traction due to their accessibility and scalability, with demonstrated utility in identifying beneficial or deleterious mutations across large mutational landscapes [4, 17, 26, 27]. KP predictors aim to estimate *kcat* and/or other related kinetic quantities (e.g., *Km*, *kcat/Km*) from sequence and substrate representations and are increasingly used to triage variants prior to experiments []. Apart from the zero-shot fitness predictors and KP predictors, docking- and stability-based predictions are also used in practice[15, 28, 29] and are thereby included as supplementary analyses to provide a transparent comparison (Supplementary Information).

Table 3 reports the computational time required to run the main-text models on a compute node (NVIDIA A10 GPU with 32 CPUs). All zero-shot predictors completed the full dataset in under 5 hours, and most KP predictors completed within hours, enabling routine use for variant triage.

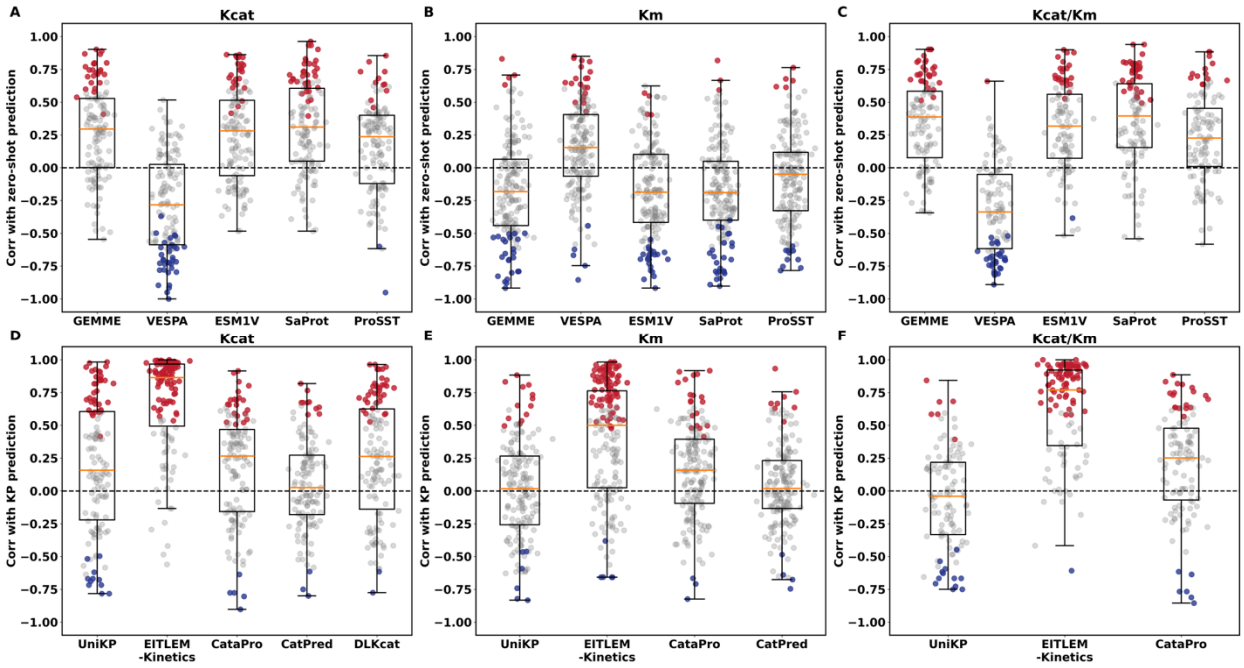

Figure 2. Performance overview of computational predictors on condition-matched subsets derived from BRENDA and SABIO-RK. Performance is measured by Spearman correlation.

In each panel, red, blue, and grey dots indicate significantly (p-value < 0.05) positive, significantly negative, and non-significant Spearman correlations, respectively.

Benchmark performance on BRENDA- and SABIO-RK–derived datasets

We first evaluated zero-shot fitness predictors and supervised kinetic-parameter (KP) predictors on subsets derived from BRENDA and SABIO-RK (Figure 2). On these database-derived datasets, both model families achieve relatively strong performance. KP predictors show particularly high correlations, with median Spearman correlations ranging from -0.04 to 0.864 across kcat, Km, and kcat/Km (Figure 2D–F).

Interestingly, zero-shot fitness predictors display consistent positive correlations with kcat and kcat/Km but not with Km (Figure 2A–C). The biological meaning of protein language model–derived fitness scores remains implicit, as these models are trained on evolutionary sequence statistics rather than kinetic measurements. Prior work has shown that such scores correlate with multiple biophysical properties, including expression, growth, stability, and activity [26]. Our results further suggest that these scores correlate positively with catalytic turnover and catalytic efficiency, while showing a negative association with Km. Km is related to enzyme–substrate affinity and often decreases as binding becomes tighter, although it also depends on catalytic turnover [30]. These trends suggest that higher predicted fitness is generally associated with improved substrate binding and catalytic performance.

One notable exception is VESPA, which exhibits trends opposite to those of other zero-shot predictors. This behavior likely reflects its training objective: VESPA is trained to distinguish effect versus no-effect mutations using ProtT5-derived embeddings, under the assumption that most mutations are deleterious [31]. Consequently, higher VESPA scores indicate a higher likelihood of deleterious effects, which explains its inverse relationship with catalytic performance. This observation highlights the importance of carefully interpreting the directionality of zero-shot fitness scores.

Taken together, these results indicate that the direction of correlation must be considered when applying zero-shot fitness predictors to different experimental readouts. Based on the observed trends across kinetic parameters, we established a directionality adjustment rule for zero-shot

predictors (Table S1). After applying this rule, the absolute values of the median correlations of zero-shot predictors range from 0.05 to 0.394, confirming their strong performance on database-derived datasets.

Among all models, EITLEM-Kinetics shows exceptionally high performance, exceeding both other KP predictors and zero-shot models, with median correlations ranging from 0.5 to 0.864. A likely explanation is that EITLEM-Kinetics is trained primarily on mutational datasets [8], whereas other KP predictors are trained on a mixture of mutant and wild-type ortholog data [9, 10]. This training focus may provide an advantage when evaluating mutational effects within the same enzyme families.

However, it is important to interpret these results cautiously. The database-derived subsets used here substantially overlap with the public resources commonly used to train KP predictors, making it difficult to identify a clean hold-out set. As a result, the performance of KP predictors in Figure 2 is likely inflated by training-test overlap. In contrast, zero-shot fitness predictors are trained solely on sequence data and do not use kinetic measurements, which eliminates this source of bias. This difference in training data introduces an inherent imbalance in retrospective benchmarking. To address this limitation and assess model generalizability, we next evaluated performance on independent mutagenesis datasets.

