Review

Machine Learning methods for Small Wet-Lab Data Challenges in Enzyme Engineering

Yumeng Zhang [1,*]

[1] Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang, Liaoning, China.
[*] Correspondence: GeneralSequence@outlook.com

Abstract

Enzyme engineering is fundamentally constrained by the scarcity of high-quality sequencefunction data, and the vastness of protein sequence space is severely mismatched with the limited throughput of experimental characterization. Traditional machine learning models, which rely heavily on massive labeled datasets, further exacerbate this dilemma. This review surveys machine learning methods specifically designed to address the small-data challenge in enzyme engineering, providing an in-depth analysis of each approach and a detailed discussion of recent progress. Furthermore, we consolidate existing achievements in this field and evaluate the effectiveness of these methods in tackling small-data problems from the perspective of specific enzymatic properties, while also summarizing the remaining challenges. This review serves as a practical methodology reference for enzyme engineers working under limited experimental data, and offers a clear and comprehensive theoretical framework for computer scientists to develop more powerful solutions. Looking forward, it is foreseeable that machine learning will propel enzyme engineering toward achieving superior results with fewer experimental requirements.

Keywords: Enzyme engineering; Data scarcity; Protein language models; Transfer learning; Zero-shot prediction; Active learning

---

1. Introduction

Enzymes, as nature's highly efficient and exquisitely selective biocatalysts, play a pivotal role in green chemistry, biomanufacturing, and the pharmaceutical industry [1]. Through rational design or directed evolution, natural enzymes can be engineered to significantly enhance their catalytic efficiency, stability, and substrate adaptability, thereby driving the development of novel biocatalytic systems. In recent years, machine learning (ML) methods have provided a new paradigm for enzyme engineering [2,3]. By capturing nonlinear interactions among residues, cofactors, and substrates, ML approaches can model complex enzymatic behaviors, propose rational mutations, and predict beneficial variants with unprecedented accuracy, substantially outperforming traditional methods [4–6].

However, the superior performance of mainstream ML heavily relies on massive, high-quality sequencefunction data, which stands in sharp contrast to the pervasive "small-data" reality in enzyme engineering [7,8]. Unlike data-rich fields such as natural language processing or computer vision, enzyme functional data are predominantly obtained through time-consuming and expensive wetlab experiments. A typical enzyme engineering workflow follows the designbuildtestlearn (DBTL) cycle, where each iteration involves gene construction, protein expression, purification, and functional assayssteps that are not only lengthy but also limited in throughput [9]. Moreover, many key enzymatic phenotypes (e.g., catalytic efficiency, substrate selectivity, or kinetic parameters) often rely on complex biochemical assays or chromatographic analyses, further restricting large-scale data acquisition [10]. Consequently, most enzyme engineering datasets in practice consist of only tens to a few thousand variants, far below the scale typically required for ML models.

To alleviate this issue, researchers have recently proposed a series of few-shot learning (FSL) strategiesML approaches that achieve high predictive performance even with limited labeled data [11,12], as illustrated in Figure 1. On one hand, pre-trained models leveraging large-scale public data or low-fidelity information, such as protein language models (PLMs), can learn generic representations of protein sequences to enable zero-shot prediction [13], and further improve performance on few-shot tasks through fine-tuning [14]. On the other hand, active learning-guided directed evolution treats enzyme engineering as a supervised learning problem, iteratively improving experimental information efficiency across rounds, thereby enabling more reliable model training under data-constrained conditions [10].

Given the importance of addressing the small-data problem, developing specialized ML methods that can effectively learn from limited data has been a prominent topic within the ML research community [15,16]. Recent years have witnessed numerous surveys and reviews on this theme,

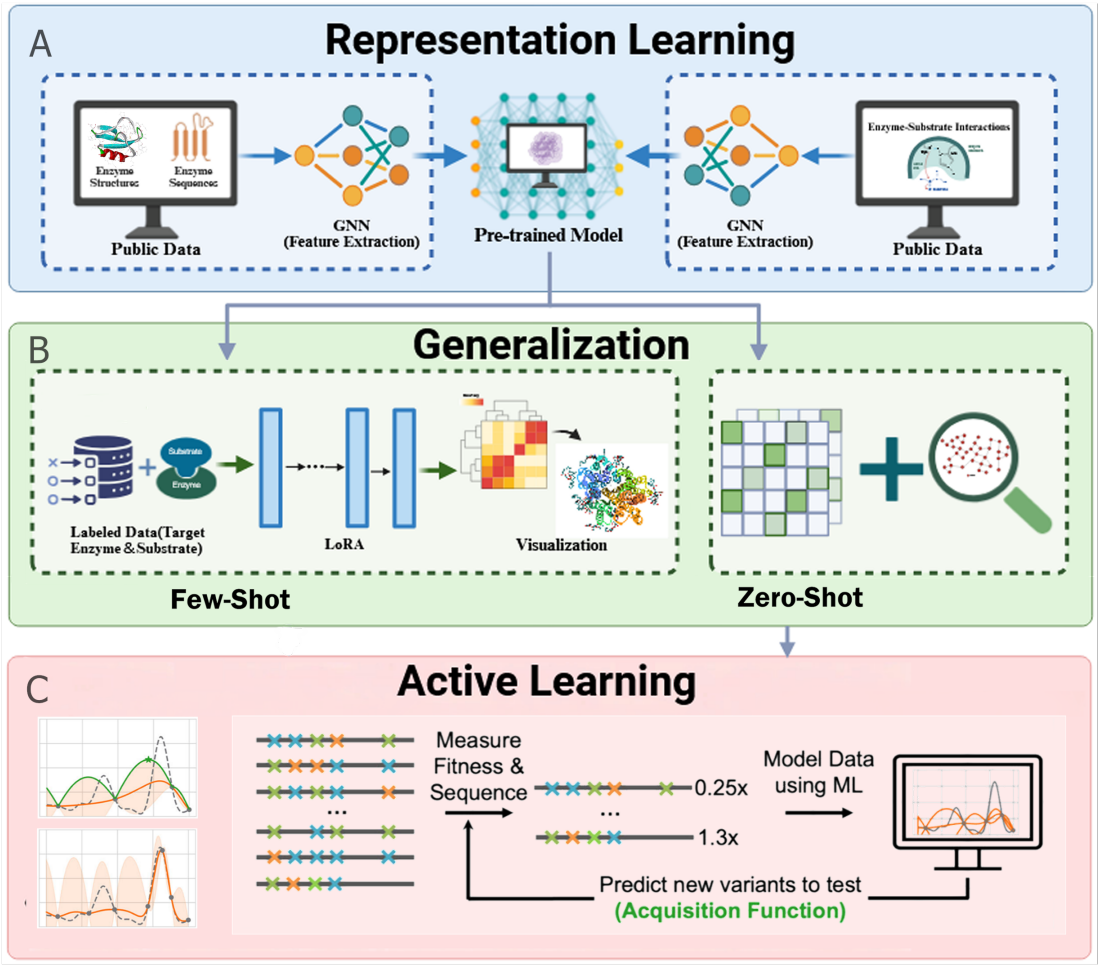

Figure 1. Schematic of few-shot machine learning strategies for enzyme engineering.(A) Pre-trained models (e.g., PLMs) learn generic representations from massive unlabeled or low-fidelity sequences, capturing sequence, structure, evolutionary, and functional information transferable to downstream tasks. (B) Pre-trained models can be used either directly for zero-shot prediction by assessing mutant consistency with natural patterns under evolutionary constraints (enabling preliminary screening), or adapted with labeled data via fine-tuning (e.g., PEFT, meta-learning, or in-context learning) to significantly boost predictive accuracy. (C) Randomly sample a small variant library for wet-lab testing, train a supervised model with uncertainty quantification, and use an acquisition function to balance exploration and exploitation for iterative variant proposal and retraining until the target fitness is achieved [125].

including applications of FSL in automatic speech processing [17], skin disease analysis [18], molecular science [19], and upstream bioprocessing [20]. This review effectively fills the gap in enzyme engineering, offering practical guidance for tackling the small-data challenge in this field.

The remainder of this paper is organized as follows: (1) we discuss the current status of small-data issues in enzyme engineering; (2) we introduce data processing and ML methods capable of addressing the small-data challenge; (3) we summarize recent research progress in applying few-shot ML to enzyme engineering and the development of integrated strategies; and (4) we conclude with the challenges faced by current approaches and future perspectives.

2. Current Status of the Small-Data Problem in Enzyme Engineering

Enzyme engineering focuses on optimizing various enzymatic properties. Classical machine learning methods rely on sequencefunction data derived exclusively from wetlab experiments, which serve as training and test sets for supervised learning to predict and evolve enzymes. However, the limited availability of high-quality sequencefunction data remains a core bottleneck [21]. For a specific enzymesubstrate pair, functional data are almost entirely dependent on experimentally generated data, which are inherently costly, lowthroughput, and noisy [22]. Consequently, typical datasets consist of only tens to a few thousand variants, a scale far below that typically required for training machine learning models, severely restricting model training and generalization capabilities [23–25].

The intrinsic complexity of protein fitness landscapes further exacerbates this challenge. Owing to pervasive epistasis [26], the functional effects of mutations are highly contextdependent, leading to rugged and nonlinear fitness landscapes that are difficult to approximate from limited observations [27]. As a result, models trained on sparse datasets may achieve reasonable interpolation within the training distribution but fail to generalize to unseen sequence combinations. Moreover, due to the difficulty of data acquisition, datasets are often imbalanced [28]. For example, when targeting a specific substrate, the obtained variants may exhibit only lowtomoderate fitness, resulting in a lack of highfitness data points and thereby limiting the model's ability to predict superior variants. These factors collectively constrain the ability of traditional machine learning models to learn accurate and generalizable representations from small datasets.

Nevertheless, it is worth emphasizing that although truly supervised sequencefunction data remain extremely scarce [29], protein science databases such as UniProt, PDB, and MGnify have accumulated hundreds of millions of unlabeled sequences [30], and an increasing number of datasets targeting specific enzyme classes or properties are being established. Most of these resources are publicly accessible and are summarized in Table 1. Meanwhile, lowfidelity data, i.e., pseudo-data obtained from molecular mechanics or quantum mechanics simulations, have been widely used to train machine learning models in other fields [31]. In enzyme engineering, advanced zero-shot predictors based on protein language models can be leveraged to efficiently generate lowfidelity data, which will be detailed in subsequent sections. A more aggressive strategy involves using data augmentation algorithms to expand sequence datasets, such as exploiting the redundancy of the genetic code [32], replacing noncritical residues [33], or generating evolutionarily plausible mutations based on natural occurrence patterns [34]. These approaches expand the effective training set of existing sequences through biologically meaningful perturbations. Promising studies have also employed generative models for sequence augmentation [35,36]. Fully exploiting these nonexperimental data can greatly aid the engineering of specific enzymesubstrate properties.

In summary, whether by extracting information from public data or lowfidelity sources and transferring generic representations to target tasks, or by maximizing the information efficiency of limited wetlab data to alleviate sparsity and imbalance and ensure exposure to diverse data points, both strategies will facilitate enzyme optimization under constrained experimental data.

Table 1. Public databases and datasets for enzyme engineering.

