Review

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

Yumeng Zhang [1], Nian Li [1], Zhiqi Ye [1], Jiaxuan Li [1], Bin Qin [1,*]

[1] Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang, Liaoning, China.
* Correspondence: binqin@syphu.edu.cn

Abstract

Enzyme engineering is fundamentally constrained by the scarcity of high-quality sequence–function data. The reliance of traditional machine learning models on massive labeled data further exacerbates this dilemma. This review systematically surveys machine learning methods under small-sample wet-lab data conditions, covering three technical routes: zero-shot prediction of mutation effects, few-shot transfer learning (parameter-efficient fine-tuning, meta-learning, and in-context learning), and active learning-guided directed evolution. On this basis, we summarize the applications of these methods in fitness, enzyme kinetic parameters, thermostability, selectivity and substrate specificity, and other tasks, categorized by enzyme properties, and evaluate their effectiveness from the perspectives of data efficiency, optimization magnitude, and experimental validation scale. We further discuss the challenges of collaboration, tool accessibility, and computational cost in practical workflows, and finally outline future directions. This review provides a methodological reference for enzyme engineering researchers working under limited experimental data, and offers a clear and comprehensive theoretical framework for methodology developers to build more powerful solutions.

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

---

Contents

1. Introduction **2**

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

3. Methodology and Research Progress of Few-Shot Machine Learning **5**
3.1. Protein Pre-trained Models 5
3.2. Zero-Shot Prediction of Mutation Effects 8
3.3. Few-Shot Transfer Learning of Pre-trained Models 8
3.4. Active Learning-Guided Directed Evolution 10

4. Applications of Few-Shot Machine Learning in Different Enzyme Engineering Tasks **11**
4.1. Fitness 11
4.2. Activity and Thermostability 13
4.3. Enzyme Kinetic Parameters 14
4.4. Thermostability 14
4.5. Selectivity and Substrate Specificity 15
4.6. Other Tasks 15

5. Challenges **16**
5.1. Collaboration and Tool Accessibility Challenges 16
5.2. High Computational Cost Challenges 17

6. Conclusion and Future Directions **17**

7. References **19**

GENERAL SEQUENCE

General Sequence

1. Introduction

Enzymes are highly efficient and selective biocatalysts in nature, playing a central role in green chemistry, biomanufacturing, and the pharmaceutical industry [1]. Engineering natural enzymes through rational design or directed evolution can enhance their catalytic efficiency, stability, and substrate adaptability, thereby advancing the development of novel biocatalytic systems. In recent years, machine learning (ML) has provided a new research paradigm for enzyme engineering [2,3]. By capturing nonlinear interactions among residues, cofactors, and substrates, ML can model complex enzyme behaviors, propose rational mutations, and predict beneficial variants, outperforming traditional methods [4–6].

However, the performance of mainstream ML methods relies on massive, high-quality sequence–function data, which conflicts with the small-data reality in enzyme engineering [7,8]. Unlike data-rich fields such as natural language processing or computer vision, enzyme functional data mainly come from time-consuming and expensive wet-lab experiments. A typical enzyme engineering workflow follows the design–build–test–learn (DBTL) cycle, where each iteration involves gene construction, protein expression, purification, and functional assays, resulting in long experimental cycles and limited throughput [9]. Moreover, key phenotypes such as catalytic efficiency, substrate selectivity, and kinetic parameters often depend on complex biochemical assays or chromatographic analysis, further limiting large-scale data acquisition [10]. Consequently, practical enzyme engineering datasets typically contain only tens to thousands of variants, far below the scale required by ML models.

To alleviate this issue, researchers have recently proposed a series of few-shot learning (FSL) strategies, i.e., machine learning strategies that can achieve high predictive performance with limited labeled data [11,12], as shown in Figure 2. On one hand, protein models pre-trained on large-scale public data or low-fidelity data, such as protein language models (PLMs), can learn universal sequence representations for zero-shot prediction [13], and can be transferred to the current task with a small amount of experimental data to improve predictive performance [14]. On the other hand, active learning-guided directed evolution treats enzyme engineering as a supervised learning problem, iteratively improving the information efficiency of experimental data, thereby enabling more reliable model training under limited data conditions [10].

Given the importance of addressing the small-data problem, effective learning from limited data has long been a key topic in the machine learning community [15,16]. In recent years, many related reviews have emerged, covering applications of FSL in automatic speech processing [17], skin disease analysis [18], molecular science [19], and upstream bioprocessing [20]. However, a systematic review specifically targeting the small-data problem in enzyme engineering is still lacking. This review fills this gap and provides practical guidance for solving the small-data problem in enzyme engineering.

The remainder of this paper is organized as follows: (1) we discuss the current status of small data in enzyme engineering; (2) we introduce the methodologies and research progress of few-shot machine learning; (3) we summarize recent applications of few-shot machine learning in enzyme engineering; and (4) we conclude with current challenges and future perspectives.

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

The core of enzyme engineering lies in optimizing enzyme properties. As shown in Figure 1, classical machine learning methods typically use sequence–function data obtained from wet-lab experiments as datasets and train models through supervised learning to achieve enzyme prediction and evolution. However, the limited availability of high-quality sequence–function data remains a core bottleneck [21]. Functional data for a specific enzyme–substrate pair almost entirely depend on experimental generation, and such data are inherently costly, low-throughput, and noisy [22]. Consequently, typical datasets usually contain only tens to thousands of variants, far below the training data scale typically required by machine learning models, severely limiting model training and generalization capabilities [23–25].