Generalization on independent mutagenesis datasets

We next evaluated model performance on independent mutagenesis datasets that are not present in BRENDA or SABIO-RK. Per-dataset Spearman correlations are highly variable and show substantial overlap between model families (Figure 3A; median correlations: 0.086-0.2 vs. -0.283-0.086), which is expected given the limited size of most mutagenesis campaigns. Many curated datasets contain only 6–10 variants (Table 1), and such small sample sizes amplify noise and make rank-based metrics such as Spearman correlation unstable [32]. To obtain a more robust summary that accounts for differences in dataset size, we combined correlation estimates using Fisher z meta-analysis [33]. In this framework, correlations are transformed to stabilize variance and then pooled using inverse-variance weighting so that larger and more informative datasets contribute more strongly to the overall estimate (Eqn. 3; Materials and Methods). This approach

is widely used to aggregate correlation coefficients across heterogeneous studies and is particularly appropriate when sample sizes differ substantially between datasets [34].

The pooled analysis reveals a consistent advantage for zero-shot fitness predictors. All zero-shot models achieve higher size-weighted pooled correlations (0.243-0.487) than any supervised kinetic-parameter predictor (-0.241-0.179) (Figure 3B). Together with the results from database-derived subsets, these findings indicate that zero-shot fitness predictors provide more generalizable signals for enzyme mutant prioritization.

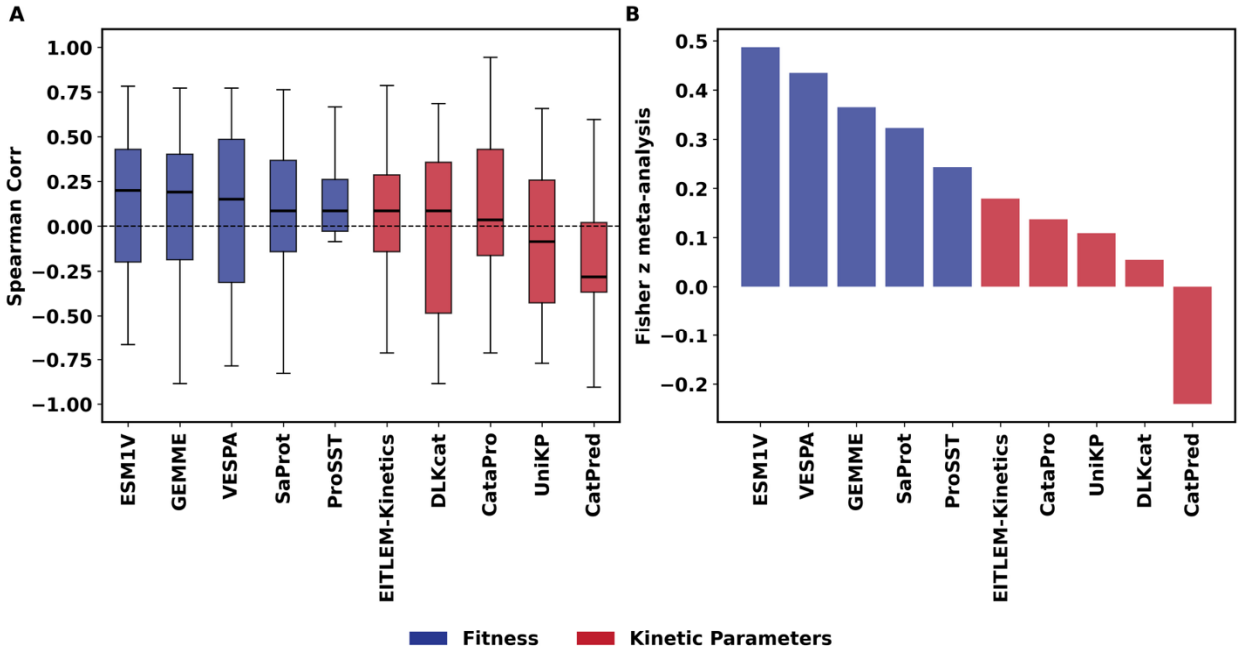

Figure 3. Performance comparison on the independent mutagenesis datasets. (A) Distribution of Spearman correlations between experimental measurements and model predictions across independent enzyme mutagenesis datasets absent from BRENDA or SABIO-RK. Models are ordered by the median correlation. Correlations for zero-shot fitness predictors were adjusted according to the directionality rules described in Table S1. Box plots show medians and interquartile ranges, and individual points represent individual datasets. (B) Fisher z meta-analysis combining correlation coefficients across datasets while accounting for differences in sample size. Zero-shot fitness predictors show higher pooled performance than KP predictors.

Prospective wet-lab validation using patchoulol synthase

To experimentally evaluate model-guided mutation prioritization, we selected patchoulol synthase (PTS) as a test enzyme. PTS catalyzes the cyclization of farnesyl pyrophosphate to

patchoulol, which is the major bioactive component of patchouli oil and has attracted increasing attention due to its reported anticancer, neuroprotective, anti-inflammatory, and gastroprotective activities [35]. Despite its industrial relevance, PTS is a multiproduct terpene synthase that produces patchoulol at relatively low selectivity [36], limiting its biotechnological application and makes it a challenging target for enzyme engineering.

We performed saturated single-site mutagenesis on PTS and used computational models to prioritize variants for experimental testing. Based on the benchmarking results, ESM-1v [4] and UniKP [10] were selected to represent the zero-shot fitness strategy and KP strategy, respectively. Each model prioritized 50 mutants, and an additional 50 mutants were selected randomly as a baseline, resulting in 150 variants in total. All variants were constructed and experimentally evaluated by measuring patchoulol production relative to the wild type (Figure 4A).

The experimental results show a clear separation between strategies. Variants prioritized by UniKP do not outperform the random baseline, whereas ESM-1v–prioritized variants show significantly higher activity than both random and UniKP-selected mutants (Figure 4A).

To further examine model–experiment agreement, we pooled all 150 variants and performed a retrospective correlation analysis between model predictions and experimental measurements. UniKP predictions show a negative correlation with experimental activity ($\rho = -0.195$; Figure 4B), whereas ESM-1v predictions show a positive correlation ($\rho = 0.394$; Figure 4C). Notably, one ESM-1v–prioritized mutant (C23S) achieved an approximately threefold higher activity over the wild type.

Overall, this prospective validation supports the conclusions from the computational benchmark and demonstrates that zero-shot fitness prediction provides a more robust and generalizable strategy for prioritizing enzyme mutations in practice.