Name	Data type	Size	Applicable tasks	Use in few-shot scenarios	Access
BRENDA	Enzyme kinetic parameters (k_{cat}, K_m, optimal pH/T)	Tens of thousands of experimentally determined data points	Various enzymes (e.g., transaminases, dehydrogenases)	Provides prior distribution of kinetic parameters; serves as supervision signal for pre-trained models	https://www.brenda-enzymes.org
SABIO-RK	Enzyme kinetic parameters (reaction rate constants, K_m, etc.)	Thousands of annotated kinetic records	Various enzymes	Provides structured kinetic parameter repository; assists in setting Bayesian priors	https://sabio.h-its.org
CAZy	Glycoside hydrolase family protein sequences	119 GH families	Glycoside hydrolases (GHs)	Used for developing DeepGH models; provides supervised training data for family classification	http://www.cazy.org
UniRef50	Protein sequences (clustered at 50% identity)	∼48 million sequences	General	Pre-training protein language models (UniRep, ESM, RITA, etc.); constructing homologous sequence alignments	https://www.uniprot.org/help/uniref
UniRef90/100	Protein sequences	∼200 million sequences	General	Pre-training larger PLMs (e.g., RITA 1.2B)	https://www.uniprot.org/help/uniref
PDB	Experimentally resolved protein structures	∼200,000 structures	Enzymes with known structures	Provides structural constraints to narrow mutation space; training structure encoders	https://www.rcsb.org
Unbiased Kinetic Datasets (CataPro)	k_{cat}, K_m, k_{cat}/K_m (deduplicated)	Strictly deduplicated	General	Evaluating model generalization; serving as test benchmark for pre-trained models	https://github.com/zchwang/CataPro
ProteinGym	Deep mutational scanning (DMS) datasets	Multiple DMS datasets	General	Evaluating generalization of mutation effect prediction models	https://proteingym.org
CAZy	Glycoside hydrolase family protein sequences	119 GH families	Glycoside hydrolases, cyclodextrinases	Pre-training or transfer learning for CAZy family classification tasks	http://www.cazy.org
UniProt	Protein sequences and annotations	Massive	General	Foundation data for pre-training language models (e.g., ESM, ProtBERT)	https://www.uniprot.org

RiPP biosynthetic gene cluster databases (e.g., MIBiG)	RiPP precursor peptide sequences and product structures	Thousands of clusters	RiPP modifying enzymes	Pre-training masked language models and substrate prediction	https://mibig.secondarymetabolites.org
FOSrelated enzyme datasets	Affinity of carbohydrate-active enzymes toward FOS	Small scale (constructed in literature)	Glycoside hydrolases, fructosyltransferases	Benchmark dataset for few-shot learning	[48]
PET hydrolase datasets (e.g., IsPETase, LCC)	Mutant sequences and PET degradation activity	Tens to hundreds of mutants	PET hydrolases	Testbed for active learning, meta-learning, and rational evolution	[49–51]
GPCR ligand binding affinity datasets	GPCR sequences, ligand structures, and binding affinities	Tens of thousands	GPCRs	Training and evaluation of dynamic deep transfer learning (TrGPCR)	https://www.ebi.ac.uk/chembl/
Temperature stability segment dataset	Enzyme sequence fragments and temperature stability labels	Small sample	Various enzymes	Modeling temperature stability from sequence segment perspective	[52]
Zearalenone hydrolase mutant datasets	Mutant sequences and thermostability (T_m)	Small scale (validated after zero-shot optimization)	Zearalenone hydrolase, xylanase	Test cases for zero-shot deep learning + multi-objective optimization	[53]
DLKcat	Enzyme kinetic parameters (k_{cat}), protein sequences, substrate SMILES strings	Over 16,000 experimentally measured k_{cat} values	Enzyme kinetic parameter prediction, enzyme-constrained model (ecGEM) construction	High-quality labeled data for fine-tuning other deep learning models	https://github.com/SizheQiu/DLKcat
Megascale	Thermodynamic folding stability measurements	Contains 1,841,285 measurement data points	Mutation effect prediction, training and evaluation of protein stability prediction models	Can serve as source data for transfer learning or pre-training	https://huggingface.co/datasets/RosettaCommons/MegaScale

3. Machine Learning Methods for the Small-Data Challenge

3.1. Representation Learning: Extracting Universal Protein Representations

Pre-trained protein models learn generic representations from large-scale public data via representation learning. These representations can be efficiently transferred to downstream tasks such as mutation effect prediction and enzyme function optimization [5], significantly improving predic-

tive accuracy and generalization under limited wet-lab data. This makes them a key technology for overcoming the sequencefunction data sparsity bottleneck in enzyme engineering. Representative pre-trained protein models are listed in Table 2, where general-purpose models are trained on massive public sequence and structure data, while task-specific models are fine-tuned on domain-specific datasets.

The mainstream tools are protein language models (PLMs), whose training methods are adapted from natural language processing to protein sequence characteristics. In this framework, protein sequences are treated as a language, with amino acids as discrete tokens and sequence patterns corresponding to biological syntax and semantics. By leveraging billions of unlabeled sequences in protein databases, these models learn statistical dependencies among residues via self-supervision [37]; hence, PLMs are mostly general-purpose and applicable to various enzyme tasks. Importantly, these representations have been shown to implicitly encode structural, functional, and evolutionary information [38].

Masked language modeling (MLM) is one of the most widely adopted pre-training strategies, with representative models including the ESM series [39] and the ProtTrans framework [40]. In MLM, a fraction (typically 15%) of amino acid residues are randomly masked, and the model is trained to predict the masked tokens based on their surrounding context. This objective encourages the model to learn contextual dependencies and latent sequence patterns, thereby capturing evolutionary constraints embedded in protein sequences [41]. Autoregressive modeling represents another major category of protein pre-training. In this framework, the model sequentially predicts each amino acid given all preceding residues, thus learning the joint probability distribution over protein sequences. Compared with MLM, autoregressive models are particularly suited for generation tasks as they naturally support sequence generation. The ProGen series exemplifies this approach, demonstrating the ability to generate de novo protein sequences with experimentally validated functionality [42,43]. By learning from massive sequence corpora, such models can explore regions of sequence space beyond natural proteins, offering new opportunities for protein design. Extended studies have shown that increasing model size and training data significantly boosts performance in both generation and prediction tasks, indicating strong scaling effects in protein language modeling [44].

Recent advances have extended protein pre-training to multimodal settings, integrating sequence, structure, and functional annotations into unified learning frameworks [45]. Models such as ESM-3 exemplify this trend, enabling joint representation learning across multiple biological modalities [39], capturing higher-order relationships that are difficult to infer from sequence alone. Emerging approaches incorporating long-context modeling or graph-based representations further enhance the ability to model complex biological systems [46].

Building custom pre-trained models is also promising. On one hand, one can selectively screen evolutionarily related or structurally similar protein family sequences from public databases (e.g., UniProt, PDB) according to the target enzyme properties, constructing a tailored pre-training corpus or directly using public datasets. This task-oriented pre-training allows the model to be exposed to sequence patterns highly homologous to the downstream task, achieving superior performance. For example, ZymCTRL was pre-trained on over 37 million enzyme sequences with Enzyme Commission (EC) number annotations in UniProt and validated experimentally [47]. On the other hand, low-fidelity data can also serve as effective pre-training resources (semi-supervised learning). Gelman et al. used Rosetta simulations to generate large-scale datasets containing 55 biophysical properties (e.g., molecular surface area, solvation energy, van der Waals interactions) for pre-training, and showed that 1,000 simulated data plus 320 experimental data performed comparably to 8,000 simulated plus 80 experimental data, confirming the partial substitutability of simulated data for experimental data [37].

Table 2. Publicly available pre-trained models for protein and enzyme engineering.

Model	Type	Training objective	Training data	Input	Zero-shot	Reference
SeqVec	biLSTM (ELMo)	Bidirectional LM	UniRef50	Sequence	No	https://github.com/mheinzinger/SeqVec
TAPE	Transformer	MLM + next-token prediction	Pfam + UniRef50	Sequence	No	https://github.com/songlab-cal/tape
ESM-1b	Transformer	Masked language modeling	UniRef50	Sequence	No	https://github.com/facebookresearch/esm
ESM-1v	Transformer	Variant effect prediction	UniRef90	Sequence	Yes	
ESM-2	Transformer	MLM	UniRef50	Sequence	Yes	
ProtBERT	BERT	MLM	BFD (2.1B sequences)	Sequence	No	https://github.com/agemagician/ProtTrans
ProtAlbert	ALBERT	MLM	BFD	Sequence	No	
ProtT5	T5	Span masking	BFD + UniRef50	Sequence	No	
ProteinBERT	Transformer	MLM + GO prediction	UniRef90	Sequence	No	https://github.com/nadavbra/protein_bert
xTrimoPGLM	Protein foundation model	MLM + generative tasks	UniRef + metagenomic datasets	Sequence	No	https://github.com/biomap-research/xTrimoPGLM
ProSST	Sequence-structure transformer	Multimodal learning	UniRef + PDB + AlphaFold DB	Sequence	Yes	https://github.com/ai4protein/ProSST
ProGen	GPT-like Transformer	Autoregressive LM	UniRef50	Sequence	Yes	https://github.com/salesforce/progen
ProtGPT2	GPT-2 Transformer	Autoregressive LM	UniRef50	Sequence	Yes	https://github.com/TeletcheaLab/protGPT2
ProGen2	Large autoregressive Transformer	Autoregressive LM	UniRef + metagenomic datasets	Sequence	Yes	https://github.com/salesforce/progen
CARP	Convolutional LM	Autoregressive LM	UniRef50	Sequence	No	https://github.com/microsoft/protein-sequence-models
RITA	Transformer	Autoregressive LM	UniRef50	Sequence	No	https://github.com/lightonai/RITA
UniRep	mLSTM	Autoregressive LM	UniRef50	Sequence	No	https://github.com/churchlab/UniRep
MSA Transformer	Axial Transformer	MLM on MSA	UniRef + MSA alignments	Sequence	No	https://github.com/facebookresearch/esm
ESM-MSA-1b	Transformer	MSA masked LM	UniRef + MSA	Sequence	No	https://github.com/facebookresearch/esm
OmegaFold	Transformer + geometric module	Structure-aware learning	UniRef + PDB	Sequence	No	https://github.com/HeliXonProtein/OmegaFold

ESMFold	Transformer + folding module	MLM + structure prediction	UniRef50	Sequence	No	https://github.com/facebookresearch/esm
SaProt	Structure-aware transformer	Sequence-structure MLM	UniRef + AlphaFold DB	Sequence	Yes	https://github.com/westlake-repl/SaProt
ProteinMPNN	Inverse folding (GNN)	Autoregressive LM (structure-conditioned)	PDB	Structure	Yes	https://github.com/dauparas/ProteinMPNN
ESM-IF1	Sequence-to-sequence Transformer / inverse folding	Structure-conditioned masked LM (span masking) / negative log-likelihood	PDB	Structure	Yes	https://github.com/facebookresearch/esm
GearNet	Graph neural network	Structure representation learning	PDB	Sequence	No	https://github.com/DeepGraphLearning/GearNet
MIF-ST	Graph Transformer	Structure-conditioned MLM	PDB	Sequence	No	https://github.com/microsoft/protein-sequence-models
PRIME	Transformer	MLM + sequence-level regression	96 million sequences with optimal growth temperature labels	Sequence	Yes	https://github.com/HySonLab/PRIME
DeepEnzyme	Transformer + GCN	Regression (k_{cat} prediction)	k_{cat} data from BRENDA, SABIO-RK, etc.	Sequence + structure	No	[73]
IECata	Bilinear attention network + evidential deep learning	Regression + uncertainty estimation	k_{cat}/K_m entries from BRENDA and SABIO-RK	Sequence + substrate structure	No	https://github.com/zhaoyanpeng208/IECata
ProTKcat	ProtT5 + MolGNet + Extra-Trees	Regression (two-stage)	BRENDA, SABIO-RK, etc.	Sequence + substrate molecular graph + temperature	No	https://github.com/bozhenhhu/ProKcat
DeepGH	Deep neural network	Multi-class classification (GH family)	CAZy database	Sequence	No	https://huggingface.co/deepghs

3.2. Zero-Shot Prediction of Mutation Effects

Zero-shot prediction is an important application of pre-trained protein models in enzyme engineering [54,55]. Unlike traditional supervised learning that relies on experimentally measured sequencefunction data, zero-shot prediction directly estimates the functional impact of mutations using pre-trained models without requiring labeled data for training. This paradigm is a key enabler of few-shot enzyme engineering. By leveraging knowledge learned from large-scale protein sequence

databases, pre-trained models can infer mutation likelihoods and prioritize promising variants before any laboratory experiments. Liu et al. have comprehensively reviewed online tools and methods for zero-shot prediction in enzyme engineering [56].

The core principle of zero-shot prediction is to assess the consistency of a mutant sequence with the statistical patterns learned during pre-training. In language-model-based approaches, this is typically achieved by comparing the likelihood of the wild-type and mutant sequences. Mutations that are more consistent with the evolutionary constraints encoded in the model receive higher scores, while disruptive mutations are assigned lower probabilities [57,58]. Importantly, this method does not require any experimental labels; instead, it relies entirely on the evolutionary information implicit in natural protein sequences [59]. This makes zero-shot prediction particularly suitable for early-stage enzyme engineering when no experimental screening data are available.

Large-scale PLMs such as the ESM series and ProtTrans have demonstrated excellent performance in zero-shot mutation effect prediction [60,61]. One study developed a fast and flexible in silico protein design tool by integrating multiple PLMs [62]. By directly scoring mutant sequences, these models can identify beneficial mutations with reasonable accuracy even without task-specific fine-tuning. These results highlight the potential of universal predictors of protein mutation effects, substantially reducing reliance on large external experimental datasets.

Beyond general PLMs, several methods have been specifically designed to enhance zero-shot mutation prediction, often by refining scoring strategies or incorporating additional biological priors. For instance, some studies integrate structural information, functional text descriptions [63], or evolutionary information [58] to improve predictive performance. Others develop task-specific strategies for particular enzymatic properties [64] or enzyme classes [65]. These task-oriented adaptations indicate that zero-shot prediction is not merely a byproduct of pre-training but a rapidly evolving research direction in its own right.