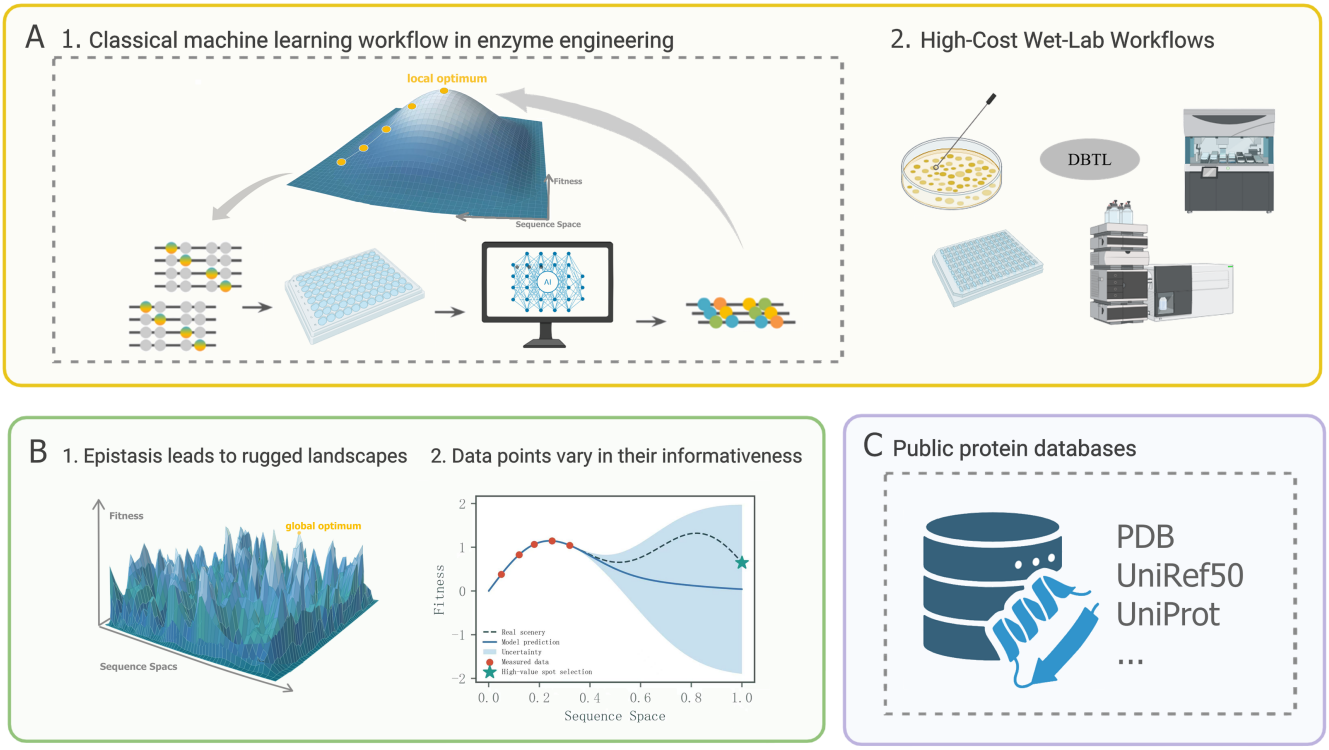


Figure 1. The small-data landscape in enzyme engineering. (A) Classical machine learning methods typically use sequence–function data obtained from wet-lab experiments as datasets and train models through supervised learning to achieve enzyme prediction and evolution. However, functional data for a specific enzyme–substrate pair rely almost entirely on experimental generation, requiring substantial time and cost. (B) Moreover, pervasive epistasis causes fitness landscapes to exhibit rugged and nonlinear features that are difficult to approximate from limited observations. Due to difficulties in data acquisition, datasets often fail to contain high-fitness and high-information data points. (C) However, it is worth emphasizing that large amounts of unlabeled sequence and structural data are available in many protein databases.

The intrinsic complexity of protein fitness landscapes further exacerbates this challenge. Due to pervasive epistasis [26], the functional effects of mutations are highly dependent on the sequence context, causing fitness landscapes to exhibit rugged and nonlinear features that are difficult to approximate from limited observations [27]. Therefore, models trained on sparse datasets can achieve reasonable interpolation within the training distribution but cannot generalize to unseen sequence combinations. In addition, due to difficulties in data acquisition, datasets are often imbalanced [28]. For example, when targeting a specific substrate, the obtained variants may exhibit only low-to-moderate fitness, leading to a lack of high-fitness data points, which limits the model's ability to predict high-fitness variants. These factors collectively constrain the ability of traditional machine learning models to learn accurate and generalizable representations from small datasets.

However, it is worth emphasizing that although truly supervised sequence–function 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 available and are summarized in Table 1. Meanwhile, low-fidelity 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, this strategy can also be used to efficiently obtain low-fidelity data from protein zero-shot predictors, which will be detailed later. A more aggressive strategy is to use data augmentation algorithms to expand sequences, such as exploiting the redundancy of the genetic code [32], replacing non-critical residues [33], and generating evolutionarily plausible mutations based on natural occurrence [34]. These methods expand the effective training

set of existing sequences through biologically meaningful perturbations, and some promising studies have also employed generative models for sequence augmentation [35,36]. Fully utilizing these non-experimental data is of great help for engineering specific enzyme–substrate properties.

In summary, whether extracting information from public data or low-fidelity data and generalizing universal representations to target tasks, or maximizing the information efficiency of limited wet-lab data to alleviate enzyme data sparsity and imbalance and ensure that models are exposed to diverse data points, both will help achieve enzyme optimization under limited wet-lab 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	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
PDB	Experimentally resolved protein structures	∼200,000 structures	General (especially for thermostability, hydrophobic patch prediction)	Provides structural constraints; training structure encoders (e.g., PatchProt, GeoEvoBuilder)	https://www.rcsb.org
BRENDA	Enzyme kinetic parameters, optimal temperature, stability	Tens of thousands	Various enzymes	Provides temperature stability priors; assists zero-shot Hamiltonian models	https://www.brenda-enzymes.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
FOS-related enzyme datasets	Affinity of carbohydrate-active enzymes toward FOS	Small scale (constructed in literature)	Glycoside hydrolases, fructosyltransferases	Benchmark dataset for few-shot learning	[142]
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	[106,110,111]
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://gpcrdb.org
Temperature stability segment dataset (constructed in literature)	Enzyme sequence fragments and temperature stability labels	Small sample	Various enzymes	Modeling temperature stability from sequence segment perspective	[112]
Zearalenone hydrolase / Xylanase mutant datasets	Mutant sequences and thermostability (T_m)	Small scale	Zearalenone hydrolase, xylanase	Test cases for zero-shot deep learning + multi-objective optimization	[143]
DLKcat	Enzyme kinetic parameters (k_{cat}), protein sequences, substrate SMILES strings	∼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	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