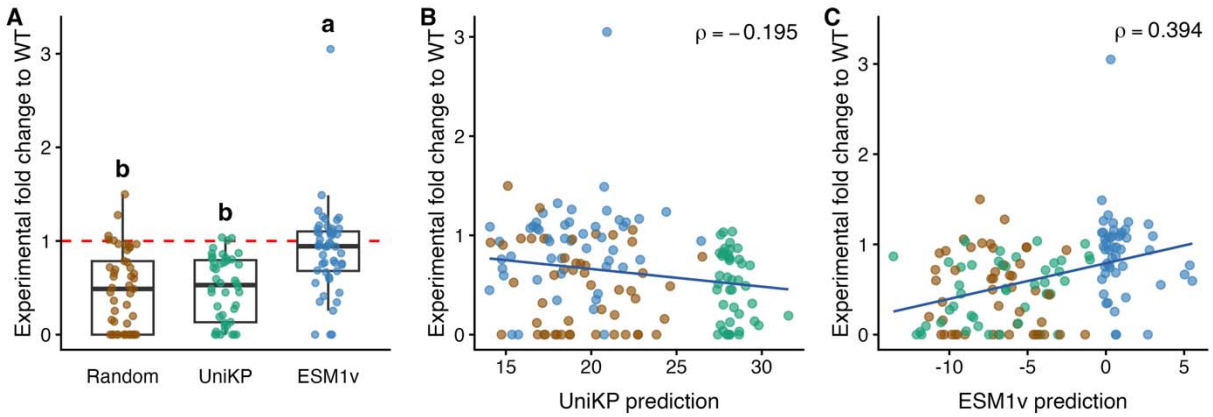

Figure 4. Model performance on a wet-lab validation dataset for model-guided mutagenesis of the PTS enzyme. (A) Distribution of experimentally measured fold change values for mutants prioritized by Random, UniKP, and ESM1v models. The red dashed line marks the wildtype (WT) level (fold change = 1). Letters above the boxes denote statistical groupings from a one-way ANOVA followed by Tukey's post-hoc test; groups that do not share a letter differ significantly ($P < 0.05$). (B) Relationship between UniKP model predictions and experimentally measured fold change relative to wild type (WT). (B) Relationship between ESM1v model predictions and experimental fold change to WT. Each point represents a mutant, colored by model category; the Spearman correlation coefficient (ρ) is indicated. The regression line shows the least-squares fit. Together, these results show that ESM1v predictions exhibit a stronger monotonic relationship with experimental measurements compared with UniKP, and that model-guided designs yield distinct distributions of functional outcomes relative to random variants.

Inter-Model Correlation Analysis Reveals Strategy-Specific Clustering

To better understand relationships among predictors, we analyzed pairwise correlations across all model outputs (Figure 5). Models within the same computational paradigm formed clear clusters, with intra-group Pearson correlations ranging from -0.812 to 0.76, while inter-group correlations were generally weak (-0.256 to 0.329).

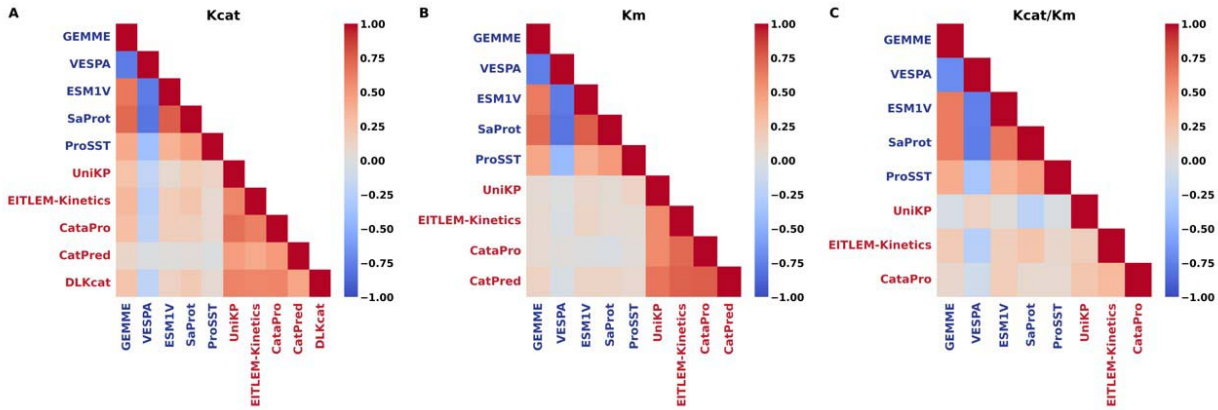

Figure 5. Heatmap of pairwise correlation among the predictors. The heatmaps display Pearson correlation coefficients between model predictions, with values ranging from +1 (red, perfect positive correlation) to –1 (blue, strong negative correlation). Models are grouped and color-coded based on computational strategy: purple (stability predictors), red (zero-shot sequence-based models), orange (kinetic parameter–specific models), and cyan (binding affinity models).

Combination of Zero-Shot Predictors Improves the Precision of Identifying Superior Enzyme Mutants

Given the strong individual performance of zero-shot fitness predictors demonstrated above, we next investigated whether combining multiple predictors could further improve the success rate of identifying mutants with enhanced catalytic activity. Our analysis focused on precision (Eqn. 1), defined as the proportion of predicted beneficial mutants that are experimentally confirmed to outperform the reference. Because "wildtype" is context-dependent, we generated an exhaustive pairwise comparison dataset in which every mutant in turn served as the reference, and all other mutants were evaluated relative to it. This strategy greatly expanded the evaluation space and enabled a robust assessment of prediction performance across diverse mutational landscapes.

For each pairwise comparison, we evaluated whether a given predictor or combination of predictors correctly identified the mutant with superior experimental activity. To assess combinations, we implemented a consensus strategy, where a mutant was predicted to be superior only if all zero-shot models in the combination ranked the mutant higher than the reference. All 31 possible combinations of the five predictors (ranging from single models to combinations of all five) were evaluated (Figure 6A; Figure S1). Importantly, predictor outputs were directionally standardized so that higher scores consistently indicated higher predicted fitness.

Across datasets containing at least 20 mutants, the combination of GEMME, ProSST, SaProt, and VESPA achieved a significantly higher average precision (0.74) compared to the best-performing single predictor, SaProt (0.65) (Wilcoxon rank-sum test, $p = 0.044$; Figure 6B). In contrast, for datasets with fewer than 20 mutants, all predictors, whether alone or in combination, showed comparable performance (Figure S3). These results suggest that consensus modeling using multiple zero-shot predictors offers a practical strategy to substantially increase the success

rate of identifying beneficial mutations, particularly when the mutational search space is moderately large (n ≥ 20).