To systematically evaluate zero-shot prediction methods, a benchmark study compared different pre-trained models on multiple protein engineering datasets. It showed that large-scale language models outperform traditional unsupervised methods, especially in data-scarce settings [66]. However, the accuracy of zero-shot prediction still has notable limitations [67]. On one hand, while pre-trained models capture general evolutionary constraints encoded in sequences, they cannot fully account for protein-specific functional mechanisms, which somewhat undermines the advantage of zero-shot methods in automated enzyme engineering [68,69]. On the other hand, zero-shot prediction often falls short in fine optimization stages, e.g., its predictive reliability significantly decreases for multiple mutations [70] or for sequences far from the natural sequence space [71].

Therefore, in practical engineering applications, zero-shot prediction typically needs to be combined with additional information to achieve usable performance [72]. This has also driven the development of fine-tuning strategies that calibrate PLMs with a small amount of experimental data, laying the groundwork for subsequent few-shot modeling approaches.

3.3. Fine-Tuning Pre-trained Models and Few-Shot Learning

While zero-shot prediction provides a powerful starting point for few-shot enzyme engineering, its performance is often limited by a lack of task-specific adaptation. In many real-world scenarios, the amount of available experimental data is very smalltypically ranging from tens to a few hundred variants. Effectively leveraging these limited data becomes crucial for improving prediction accuracy. Fine-tuning adapts a pre-trained protein model to the target task using a small set of labeled data, enabling the model to capture protein-specific fitness landscapes with significantly better accuracy than zero-shot methods [67].

3.3.1. Parameter-Efficient Fine-Tuning (PEFT)

PEFT addresses the issue by updating only a small fraction of parameters while keeping most of the pre-trained model fixed. This approach not only reduces computational cost but also mitigates overfitting, making it particularly suitable for data-scarce enzyme engineering tasks [74].

Various PEFT techniques have been successfully applied to PLMs. For example, adapter-based methods insert small trainable modules into each Transformer layer, enabling task-specific adaptation with minimal parameter updates [75,76]. Similarly, low-rank adaptation (LoRA) decomposes weight updates into low-rank matrices, significantly reducing the number of trainable parameters while preserving model expressiveness. This method requires training only a very small number of added parameters, greatly lowering training costs, and has been validated across various protein tasks [77,78]. Sledzieski et al. demonstrated that LoRA achieves comparable or even superior performance to full fine-tuning in tasks such as proteinprotein interaction prediction [14]. Additionally, various innovative fine-tuning approaches [79] and auxiliary strategies [7] have been proposed.

In practical applications, PEFT has been widely adopted in diverse enzyme engineering tasks. One study leveraged transfer features from large-scale tasks such as ProteinMPNN and ESM2, performing lightweight fine-tuning for stability prediction, significantly boosting predictive performance in few-shot scenarios [80]. The same idea is embodied in the SPURS framework proposed by Li et al. PEFT strategies have also shown promising results in specific enzyme engineering tasks [81,82]. These methods follow a common principle: decoupling general biological knowledge from task-specific adaptation, allowing large pre-trained models to be efficiently reused across various protein engineering tasks.

More broadly, these case studies suggest that combining pre-trained representations with task-specific adaptation enables models to effectively capture local fitness landscapes, which is crucial for guiding enzyme optimization. These findings underscore the role of PEFT as a key enabler of scalable and few-shot enzyme engineering workflows.

3.3.2. Meta-Learning

Meta-learning offers another powerful strategy for few-shot adaptation by adjusting model parameters, architecture, or optimization methods based on experience gained during training. In specific applications, this enables pre-trained models to rapidly adapt to the current task with very little data from new enzymes [83]. Similarly, multi-task meta-learning frameworks have significantly improved prediction accuracy in data-scarce environments [84].

3.3.3. In-Context Learning

While fine-tuning-based methods rely on parameter updateswhich allow precise task fitting but still risk overfitting under extremely small sample sizesan alternative paradigm adapts the model at inference time without modifying its parameters. Such approaches, known as in-context learning, leverage the inherent flexibility of large pre-trained models to generalize across tasks via contextual information [67]. By avoiding explicit training, these methods enable rapid exploration of protein sequence space with minimal computational overhead even with very limited labeled datasets [85,86].

In this approach, a small number of labeled examples are provided as part of the input, allowing the model to infer task-specific patterns during inference. Recent studies have shown that large protein models can utilize this mechanism for few-shot mutation effect prediction, significantly outperforming zero-shot baselines [87]. This also indicates that pre-trained models not only encode static biological knowledge but also possess the capacity for dynamic task adaptation. Furthermore, pre-trained models integrating multiple types of information can more accurately map sequence to function in task-adaptive inference [88].

Overall, task-adaptive inference methods reduce or eliminate the need for parameter updates, further decoupling transfer learning from the scale of experimental data, providing an important tool for early-stage exploration in enzyme engineering. However, their performance is highly dependent on the representational capacity of the pre-trained model [89]. When the target task deviates significantly from the pre-training distribution, such methods often struggle to capture functionally relevant fine-grained differences. Moreover, due to the lack of explicit parameter updates, these approaches may be limited in modeling multiple mutations or complex fitness landscapes [90,91].

3.4. Active Learning-Guided Directed Evolution

Another category of strategies improves data efficiency by selecting the most informative samples, thereby achieving maximal model performance gains with minimal data [92,93]. This paradigm is formalized as active learning (AL) and is particularly well suited for enzyme engineering [94]. Unlike random sampling, active learning employs model-driven criteria to guide data acquisition. A core concept is that not all data points are equally informativesome sequences can provide disproportionately more information about the underlying fitness landscape [95].

Most active learning frameworks rely on uncertainty estimation to identify informative samples. Bayesian optimization (BO) is a widely used approach in which a probabilistic model (e.g., a Gaussian process or a Bayesian neural network) estimates both the predictive mean and uncertainty, and an acquisition function is used to balance exploration and exploitation. Extensions such as BO-EVO combine Bayesian optimization with evolutionary strategies to efficiently search combinatorial mutation spaces. These methods have demonstrated remarkable efficiency in identifying high-performance variants by experimentally testing only a small fraction of candidates [96]. Alternative strategies focus on diversity or information gain rather than uncertainty alone. For example, Thompson sampling and Monte Carlo-based methods aim to improve exploration efficiency in high-dimensional sequence spaces [97], while the Most Informative Positive strategy prioritizes samples that are likely to yield functional variants in discovery tasks [98].

In enzyme engineering, active learning integrates naturally into the designbuildtestlearn (DBTL) cycle, forming a closed-loop optimization framework. In each iteration, the model proposes candidate sequences, experiments provide new measurements, and the updated data further refine the model. Recent studies have shown that such closed-loop systems can significantly enhance engineering efficiency [91]. By iteratively selecting informative variants, these frameworks can achieve substantial performance improvements within limited experimental rounds, effectively exploring vast combinatorial sequence spaces [99]. More importantly, active learning shifts the role of machine learning from passive prediction to active experimental design. By constructing small but information-rich training datasets, reliable models can be built with very few experimental samples, transforming enzyme engineering into a data-driven optimization process [100].

However, recent benchmarking studies have shown that when uncertainty estimates are unreliable or the protein fitness landscape is highly complex, uncertainty-based methods may not consistently outperform simple greedy sampling [101]. Furthermore, active learning performance is strongly dependent on the underlying predictive model. Poor model initialization or distributional shifts can lead to suboptimal sampling decisions, resulting in inefficient exploration. These findings highlight the need for more robust uncertainty quantification methods and a deeper understanding of when active learning is likely to succeed in the protein engineering context.

Integrating pre-trained protein language models into this pipeline can significantly improve information efficiency. The zero-shot scoring capability of language models has proven powerful for nominating high-potential lead mutants. Fan et al. innovatively used the minimum probability of masked positions as an indicator of mutational effects on activity to rank candidate sequences, achieving 62.5% protease mutants with enhanced thermostability and 50% lysozyme mutants with improved activity within only two rounds of iterative design [102]. Similarly, on the nomination strategy front, the ProSpero framework combines a pre-trained generative model with a surrogate model updated from experimental feedback. By adaptively integrating fitness-relevant residue selection and biologically constrained sampling at each active learning round, it explores sequence space beyond the wild-type neighborhood while maintaining biological plausibility, consistently outperforming or matching state-of-the-art methods across various protein engineering tasks [103]. The Venus-MAXWELL framework integrates the zero-shot scoring capability of PLMs with sequencelandscape modeling. After incorporating ESM-IF into the framework, it not only achieves higher prediction accuracy than ThermoMPNN but also delivers inference speeds 10 times faster [104]. Pre-trained models are also effective at filtering out deleterious mutations, excluding candidates that are likely to

cause inactivation or instability. Deguchi et al. proposed a data augmentation method termed weakly supervised data, which combines molecular simulations with PLM zero-shot predictions. When experimental training data are insufficient, weakly labeled data generated by computational estimates (simulations or PLM predictions) are incorporated into training. The weight and participation of weak training data are dynamically adjusted according to the scale of available experimental data, thereby extending the applicability of the method to various protein propertiesincluding binding affinity and enzymatic activitywhile minimizing potential negative impacts [72].

4. Applications of Few-Shot Learning in Enzyme Engineering Tasks

Building upon the various few-shot machine learning methods discussed above, a series of integrated strategies targeting specific enzymatic properties have been developed in recent years and applied to enzyme engineering. However, it should be noted that most studies primarily aim to develop a general architecture or strategy, using wet-lab experiments or publicly available labeled data to demonstrate their effectiveness; only a handful of studies have completed a full few-shot engineering campaign for a single enzyme.

Moreover, given the vast differences in complexity across enzyme engineering tasks, few-shot does not have a unified quantitative standard. This review adopts the flexible definition used in previous studies: for a machine learning task, small data refers to datasets with fewer data points than typically required to accomplish that task. All studies summarized below explicitly list achieving enzyme engineering tasks with few-shot or limited data as one of the core objectives in their method development and enzyme optimization processes, and are compiled in Table 3. This section aims to inspire researchers to generalize these methods to their own target tasks (Figure 2 provides an overview of few-shot machine learning strategies applied in enzyme engineering).

4.1. Fitness

In protein engineering, fitness is a broad concept that generally refers to the comprehensive performance of enzyme variants under specific selection pressures, encompassing multiple properties such as activity, stability, and expression level. Directly modeling and optimizing the fitness landscape is one of the core application directions of few-shot machine learning in enzyme engineering.

Pioneering work demonstrated the powerful modeling capability of active learning under limited data. Romero et al. used Gaussian processes to probabilistically model protein fitness landscapes. Their model not only predicted the fitness of unseen sequences but also explicitly quantified predictive uncertainty to guide experimental selection. With only 40 training data points, it achieved accuracy comparable to that of traditional fragment-based regression models requiring 218 data points [107]. Building on this foundation, Roberts et al. developed the FolDE framework, which in the first round leverages the ESM protein language model for zero-shot naturalness ranking to nominate mutants, and then uses neural networks to predict activity in subsequent rounds. Over three rounds (16 variants per round), this method increased the probability of discovering top-1% mutants by 55% [127].

In recent years, fine-tuning strategies based on pre-trained models have shown great potential in few-shot fitness prediction. The UniRep model proposed by Biswas et al. performed global unsupervised pre-training on over 20 million raw amino acid sequences, and then fine-tuned with functional characterization data from only 24 random mutants. Through in silico directed evolution, it generated over 1,000 new designs superior to the wild type [23]. In an extreme exploration of data efficiency, the METL model developed by Gelman et al. combined biophysical simulation with transfer learning: a Transformer encoder was pre-trained on millions of Rosetta-simulated variants and then fine-tuned on a small amount of experimental data. Their study showed that adding approximately 29 simulated data points yielded a performance improvement equivalent to adding one experimental data point, and they successfully obtained high-fluorescence GFP variants with only 64 experimental data points [37]. Similarly, SESNet integrated sequence, co-evolutionary, and structural information for pre-training, and with fine-tuning on only 40 high-quality data points, achieved accurate prediction

of 58-site multi-mutants (Spearman correlation coefficient 0.50.7) [105]. In addition, the active learning framework EVOLVEpro, built upon ESM-2, iteratively selected top-ranked sequences for experimentation over multiple rounds. It reached the performance level of traditional methods pre-trained with 160 variants in just 5 rounds (16 variants per round) [116].

Table 3. Summary of few-shot machine learning models applied to fitness optimization tasks.