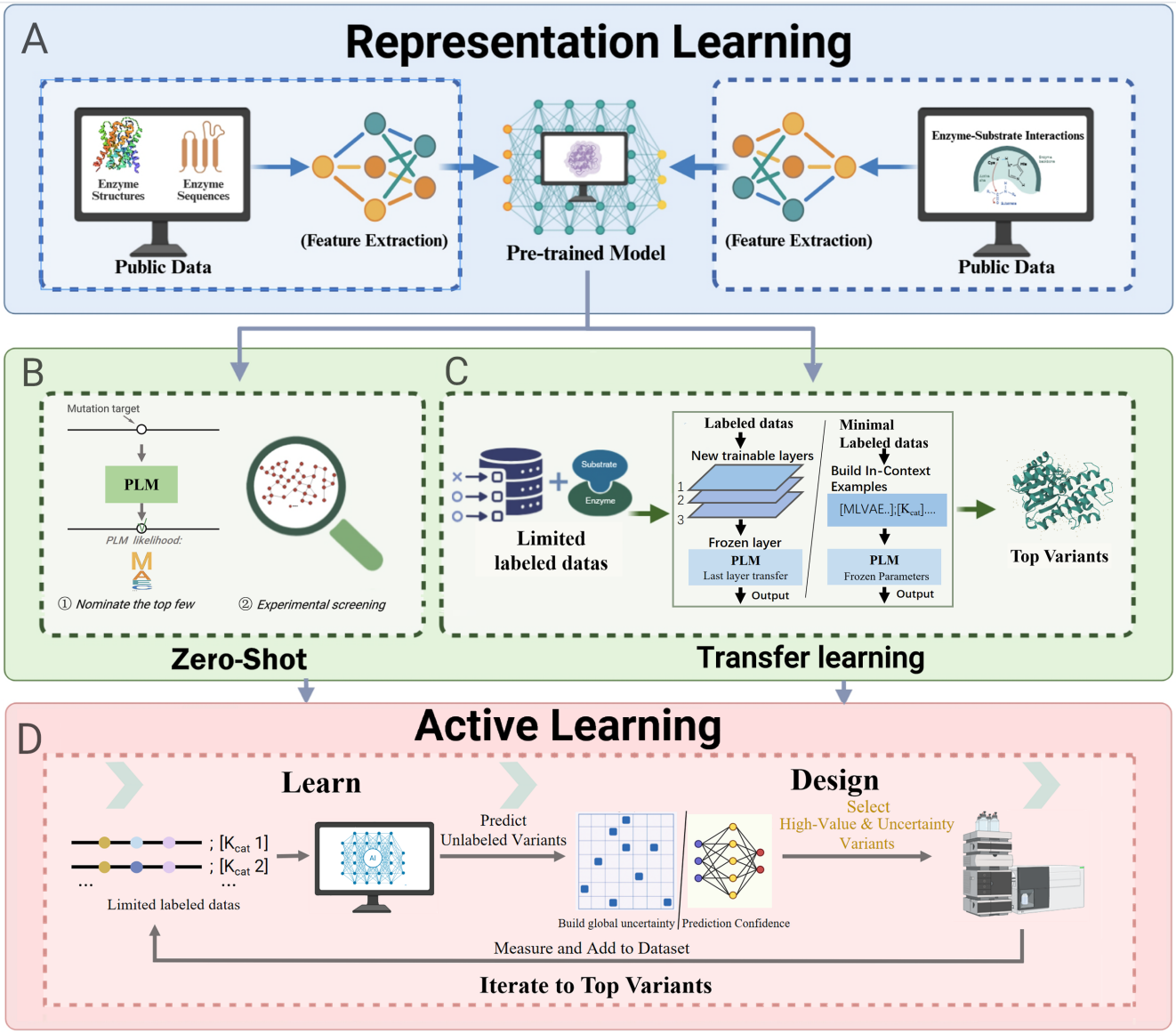


Figure 2. Architecture of few-shot machine learning methods. (A) Pre-trained models (e.g., protein language models) learn general feature representations from massive unlabeled sequence data or low-fidelity data, encoding sequence, structure, evolutionary, and functional information that can be transferred to downstream enzyme engineering tasks. (B) Pre-trained models are directly used as zero-shot predictors to assess the consistency of mutant sequences with the statistical patterns of natural sequences, enabling inference of mutational functional effects without experimental labeled data and allowing preliminary screening of promising mutants. (C) Transfer learning adapts pre-trained models to a target task using a small amount of experimentally labeled data, including parameter-efficient fine-tuning (PEFT) and in-context learning without parameter modification, significantly improving prediction accuracy compared with zero-shot methods. (D) A small subset is randomly sampled from a variant library for experiments, and a supervised model with uncertainty quantification is trained on the experimental data to predict the remaining unmeasured variants. An acquisition function is then used to nominate variants with both high uncertainty and high value, which are experimentally measured and fed back for retraining, iterating until top-tier variants are reached.

3. Methodology and Research Progress of Few-Shot Machine Learning

3.1. Protein Pre-trained Models

Protein pre-trained models extract universal protein representations from large-scale public data via representation learning. The value of such models lies in the fact that their learned representations can be efficiently transferred to downstream tasks such as mutation effect prediction and enzyme function optimization [5]. In scenarios where wet-lab data are extremely limited, these models

can significantly improve prediction accuracy and generalization, making them a key technology for overcoming the sequence–function data sparsity bottleneck in enzyme engineering. Representative protein pre-trained models are listed in Table 2. Models trained on large amounts of public sequence and structure data are general-purpose models, whereas those trained on specific datasets focus on particular tasks.

The mainstream tools are protein language models (PLMs), whose training methods are mainly derived from language modeling techniques in natural language processing and adapted to protein sequence characteristics. In this framework, protein sequences are analogized to natural language, with amino acids as discrete symbols 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], so most PLMs are general-purpose models applicable to various enzyme tasks. Importantly, this representation has been shown to implicitly encode structural, functional, and evolutionary information [38].

Masked language modeling is one of the most widely adopted pre-training strategies, with representative models including the ESM series [39,40] and the ProtTrans framework [41]. In this approach, a fraction of amino acid residues (typically about 15%) are randomly masked, and the model predicts these 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 [42]. Autoregressive modeling represents another major category of protein pre-training strategies. In this framework, the model sequentially predicts each amino acid given all preceding residues, thereby learning the joint probability distribution of protein sequences. Compared with masked language models, autoregressive models are particularly suited for generation tasks because they naturally support sequence generation. The ProGen series exemplifies the capability of this approach, demonstrating the ability to generate entirely new protein sequences with experimentally validated functions [43,44]. By learning from large-scale 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 can significantly improve performance on both generation and prediction tasks, indicating a strong scaling effect in protein language modeling [45].

With the continuous expansion of model architectures, recent work has extended protein pre-training to multimodal settings, integrating sequence, structure, and functional annotations into unified learning frameworks [46]. Models such as ESM-3 embody this trend, enabling joint representation learning across multiple biological modalities [47], and can capture higher-order relationships that are difficult to infer from sequence alone. In addition, emerging methods such as long-context modeling and graph-based representations further enhance the ability to simulate complex biological systems [48].

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 such as UniProt and PDB according to the properties of the target enzyme, constructing a customized pre-training corpus, or directly using public datasets. This "task-oriented pre-training" enables the model to be exposed to sequence patterns highly homologous to downstream tasks during pre-training, achieving better performance on specific tasks. For example, the ZymCTRL model was pre-trained specifically on over 37 million enzyme sequences with Enzyme Commission (EC) number annotations in UniProt, and its effectiveness was experimentally validated [49]. On the other hand, low-fidelity data can also serve as an effective pre-training resource (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 experiments showed that 1,000 simulated data plus 320 experimental data achieved comparable performance to 8,000 simulated data plus 80 experimental data, validating 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 data	Zero-shot	Reference
General protein language models						
SeqVec	biLSTM (ELMo)	Bidirectional LM	UniRef50	Sequence	No	https://github.com/mheinzinger/SeqVec
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
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	MLM	Autoregressive LM	UniRef50	Sequence	No	https://github.com/microsoft/protein-sequence-models
RITA	Transformer	Autoregressive LM	UniRef-100	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
Structure-aware models						
ProSST	Sequence-structure transformer	Multimodal learning	UniRef + PDB + AlphaFold DB	Sequence + structure	Yes	https://github.com/ai4protein/ProSST
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 + structure	Yes	https://github.com/westlake-repl/SaProt
ProteinMPNN	Inverse folding (graph neural network)	Autoregressive LM (structure-conditioned)	PDB	Structure	Yes	https://github.com/dauparas/ProteinMPNN
ESM-IF1	Transformer inverse folding model	Structure-conditioned masked language modeling (span masking)	PDB + AlphaFold	Structure	Yes	https://github.com/facebookresearch/esm
GearNet	Graph neural network	Structure representation learning	PDB + AlphaFold	Structure (residue relation graph)	No	https://github.com/DeepGraphLearning/GearNet
MIF-ST	Graph Transformer	Structure-conditioned MLM	PDB + AlphaFold2	Sequence + structure	No	https://github.com/microsoft/protein-sequence-models
Enzyme function prediction 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	[144]
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
PreTKcat	ProtT5 + MolGNet + ExtraTrees	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 one of the important applications of pre-trained protein models in enzyme engineering [50,51]. Unlike traditional supervised learning methods that rely on experimentally measured sequence–function data, zero-shot prediction uses pre-trained models to directly estimate the functional impact of mutations without relying on labeled data for training. This paradigm is an important means of achieving few-shot enzyme engineering. By leveraging knowledge learned from large-scale protein sequence databases, pre-trained models can infer the likelihood of mutations and prioritize promising variants before wet-lab experiments. Liu et al. have comprehensively summarized online tools and methods for zero-shot prediction in enzyme engineering [52].

Zero-shot prediction is typically implemented by comparing the likelihood of wild-type and mutant sequences. Mutations that are more consistent with the evolutionary constraints encoded in the model tend to receive higher scores, while disruptive mutations are assigned lower probabilities [53,54]. Importantly, this method does not require any experimental labels; instead, it relies entirely on evolutionary information implicit in natural protein sequences [55]. This makes zero-shot prediction particularly suitable for early-stage enzyme engineering.

Large-scale PLMs such as the ESM series and ProtTrans have demonstrated excellent performance in zero-shot mutation effect prediction [56,57]. One study developed a fast and flexible in silico protein design tool by integrating multiple PLMs [58]. By directly scoring mutant sequences, these models can identify beneficial mutations with reasonable accuracy even without any task-specific fine-tuning. These results highlight the potential of universal predictors of protein mutation effects, significantly reducing reliance on large external experimental datasets.