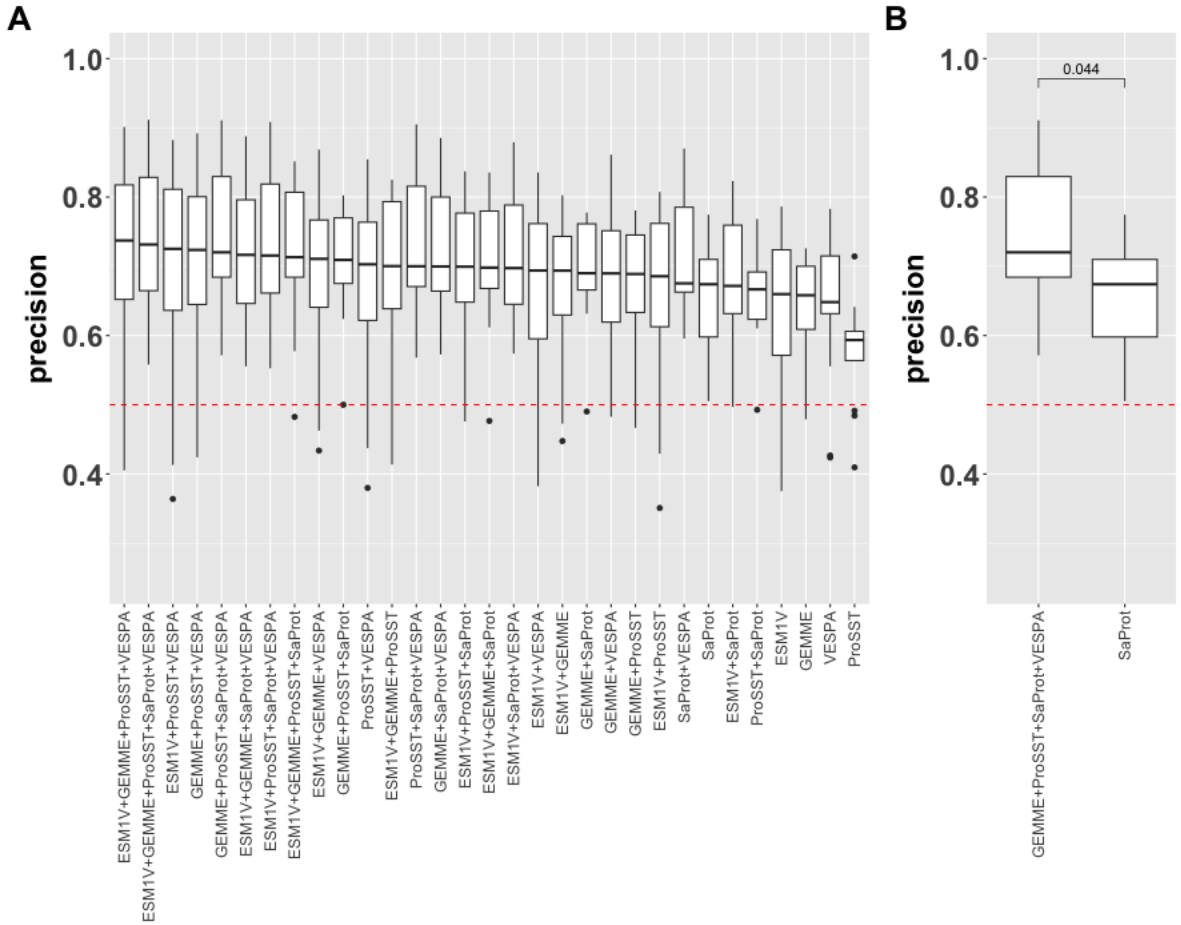

Figure 6. Combination of zero-shot predictors improves precision in identifying beneficial enzyme mutants. (A) Precision of all 31 combinations of five zero-shot predictors, evaluated using exhaustive pairwise comparisons across datasets with at least 20 mutants. Each box plot represents the distribution of precision values across datasets, with the dashed red line indicating random baseline performance. (B) Comparison of the best-performing single predictor (SaProt) and the top-performing combination (GEMME + ProSST + SaProt + VESPA). The consensus predictor achieves a significantly higher precision (mean 0.74) than SaProt alone (mean 0.65), demonstrating the benefit of integrating multiple predictive models (p = 0.044, Wilcoxon rank-sum test).

Other considerations and concluding remarks

This study highlights several inherent pitfalls of KP predictors that must be considered when applying them to enzyme-directed evolution. First, as previously mentioned, KP data from public

databases are often pooled together without aligning the experimental conditions, which can obscure the primary goal of enzyme-directed evolution: identifying sequence-level differences. In this study, we addressed this issue by sorting noisy data into subsets where experimental conditions were consistent across samples, a critical step for accurately evaluating the performance of computational tools.

Second, KP predictors typically use standard loss functions (such as root mean squared error or mean squared error) during training. However, in practice, the relative ranking of mutants is often of greater interest rather than their absolute KP values. Researchers typically aim to predict the relative ranking among mutants to identify variants that outperform the wild type. To address this, we focused on Spearman correlation as a metric to evaluate the ranking performance of computational tools in guiding enzyme-directed evolution.

Additionally, published studies on computationally guided enzyme evolution may be biased toward successful cases, while failures are often underreported. This selective reporting can mislead future researchers, who may overestimate the utility of these tools and continue to invest in computational methods with limited effectiveness. Our study aims to provide a more balanced perspective, revealing both the strengths and limitations of these computational tools.

It is also important to note that, despite our efforts to ensure maximal consistency within subsets, some inconsistencies may still exist in the curated data, such as variations in codon optimization, culture conditions, and experimental protocols. These factors are difficult to resolve without more comprehensive reporting in publications and the development of advanced text processing tools. Another key limitation of our study is that it did not address epistatic interactions, where multiple mutations co-occur and affect each other's effects. Predicting such epistatic interactions remains a significant challenge, as many predictors (Table 2) are not compatible with these scenarios.

In summary, we evaluated the performance of various commonly used and emerging computational tools for guiding enzyme evolution toward improved catalytic activity. The datasets in this study were carefully curated to ensure consistency, which is a crucial prerequisite

for meaningful comparison. Our results highlight the promising potential of large language model-based zero-shot fitness predictors and suggest that mutagenesis data should be sufficiently large to fully leverage the capabilities of computational tools.

Materials and Methods

Data Collection from Public Databases

Data were primarily obtained from the BRENDA database (https://www.brenda-enzymes.org/index.php) and the SABIO-RK database (https://sabio.h-its.org/newSearch/index). From BRENDA, entries were collected based on the "Classic view" categories of "Turnover Number," "*Km* Value," and "*kcat*/*Km* Value." For SABIO-RK, data were retrieved by setting the search condition Parameter type to ("*kcat*" OR "*Km*" OR "*kcat*/*Km*").

Curation of Independent Datasets

We manually curated 25 additional mutagenesis datasets. The inclusion criteria were:

a) identifiable sequences;
b) small molecule substrate (to ensure compatibility with KP predictors);
c) at least six mutant sequences (including wild type);
d) for *kcat*, *Km*, or *kcat*/*Km* datasts, publication date must be within the latest year (2025) to avoid possible overlap with KP predictor's training data.

As a result, we managed to curate datasets with a variety of experimental measurements, including relative activity, *kcat*, *Km*, *kcat*/*Km*, and deep mutational scanning (DMS) scores. For the DMS data, the reported fitness scores were averaged across three replicates. Since our downstream evaluation is based on Spearman correlation, which depends on the relative ordering of data points, we first ranked all DMS measurements from low to high, then selected 100 data points at fixed intervals. This strategy reduces the influence of experimental noise and increases the likelihood that adjacent points reflect true biological differences rather than measurement variability, thereby producing a more robust correlation assessment.