Few-shot strategy	Target enzyme/benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fine-tuning	TEM-1 β-lactamase	Global unsupervised pre-training of UniRep on >20 million raw amino acid sequences; unsupervised fine-tuning; *in silico* directed evolution	24	Generated over 1,000 new designs with >wild-type performance	[23]
Fine-tuning	GFP dataset	METL: Transformer encoder trained on 55 biophysical properties of millions of sequence variants generated by Rosetta simulation; fine-tuned with small experimental data	64	METL learned and generalized from only 64 training examples across sequence space	[37]
Fine-tuning	GFP dataset	SESNet: Pre-training with large amounts of low-quality results from unsupervised models; fine-tuning with small high-quality experimental data	40	Achieved Spearman correlation of 0.5–0.7 for multi-site mutants	[105]
Fine-tuning	Glycoside hydrolase	MECE: Pre-trained DeepGH model on 119 GH family sequences; used Grad-CAM to extract functionally important residues and combined with evolutionary information to construct mutation scoring metrics for beneficial mutant design	9 single-point mutants; 4 multi-site mutants	Identified 7-site mutant CHIS1754-MUT7 from CHIS1754 with 23.53-fold improvement in k_{cat}/K_m	[106]
Active learning	P450 enzymes	Probabilistic modeling of protein fitness landscape using Gaussian processes; predicted fitness of unseen sequences and explicitly quantified uncertainty (variance) to guide experimental selection	40	With only 40 training points, Gaussian process achieved accuracy comparable to traditional fragment-based regression requiring 218 data points	[107]
Zero-shot	ProteinGym benchmark; Mg-pCSO	GEMS: Integrated multiple zero-shot predictors including ESM-IF, SaProt, MSA Transformer, and GEMME	/	Achieved excellent performance on ProteinGym benchmark, significantly outperforming existing SOTA methods; showed 1.1- to 3.2-fold improvement in catalytic efficiency for MgpCSO single mutants	[108]
Zero-shot	Gene editing fusion proteins	AiCE (AI-informed Constraints for Protein Engineering): Employed zero-shot prediction by combining inverse folding sampling with structural/evolutionary constraints to nominate single and multi-point mutants	10–100	Achieved high success rates (11%–88%) in functional optimization with only tens to hundreds of experimentally validated variants	[109]

4.2. Activity and Thermostability

Catalytic activity and thermostability are two key properties for the industrial application of enzymes, and have been major targets of few-shot machine learning methods. Researchers have

achieved significant progress across various enzyme systems through fine-tuning, zero-shot prediction, and active learning strategies.

In fine-tuning-based approaches, ESM-Ezy fine-tuned ESM-1b to build a binary classifier trained on only 147 known multicopper oxidases (MCOs), and successfully mined 18 candidate enzymes from the UniRef50 database, of which 44% outperformed the query enzyme in at least one catalytic property. The best variant, Sulfur, exhibited a half-life of 156.9 minutes at 80řC and a 32.9-fold improvement in catalytic efficiency over the wild type [110]. GeoEvoBuilder employed geometric deep learning to extract features from target protein structures, combined with ESM2 to capture evolutionary information, and resolved structureevolution conflicts through iterative optimization. In GPX4 optimization, the best variant achieved an 8.49-fold improvement in catalytic efficiency with only 19 data points; in DHFR optimization, a 20-fold improvement in catalytic efficiency and a 10řC increase in T_m were achieved with 23 data points [49].

In zero-shot prediction, the PRIME model was pre-trained on 96 million sequences annotated with optimal growth temperature (OGT), enabling zero-shot inference of protein thermostability and mutation effects. For LbCas12a, experimental validation of only 60 mutants yielded an 8-site mutant R2-26 with a 6.25řC increase in T_m while retaining activity [111]. GEMS, a multimodal ensemble framework integrating multiple zero-shot predictors including ESM-IF, SaProt, and MSA Transformer, achieved excellent performance on the ProteinGym benchmark and successfully identified MgpCSO mutants with 1.1- to 3.2-fold improvements in catalytic efficiency [108]. AiCE employed inverse folding models to generate structurally compatible sequences and combined evolutionary coupling scores to nominate multi-mutant combinations. With only 10100 experimentally validated variants, it achieved 11%88% functional optimization success rates, including a 14.3-fold improvement in mitochondrial base editing efficiency [109].

In active learning, Jiang et al. proposed a rationality-informed protein evolution (RIPE) strategy for the PET hydrolase BhrPET. They first used 150 rationally designed single-point mutants as the initial dataset, followed by four rounds of iterative evolution with EVOLVEpro. The final optimal variant IYLL exhibited 6.91-fold higher hydrolytic activity toward amorphous PET at 72řC compared to the wild type [50]. Chen et al. combined graph learning with meta-learning, first using GVP-GNN for structure-aware optimization and then MAML to learn cross-temperature general knowledge. By exploring only 1040 variants, the BHETase variant Y109R achieved a 5.5-fold increase in product yield at 60řC [51]. Zheng et al. developed a Segment Transformer for cutinase thermostability optimization, trained on only 3,454 data points from the BRENDA database. In guiding cutinase engineering, only 17 mutations were needed to achieve a 1.64-fold increase in relative activity and a 3.9-fold extension of half-life [52].

Table 4. Summary of few-shot machine learning models applied to activity and thermostability optimization.

Few-shot strategy	Target enzyme/benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fine-tuning	Novel multi-copper oxidases (MCOs)	ESM-Ezy: Fine-tuned ESM-1b to build binary classifier for MCO screening; selected nearby sequences in semantic space via Euclidean distance to the query enzyme (QE), enabling discovery of novel enzymes with low sequence similarity but high functional activity	147	With only 147 known MCOs for fine-tuning, screened 18 candidates from UniRef50; 44% outperformed the query enzyme in at least one catalytic property	[110]
Zero-shot	LbCas12a	PRIME: Pre-trained on 96 million sequences with OGT annotations; combined unsupervised MLM with supervised OGT prediction for zero-shot and fine-tuned prediction of thermostability and mutation effects	60	Among 30 single-point mutants, 9 had T_m not lower than wild type; final 8-site mutant R2-26 achieved 6.25řC increase in T_m (reaching 48.15řC) while maintaining or improving activity	[111]
Zero-shot	GPX4; DHFR	GeoEvoBuilder: Geometric deep learning to generate sequence designs from target protein structures; ESM2 to capture evolutionary information; AlphaFold2 for structure prediction and filtering	19; 23	2OBI-10 achieved 8.49-fold catalytic efficiency improvement with 4.74-fold lower K_m; 3D80-15 achieved 20-fold catalytic efficiency improvement with 10řC higher T_m than wild type	[49]
Active learning	BhrPET (PET hydrolase)	RIPE: Used rationally designed mutants as initial EVOLVEpro dataset; iterative optimization starting from functional mutants rather than wild type	150 single-point mutants	After four rounds, IYLL (F208I/H183Y/Q142L/S27L) showed 6.91-fold and 2.58-fold higher hydrolytic activity toward amorphous and post-consumer PET at 72řC vs. BhrPET	[50]
Active learning	BHETase	GVP-GNN for structure-aware sequence optimization; MAML with Gaussian process regression using limited multi-temperature activity data to learn cross-temperature general knowledge	$N_1 = 10$; $N_2 = 10$–40	Optimal variant Y109R achieved 5.5-fold higher TPA yield at 60řC than wild type with only 10–40 variants explored across two rounds	[51]
Zero-shot	RmZHD	MSA Transformer with ABACUS-R-driven zero-shot prediction; axial attention captures co-evolutionary signals to evaluate mutational effects	0 (validated 6 variants)	Achieved 8.1řC increase in T_m with >95% activity retention using only 6 mutations for validation	[53]
Zero-shot	Cutinase	Segment Transformer: Sequence segment representation with dual-group segment attention for multi-scale dependencies; weighted RMSE loss to address temperature distribution imbalance	0 (validated 17 variants)	Achieved 1.64-fold relative activity increase and 3.9-fold half-life extension with only 17 mutations while maintaining catalytic activity	[52]

4.3. Enzyme Kinetic Parameters

Accurate prediction of enzyme kinetic parameters (e.g., k_{cat}, K_m, k_{cat}/K_m) is a core prerequisite for rational enzyme design, yet experimental determination of these parameters is extremely time-

consuming and labor-intensive. To address this bottleneck, various few-shot prediction methods have emerged.

In fine-tuning-based approaches, BioStructNet was pre-trained on a hydrolase k_{cat} database and compoundprotein interaction data, and achieved high-accuracy prediction on Candida antarctica lipase B (CalB) through multiple fine-tuning strategies. The high-attention residues identified by the model (e.g., W104, T42, S47) were highly consistent with experimentally validated catalytic key residues [81]. CataPro utilized ProtT5 and molecular fingerprint techniques with approximately 27,658 k_{cat} data points, achieving a Pearson correlation coefficient of 0.497 on an unbiased test set. In the mining and engineering of SsCSO, it successively improved activity by 19.53-fold and 3.34-fold [112]. PMAK employed a residue-aware attention mechanism to capture feature interactions between enzyme and substrate representations, achieving lower mean squared error than DLKcat and TurNuP models in k_{cat} prediction across 16 environmental conditions [113].

In data augmentation strategies, Luo et al. integrated Rosetta molecular simulations and ESM-2 predicted mutation effects as weakly supervised data with experimental data, and developed algorithms to dynamically adjust their weights. When experimental data were fewer than 50, computational estimates alone were superior; however, when experimental data ranged from 50 to 200, the weakly supervised approach showed clear advantages [72]. KinForm combined intermediate layer embeddings from multiple PLMs, binding-site weighted pooling, PCA dimensionality reduction, and similarity-based oversampling. With only 16,775 k_{cat} data points, it improved R^2 from 0.077 to 0.15 (approximately 100% relative improvement) on a strict test set with sequence identity below 20% [114]. OšDENet employed pseudo-data-guided invariant representation learning with mask augmentation on enzyme sequences and substrate molecules, improving k_{cat} prediction R^2 from 0.342 to 0.410 on test sets with sequence identity below 40%, significantly enhancing out-of-distribution generalization [115].

Active learning has also been successfully applied to kinetic parameter-related enzyme optimization. Greenhalgh et al. targeted acyl-ACP reductase (FAR) using Gaussian process modeling with the upper confidence bound criterion to balance exploration and exploitation. After 10 iterative rounds testing only 96 data points, they identified the top-performing variant ATR-83, which exhibited a 4.9-fold increase in in vivo fatty acid production compared to the wild type [124].

Table 5. Summary of few-shot machine learning models applied to enzyme kinetic parameter optimization.

Few-shot strategy	Target enzyme/benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fine-tuning	CalB	BioStructNet: Pre-trained on hydrolase k_{cat} and compound-protein interaction data; multi-strategy fine-tuning for CalB	233 wild-type; 65 variants	High-attention residues (W104, T42, S47) matched experimentally validated catalytic residues	[81]
Fine-tuning	SsCSO	CataPro: ProtT5 + molecular fingerprints for k_{cat}, K_m, k_{cat}/K_m prediction	~27,658 k_{cat} data points	Achieved PCC of 0.497 on unbiased test set; identified new enzyme with 19.53-fold higher activity, further improved 3.34-fold	[112]
Fine-tuning	Various enzymes	PMAK: ProtT5 + RXNFP embeddings; residue-aware attention for enzymesubstrate feature interaction	/	Lower MSE than DLKcat and TurNuP across 16 environmental conditions	[113]
Data augmentation	Various enzymes	Weak supervision: Integrated Rosetta/ESM-2 estimates as weakly labeled data; dynamic weight adjustment	/	Outperformed computational estimates alone when experimental data 50200; superior to supervised learning with <50 data points	[72]
Data augmentation	Various enzymes	KinForm: Multi-PLM embeddings + binding-site pooling + PCA + similarity oversampling	16,775 k_{cat} data points	Improved R^2 from 0.077 to 0.15 (100% relative) on test set with sequence identity <20%	[114]
Data augmentation	Various enzymes	OšDENet: Pseudo-data-guided invariant learning with residue/SMILES mask augmentation	/	Improved R^2 from 0.342 to 0.410 on test set with sequence identity <40%; enhanced OOD generalization	[115]
Active learning	LSRs	EVOLVEpro: ESM-2-based active learning; iterative ranking and validation	N_1 = 11; 9 rounds	5 rounds (16 mutants/round) matched performance of 160-mutant pre-training; 10 rounds matched 500-mutant pre-training; >2.6-fold activity improvement	[116]
Zero-shot	Glutamate decarboxylase	Ensemble of four zero-shot predictors (UniKP, DLKcat, Pythia, ESM-1v) for consensus screening of 8,911 single-point mutants	0 (validated 107 variants)	Y39T/Δ460-469 variant achieved 300 g/L GABA in 5 L fermenter, 1.3-fold improvement over wild type	[117]

4.4. Thermostability

Thermostability is one of the key factors limiting the industrial application of enzymes, and has become a major target of few-shot machine learning methods.

In fine-tuning-based approaches, the FSFP framework combined model-agnostic meta-learning (MAML) with low-rank adaptation (LoRA). On Phi29 DNA polymerase, using only 20 data points for few-shot learning, it achieved an average T_m increase of over 1řC among the top-20 predicted variants and a 25% improvement in positive rate [67]. ActiMut-XGB integrated ESM2 sequence embeddings, physicochemical properties, evolutionary features (PSSM/PsePSSM), and positional features to predict thermostability changes in CALB single-point mutants using an XGBoost classifier. With

limited fine-tuning data, it achieved a Matthews correlation coefficient (MCC) of 0.2511, substantially outperforming traditional tools such as PremPS (MCC: -0.0105) and DeepDDG (MCC: 0.0305) [118]. The SPURS model employed a parameter-efficient fine-tuning strategy, integrating ESM2 and ProteinMPNN via lightweight adapters, fine-tuning only approximately 1% of parameters to achieve efficient and generalizable prediction of point mutation stability [119].