Several methods have been specifically designed to enhance zero-shot mutation prediction. These methods typically improve prediction accuracy by refining scoring strategies or incorporating additional biological priors. For example, some studies integrate structural information, functional text descriptions [59], or evolutionary information [54] to better improve predictive performance. Others develop dedicated strategies for specific enzymatic properties [60] or specific enzyme classes [61]. These task-oriented adaptive methods 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, Li et al. compared different pre-trained models on benchmarks such as ProteinGym, showing that large-scale language models outperform traditional unsupervised methods, especially in data-scarce settings [62]. However, the accuracy of zero-shot prediction still has notable limitations [63]. On one hand, although pre-trained models can capture general evolutionary constraints embedded in sequences, they cannot fully account for protein-specific functional mechanisms, which somewhat weakens the advantage of zero-shot methods in automated enzyme engineering [64,65]. On the other hand, zero-shot prediction often falls short in fine optimization stages; for example, its predictive reliability significantly decreases in tasks involving multiple mutations [66] or sequences far from the natural sequence space [67]. Therefore, in practical engineering applications, zero-shot prediction typically needs to be combined with additional information to reach a usable level [68]. This has also driven the development of fine-tuning strategies that calibrate PLMs with a small amount of experimental data.

3.3. Few-Shot Transfer Learning of Pre-trained Models

Zero-shot prediction requires no experimental labels but struggles to finely adapt to specific enzyme–substrate tasks. When a small number of labeled variants are available from experiments, using these data to transfer the pre-trained model to the target task can usually significantly improve prediction accuracy [63]. In this review, we collectively refer to such strategies as few-shot transfer learning, i.e., under the condition of tens to hundreds of labeled samples, transferring the general biological knowledge encoded in pre-trained models to downstream enzyme engineering tasks through parameter updates or non-parametric adaptation.

Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning (PEFT) addresses this 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 enzyme engineering tasks under data-scarce conditions [69].

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 [70,71]. Similarly, low-rank adaptation (LoRA) decomposes weight updates into low-rank matrices, significantly reducing the number of trainable parameters while preserving the model's expressiveness. This method requires training only a very small number of newly added parameters, greatly reducing training costs, and has been validated in various protein tasks [72,73]. Sledzieski et al. demonstrated that the LoRA strategy achieves performance comparable to or even better than traditional fine-tuning in tasks such as protein interaction prediction [14]. In addition, many innovative fine-tuning methods [74] and auxiliary strategies [7] have been proposed.

In practical applications, parameter fine-tuning has been widely used in various enzyme engineering tasks. One study leveraged transfer features from large-scale tasks such as ProteinMPNN and ESM2, and through lightweight fine-tuning for stability prediction tasks, significantly improved model performance in few-shot scenarios [75]; the same idea is also reflected in the SPURS framework proposed by Li et al. PEFT strategies have also shown good results in specific enzyme engineering tasks [76,77]. These methods follow a common principle: decoupling general biological knowledge from task-specific adaptation, enabling large pre-trained models to be efficiently reused across various protein engineering tasks.

More broadly, these case studies show 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 emphasize the role of parameter fine-tuning as a key enabler of scalable and few-shot enzyme engineering workflows.

Meta-Learning

Meta-learning can provide another powerful strategy for few-shot adaptation by adjusting model parameters, architecture, or optimization methods using experience gained during learning. In specific applications, this strategy enables pre-trained models to rapidly adapt to the current task with very little data from new enzymes [78]. Similarly, multi-task meta-learning frameworks have significantly improved prediction accuracy in data-scarce environments [79].

In-Context Learning

Fine-tuning-based methods rely on parameter updates; although they can fit specific tasks more precisely, they still carry a risk of overfitting under extremely small sample conditions. In contrast, another paradigm adapts the model at inference time without modifying model parameters. Such methods are called in-context learning, which leverages the inherent flexibility of large pre-trained models to generalize across tasks through contextual information [63]. By avoiding explicit training, these methods can rapidly explore protein sequence space with low computational overhead even with small labeled datasets [80,81].

In this approach, a small number of labeled examples are provided as part of the input, enabling 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 methods [82]. This also indicates that pre-trained models not only encode static biological knowledge but also possess the ability to dynamically adapt to new tasks. Moreover, pre-trained models integrating multiple types of information can more accurately map sequence to function in task-adaptive inference [83].

Overall, task-adaptive inference methods reduce or avoid parameter updates, further decoupling transfer learning from dependence on 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 [84]. When the target task deviates significantly from the pre-training distribution, such methods often struggle to accurately capture functionally relevant fine-grained differences. In addition, due to the lack of explicit parameter updates, these methods may be limited in modeling multiple mutations or complex fitness landscapes [85,86].

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 improvement with minimal data [87,88]. This paradigm is called active learning (AL) and is particularly suitable for enzyme engineering [89]. Unlike random sampling, active learning models guide data acquisition. A core concept is that not all data points are equally informative—some sequences can provide more information about the underlying fitness landscape [90].

Most active learning frameworks rely on uncertainty estimation to identify informative samples. Bayesian optimization (BO) is a widely used method in which a probabilistic model (e.g., a Gaussian process or a Bayesian neural network) estimates 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, identifying high-performance variants by experimentally testing only a small fraction of candidates [91]. 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 [92], while the "Most Informative Positive" strategy prioritizes samples likely to yield functional variants in discovery tasks [93].

In enzyme engineering, active learning integrates naturally into the design–build–test–learn (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 improve engineering efficiency [86]. By iteratively selecting informative variants, these frameworks can achieve significant performance improvements within limited experimental rounds, effectively exploring vast combinatorial sequence spaces [94]. More importantly, active learning shifts the role of machine learning from passive prediction to active experimental design. By constructing small but information-dense training datasets, reliable models can be established with very few experimental samples, transforming enzyme engineering into a data-driven optimization process [95].

However, recent benchmarking studies have shown that when uncertainty estimates are unreliable or protein fitness landscapes are highly complex, uncertainty-based methods may not always outperform simple greedy sampling [96]. In addition, active learning performance strongly depends on the underlying predictive model. Poor model initialization or distributional shifts can lead to suboptimal sampling decisions, resulting in inefficient exploration. These findings emphasize 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.

Introducing pre-trained PLMs into this pipeline can significantly improve information efficiency. In nominating high-potential candidate mutants, the zero-shot scoring capability of language models has demonstrated powerful lead mutation identification. Fan et al. innovatively used the minimum probability prediction at all 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 in only two rounds of iterative design [97]. Similarly, regarding nomination strategies, 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 round of active learning, it explores sequence space beyond the wild-type neighborhood while maintaining biological plausibility, consistently outperforming or matching the best existing methods across different protein engineering tasks [98]. The Venus-MAXWELL framework combines the zero-shot scoring capability of PLMs with sequence–landscape modeling. After integrating ESM-IF into the framework, it not only achieved higher prediction accuracy than ThermoMPNN but also delivered 10-fold faster inference [99]. In excluding disruptive mutations, pre-trained models can also effectively filter out candidates likely to cause inactivation or instability. Deguchi et al. proposed a data augmentation method called "weakly supervised data" that combines molecular simulation with PLM zero-shot prediction. When experimental training data are insufficient, weakly labeled data generated by computational estimates (simulations or PLM predictions) are incorporated into training, and the weight and participation of weak training data are dynamically adjusted according to the available scale of experimental data, thereby extending the method's applicability to various protein properties such as binding affinity and enzyme activity while minimizing potential negative impacts [68].