Data Preprocessing

To ensure that the curated datasets were accurate and comparable, strict filtering criteria were applied to the raw entries from BRENDA and SABIO-RK. Entries lacking a UniProt ID, containing multiple UniProt IDs, or missing information on substrate, temperature, pH, or literature reference were removed. Because standardized mutation annotations were not provided in these databases, mutation data were extracted from textual annotations.

We focused on single-point mutations and excluded entries containing the keywords 'insert', 'insertion', 'inserted', 'delete', 'deletion', 'deleted', 'double', 'triple', 'multiple', '/', 'hybrid', 'mixture', 'delta', 'truncate', 'truncated', or 'fragment' (case-insensitive). For SABIO-RK, enzyme types were annotated based on the “Enzyme Variant” field: entries containing “wildtype” were labeled as wildtype enzymes, and those containing “mutate,” “mutated,” “mutation,” “mutant,” “mutagenesis,” or “variant” were labeled as mutated enzymes. For BRENDA, due to the absence of explicit annotations, entries without mutation keywords were considered wildtype.

For entries annotated as mutated enzymes, mutation annotations were retained only if they followed the format of “single amino acid + position number + single amino acid” and involved exactly one mutation. Small datasets (referred to as “subsets”) were constructed as follows: for BRENDA, entries were grouped by EC number, enzyme name, UniProt ID, substrate, organism, reference, temperature, and pH. For SABIO-RK, explicit reaction information was also included in grouping.Only subsets containing one wildtype enzyme and at least nine single-point mutations were retained.

All subsets were then manually validated. Using the UniProt ID, wildtype protein sequences were retrieved from the UniProt database (https://www.uniprot.org/). For each mutation, the amino acid at the specified position was verified against the wildtype sequence. Mutated sequences were generated accordingly. Subsets involving enzymes longer than 1000 amino acids were excluded due to downstream model constraints. For substrate preprocessing, Substrate SMILES representations were obtained as following. InChI identifiers were retrieved under the “Ligands” section and converted to SMILES using RDKit for the BRENDA entries. For SABIO-RK entries, SMILES were obtained from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/).

In total, 441 validated subsets were obtained, including 293 from BRENDA, 121 from SABIO-RK, and 25 from the literature.

Model Deployment and Execution

Data were analyzed using two categories of models: fitness prediction and enzyme kinetic parameter prediction. We also added stability prediction and docking affinity prediction for reference (Supplementary Information). In total, 20 computational models were employed (5 models per category).

Fitness Prediction Models. Mutation effects on protein fitness were evaluated using five models based on sequence and structure representations. GEMME derives fitness scores from multiple sequence alignments produced via MMseqs2. VESPA predicts saturation mutagenesis fitness directly from the wildtype sequence. ESM1V estimates site-specific mutation fitness using transformer-based sequence modeling. SaProt incorporates predicted protein structure and mutation site information. ProSST integrates wildtype sequence, structure, and mutation positions to determine mutation-induced functional changes. These models capture evolutionary constraints and mutation impact on enzyme functionality.

Enzyme Kinetic Parameter Prediction Models. Enzymatic parameters were predicted from protein sequences and substrate SMILES using five models. UniKP, EITLEM-Kinetics, and CataPro each predict *kcat*, *Km*, and catalytic efficiency (*kcat*/*Km*). CatPred outputs only *kcat* and *Km*, while DLKcat exclusively predicts *kcat*. All models operate without structural preprocessing, enabling efficient evaluation of catalytic performance from sequence-level inputs.

Adjustment of model prediction directionality

Because different predictors are expected to correlate with experimental measurements in different directions, it is essential to standardize the directionality of the correlations so that higher (positive) correlation values consistently indicate better predictive performance. To enable consistent comparison across models, we therefore established the following working

assumptions: a) fitness is assumed to positively correlate with *kcat* and *kcat/Km*, but negatively with *Km*, except in cases where model outputs follow the opposite trend (e.g., VESPA).

Using these assumptions, we adjusted the sign of each zero-shot fitness prediction–measurement correlation for the so that, after transformation, higher correlation values uniformly represent better performance. The adjusted results are shown in Figure 4.

Metric calculation

Spearman correlation, precision, and size-weighted pooled correlation are computed as in Eqn. 1, Eqn. 2, and Eqn. 3, respectively.

$$corr = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \text{(Eqn. 1)}$$

Where d_i= difference between the ranks of each pair of observations; n=number of pairs.

$$precision = \frac{TP}{TP+FP} \qquad \text{(Eqn. 2)}$$

Where TP = True Positives (correctly predicted positive cases); FP = False Positives (incorrectly predicted positive cases).

$$size - weighted\ pooled\ correlation = \tanh\left(\frac{\sum(n_i-3)\ tanh^{-1}(\rho_i)}{\sum(n_i-3)}\right) \qquad \text{(Eqn. 3)}$$

Where n_i and ρ_i the number of mutants and correlation for the i-*th* dataset, respectively.

Wet-lab validation for the PTS mutants

Strains and Medium. *Escherichia coli* DH5α was used for plasmid construction and cultured in LB medium supplemented with 100 mg/L ampicillin at 37 °C. The *Saccharomyces cerevisiae* strain yCAN10, previously engineered to enhance the supply of farnesyl pyrophosphate, was used for PTS expression [37]. Synthetic complete uracil dropout medium (SC-Ura; Coolaber, Beijing, China) was used for yeast selection and cultivation. Yeast extract peptone dextrose (YPD) medium was used for routine yeast culture, and yeast extract peptone galactose (YPG), in which galactose replaced dextrose, was used for fermentation.

Mutant recommendation. Saturation mutagenesis of the PTS protein sequence was performed, which were scored and ranked by UniKP (*kcat/Km model*) and ESM-1v—we selected the top 50 mutans recommended by each of these two models. In addition, 50 mutants were randomly selected as the control group. The mutant information (mutation, sequence, and model prediction can be found in Table S4) Primers for these mutants were designed by the Primer3 program [38].

Plasmids and Strains Construction. The patchoulol synthase gene (GenBank ID: AAS86323.1) was synthesized by GenScript (Nanjing, China) with codon optimization for yeast expression. The gene was inserted into plasmid pESC-Ura under the control of the GAL1 promoter and CYC1 terminator using Gibson assembly to generate pESC-Ura-PTS (pGAL1-PTS-tCYC1 expression cassette). Using this plasmid as a template, mutated PTS fragments were amplified by PCR with primers designed to introduce the desired mutations (Table S4). The mutated PTS genes were integrated into the ARS1114a locus of the host strain using CRISPR/Cas9 genome editing [39]. Briefly, PTS mutational fragments together with ~1000 bp upstream and downstream homologous arms (amplified from yCAN10 genomic DNA) and the Cas9/sgRNA plasmid p426_Cas9_gRNA-ARS1114a (Addgene #87397) were transformed into the host strain using the lithium acetate method. Transformants were selected on SD-Ura plates and verified by colony PCR and sequencing. All primer sequences used in this study are listed in Table S5.