In zero-shot prediction, the ZSH model generated residue contact maps using MSA Transformer and scored all possible double mutations of highly interacting residue pairs via a Hamiltonian scoring scheme. On α-amylase, experimental validation of only 20 double mutants (17 of which were successfully expressed) identified the optimal variant F237R/S240G with an 8.5řC increase in T_m [64]. The zero-shot strategy combining MSA Transformer with ABACUS-R achieved an 8.1řC increase in T_m with over 95% activity retention on RmZHD using only 6 validated mutations [59].

Table 6. Summary of few-shot machine learning models applied to thermostability optimization.

Few-shot strategy	Target enzyme/benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fine-tuning	Phi29 DNA polymerase	FSFP: ESM-1 backbone; MAML for initial parameters + LoRA for parameter-efficient adaptation	20	Top-20 predicted variants showed >1řC average T_m increase and 25% higher positive rate vs. ESM-1v	[67]
Fine-tuning	CALB	ActiMut-XGB: ESM2 embeddings + physicochemical + evolutionary (PSSM/PsePSSM) + positional features; XGBoost classifier	Low-N (not specified)	Achieved MCC of 0.2511, significantly outperforming PremPS (-0.0105) and DeepDDG (0.0305)	[118]
Fine-tuning	Various enzymes	SPURS: ESM2 + ProteinMPNN integrated via lightweight adapters; ∼1% parameters fine-tuned; single forward-pass stability prediction	/	$\Delta\Delta G$ predictions served as effective priors for low-N fitness prediction models	[119]
Zero-shot	α-Amylase	ZSH: MSA Transformer contact maps + Hamiltonian scoring for double mutants of high-interaction residue pairs	0 (validated 20 variants)	17 of 20 double mutants successfully expressed; optimal F237R/S240G achieved 8.5řC T_m increase	[64]

4.5. Selectivity and Substrate Specificity

Enzyme stereoselectivity, regioselectivity, and substrate specificity are core properties determining their synthetic application value. Actual enzyme engineering tasks often require simultaneous co-optimization of multiple properties including activity and selectivity. Given that selectivity-related experimental data are typically even scarcer than activity data, few-shot machine learning methods offer unique advantages in this domain.

In fine-tuning and transfer learning, Pro-PRIME evaluated mutation effects using a protein language model and integrated prior knowledge from a dataset of 68 single-point mutants. For cyclodextrinase, the optimal mutant improved the transglycosylation-to-hydrolysis (T/H) ratio by 12-fold while increasing the target product pNP-G7 yield from 63% to 98% [120]. UniESA fused embeddings from multiple pre-trained PLMs (ProteinBERT, ESM-1v, ESM-2) via ensemble learning to predict stereoselectivity and activity. On carbonyl reductase, virtual screening of only 30 mutants successfully identified 2 mutants with strict stereoselectivity (enantiomeric excess >99.5%) and 2 mutants with significantly improved activity [121]. To address substrate specificity data scarcity, a masked language modeling transfer learning framework was first pre-trained on large amounts of unlabeled LazBF substrate sequences via MLM, then transferred to LazDEF substrate classification. Using only 200 target data points, accuracy improved from approximately 72% to over 86% [122].

FusionESP employed contrastive learning with parameter-efficient fine-tuning, freezing pre-trained ESM-2 and MoLFormer while training only lightweight projection heads, achieving 93.57% accuracy in enzymesubstrate pairing prediction [123].

In active learning, Greenhalgh et al. used Gaussian processes as the sequencefunction model with the upper confidence bound (UCB) criterion to balance exploration and exploitation for iterative engineering of acyl-ACP reductase (FAR). After 10 rounds testing only 96 chimeric sequence data points, the optimal variant ATR-83 was identified, exhibiting 4.9-fold higher in vivo fatty acid production than the native sequence, demonstrating simultaneous improvement in activity and substrate selectivity [124]. pFSLNN targeted carbohydrate-active enzymes (CAZymes) by encoding structural features from a few anchor proteins and augmenting samples via Poisson distribution. With only approximately 60 samples, it achieved affinity prediction toward oligosaccharides with an F1 score of 78.67% [128].

Table 7. Summary of few-shot machine learning models applied to selectivity and substrate specificity optimization.

Few-shot strategy	Target enzyme/benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fine-tuning	Cyclodextrinase	Pro-PRIME: PLM-based mutation effect estimation; integrated prior knowledge from 68 single-point mutants	68 single-point mutants	Best mutant increased T/H ratio by 12-fold (8× transglycosylation, 33.3% less hydrolysis, 98.4% selectivity); pNP-G7 yield from 63% to 98%	[120]
Fine-tuning	Carbonyl reductase	UniESA: Multi-PLM ensemble (ProteinBERT, ESM-1v, ESM-2) for stereoselectivity and activity prediction	139	Virtual screening of 30 mutants identified 2 with strict stereoselectivity (ee >99.5%) and 2 with significantly improved activity	[121]
Fine-tuning	LazDEF	MLM pre-training on unlabeled LazBF substrates; transfer to LazDEF classification	200	Accuracy improved from ∼72% to >86% with only 200 target data points	[122]
Fine-tuning	Various enzymes	FusionESP: Frozen ESM-2 + MoLFormer with lightweight projection heads; cosine similarity for enzymesubstrate pairing	/	Achieved 93.57% accuracy; improved to 94.77% with additional phylogenetic data	[123]
Active learning	CAZymes	pFSLNN: Few-shot learning with Poisson noise augmentation; prototype network on structural features of anchor proteins	∼60	Achieved F1 score of 78.67% for oligosaccharide affinity prediction, significantly outperforming traditional DL (58.70%)	[48]
Active learning	Acyl-ACP reductase (FAR)	Gaussian process with UCB acquisition function; iterative designtestlearn cycles	96 chimeric sequences	Top variant ATR-83 showed 4.9-fold higher in vivo fatty acid production (from 11 to 54 mg/L); k_{cat} increased from 0.047 to 0.067 s^{-1}	[124]

4.6. Other Tasks

Beyond direct optimization of core catalytic properties, few-shot machine learning methods have also been widely applied to supporting tasks including product yield improvement, binding affinity prediction, accelerated discovery of superior variants, and enzyme function annotation.

For yield optimization, the ALDE framework was designed for low-sample batch protein engineering. On a non-native cyclopropanation reaction catalyzed by Pyrobaculum arsenaticum protein, ALDE started with a random sample of only 9 variants and, after four active learning rounds (96 variants each), increased the target product yield from 12% to 93%, substantially reducing time and screening costs compared to conventional directed evolution [125].

For binding affinity prediction, TrGPCR employed a dynamic TrAdaBoost algorithm to transfer samples from BindingDB (source domain) with distributions similar to GPCR data to augment the target dataset. In GPCRligand binding affinity prediction, TrGPCR achieved 5.2% lower RMSE and 4.5% lower MAE than existing models such as DeepDTA, demonstrating strong generalization capability [126].

For the efficiency of discovering superior mutants, the FolDE framework increased the probability of finding top-1% mutants by 55% across three rounds (16 variants each) [127], while EVOLVEpro demonstrated that 5 rounds of evolution matched the performance of traditional methods pre-trained with 160 variants [116]. These studies collectively demonstrate that intelligent nomination strategies can transform protein optimization into a routine and accessible task.

In supporting tasks such as enzyme function annotation and physicochemical property prediction, active learning and parameter-efficient fine-tuning have also made progress. An active learning framework integrating multiple uncertainty sampling strategies achieved comparable performance to standard training in an average of 70 epochs versus 86 epochs for EC number prediction [93]. PatchProt combined LoRA with multi-task learning, training only a minimal number of low-rank parameters, and achieved an R^2 of 0.54 for predicting the largest hydrophobic patch (LHP) of proteins, significantly outperforming traditional methods [82].

Table 8. Summary of few-shot machine learning models applied to other optimization tasks.

Optimization target	Few-shot strategy	Target enzyme	Method development	Wet-lab data size	Optimization outcome	Reference
Yield	Active learning	Pyrobaculum arsenaticum (ParPgb)	ALDE: 4-site landscape; acquisition function balancing fitness and uncertainty	$N_1 = 9$; 4×96	Yield improved from 12% to 93% for non-native cyclopropanation	[125]
Affinity	Fine-tuning	GPCR	TrGPCR: Dynamic TrAdaBoost; sample transfer from BindingDB to GLASS	/	5.2% lower RMSE and 4.5% lower MAE than DeepDTA; strong generalization across datasets	[126]
Mutant discovery	Active learning	ProteinGym benchmark	FolDE: Zero-shot naturalness ranking (Round 1); ESM embeddings + neural network (subsequent rounds)	3×16	55% increase in probability of discovering top-1% mutants	[127]
EC annotation	Active learning	Various enzymes	Ensemble uncertainty sampling (entropy, margin, BALD); stratified dataset by sequence similarity	/	Active learning reached comparable performance in ∼70 epochs vs. 86 epochs for standard training	[93]
Hydrophobic patch prediction	Fine-tuning	Various proteins	PatchProt: ESM-2 backbone; LoRA + multi-task learning with uncertainty-weighted loss	/	Achieved $R^2 = 0.54$ for LHP prediction, significantly outperforming XGBoost ($R^2 = 0.12$)	[82]

5. Challenges

While few-shot machine learning offers cost-effective opportunities for enzyme engineering, its application presents practical and computational challenges that must be addressed prior to full integration into experimental workflows. Additionally, the application of these methods to selectivity-related enzyme engineering remains relatively underexplored.

5.1. Challenges in Practical Workflows

Accessibility remains a major barrier. While some tools are available via web servers or Colab notebooks, others are only on GitHub and require substantial programming effort to adapt to target tasks. Data processing, parameter selection, model interpretation, and results analysis also demand specialized expertise. Public funding agencies should support the development and maintenance of accessible tool implementations and provide necessary training to lower these barriers.

The complexity of few-shot ML algorithms requires close collaboration with wet-lab workflows. However, limited cross-disciplinary training often leads to misunderstandings and unrealistic expectations between experimentalists and computational scientists. For instance, ideal datasets for ML may differ significantly from those readily generated in the laboratory, and proper data handling requires understanding of the experimental context. Experimental conditions often vary across screening rounds, posing challenges for computational analysis. Early discussion of requirements and expectations, including pilot experiments to assess data suitability, is critical. University curricula should be updated to reflect the growing importance of ML in life sciences and provide the interdisciplinary training needed.

5.2. Computational Cost

Despite the remarkable progress in pre-training strategies for improving data efficiency, they impose substantial computational demands. Training large-scale PLMs on datasets such as UniRef or metagenomic corpora requires high-performance computing infrastructure, typically involving distributed GPU or TPU clusters and extended training times. Although this cost is typically amortized across multiple downstream applications, it remains a significant barrier to entry [38,129]. Computational costs also arise during inference and adaptation, even with PEFT methods, when evaluating large numbers of candidate sequences. Thus, overall costs may remain high despite reduced experimental efforts.

Recent research has focused on improving computational efficiency at multiple levels. Model compression techniques, such as knowledge distillation [130] and pruning, aim to reduce model size while maintaining performance. Lightweight architectures [131] and retrieval-based methods have also been proposed. PEFT methods such as Adapters, LoRA, and Prefix Tuning are inherently computationally efficient [77,135], and some studies have enhanced transfer efficiency at the architectural level [136]. At the system level, activation checkpointing reduces memory usage by recomputing intermediate activations during backpropagation, enabling training of large models on limited GPU resources [137]. In practical applications, mechanisms such as FlashAttention have improved ESM inference efficiency by approximately 16-fold with significantly reduced memory footprint; low-bit quantization (e.g., 4-bit) can reduce memory requirements by 23-fold while maintaining performance [138].

More broadly, these developments point to a shift from maximizing predictive performance to optimizing overall resource efficiency. Future progress may depend not only on more accurate models but also on more efficient ones that operate under realistic constraints. We call for integrated evaluation frameworks that consider data, computational, and experimental costs simultaneously, rather than optimizing each dimension in isolation.

6. Conclusion and Future Directions

Few-shot machine learning is transforming enzyme engineering from a high-throughput screening-driven, experience-dependent approach to a rational design paradigm centered on computational prediction. By leveraging evolutionary and structural priors encoded in large-scale protein pre-trained models, strategies including zero-shot prediction, transfer learning, and PEFT enable reliable mutation effect prediction with minimal experimental data, significantly reducing laboratory costs. Data augmentation, generative models, and active learning further enhance the efficiency of limited data, maintaining robust generalization in few-shot settings. Recent studies have demonstrated substantial

progress in optimizing catalytic efficiency, substrate specificity, and stability, establishing few-shot ML as a foundational methodology for modern enzyme engineering.