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

Building on the various few-shot machine learning methods described above, a series of integrated strategies targeting specific enzyme properties have been developed in recent years and applied to enzyme engineering. However, it should be noted that the primary goal of most studies is to develop a general architecture or strategy and demonstrate its effectiveness using certain wet-lab or publicly labeled data; only a small number of studies have truly completed a full few-shot engineering campaign for a single enzyme.

In addition, considering the enormous differences in complexity across enzyme engineering tasks, "few-shot" does not have a unified quantitative standard. This review follows 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 complete that task. All studies summarized below explicitly take "achieving enzyme engineering tasks under few-shot or limited data" as one of their core objectives in method development and enzyme optimization, and representative studies are listed in Table 3. This section aims to inspire researchers to generalize these methods to their own target tasks.

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 fitness landscapes is one of the core application directions of few-shot machine learning in enzyme engineering.

Early 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. The model not only predicted the fitness of unknown sequences but also explicitly quantified predictive uncertainty to guide experimental selection. With only 40 training data points, its accuracy reached the level that traditional fragment-based regression models required 218 data points to achieve [100]. Building on this, Roberts et al. developed the FolDE framework, which in the first round uses the ESM PLM for zero-shot "naturalness" ranking to nominate mutants, and then uses neural networks to predict activity in subsequent rounds. In three rounds of experiments (16 variants per round), this method increased the probability of discovering top-1% mutants by 55% [101].

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. was globally unsupervised pre-trained on over 20 million raw amino acid sequences, and then fine-tuned using functional characterization of 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,

Table 3. Representative studies of few-shot machine learning applications in enzyme engineering.

Few-shot strategy	Target / benchmark	Method development	Wet-lab data size	Optimization outcome	Reference
Fitness optimization					
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]
	GFP dataset	SESNet: Pre-training with large amounts of low-quality results from unsupervised models; fine-tuning with a small amount of high-quality experimental data	40	Achieved Spearman correlation of 0.5–0.7 for multi-site mutants	[102]
	Glycoside hydrolase	MECE: Pre-trained DeepGH model on sequence data from 119 glycoside hydrolase (GH) families; used Grad-CAM to extract key residues for mutation design	9 single-point mutants; 4 multi-site mutants	Screened 7-site mutant CHIS1754-MUT7 from CHIS1754, with 23.53-fold improvement in k_{cat}/K_m	[104]
Active learning	P450 enzymes	Probabilistic modeling of protein fitness landscape using Gaussian processes; predicted fitness of unseen sequences and quantified uncertainty to guide experimental selection	40	With only 40 training data points, achieved prediction accuracy comparable to traditional models requiring 218 data points	[100]
Zero-shot prediction	ProteinGym benchmark; MgpCSO	GEMS: Integrated multiple zero-shot prediction models including ESM-IF, SaProt, MSA Transformer, and GEMME	/	Significantly outperformed existing SOTA methods; showed 1.1- to 3.2-fold improvement in catalytic efficiency for MgpCSO single mutants	[108]
	Gene editing fusion proteins	AiCE: Employed zero-shot prediction by combining inverse folding model sampling with structural/evolutionary constraints for nominating single and multi-point mutations	10–100	Achieved high success rates (11%–88%) in functional optimization with only tens to hundreds of experimentally validated variants	[109]
Activity and stability optimization					
Fine-tuning	Novel multicopper oxidases	ESM-Ezy: Fine-tuned ESM-1b to build a binary classifier; combined sequence distance to screen for novel enzymes with low sequence similarity but high functional activity	147	With only 147 known MCOs for fine-tuning, screened 18 candidate enzymes from UniRef50, of which 44% outperformed the query enzyme in catalytic performance	[105]
Zero-shot prediction	RmZHD	MSA Transformer: MSA combined with ABACUS-R-driven zero-shot prediction; captured co-evolutionary signals via axial attention	0 (validated 6 variants)	Achieved zero-shot prediction with only 6 mutations, increasing RmZHD T_m by 8.1ř C with >95% activity retention	[143]
	Cutinase enzyme	Segment Transformer: Combined zero-shot prediction with supervised learning; segment embedding and dual-group segment attention with weighted RMSE for temperature distribution	0 (validated 17 variants)	Thermostability optimization of cutinase with only 17 mutants achieved 1.64-fold relative activity increase and 3.9-fold half-life extension	[112]
Enzyme kinetic parameter optimization					
Active learning	Large serine recombinases (LSRs)	EVOLVEpro: Active learning of fitness landscape over multiple rounds; used ESM-2 as base PLM for ranking and iterative cycles	$N_1 = 11$ and 9 rounds	5 rounds (16 mutants each) matched performance of pre-training with 160 mutants; 10 rounds matched 500 mutants; variant activity >2.6-fold higher than wild type	[103]
Zero-shot prediction	Glutamate decarboxylase	Integrated four zero-shot predictors (UniKP, DLKcat, Pythia, ESM-1v) for virtual screening of 8,911 single-point mutations of EfGAD; consensus 107 mutations validated experimentally combined with C-terminal truncation	0 (validated 107 variants)	With only 107 single-point mutations validated, identified Y39T/Δ460-469 variant with 300 g/L GABA production in 5 L fermenter, 1.3-fold higher than wild type	[118]
Thermostability optimization					
Fine-tuning	Phi29 DNA polymerase	FSFP: ESM-1 as base pre-trained model; fine-tuned with LoRA and transferred to target task	20	Top-20 predictions after FSFP training showed >1ř C average T_m increase and 25% higher positive rate vs. ESM-1v	[63]
Zero-shot prediction	α-Amylase	ZSH (Zero-Shot Hamiltonian): Generated residue contact maps via MSA Transformer; Hamiltonian scoring for all possible double mutations	0 (validated 20 variants)	For α-amylase "short board" repair variant, validated 20 double mutants, 17 successfully expressed; optimal F237R/S240G increased T_m by 8.5ř C	[60]
Selectivity and substrate specificity optimization					
Fine-tuning	Cyclodextrinase	Pro-PRIME: Integrated prior knowledge from a dataset of 68 single-point mutants into the Pro-PRIME model to estimate sequence variation	68 single-point mutants	Optimal mutant increased transglycosylation-to-hydrolysis (T/H) ratio by 12-fold (8× transglycosylation, 33.3% less hydrolysis, 98.4% selectivity); pNP-G7 yield from 63% to 98%	[121]
	Carbonyl reductase	UniESA: Used pre-trained protein language models (ProteinBERT, ESM1v, ESM2) to generate sequence embeddings; fused multi-model features to predict stereoselectivity and activity	139	Virtual screening of only 30 mutants identified 2 mutants with dep >99.5% and 2 with significantly improved activity	[122]
Active learning	Acyl-ACP reductase	Gaussian process as sequence-function model; upper confidence bound criterion to balance exploration and exploitation; iterative experimental guidance	96 chimeric sequence-fatty acid titer data points	Tested only 96 chimeric sequences; optimal variant ATR-83 showed 4.9-fold higher in vivo fatty acid production than native sequence, with k_{cat} increased from 0.047 s^{-1} to 0.067 s^{-1}	[117]

the METL model developed by Gelman et al. combines biophysical simulation with transfer learning: pre-training a Transformer encoder on millions of Rosetta-simulated variants, then fine-tuning on a small amount of experimental data. Studies showed that adding approximately 29 simulated data points yielded a performance improvement equivalent to adding 1 experimental data point, and high-fluorescence variants were successfully obtained in GFP design using only 64 experimental data points [37]. Similarly, SESNet pre-trains by fusing sequence, co-evolutionary, and structural information, and with fine-tuning on only 40 high-quality data points, achieves high-precision prediction of 5–8-site multi-mutants (Spearman correlation coefficient 0.5–0.7) [102]. In addition, the active learning framework EVOLVEpro, based on ESM-2, iteratively selects top-ranked sequences for experiments over multiple rounds of evolution, reaching the performance level of traditional methods pre-trained with 160 variants in just 5 rounds (16 variants per round) [103].