Fermentation of PTS Mutant Strains. Single colonies of mutant strains were inoculated into 96-deep-well plates containing 500 µL YPD per well and cultured at 30 °C and 800 rpm for 18 h to obtain primary seed cultures. Cultures were then transferred to 24-deep-well plates containing 2 mL YPG per well at a 2% inoculum and incubated under the same conditions. All strains were cultured in three biological replicates. After 72 h of fermentation, 500 µL of broth was collected for patchoulol quantification.

Extraction and detection methods. Patchoulol was extracted with 200 µL *n*-hexane using bead beating (70 Hz; 90 s on, 30 s off; 10 cycles), followed by centrifugation at 12,000 ×g for 5 min. The upper organic phase was collected for analysis. Patchoulol was quantified by GC-MS (8890-7010B, Agilent, USA) equipped with an HP-5 column (30 m × 250 µm × 0.25 µm, Agilent). The

oven temperature program was: 150 °C for 1 min, ramp to 180 °C at 20 °C/min, then to 300 °C at 80 °C/min and held for 4 min. A 1 µL sample was injected with helium as the carrier gas. Selected ion monitoring (SIM) was used for quantification with *m/z* 137.9, 160.9, and 222. The mutants' patchoulol yield relative to wildtype can be found in Table S4.

Data and code availability

Data in this study, including curated subsets, experimental conditions, experimental measurement, and pre-computed predictions, are deposited in our GitHub repository. (https://github.com/zishuozeng/EnzyArena).

Ethical statement

Not applicable.

Acknowledgements

We would like to thank all researchers who made the data and tools relevant to this study available.

Authorship contribution

Z.Z. and X.L. conceived the ideas, analyzed the data, and wrote the manuscript; J.J. collected the data and ran the predictions; R.X. performed the wet-lab experiments; X.L. revised the manuscript.

Funding

This project is supported by National Key R&D Program of China (2025YFA0922700), Shenzhen Medical Research Fund (D2301005), Guangdong S&T Program (2024B1111160007), National Natural Science Foundation of China (32570096), Shenzhen Science and Technology Program (KJZD20240903102304006 and ZDSYS20210623091810032), SlAT Distinguished Young Scholars (E4G021).

Declaration of competing interest

X.L. has a financial interest in Demetrix and Synceres.

Table 1. Summary of the manually curated datasets.

Dataset No.	Ref.	Enzyme	Substrate	Measurement	No. mutants
Dataset 1	[29]	D-amino acid oxidase	D-alanine	DMS score	100 out of 6396
Dataset 2	[40]	Aliphatic amide hydrolase	Acetamide	DMS score	100 out of 7161
Dataset 3	[40]	Aliphatic amide hydrolase	Isobutyramide	DMS score	100 out of 7161
Dataset 4	[40]	Aliphatic amide hydrolase	Propionamide	DMS score	100 out of 7161
Dataset 5	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	10
Dataset 6	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	8
Dataset 7	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	8
Dataset 8	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	7
Dataset 9	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	6
Dataset 10	[41]	Adenylate kinase	Adenosine diphosphate	*kcat*	6

Dataset 11	[42]	Polyphenol oxidases	Catechol	*kcat*	6
Dataset 12	[42]	Polyphenol oxidases	Dopamine	*kcat*	6
Dataset 13	[42]	Polyphenol oxidases	Tyramine	*kcat*	6
Dataset 14	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	10
Dataset 15	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	8
Dataset 16	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	8
Dataset 17	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	7
Dataset 18	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	6
Dataset 19	[41]	Adenylate kinase	Adenosine diphosphate	*Km*	6
Dataset 20	[42]	Polyphenol oxidases	Catechol	*Km*	6
Dataset 21	[42]	Polyphenol oxidases	Dopamine	*Km*	6
Dataset 22	[42]	Polyphenol oxidases	Tyramine	*Km*	6
Dataset 23	[42]	Polyphenol oxidases	Catechol	*kcat*/*Km*	6
Dataset 24	[42]	Polyphenol oxidases	Dopamine	*kcat*/*Km*	6
Dataset 25	[42]	Polyphenol oxidases	Tyramine	*kcat*/*Km*	6

Table 2. Summary of the selected predictors in this study.

Group	Name	Input format	Output format	Support multiple mutations	Key characteristics
Zero-shot fitness prediction	GEMME [27]	A wildtype seq.	Scores of all saturated mutants	No	Computational model that considers conservation and epistatic effects
	VESPA [43]	A wildtype seq.	Scores of all saturated mutants	No	Transfer learning based on ProtT5, predicts effect vs. neutral
	ESM-1v [4]	A wildtype seq.	Scores of input mutants	No	LLM that quantifies benefits/deleteriousness of mutations
	SaProt [6]	Struct. of wildtype, mutation descriptor	Scores of input mutants	Yes	LLM with both sequence and structure information
	ProSST [7]	Struct. of wildtype, wildtype seq,mutation descriptor	Scores of input mutants	Yes	LLM with both sequence and structure information
Kinetic parameter (KP)	DLkcat[9]	Seq. of mutants, substrate	Scores of input mutants	Yes	N-gram (protein features), graph neural network (substrate features), deep learning (algorithm); predicts *kcat*

prediction		SMILES			
	UniKP [10]	Seq. of mutants, substrate	Scores of input mutants	Yes	ProtT5 (protein features), SMILES transformers (substrate features), ExtraTree (algorithm); predicts *kcat*, *Km*, and *kcat*/*Km*
	EITLEM-kinetics [8]	Seq. of mutants, substrate	Scores of input mutants	Yes	ESM-1v (protein features), fingerprints (substrate features), neural network (algorithm) ; predicts *kcat*, *Km*, and *kcat*/*Km*
	CataPro [12]	Seq. of mutants, substrate SMILES	Scores of input mutants	Yes	ProtT5 (protein features), MolT5 + MACCS fingerprints (substrate features); predicts *kcat*, *Km*, and *kcat*/*Km*
	CatPred [13]	Seq. of mutants, substrate SMILES	Scores of input mutants	Yes	ESM2 (protein features), D-MPNN (substrate), deep learning (algorithm); predicts *kcat*, *Km*, and Ki

Table 3. Model time consumption to run all data in this study.