The field is converging toward a new paradigm centered on protein foundation models, enabled by few-shot learning strategies, and operated through the DBTL cycle. In this framework, pre-trained models provide general biological knowledge representations, few-shot methods enable rapid task adaptation, and active learning with generative models efficiently explores vast sequence spaces for functional variants. As computational and experimental technologies continue to converge, ML is evolving from a predictive tool into an integrated intelligent system for enzyme design, optimization, and discovery.

Key future directions include: (1) tight integration of PLMs with automated biofoundries to realize truly closed-loop DBTL workflows [109,139]; (2) open platforms and community-driven ecosystems, such as the Open Protein Modeling Consortium, SaprotHub, and ColabSaprot, to lower technical barriers and enable broader adoption [14,141]; and (3) algorithms that jointly optimize predicted fitness and sequence diversity, such as MODIFY [91], in combination with automated high-throughput screening platforms [96,142], to improve experimental hit rates and drive efficient closed-loop optimization.

As protein foundation models, automated platforms, and open computational ecosystems continue to mature, enzyme engineering is advancing toward a highly automated and data-driven intelligent design era. In this new paradigm, ML will not only reduce experimental costs and development cycles but also deepen our understanding of the complex sequencestructurefunction relationships, accelerating the discovery and industrial application of novel biocatalysts.

References

1. Farhan, M.; Hasani, I.W.; Khafaga, D.S.R.; Ragab, W.M.; Ahmed Kazi, R.N.; Aatif, M.; Muteeb, G.; Fahim, Y.A. Enzymes as Catalysts in Industrial Biocatalysis: Advances in Engineering, Applications, and Sustainable Integration. *Catalysts* **2025**, *15*, 891, doi:10.3390/catal15090891.
2. Sawhney, R.; Ferrell, B.; Dejean, T.; Schreiber, Z.; Harrigan, W.; Polson, S.W.; Wommack, K.E. Fine-Tuning Protein Language Models Unlocks the Potential of Underrepresented Viral Proteomes.
3. 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. *Nat. Commun.* **2025**, *16*, 604, doi:10.1038/s41467-025-55944-5.
4. Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal* **2020**.
5. Yang, J.; Li, F.-Z.; Arnold, F.H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **2024**, *10*, 226–241, doi:10.1021/acscentsci.3c01275.
6. Mao, S.; Jiang, J.; Xiong, K.; Chen, Y.; Yao, Y.; Liu, L.; Liu, H.; Li, X. Enzyme Engineering: Performance Optimization, Novel Sources, and Applications in the Food Industry. *Foods* **2024**, *13*, 3846, doi:10.3390/foods13233846.
7. Zhao, J.; Zhang, C.; Luo, Y. Contrastive Fitness Learning: Reprogramming Protein Language Models for Low-N Learning of Protein Fitness Landscape.
8. Yuan, L.; Shafaei, S.; Zhao, H. Enzyme Property Prediction Using Artificial Intelligence. *Curr. Opin. Chem. Eng.* **2026**, *51*, 101208, doi:10.1016/j.coche.2025.101208.
9. Siedhoff, N.E. Machine Learning-Assisted Enzyme Engineering.
10. Vornholt, T.; Mutný, M.; Schmidt, G.W.; Schellhaas, C.; Tachibana, R.; Panke, S.; Ward, T.R.; Krause, A.; Jeschek, M. Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. *ACS Cent. Sci.* **2024**, *10*, 1357–1370, doi:10.1021/acscentsci.4c00258.
11. Adadi, A. A Survey on Dataefficient Algorithms in Big Data Era. *J. Big Data* **2021**, *8*, 24, doi:10.1186/s40537-021-00419-9.
12. Shu, J.; Xu, Z.; Meng, D. Small Sample Learning in Big Data Era.
13. Lv, F.; Zhang, J.; You, S.; Qi, W. From Traditional to AI-Driven: The Evolution of Intelligent Enzyme Engineering for Biocatalysis. *Biotechnol. Adv.* **2026**, *87*, 108788, doi:10.1016/j.biotechadv.2025.108788.
14. Sledzieski, S.; Kshirsagar, M.; Baek, M.; Dodhia, R.; Lavista Ferres, J.; Berger, B. Democratizing Protein Language Models with Parameter-Efficient Fine-Tuning. *Proc. Natl. Acad. Sci.* **2024**, *121*, e2405840121, doi:10.1073/pnas.2405840121.

15. Cao, X.; Bu, W.; Huang, S.; Zhang, M.; Tsang, I.W.; Ong, Y.S.; Kwok, J.T. A Survey of Learning on Small Data: Generalization, Optimization, and Challenge **2023**.
16. Lu, J.; Gong, P.; Ye, J.; Zhang, J.; Zhang, C. A Survey on Machine Learning from Few Samples. *Pattern Recognit.* **2023**, *139*, 109480, doi:10.1016/j.patcog.2023.109480.
17. Roger, V.; Farinas, J.; Pinquier, J. Deep Neural Networks for Automatic Speech Processing: A Survey from Large Corpora to Limited Data. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 19, doi:10.1186/s13636-022-00251-w.
18. Mohanty, A.; Sutherland, A.; Bezbradica, M.; Javidnia, H. Skin Disease Analysis With Limited Data in Particular Rosacea: A Review and Recommended Framework. *IEEE Access* **2022**, *10*, 39045–39068, doi:10.1109/ACCESS.2022.3165574.
19. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123*, 8736–8780, doi:10.1021/acs.chemrev.3c00189.
20. Peng, J.; Khuat, T.T.; Musial, K.; Gabrys, B. Machine Learning Methods for Small Data and Upstream Bioprocessing Applications: A Comprehensive Review. *Biotechnol. Adv.* **2026**, *87*, 108749, doi:10.1016/j.biotechadv.2025.108749.
21. Wittmann, B.J.; Johnston, K.E.; Wu, Z.; Arnold, F.H. Advances in Machine Learning for Directed Evolution. *Curr. Opin. Struct. Biol.* **2021**, *69*, 11–18, doi:10.1016/j.sbi.2021.01.008.
22. Minot, M.; Reddy, S.T. Meta Learning Addresses Noisy and Under-Labeled Data in Machine Learning-Guided Antibody Engineering. *Cell Syst.* **2024**, *15*, 4-18.e4, doi:10.1016/j.cels.2023.12.003.
23. Biswas, S.; Khimulya, G.; Alley, E.C.; Esvelt, K.M.; Church, G.M. Low-N Protein Engineering with Data-Efficient Deep Learning. *Nat. Methods* **2021**, *18*, 389–396, doi:10.1038/s41592-021-01100-y.
24. Freschlin, C.R.; Fahlberg, S.A.; Romero, P.A. Machine Learning to Navigate Fitness Landscapes for Protein Engineering. *Curr. Opin. Biotechnol.* **2022**, *75*, 102713, doi:10.1016/j.copbio.2022.102713.
25. Tan, Y.; Wang, R.; Wu, B.; Hong, L.; Zhou, B. From High-Throughput Evaluation to Wet-Lab Studies: Advancing Mutation Effect Prediction with a Retrieval-Enhanced Model. *Bioinformatics* **2025**, *41*, i401–i409, doi:10.1093/bioinformatics/btaf189.
26. Bank, C. Epistasis and Adaptation on Fitness Landscapes. *Annu. Rev. Ecol. Evol. Syst.* **2022**, *53*, 457–479, doi:10.1146/annurev-ecolsys-102320-112153.
27. Trinquier, J.; Uguzzoni, G.; Pagnani, A.; Zamponi, F.; Weigt, M. Efficient Generative Modeling of Protein Sequences Using Simple Autoregressive Models. *Nat. Commun.* **2021**, *12*, 5800, doi:10.1038/s41467-021-25756-4.
28. Raghavan, P.; Haas, B.C.; Ruos, M.E.; Schleinitz, J.; Doyle, A.G.; Reisman, S.E.; Sigman, M.S.; Coley, C.W. Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023**, *9*, 2196–2204, doi:10.1021/acscentsci.3c01163.
29. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531, doi:10.1093/nar/gkac1052.
30. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Adesina, A.; Ahmad, S.; Bowler-Barnett, E.H.; Bye-A-Jee, H.; Carpentier, D.; et al. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53*, D609–D617, doi:10.1093/nar/gkae1010.
31. Keith, J.A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872, doi:10.1021/acs.chemrev.1c00107.
32. Minot, M.; Reddy, S.T. Nucleotide Augmentation for Machine Learning-Guided Protein Engineering. *Bioinforma. Adv.* **2023**, *3*, vbac094, doi:10.1093/bioadv/vbac094.
33. Sun, R.; Wu, L.; Lin, H.; Huang, Y.; Li, S.Z. Enhancing Protein Predictive Models via Proteins Data Augmentation: A Benchmark and New Directions **2024**.
34. Aledo, P.; Aledo, J.C. Proteome-Wide Structural Computations Provide Insights into Empirical Amino Acid Substitution Matrices. *Int. J. Mol. Sci.* **2023**, *24*, 796, doi:10.3390/ijms24010796.
35. Han, X.; Zhang, L.; Zhou, K.; Wang, X. Deep Learning Framework DNN with Conditional WGAN for Protein Solubility Prediction.
36. Han, X.; Zhang, L.; Zhou, K.; Wang, X. ProGAN: Protein Solubility Generative Adversarial Nets for Data Augmentation in DNN Framework. *Comput. Chem. Eng.* **2019**, *131*, 106533, doi:10.1016/j.compchemeng.2019.106533.

37. Gelman, S.; Johnson, B.; Freschlin, C.; Sharma, A.; D'Costa, S.; Peters, J.; Gitter, A.; Romero, P.A. Biophysics-Based Protein Language Models for Protein Engineering **2024**.
38. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model.
39. https://www.emergentmind.com/topics/esm-series;
40. https://www.emergentmind.com/topics/prottrans;
41. Vitale, R.; Bugnon, L.A.; Fenoy, E.L.; Milone, D.H.; Stegmayer, G. Evaluating Large Language Models for Annotating Proteins.
42. Nijkamp, E.; Ruffolo, J.A.; Weinstein, E.N.; Naik, N.; Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. *Cell Syst.* **2023**, *14*, 968-978.e3, doi:10.1016/j.cels.2023.10.002.
43. Madani, A.; Krause, B.; Greene, E.R.; Subramanian, S.; Mohr, B.P.; Holton, J.M.; Olmos, J.L.; Xiong, C.; Sun, Z.Z.; Socher, R.; et al. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat. Biotechnol.* **2023**, *41*, 1099–1106, doi:10.1038/s41587-022-01618-2.
44. Cheng, X.; Chen, B.; Li, P.; Gong, J.; Tang, J.; Song, L. Training Compute-Optimal Protein Language Models **2024**.
45. Min, S.; Park, S.; Kim, S.; Choi, H.-S.; Lee, B.; Yoon, S. Pre-Training of Deep Bidirectional Protein Sequence Representations with Structural Information **2021**.
46. Wang, Y.; Wang, Z.; Sadeh, G.; Zancato, L.; Achille, A.; Karypis, G.; Rangwala, H. Long-Context Protein Language Modeling Using Bidirectional Mamba with Shared Projection Layers **2025**.
47. Munsamy, G.; Illanes-Vicioso, R.; Funcillo, S.; Nakou, I.T.; Lindner, S.; Ayres, G.; Sheehan, L.S.; Moss, S.; Eckhard, U.; Lorenz, P.; et al. Conditional Language Models Enable the Efficient Design of Proficient Enzymes **2024**.
48. Liu, S.; Kou, Y.; Chen, L. Novel Few-Shot Learning Neural Network for Predicting Carbohydrate-Active Enzyme Affinity Toward Fructo-Oligosaccharides. *J. Comput. Biol.* **2021**, *28*, 1208–1218, doi:10.1089/cmb.2021.0091.
49. Liu, J.; You, H.; Guo, Z.; Xu, Q.; Zhang, C.; Lai, L. GeoEvoBuilder: A Deep Learning Framework for Efficient Functional and Thermostable Protein Design. *Proc. Natl. Acad. Sci.* **2025**, *122*, e2504117122, doi:10.1073/pnas.2504117122.
50. Lin, W.; Zhang, J.; Zhou, Y.; Wang, M.; You, S.; Su, R.; Qi, W. Rationality-Informed Protein Evolution Enables Enhancement of PET Hydrolase Activity. *J. Hazard. Mater.* **2026**, *503*, 141130, doi:10.1016/j.jhazmat.2026.141130.
51. Li, A.; Song, Y.; Hu, J.; Zhang, H.; Li, Z.; Tang, J.; Huang, H.; Li, S.; Li, X. Integrating Graph Learning and Meta-Learning to Enhance PET Hydrolase Activity at Elevated Temperatures. *Cell Rep. Phys. Sci.* **2026**, *7*, 103037, doi:10.1016/j.xcrp.2025.103037.
52. Zhang, Z.; Chen, S.; Yang, R.; Wei, Z.; Zhang, W.; Wang, L.; Liu, Z.; Zhang, F.; Wu, J.; Pan, X.; et al. Modeling Enzyme Temperature Stability from Sequence Segment Perspective. *J. Chem. Inf. Model.* **2025**, *65*, 10932–10944, doi:10.1021/acs.jcim.5c01674.
53. Wu, F.; Wu, R.; Chen, L.; Chen, Q.; Liu, H. Zero-Shot Deep Learning with Multi-Objective Optimization Improves Thermostability of Zearalenone Hydrolase and Xylanase. *New Biotechnol.* **2026**, *92*, 51–61, doi:10.1016/j.nbt.2026.01.009.
54. Reeves, S.; Kalyaanamoorthy, S. Zero-Shot Transfer of Protein Sequence Likelihood Models to Thermostability Prediction.
55. 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 **2021**.
56. 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*, e202401542, doi:10.1002/cctc.202401542.
57. Ertelt, M.; Meiler, J.; Schoeder, C.T. Combining Rosetta Sequence Design with Protein Language Model Predictions Using Evolutionary Scale Modeling (ESM) as Restraint. *ACS Synth. Biol.* **2024**, *13*, 1085–1092, doi:10.1021/acssynbio.3c00753.
58. Tan, Y.; Wang, R.; Wu, B.; Hong, L.; Zhou, B. Retrieval-Enhanced Mutation Mastery Augmenting Zero-Shot Prediction of Protein Language Model **2024**.
59. Cheng, P.; Mao, C.; Tang, J.; Yang, S.; Cheng, Y.; Wang, W.; Gu, Q.; Han, W.; Chen, H.; Li, S.; et al. Zero-Shot Prediction of Mutation Effects with Multimodal Deep Representation Learning Guides Protein Engineering. *Cell Res.* **2024**, *34*, 630–647, doi:10.1038/s41422-024-00989-2.