In addition to the above work, MECE pre-trains the DeepGH model using sequence data from 119 glycoside hydrolase families, extracts the functional residues contributing most to classification via Grad-CAM, and constructs mutation scoring metrics combined with evolutionary information. This method screened a 7-site mutant CHIS1754-MUT7 from 9 single-point mutants and 4 multi-site mutants, with a 23.53-fold improvement in k_{cat}/K_m [104].

4.2. Activity and Thermostability

Catalytic activity and thermostability are two key properties for the industrial application of enzymes, and are also targets that few-shot machine learning methods focus on. Researchers have achieved significant progress across various enzyme systems through different strategies such as fine-tuning, zero-shot prediction, and active learning.

In fine-tuning-based prediction, ESM-Ezy constructs a binary classifier by fine-tuning ESM-1b, 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 optimal 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 [105]. GeoEvoBuilder employs geometric deep learning to extract features from target protein structures, combined with ESM2 to capture evolutionary information, and resolves structure–evolution conflicts through iterative optimization. In GPX4 optimization, the optimal variant obtained with only 19 data points achieved an 8.49-fold improvement in catalytic efficiency; in DHFR optimization, the optimal variant obtained with 23 data points achieved a 20-fold improvement in catalytic efficiency and a 10řC increase in T_m [106].

In the field of zero-shot prediction, the PRIME model was pre-trained on 96 million sequences annotated with optimal growth temperature, achieving 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 [107]. GEMS, as a multimodal ensemble framework, integrates multiple zero-shot predictors including ESM-IF, SaProt, and MSA Transformer, achieving excellent performance on the ProteinGym benchmark and successfully identifying MgpCSO mutants with 1.1- to 3.2-fold improvements in catalytic efficiency [108]. AiCE employs inverse folding models to generate structurally compatible sequences, combined with evolutionary coupling scores to nominate multi-mutant combinations. With only 10–100 experimentally validated variants, it achieved an 11%–88% functional optimization success rate, for example, improving mitochondrial base editing efficiency by 14.3-fold [109].

In active learning, Jiang et al. proposed a rationality-informed protein evolution (RIPE) strategy for the PET hydrolase BhrPET: first using 150 rationally designed single-point mutants as the initial dataset, then performing four rounds of iterative evolution via EVOLVEpro. The final optimal variant IYLL exhibited 6.91-fold higher hydrolytic activity toward amorphous PET at 72řC compared to the wild type [110]. Chen et al. combined graph learning with meta-learning, first using GVP-GNN for structure-aware optimization, then MAML to learn cross-temperature general knowledge. By exploring only 10–40 variants, the BHETase variant Y109R achieved a 5.5-fold increase in product yield at 60řC [111]. Zhang 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 [112].

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, but 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 prediction, BioStructNet was pre-trained on a hydrolase k_{cat} database and compound–protein 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 [76]. 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 [113]. PMAK captures feature interactions between enzyme and substrate representations through a residue-aware attention mechanism, achieving lower mean squared error than DLKcat and TurNuP models in k_{cat} prediction across 16 environmental conditions [114].

In data augmentation strategies, Deguchi 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 method showed clear advantages [68]. KinForm combines 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% [115]. O^2DENet employs 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 [116].

In addition, active learning has 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 [117].

In zero-shot approaches, Miao et al. used four zero-shot predictors (UniKP, DLKcat, Pythia, and ESM-1v) for virtual screening of 8,911 single-point mutations of EFGAD, selecting 107 mutations with consensus from at least three predictors for experimental validation, combined with C-terminal truncation optimization. The final Y39T/Δ460-469 variant achieved 300 g/L GABA production in a 5 L fermenter, a 1.3-fold improvement over the wild type [118].

4.4. Thermostability

Thermostability is one of the key factors limiting the industrial application of enzymes, and is also a property that few-shot machine learning methods focus on.

In fine-tuning-based prediction, the FSFP framework combines 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 [63]. ActiMut-XGB combines ESM2 sequence embeddings, physicochemical properties, evolutionary features, and positional features, using an XGBoost classifier to predict thermostability changes in CALB single-point mutants. With limited fine-tuning

data, it achieved a Matthews correlation coefficient of 0.2511, far exceeding traditional tools such as PremPS (-0.0105) and DeepDDG (0.0305) [119]. The SPURS model employs 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 [120].

In zero-shot prediction, the ZSH model generates residue contact maps based on MSA Transformer and scores 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 [60]. 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 [55].

4.5. Selectivity and Substrate Specificity

Enzyme stereoselectivity, regioselectivity, and substrate specificity are core properties determining their synthetic application value, and 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 evaluates sequence mutation effects through protein language models and integrates 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% [121]. UniESA fuses 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 [122]. 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% [123]. FusionESP employs a contrastive learning and parameter-efficient fine-tuning paradigm, freezing pre-trained ESM-2 and MoLFormer while training only lightweight projection heads, achieving 93.57% accuracy in enzyme–substrate pairing prediction [124].

In active learning, Greenhalgh et al. used Gaussian processes as the sequence–function 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, achieving simultaneous improvement in activity and substrate selectivity [117]. pFSLNN targets 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% [125].

4.6. Other Tasks

Beyond direct optimization of core catalytic properties, few-shot machine learning methods have also been widely applied to supporting tasks such as improving target product yield, predicting binding affinity, accelerating the discovery of superior mutants, and enzyme function annotation.

For yield optimization, the ALDE framework is 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 [126].

For binding affinity prediction, TrGPCR employs a dynamic TrAdaBoost algorithm to transfer samples from BindingDB (source domain) with distributions similar to GPCR data to augment the target dataset. In GPCR–ligand binding affinity prediction, TrGPCR achieved 5.2% lower RMSE and 4.5% lower MAE than existing models such as DeepDTA, demonstrating strong generalization capability [127].

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) [101], while EVOLVEpro demonstrated that 5 rounds of evolution matched the performance of traditional methods pre-trained with 160 variants [103]. 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 [88]. 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 [77].

In summary, the above application cases clearly demonstrate the broad applicability and remarkable effectiveness of current few-shot machine learning methods across various subfields of enzyme engineering. From fine-tuning of pre-trained models to zero-shot prediction, from intelligent nomination by active learning to data augmentation and transfer learning, diverse strategies provide researchers with a powerful toolkit to address data scarcity challenges.

5. Challenges

Few-shot machine learning provides more cost-effective approaches for prediction and optimization in enzyme engineering, but its application still faces obstacles in experimental collaboration, computational cost, and model reliability. If these issues are not resolved before models enter experimental and translational workflows, they will limit their practical value. This section analyzes the technical and conceptual barriers limiting their application and outlines potential solutions.

5.1. Collaboration and Tool Accessibility Challenges

Experimentalists face tool accessibility issues when integrating machine learning into their workflows. Some tools offer web servers or Colab implementations, but more are released only as GitHub code, and pre-trained models often require substantial programming work to adapt to target tasks. High experience is also required for data processing, parameter selection, model understanding, and result interpretation. Public funding agencies should strive to build and maintain accessible implementations of popular tools and provide necessary training to eliminate these barriers.