Category	Model	Data processed	Model run time (h)	Run time for pdb/pdbqt preparation (h)	Total run time (h)
Kinetic Parameters	DLKcat	2886 protein sequence - substrate pairs	0.08	-	0.08
	UniKP		0.5	-	0.5
	CataPro		0.5	-	0.5
	CatPred		0.5	-	0.5
	EITLEM-kinetics		8	-	8
Fitness	SaProt	126 wildtype protein pdb files	0.33	1	1.33
	ESM1V		2	-	2
	ProSST		1	1	2
	GEMME		2.5	-	2.5
	VESPA		4.33	-	4.33

References

(1) Radley, E.; Davidson, J.; Foster, J.; Obexer, R.; Bell, E. L.; Green, A. P. Engineering enzymes for environmental sustainability. *Angewandte Chemie International Edition* **2023**, *62* (52), e202309305.
(2) Katsimpouras, C.; Stephanopoulos, G. Enzymes in biotechnology: Critical platform technologies for bioprocess development. *Current opinion in biotechnology* **2021**, *69*, 91-102.
(3) Markus, B.; C, G. C.; Andreas, K.; Arkadij, K.; Stefan, L.; Gustav, O.; Elina, S.; Radka, S. Accelerating biocatalysis discovery with machine learning: a paradigm shift in enzyme engineering, discovery, and design. *ACS catalysis* **2023**, *13* (21), 14454-14469.
(4) 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. *Advances in Neural Information Processing Systems* **2021**, *34*, 29287-29303.
(5) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* **2020**.
(6) Su, J.; Han, C.; Zhou, Y.; Shan, J.; Zhou, X.; Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv* **2023**, 2023.2010. 2001.560349.
(7) Li, M.; Tan, Y.; Ma, X.; Zhong, B.; Yu, H.; Zhou, Z.; Ouyang, W.; Zhou, B.; Tan, P.; Hong, L. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems* **2024**, *37*, 35700-35726.
(8) Shen, X.; Cui, Z.; Long, J.; Zhang, S.; Chen, B.; Tan, T. EITLEM-Kinetics: A deep-learning framework for kinetic parameter prediction of mutant enzymes. *Chem Catalysis* **2024**, *4* (9).
(9) Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M. K.; Kerkhoven, E. J.; Nielsen, J. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* **2022**, *5* (8), 662-672.
(10) Yu, H.; Deng, H.; He, J.; Keasling, J. D.; Luo, X. UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nature communications* **2023**, *14* (1), 8211.
(11) Wang, J.; Yang, Z.; Chen, C.; Yao, G.; Wan, X.; Bao, S.; Ding, J.; Wang, L.; Jiang, H. MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. *Briefings in Bioinformatics* **2024**, *25* (5), bbae387.
(12) Wang, Z.; Xie, D.; Wu, D.; Luo, X.; Wang, S.; Li, Y.; Yang, Y.; Li, W.; Zheng, L. Robust enzyme discovery and engineering with deep learning using CataPro. *Nature communications* **2025**, *16* (1), 2736.
(13) Boorla, V. S.; Maranas, C. D. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature communications* **2025**, *16* (1), 2072.
(14) Cui, Y.; Sun, J.; Wu, B. Computational enzyme redesign: large jumps in function. *Trends in Chemistry* **2022**, *4* (5), 409-419. Wang, H.-Y.; Qian, Z.-H.; Zhang, C.; Li, J.-C.; Xu, J.-Z. Design Glycosyltransferases with High Glycosyl Transfer Efficiency to Efficiently Produce Isoquercetin from Quercetin. *Journal of Agricultural and Food Chemistry* **2026**. Ding, H.; Jiang, L.; Song, Y.; Pu, Z.; Yang, L.; Yu, H. Protein Language Model□Guided Engineering of a 2, 3□Butanediol

Dehydrogenase for the Enantioselective Synthesis of Cyclic α□Hydroxy Ketones. *Advanced Science* **2026**, e09314. Zhang, Y.; Wang, X.; Yin, X.; Huang, T.; Lin, S. Engineered ω-Transaminase Enables Remote Stereocontrol in the Synthesis of Chiral Pharmaceutical N-Heterocyclic Amines with Nonadjacent Stereocenters. *ACS Catalysis* **2025**, *16* (1), 587-598. Li, T.; Wang, M.; Ren, H.; Yu, H.; Luo, X.; Lv, L.; Jin, H.; Yin, X.; Chen, R. Methylation of isopentenyl pyrophosphate by C□methyltransferase and optimizing irregular β□elemene yield using UniKP, an enzyme kinetic parameter prediction tool. *Journal of Integrative Plant Biology* **2025**. Yu, J.; Cheng, K.; Qu, S.; Wang, J.; Wang, X.; Zhao, C.; Li, W. Genome□scale Metabolic Modeling Guided Escherichia coli Engineering for De Novo Biosynthesis of Chrysanthemic Acid. *Advanced Science* **2026**, *13* (2), e12736. Mutailifu, M.; Lin, N.-Q.; Liu, F.; Liu, J.-Z. Enhancing Lactulose Production by An Engineered Escherichia coli via Enzymatic Modification and Compartmentalization. *Journal of Agricultural and Food Chemistry* **2025**, *73* (30), 18876-18888.

(15) Jiang, F.; Li, M.; Dong, J.; Yu, Y.; Sun, X.; Wu, B.; Huang, J.; Kang, L.; Pei, Y.; Zhang, L. A general temperature-guided language model to design proteins of enhanced stability and activity. *Science Advances* **2024**, *10* (48), eadr2641.

(16) Miao, L.; Zheng, Y.; Cheng, R.; Liu, J.; Zheng, Z.; Yang, H.; Zhao, J. Efficient Synthesis of γ-Aminobutyric Acid from Monosodium Glutamate Using an Engineered Glutamate Decarboxylase Active at a Neutral pH. *Catalysts* **2024**, *14* (12), 905.

(17) Zhang, Q.; Chen, W.; Qin, M.; Wang, Y.; Pu, Z.; Ding, K.; Liu, Y.; Zhang, Q.; Li, D.; Li, X. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications* **2025**, *16* (1), 1553.

(18) Vanella, R.; Kovacevic, G.; Doffini, V.; de Santaella, J. F.; Nash, M. A. High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. *Chemical Communications* **2022**, *58* (15), 2455-2467. Xiao, H.; Bao, Z.; Zhao, H. High throughput screening and selection methods for directed enzyme evolution. *Industrial & engineering chemistry research* **2015**, *54* (16), 4011-4020. Wittmann, B. J.; Johnston, K. E.; Wu, Z.; Arnold, F. H. Advances in machine learning for directed evolution. *Current opinion in structural biology* **2021**, *69*, 11-18.

(19) Yang, J.; Li, F.-Z.; Long, Y.; Arnold, F. H. Illuminating the universe of enzyme catalysis in the era of artificial intelligence. *Cell Systems* **2025**. Liu, C.; Wu, J.; Chen, Y.; Liu, Y.; Zheng, Y.; Liu, L.; Zhao, J. Advances in Zero□Shot Prediction□Guided Enzyme Engineering Using Machine Learning. *ChemCatChem* **2025**, *17* (5), e202401542. Qiu, S.; Saeed, H.; Leonard, W.; Li, F.; Yang, A. Machine learning for enzyme catalytic activity: current progress and future horizons. *Briefings in Bioinformatics* **2026**, *27* (1), bbag002.