60. Wang, X.; Wu, Z.; Wang, R.; Gao, X. UniproLcad: Accurate Identification of Antimicrobial Peptide by Fusing Multiple Pre-Trained Protein Language Models. *Symmetry* **2024**, *16*, 464, doi:10.3390/sym16040464.
61. Zhang, Q.; Chen, W.; Qin, M.; Wang, Y.; Pu, Z.; Ding, K.; Liu, Y.; Zhang, Q.; Li, D.; Li, X.; et al. Integrating Protein Language Models and Automatic Biofoundry for Enhanced Protein Evolution. *Nat. Commun.* **2025**, *16*, 1553, doi:10.1038/s41467-025-56751-8.
62. Massett, M.; Carr, A. Prodigy Protein: Python Package for Zero-Shot Protein Engineering Using Protein Language Models. *BMC Bioinformatics* **2025**, *26*, 298, doi:10.1186/s12859-025-06316-9.
63. Char, S.; Corley, N.; Alamdari, S.; Yang, K.K.; Amini, A.P. ProtNote: A Multimodal Method for Protein-Function Annotation **2024**.
64. Liao, M.; Feng, S.; Liu, X.; Xu, G.; Li, S.; Bai, Y.; Luo, H.; Yao, B.; Wang, H.; Tu, T. Novel Insights into Enzymatic Thermostability: The Short Board Theory and ZeroShot Hamiltonian Model. *Adv. Sci.* **2024**, *11*, 2402441, doi:10.1002/advs.202402441.
65. Deznabi, I.; Arabaci, B.; Koyutürk, M.; Tastan, O. DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases.
66. Li, F.-Z. Evaluation of the Generalizability of Machine Learning-Assisted Protein Engineering Methods.
67. Zhou, Z.; Zhang, L.; Yu, Y.; Wu, B.; Li, M.; Hong, L.; Tan, P. Enhancing Efficiency of Protein Language Models with Minimal Wet-Lab Data through Few-Shot Learning. *Nat. Commun.* **2024**, *15*, 5566, doi:10.1038/s41467-024-49798-6.
68. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function **2021**.
69. Tan, Y.; Yu, Y.; Wu, C.; Zhong, B.; Li, M.; Fan, G.; Zhu, J.; Liang, Y.; Dong, N.; Hong, L. Rank-and-Reason: Multi-Agent Collaboration Accelerates Zero-Shot Protein Mutation Prediction **2026**.
70. Fan, J.; Jiao, X.; Lin, S.; Liang, Z.; Mao, W.; Jing, C.; Chen, H.; Shen, C. Evolutionary Profiles for Protein Fitness Prediction **2025**.
71. Gordon, C.; Lu, A.X.; Abbeel, P. Protein Language Model Fitness Is a Matter of Preference **2024**.
72. Deguchi, T.; Ab Ghani, N.S.; Kurumida, Y.; Iida, S.; Kobayashi, K.; Saito, Y. Data-Efficient Protein Mutational Effect Prediction with Weak Supervision by Molecular Simulation and Protein Language Models. *Brief. Bioinform.* **2025**, *26*, bbaf536, doi:10.1093/bib/bbaf536.
73. Wang, T.; Xiang, G.; He, S.; Su, L.; Wang, Y.; Yan, X.; Lu, H. DeepEnzyme: A Robust Deep Learning Model for Improved Enzyme Turnover Number Prediction by Utilizing Features of Protein 3D-Structures. *Brief. Bioinform.* **2024**, *25*, bbae409, doi:10.1093/bib/bbae409.
74. Xu, J.; Shi, Y.; Lang, L.; Cui, T.; Zhang, Z.; Chen, G.; Qiu, J.; Heng, P.-A. InstructPLM-Mu: 1-Hour Fine-Tuning of ESM2 Beats ESM3 in Protein Mutation Predictions **2026**.
75. AdapterHub: A Framework for Adapting Transformers.
76. Wang, Z.; Ma, Z.; Cao, Z.; Zhou, C.; Zhang, J.; Gao, Y.Q. Prot2Chat: Protein Large Language Model with Early Fusion of Text, Sequence, and Structure. *Bioinformatics* **2025**, *41*, btaf396, doi:10.1093/bioinformatics/btaf396.
77. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models **2021**.
78. Zhang, Z.; Zhou, Y.; Zheng, J.; Feng, C.; Cui, S.; Wang, S.; Li, Z. Boost Protein Language Model with Injected Structure Information Through Parameter Efficient Fine-Tuning. *Comput. Biol. Med.* **2025**, *195*, 110607, doi:10.1016/j.compbiomed.2025.110607.
79. Guo, L.-X.; Wang, L.; You, Z.-H.; Yu, C.-Q.; Hu, M.-L.; Zhao, B.-W.; Li, Y. Likelihood-Based Feature Representation Learning Combined with Neighborhood Information for Predicting circRNAmiRNA Associations. *Brief. Bioinform.* **2024**, *25*, bbae020, doi:10.1093/bib/bbae020.
80. Dieckhaus, H.; Brocidiacono, M.; Randolph, N.Z.; Kuhlman, B. Transfer Learning to Leverage Larger Datasets for Improved Prediction of Protein Stability Changes. *Proc. Natl. Acad. Sci.* **2024**, *121*, e2314853121, doi:10.1073/pnas.2314853121.
81. Wang, X.; Zhou, J.; Mueller, J.; Quinn, D.; Carvalho, A.; Moody, T.S.; Huang, M. BioStructNet: Structure-Based Network with Transfer Learning for Predicting Biocatalyst Functions. *J. Chem. Theory Comput.* **2025**, *21*, 474–490, doi:10.1021/acs.jctc.4c01391.
82. Gogishvili, D.; Minois-Genin, E.; Van Eck, J.; Abeln, S. PatchProt: Hydrophobic Patch Prediction Using Protein Foundation Models. *Bioinforma. Adv.* **2024**, *4*, vbae154, doi:10.1093/bioadv/vbae154.
83. Zhang, L.; Luo, K.; Zhou, Z.; Yu, Y.; Jiang, F.; Wu, B.; Li, M.; Hong, L. A Deep Retrieval-Enhanced Meta-Learning Framework for Enzyme Optimum pH Prediction. *J. Chem. Inf. Model.* **2025**, *65*, 3761–3770, doi:10.1021/acs.jcim.4c02291.

84. Lu, Q.; Zhang, R.; Zhou, H.; Ni, D.; Xiao, W.; Li, J. MetaHMEI: Meta-Learning for Prediction of Few-Shot Histone Modifying Enzyme Inhibitors. *Brief. Bioinform.* **2023**, *24*, bbad115, doi:10.1093/bib/bbad115.
85. Li, Z.; Xu, Z.; Han, L.; Gao, Y.; Wen, S.; Liu, D.; Wang, H.; Metaxas, D.N. Implicit In-Context Learning **2025**.
86. Lu, J.; Teehan, R.; Yang, Z.; Ren, M. Context Tuning for In-Context Optimization **2025**.
87. Teufel, F.; Kollasch, A.W.; Huang, Y.; Winther, O.; Yang, K.K.; Notin, P.; Marks, D.S. Few-Shot Protein Fitness Prediction via In-Context Learning and Test-Time Training **2025**.
88. Luo, Y.; Jiang, G.; Yu, T.; Liu, Y.; Vo, L.; Ding, H.; Su, Y.; Qian, W.W.; Zhao, H.; Peng, J. ECNet Is an Evolutionary Context-Integrated Deep Learning Framework for Protein Engineering. *Nat. Commun.* **2021**, *12*, 5743, doi:10.1038/s41467-021-25976-8.
89. Feng, R.; Zhang, C.; Zhang, Y. Large Language Models for Biomolecular Analysis: From Methods to Applications. *TrAC Trends Anal. Chem.* **2024**, *171*, 117540, doi:10.1016/j.trac.2024.117540.
90. Reis, D.R.; Santos, B.C.; Bleicher, L.; Zárate, L.E.; Nobre, C.N. Prediction of Enzymatic Function with High Efficiency and a Reduced Number of Features Using Genetic Algorithm. *Comput. Biol. Med.* **2023**, *158*, 106799, doi:10.1016/j.compbiomed.2023.106799.
91. Ding, K.; Chin, M.; Zhao, Y.; Huang, W.; Mai, B.K.; Wang, H.; Liu, P.; Yang, Y.; Luo, Y. Machine Learning-Guided Co-Optimization of Fitness and Diversity Facilitates Combinatorial Library Design in Enzyme Engineering. *Nat. Commun.* **2024**, *15*, 6392, doi:10.1038/s41467-024-50698-y.
92. Cohn, D.A.; Ghahramani, Z.; Jordan, M.I. Active Learning with Statistical Models.
93. Babjac, A.; Hoarfrost, A. An Active Learning Framework for Data-Efficient, Human-in-the-Loop Enzyme Function Prediction **2026**.
94. Du, Q.; Wang, H.; Jiang, B.; Wang, X. Advancing Genetic Engineering with Active Learning: Theory, Implementations and Potential Opportunities. *Brief. Bioinform.* **2025**, *26*, bbaf286, doi:10.1093/bib/bbaf286.
95. Cohn, D. Improving Generalization with Active Learning.
96. Hu, R.; Fu, L.; Chen, Y.; Chen, J.; Qiao, Y.; Si, T. Protein Engineering via Bayesian Optimization-Guided Evolutionary Algorithm and Robotic Experiments. *Brief. Bioinform.* **2023**, *24*, bbac570, doi:10.1093/bib/bbac570.
97. Kakuzaki, T.; Koga, H.; Takizawa, S.; Metsugi, S.; Shiraiwa, H.; Sampei, Z.; Yoshida, K.; Tsunoda, H.; Teramoto, R. Monte Carlo Thompson Sampling-Guided Design for Antibody Engineering. *mAbs* **2023**, *15*, 2244214, doi:10.1080/19420862.2023.2244214.
98. Danziger, S.A.; Baronio, R.; Ho, L.; Hall, L.; Salmon, K.; Hatfield, G.W.; Kaiser, P.; Lathrop, R.H. Predicting Positive P53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning. *PLoS Comput. Biol.* **2009**, *5*, e1000498, doi:10.1371/journal.pcbi.1000498.
99. Chu, H.Y.; Fong, J.H.C.; Thean, D.G.L.; Zhou, P.; Fung, F.K.C.; Huang, Y.; Wong, A.S.L. Accurate Top Protein Variant Discovery via Low-N Pick-and-Validate Machine Learning. *Cell Syst.* **2024**, *15*, 193-203.e6, doi:10.1016/j.cels.2024.01.002.
100. Ding, K.; Chin, M.; Zhao, Y.; Huang, W.; Mai, B.K.; Wang, H.; Liu, P.; Yang, Y.; Luo, Y. Machine Learning-Guided Co-Optimization of Fitness and Diversity Facilitates Combinatorial Library Design in Enzyme Engineering. *Nat. Commun.* **2024**, *15*, 6392, doi:10.1038/s41467-024-50698-y.
101. Greenman, K.P.; Amini, A.P.; Yang, K.K. Benchmarking Uncertainty Quantification for Protein Engineering. *PLOS Comput. Biol.* **2025**, *21*, e1012639, doi:10.1371/journal.pcbi.1012639.
102. Fan, L.; Wang, H.; Gao, H.; Ding, Y.; Zhao, J.; Luo, H.; Tu, T.; Wu, N.; Yao, B.; Guan, F.; et al. Rational Protein Engineering Using an Omni-Directional Multipoint Mutagenesis Generation Pipeline. *iScience* **2025**, *28*, 113273, doi:10.1016/j.isci.2025.113273.
103. Kmicikiewicz, M.; Fortuin, V.; Szczurek, E. ProSpero: Active Learning for Robust Protein Design Beyond Wild-Type Neighborhoods **2025**.
104. Yu, Y.; Jiang, F.; Ma, X.; Zhang, L.; Zhong, B.; Ouyang, W.; Fan, G.; Yu, H.; Hong, L.; Li, M. Venus-MAXWELL: Efficient Learning of Protein-Mutation Stability Landscapes Using Protein Language Models.
105. Li, M.; Kang, L.; Xiong, Y.; Wang, Y.G.; Fan, G.; Tan, P.; Hong, L. SESNet: Sequence-Structure Feature-Integrated Deep Learning Method for Data-Efficient Protein Engineering. *J. Cheminformatics* **2023**, *15*, 12, doi:10.1186/s13321-023-00688-x.
106. Liu, H.; Guan, F.; Liu, T.; Yang, L.; Fan, L.; Liu, X.; Luo, H.; Wu, N.; Yao, B.; Tian, J.; et al. MECE: A Method for Enhancing the Catalytic Efficiency of Glycoside Hydrolase Based on Deep Neural Networks and Molecular Evolution. *Sci. Bull.* **2023**, *68*, 2793–2805, doi:10.1016/j.scib.2023.09.039.
107. Romero, P.A.; Krause, A.; Arnold, F.H. Navigating the Protein Fitness Landscape with Gaussian Processes. *Proc. Natl. Acad. Sci.* **2013**, *110*, doi:10.1073/pnas.1215251110.