On the other hand, few-shot machine learning enzyme engineering algorithms are generally more complex and require closer communication with wet-lab workflows. However, since most experimentalists and computer scientists have received limited training in machine learning and biochemistry, respectively, such collaboration can easily lead to misunderstandings and unrealistic expectations on both sides. For example, the ideal dataset for machine learning may differ significantly from the data readily generated in the laboratory, and proper data handling requires understanding how the data were generated. It is not uncommon for experimental conditions to change between screening rounds, and such changes pose difficulties for computational analysis. Therefore, data requirements and expectations should be clarified early in the collaboration, and preliminary experiments should be used to determine whether the generated data are suitable for computational analysis. University curricula should also reflect the importance of machine learning in the life sciences and provide interdisciplinary training for future researchers.

5.2. High Computational Cost Challenges

Pre-training strategies have significantly improved data efficiency but also bring enormous computational demands. Training large-scale protein models on UniRef or metagenomic corpora requires distributed GPU or TPU clusters and long training times. Although this cost can be amortized across multiple downstream tasks, it still constitutes a significant barrier to entry [38,128]. Inference and model adaptation also incur computational costs. Even with parameter-efficient fine-tuning, evaluating large numbers of candidate sequences can still be expensive. Therefore, although the wet-lab workload is reduced, the overall cost of model-based enzyme engineering workflows may not necessarily decrease.

To address the above issues, recent research has improved computational efficiency at multiple levels. Model compression techniques such as knowledge distillation [129] and pruning aim to reduce model size while maintaining performance; lightweight architectures [130] and retrieval-based methods can reduce dependence on large-scale models. These strategies collectively aim to balance predictive performance and computational efficiency, making machine learning more easily applicable to practical enzyme engineering. In addition, a better balance between computational efficiency and performance can be achieved by appropriately selecting model scale [131]; for example, lightweight models developed for enzyme property prediction typically have smaller parameter counts and lower training costs, making them more suitable for deployment in ordinary laboratory computing environments [132,133]. At the transfer level, parameter-efficient fine-tuning methods such as Adapters, LoRA, and Prefix Tuning are inherently computationally economical [72,134], and some architectural improvements can also improve transfer efficiency [135].

At the system level, activation checkpointing reduces memory usage by recomputing intermediate activations during backpropagation, enabling large models to be trained on limited GPU resources [136]. Mechanisms such as FlashAttention can improve ESM model inference efficiency and reduce GPU memory usage, and low-bit quantization can further compress model memory requirements while basically maintaining performance, enabling billion-parameter models to run on common GPUs [137].

These advances indicate that the research focus is shifting from solely pursuing predictive performance to optimizing overall resource efficiency. Future enzyme engineering requires not only more accurate models but also efficient models that can operate under realistic constraints. Therefore, comprehensive evaluation frameworks that simultaneously consider data, computational, and experimental costs should be established, rather than optimizing a single dimension in isolation.

6. Conclusion and Future Directions

Overall, few-shot machine learning is driving enzyme engineering from a traditional experience-driven model relying on high-throughput screening toward a rational design paradigm centered on computational prediction. Leveraging the evolutionary and structural priors embedded in large-scale protein pre-trained models, strategies such as zero-shot prediction, transfer learning, and parameter-efficient fine-tuning can achieve relatively reliable mutation effect prediction with very little experimental data, reducing experimental costs. Data augmentation, generative models, and active learning further improve the efficiency of limited data utilization. Recent studies have made progress in tasks such as catalytic efficiency, substrate specificity, and enzyme stability, indicating that few-shot machine learning has become an important methodological foundation for modern enzyme engineering. However, it should be noted that these methods are still limited by the pre-training distribution, data quality, and task complexity, and their reliability varies significantly across different enzyme systems.

The field is gradually forming a research paradigm centered on protein foundation models, with few-shot learning as the method and the DBTL cycle as the operating mechanism. Pre-trained models provide general biological knowledge representations, few-shot strategies enable rapid task adaptation, and active learning and generative models explore potential functional sequences in the vast sequence space. Future progress may come from the pipeline integration of tools and data sources:

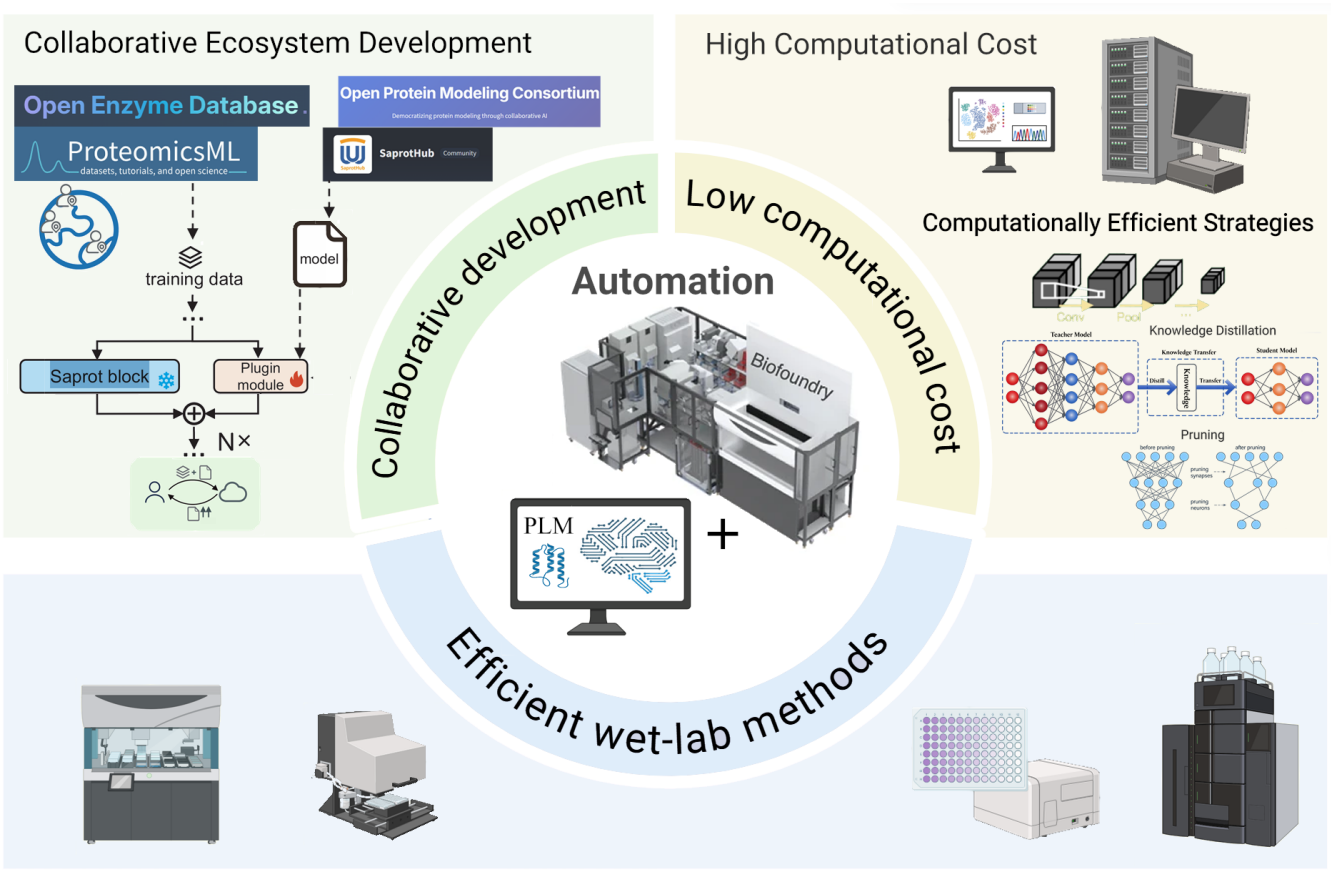


Figure 3. Architecture of few-shot machine learning methods. (A) Pre-trained models (e.g., protein language models) learn general feature representations from massive unlabeled sequence data or low-fidelity data, encoding sequence, structure, evolutionary, and functional information that can be transferred to downstream enzyme engineering tasks. (B) Pre-trained models are directly used as zero-shot predictors to assess the consistency of mutant sequences with the statistical patterns of natural sequences, enabling inference of mutational functional effects without experimental labeled data and allowing preliminary screening of promising mutants. (C) Transfer learning adapts pre-trained models to a target task using a small amount of experimentally labeled data, including parameter-efficient fine-tuning (PEFT) and in-context learning without parameter modification, significantly improving prediction accuracy compared with zero-shot methods. (D) A small subset is randomly sampled from a variant library for experiments, and a supervised model with uncertainty quantification is trained on the experimental data to predict the remaining unmeasured variants. An acquisition function is then used to nominate variants with both high uncertainty and high value, which are experimentally measured and fed back for retraining, iterating until top-tier variants are reached.

combining the global exploration capability of generative models in sequence space with the local optimization capability of engineering methods, and iteratively improving design models based on experimental feedback.