(20) Mangul, S.; Martin, L. S.; Hill, B. L.; Lam, A. K.-M.; Distler, M. G.; Zelikovsky, A.; Eskin, E.; Flint, J. Systematic benchmarking of omics computational tools. *Nature communications* **2019**, *10* (1), 1393. Camacho, D. M.; Collins, K. M.; Powers, R. K.; Costello, J. C.; Collins, J. J. Next-generation machine learning for biological networks. *Cell* **2018**, *173* (7), 1581-1592. Pleiss, J. r. Modeling Enzyme Kinetics: Current Challenges and Future Perspectives for Biocatalysis. *Biochemistry* **2024**, *63* (20), 2533-2541.

(21) Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic acids research* **2021**, *49* (D1), D498-D508.

(22) Wittig, U.; Rey, M.; Weidemann, A.; Kania, R.; Müller, W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic acids research* **2018**, *46* (D1), D656-D660.
(23) Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4* (9).
(24) Bernett, J.; Blumenthal, D. B.; Grimm, D. G.; Haselbeck, F.; Joeres, R.; Kalinina, O. V.; List, M. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods* **2024**, *21* (8), 1444-1453.
(25) Graber, D.; Stockinger, P.; Meyer, F.; Mishra, S.; Horn, C.; Buller, R. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence* **2025**, 1-13.
(26) Notin, P.; Kollasch, A.; Ritter, D.; Van Niekerk, L.; Paul, S.; Spinner, H.; Rollins, N.; Shaw, A.; Orenbuch, R.; Weitzman, R. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems* **2023**, *36*, 64331-64379.
(27) Laine, E.; Karami, Y.; Carbone, A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution* **2019**, *36* (11), 2604-2619.
(28) Siddiqui, K. S. Defying the activity–stability trade-off in enzymes: taking advantage of entropy to enhance activity and thermostability. *Critical reviews in biotechnology* **2017**, *37* (3), 309-322. Zheng, N.; Cai, Y.; Zhang, Z.; Zhou, H.; Deng, Y.; Du, S.; Tu, M.; Fang, W.; Xia, X. Tailoring industrial enzymes for thermostability and activity evolution by the machine learning-based iCASE strategy. *Nature Communications* **2025**, *16* (1), 604. Teufl, M.; Zajc, C. U.; Traxlmayr, M. W. Engineering strategies to overcome the stability–function trade-off in proteins. *ACS Synthetic Biology* **2022**, *11* (3), 1030-1039. Yu, Y.; Rué Casamajo, A.; Finnigan, W.; Schnepel, C.; Barker, R.; Morrill, C.; Heath, R. S.; De Maria, L.; Turner, N. J.; Scrutton, N. S. Structure-based design of small imine reductase panels for target substrates. *ACS catalysis* **2023**, *13* (18), 12310-12321. Qian, S.; Clomburg, J. M.; Gonzalez, R. Engineering Escherichia coli as a platform for the in vivo synthesis of prenylated aromatics. *Biotechnology and Bioengineering* **2019**, *116* (5), 1116-1127. Mai, P.; Zocher, G.; Stehle, T.; Li, S.-M. Structure-based protein engineering enables prenyl donor switching of a fungal aromatic prenyltransferase. *Organic & Biomolecular Chemistry* **2018**, *16* (40), 7461-7469. Hua, L.; Qianqian, B.; Jianfeng, Z.; Yinbiao, X.; Shengyu, Y.; Weishi, X.; Yang, S.; Yupeng, L. Directed evolution engineering to improve activity of glucose dehydrogenase by increasing pocket hydrophobicity. *Frontiers in Microbiology* **2022**, *13*, 1044226.
(29) Vanella, R.; Küng, C.; Schoepfer, A. A.; Doffini, V.; Ren, J.; Nash, M. A. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. *Nature Communications* **2024**, *15* (1), 1807.
(30) Cornish-Bowden, A. *Fundamentals of enzyme kinetics*; John Wiley & Sons, 2013.
(31) Zeng, Z.; Aptekmann, A. A.; Bromberg, Y. Decoding the effects of synonymous variants. *Nucleic acids research* **2021**, *49* (22), 12673-12691. Bromberg, Y.; Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research* **2007**, *35* (11), 3823-3835.
(32) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Springer, 2013.
(33) Fisher, R. A. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1921**, *1*, 3-32.

(34) Borenstein, M.; Hedges, L. V.; Higgins, J. P.; Rothstein, H. R. *Introduction to meta-analysis*; John wiley & sons, 2021.
(35) Lee, H.-S.; Lee, J.; Smolensky, D.; Lee, S.-H. Potential benefits of patchouli alcohol in prevention of human diseases: A mechanistic review. *International immunopharmacology* **2020**, *89*, 107056.
(36) Ekramzadeh, K.; Brämer, C.; Frister, T.; Fohrer, J.; Kirschning, A.; Scheper, T.; Beutel, S. Optimization of factors influencing enzyme activity and product selectivity and the role of proton transfer in the catalytic mechanism of patchoulol synthase. *Biotechnology Progress* **2020**, *36* (2), e2935.
(37) Luo, X.; Reiter, M. A.; d'Espaux, L.; Wong, J.; Denby, C. M.; Lechner, A.; Zhang, Y.; Grzybowski, A. T.; Harth, S.; Lin, W. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **2019**, *567* (7746), 123-126.
(38) Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B. C.; Remm, M.; Rozen, S. G. Primer3—new capabilities and interfaces. *Nucleic acids research* **2012**, *40* (15), e115-e115.
(39) Reider Apel, A.; d'Espaux, L.; Wehrs, M.; Sachs, D.; Li, R. A.; Tong, G. J.; Garber, M.; Nnadi, O.; Zhuang, W.; Hillson, N. J. A Cas9-based toolkit to program gene expression in Saccharomyces cerevisiae. *Nucleic acids research* **2017**, *45* (1), 496-508.
(40) Wrenbeck, E. E.; Azouz, L. R.; Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature communications* **2017**, *8* (1), 15695.
(41) Muir, D. F.; Asper, G. P.; Notin, P.; Posner, J. A.; Marks, D. S.; Keiser, M. J.; Pinney, M. M. Evolutionary-scale enzymology enables exploration of a rugged catalytic landscape. *Science* **2025**, *388* (6752), eadu1058.
(42) Luo, S.; Hou, Y.; Hu, S.-Q. Enhancement of preference, catalytic activity and thermostability of polyphenol oxidase from Rosa Chinensis by semi-rational engineering. *Molecular Catalysis* **2024**, *559*, 114059.
(43) Marquet, C.; Heinzinger, M.; Olenyi, T.; Dallago, C.; Erckert, K.; Bernhofer, M.; Nechaev, D.; Rost, B. Embeddings from protein language models predict conservation and variant effects. *Human genetics* **2022**, *141* (10), 1629-1647.