108. Luo, D.; Ji, H.; Hu, B.; Cai, J.; Wen, K.; Qu, X.; Cao, M.; Zhao, X.; Wang, B. A Multimodal Ensemble Framework for Optimal Mutant Prediction and Computational Enzyme Engineering. *Angew. Chem. Int. Ed.* **2026**, *65*, e21396, doi:10.1002/anie.202521396.
109. Fei, H.; Li, Y.; Liu, Y.; Wei, J.; Chen, A.; Gao, C. Advancing Protein Evolution with Inverse Folding Models Integrating Structural and Evolutionary Constraints. *Cell* **2025**, *188*, 4674-4692.e19, doi:10.1016/j.cell.2025.06.014.
110. Qian, H.; Wang, Y.; Zhou, X.; Gu, T.; Wang, H.; Lyu, H.; Li, Z.; Li, X.; Zhou, H.; Guo, C.; et al. ESM-Ezy: A Deep Learning Strategy for the Mining of Novel Multicopper Oxidases with Superior Properties. *Nat. Commun.* **2025**, *16*, 3274, doi:10.1038/s41467-025-58521-y.
111. Jiang, F.; Li, M.; Dong, J.; Yu, Y.; Sun, X.; Wu, B.; Huang, J.; Kang, L.; Pei, Y.; Zhang, L.; et al. A General Temperature-Guided Language Model to Design Proteins of Enhanced Stability and Activity. *Sci. Adv.* **2024**, *10*, eadr2641, doi:10.1126/sciadv.adr2641.
112. 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. *Nat. Commun.* **2025**, *16*, 2736, doi:10.1038/s41467-025-58038-4.
113. Cai, Y.; Ge, F.; Zhang, C.; Chen, H.; Qi, X.; Liao, X.; Wang, L. Enhancing Kcat Prediction through Residue-Aware Attention Mechanism and Pre-Trained Representations. *Commun. Biol.* **2026**, *9*, 273, doi:10.1038/s42003-026-09551-9.
114. Alwer, S.; Fleming, R. KinForm: Kinetics Informed Feature Optimised Representation Models for Enzyme k_{cat} and K_M Prediction **2025**.
115. Wu, H.; Nie, Z.; Zhang, H.; Ren, Z. Pseudodata-Guided Invariant Representation Learning Boosts the Out-of-Distribution Generalization in Enzymatic Kinetic Parameter Prediction **2026**.
116. Jiang, K.; Yan, Z.; Di Bernardo, M.; Sgrizzi, S.R.; Villiger, L.; Kayabolen, A.; Kim, B.J.; Carscadden, J.K.; Hiraizumi, M.; Nishimasu, H.; et al. Rapid in Silico Directed Evolution by a Protein Language Model with EVOLVEpro. *Science* **2025**, *387*, eadr6006, doi:10.1126/science.adr6006.
117. Miao, L.; Zheng, Y.; Cheng, R.; Liu, J.; Zheng, Z.; Yang, H.; Zhao, J. Efficient Synthesis of -Aminobutyric Acid from Monosodium Glutamate Using an Engineered Glutamate Decarboxylase Active at a Neutral pH. *Catalysts* **2024**, *14*, 905, doi:10.3390/catal14120905.
118. Jiang, Y.; Huang, S.; Chen, H.-F. ActiMut-XGB: Predicting Thermodynamic Stability of Point Mutations for CALB with Protein Language Model. *Int. J. Biol. Macromol.* **2025**, *317*, 144609, doi:10.1016/j.ijbiomac.2025.144609.
119. Li, Z.; Luo, Y. Generalizable and Scalable Protein Stability Prediction with Rewired Protein Generative Models. *Nat. Commun.* **2025**, *17*, 891, doi:10.1038/s41467-025-67609-4.
120. Nie, T.; Yan, Z.; Liu, H.; Zhang, X.; Zhong, W.; Chen, Q.; Zhu, C.; Hong, L.; Yang, G. Protein Language Model-Assisted Directed Evolution of Cyclodextrinase Enables Precision -O-Oligosaccharide Synthesis. *Bioresour. Technol.* **2025**, *438*, 133206, doi:10.1016/j.biortech.2025.133206.
121. Weng, C.-Y.; Li, J.; Chen, Q.-L.; Han, J.-Y.; Dong, Z.-T.; Liu, Z.-Q.; Zheng, Y.-G. UniESA: A Unified Data-Driven Framework for Enzyme Stereoselectivity and Activity Prediction. *Green Chem.* **2025**, *27*, 3777–3788, doi:10.1039/D4GC06253A.
122. Clark, J.D.; Mi, X.; Mitchell, D.A.; Shukla, D. Substrate Prediction for RiPP Biosynthetic Enzymes via Masked Language Modeling and Transfer Learning. *Digit. Discov.* **2025**, *4*, 343–354, doi:10.1039/D4DD00170B.
123. Du, Z.; Fu, W.; Guo, X.; Caragea, D.; Li, Y. FusionESP: Improved EnzymeSubstrate Pair Prediction by Fusing Protein and Chemical Knowledge. *J. Chem. Inf. Model.* **2025**, *65*, 2806–2817, doi:10.1021/acs.jcim.4c02357.
124. Greenhalgh, J.C.; Fahlberg, S.A.; Pfleger, B.F.; Romero, P.A. Machine Learning-Guided Acyl-ACP Reductase Engineering for Improved in Vivo Fatty Alcohol Production. *Nat. Commun.* **2021**, *12*, 5825, doi:10.1038/s41467-021-25831-w.
125. Yang, J.; Lal, R.G.; Bowden, J.C.; Astudillo, R.; Hameedi, M.A.; Kaur, S.; Hill, M.; Yue, Y.; Arnold, F.H. Active Learning-Assisted Directed Evolution.
126. Lu, Y.; Zhang, R.; Jiang, T.; Fu, Q.; Cui, Z.; Wu, H. TrGPCR: GPCRLigand Binding Affinity Prediction Based on Dynamic Deep Transfer Learning. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 1613–1624, doi:10.1109/JBHI.2023.3307928.
127. Roberts, J.B.; Ji, C.R.; Donnell, I.; Young, T.D.; Pearson, A.N.; Hudson, G.A.; Keiser, L.S.; Wesselkamper, M.; Winegar, P.H.; Ludwig, J.; et al. Low-N Protein Activity Optimization with FolDE **2025**.

128. Liu, S.; Kou, Y.; Chen, L. Novel Few-Shot Learning Neural Network for Predicting Carbohydrate-Active Enzyme Affinity Toward Fructo-Oligosaccharides. *J. Comput. Biol.* **2021**, *28*, 1208–1218, doi:10.1089/cmb.2021.0091.
129. Bileschi, M.L.; Belanger, D.; Bryant, D.H.; Sanderson, T.; Carter, B.; Sculley, D.; Bateman, A.; DePristo, M.A.; Colwell, L.J. Using Deep Learning to Annotate the Protein Universe. *Nat. Biotechnol.* **2022**, *40*, 932–937, doi:10.1038/s41587-021-01179-w.
130. Shang, J.; Peng, C.; Ji, Y.; Guan, J.; Cai, D.; Tang, X.; Sun, Y. Accurate and Efficient Protein Embedding Using Multi-Teacher Distillation Learning. *Bioinformatics* **2024**, *40*, btae567, doi:10.1093/bioinformatics/btae567.
131. Jendrusch, M.; Korbel, J.O.; Jendrusch, M.; Korbel, J.O.; Jendrusch, M.; Korbel, J.O. Efficient Protein Structure Generation with Sparse Denoising Models **2025**.
132. Vieira, L.C.; Handojo, M.L.; Wilke, C.O. Scaling Down for Efficiency: Medium-Sized Transformer Models for Protein Sequence Transfer Learning.
133. Wang, X.; Yin, X.; Jiang, D.; Zhao, H.; Wu, Z.; Zhang, O.; Wang, J.; Li, Y.; Deng, Y.; Liu, H.; et al. Multi-Modal Deep Learning Enables Efficient and Accurate Annotation of Enzymatic Active Sites. *Nat. Commun.* **2024**, *15*, 7348, doi:10.1038/s41467-024-51511-6.
134. Zaretckii, M.; Buslaev, P.; Kozlovskii, I.; Morozov, A.; Popov, P. Approaching Optimal pH Enzyme Prediction with Large Language Models. *ACS Synth. Biol.* **2024**, *13*, 3013–3021, doi:10.1021/acssynbio.4c00465.
135. Mahabadi, R.K.; Henderson, J.; Ruder, S. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers **2021**.
136. Wang, L.; Xue, Z.; Wang, Y. Exploring ProtFlash: An Efficient Approach to Protein Data Analysis. In *Large Language Models (LLMs) in Protein Bioinformatics*; Kc, D.B., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2025; Vol. 2941, pp. 59–70 ISBN 978-1-0716-4622-9.
137. https://www.deepspeed.ai/tutorials/zero-offload/;
138. Çelik, M.H.; Xie, X. Efficient Inference, Training, and Fine-Tuning of Protein Language Models. *iScience* **2025**, *28*, 113495, doi:10.1016/j.isci.2025.113495.
139. Singh, N.; Lane, S.; Yu, T.; Lu, J.; Ramos, A.; Cui, H.; Zhao, H. A Generalized Platform for Artificial Intelligence-Powered Autonomous Enzyme Engineering. *Nat. Commun.* **2025**, *16*, 5648, doi:10.1038/s41467-025-61209-y.
140. He, Y.; Zhou, X.; Yuan, F.; Chang, X. Protocol to Use Protein Language Models Predicting and Following Experimental Validation of Function-Enhancing Variants of Thymine-N-Glycosylase. *STAR Protoc.* **2024**, *5*, 103188, doi:10.1016/j.xpro.2024.103188.
141. Özsari, G.; GarcíaSoriano, D.A.; Parate, S.; El Issaoui, A.; WittungStafshede, P. CAPIM Catalytic Activity and Site Prediction and Analysis Tool in Multimer Proteins. *Protein Sci.* **2025**, *34*, e70347, doi:10.1002/pro.70347.
142. Helleckes, L.M.; Küsters, K.; Wagner, C.; Hamel, R.; Saborowski, R.; Marienhagen, J.; Wiechert, W.; Oldiges, M. High-Throughput Screening of Catalytically Active Inclusion Bodies Using Laboratory Automation and Bayesian Optimization. *Microb. Cell Factories* **2024**, *23*, 67, doi:10.1186/s12934-024-02319-y.