Looking ahead, the development of machine learning in enzyme engineering may advance along two directions: automated closed-loop platforms and open collaborative ecosystems. Integrating protein language models with automated biofoundry systems can build a truly closed-loop DBTL. Existing platforms combine large language models with automated experiments to complete enzyme performance optimization within several rounds of iteration and hundreds of variants [138]. Other studies use ESM-2 zero-shot prediction to initiate the optimization cycle, training predictors with feedback data to achieve multiple rounds of directed evolution in a relatively short time [109]. These cases demonstrate that combining PLM prediction with automated experiments can accelerate enzyme optimization with few samples. However, such systems rely on adaptability metrics that can be measured at high throughput as feedback signals, and still have limitations in complex substrate systems that are difficult to characterize quantitatively. Standardization of experimental workflows

is also important; existing work has provided a complete workflow for optimizing thymine-DNA glycosylase based on PLM zero-shot prediction to reduce trial-and-error costs [139].

Open collaboration and platform-based tools are another important direction. As the scale of protein language models continues to expand, their training and deployment typically require high machine learning expertise, which to some extent limits their use by interdisciplinary researchers. To lower technical barriers, a series of community-oriented open platforms have emerged in recent years. For example, the proposed Open Protein Modeling Consortium (OPMC) aims to establish a community-driven model-sharing ecosystem, enabling researchers to share and fine-tune existing models, thereby accelerating algorithm iteration and knowledge dissemination. As an important part of this ecosystem, SaprotHub provides an integrated platform for model training, prediction, and sharing, while the ColabSaprot framework proposed in [14] supports hundreds of protein modeling tasks through cloud-based workflows, enabling researchers without machine learning backgrounds to conveniently use PLMs for prediction and design. Meanwhile, some studies have also explored improving overall efficiency through workflow integration. For example, [140] integrated pocket prediction (P2Rank), active site search (GASS), and molecular docking (AutoDock Vina) into a unified automated workflow, achieving an efficient computational pipeline from structural analysis to function prediction through algorithmic combination rather than single-model innovation.

Furthermore, closer synergy between algorithm design and experimental strategies is also needed. Existing methods design mutation libraries by simultaneously optimizing predicted fitness and sequence diversity, achieving a balance between performance improvement and sequence space exploration [86]. Combined with automated high-throughput screening platforms (the experimental systems described in [141] and [91]), these computational strategies can significantly improve experimental hit rates and drive enzyme engineering research toward more efficient closed-loop optimization.

As protein foundation models, automated experimental platforms, and open computational ecosystems mature, enzyme engineering is gradually moving toward an automated and data-driven intelligent design stage. In this paradigm, machine learning can not only reduce experimental costs and development cycles but also help understand the complex relationships among protein sequence, structure, and function, thereby 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.D.; Dejean, T.; Schreiber, Z.; Harrigan, W.; Polson, S.W.; Wommack, K.E.; Belcaid, M. Fine-Tuning Protein Language Models Unlocks the Potential of Underrepresented Viral Proteomes. *PeerJ* **2025**, *13*, e19919, doi:10.7717/peerj.19919.
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. Yang, Q.; Yu, J.; Zheng, J. A Survey of Downstream Applications of Evolutionary Scale Modeling Protein Language Models. *Quant. Biol.* **2026**, *14*, e70013, doi:10.1002/qub2.70013.
40. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci.* **2021**, *118*, e2016239118, doi:10.1073/pnas.2016239118.
41. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127, doi:10.1109/TPAMI.2021.3095381.
42. Vitale, R.; Bugnon, L.A.; Fenoy, E.L.; Milone, D.H.; Stegmayer, G. Evaluating Large Language Models for Annotating Proteins.
43. 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.
44. 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.
45. Cheng, X.; Chen, B.; Li, P.; Gong, J.; Tang, J.; Song, L. Training Compute-Optimal Protein Language Models **2024**.
46. 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**.
47. https://www.emergentmind.com/topics/esm-series;
48. 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**.
49. 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**.
50. Reeves, S.; Kalyaanamoorthy, S. Zero-Shot Transfer of Protein Sequence Likelihood Models to Thermostability Prediction.
51. 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**.
52. 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.
53. 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.
54. Tan, Y.; Wang, R.; Wu, B.; Hong, L.; Zhou, B. Retrieval-Enhanced Mutation Mastery Augmenting Zero-Shot Prediction of Protein Language Model **2024**.
55. 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.

56. 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.
57. 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.
58. 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.
59. Char, S.; Corley, N.; Alamdari, S.; Yang, K.K.; Amini, A.P. ProtNote: A Multimodal Method for Protein-Function Annotation **2024**.
60. 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.
61. Deznabi, I.; Arabaci, B.; Koyutürk, M.; Tastan, O. DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases.
62. Li, F.-Z. Evaluation of the Generalizability of Machine Learning-Assisted Protein Engineering Methods.
63. 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.
64. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function **2021**.
65. 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**.
66. Fan, J.; Jiao, X.; Lin, S.; Liang, Z.; Mao, W.; Jing, C.; Chen, H.; Shen, C. Evolutionary Profiles for Protein Fitness Prediction **2025**.
67. Gordon, C.; Lu, A.X.; Abbeel, P. Protein Language Model Fitness Is a Matter of Preference **2024**.
68. 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.
69. 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**.
70. AdapterHub: A Framework for Adapting Transformers.
71. 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.
72. 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**.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. 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.
79. 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.
80. Li, Z.; Xu, Z.; Han, L.; Gao, Y.; Wen, S.; Liu, D.; Wang, H.; Metaxas, D.N. Implicit In-Context Learning **2025**.

81. Lu, J.; Teehan, R.; Yang, Z.; Ren, M. Context Tuning for In-Context Optimization **2025**.
82. 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**.
83. 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.
84. 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.
85. 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.
86. 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.
87. Cohn, D.A.; Ghahramani, Z.; Jordan, M.I. Active Learning with Statistical Models.
88. Babjac, A.; Hoarfrost, A. An Active Learning Framework for Data-Efficient, Human-in-the-Loop Enzyme Function Prediction **2026**.
89. 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.
90. Cohn, D. Improving Generalization with Active Learning.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. Kmicikiewicz, M.; Fortuin, V.; Szczurek, E. ProSpero: Active Learning for Robust Protein Design Beyond Wild-Type Neighborhoods **2025**.
99. 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.
100. 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.
101. 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**.
102. 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.
103. 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.

104. 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.
105. 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.
106. 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.
107. 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.
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. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. Alwer, S.; Fleming, R. KinForm: Kinetics Informed Feature Optimised Representation Models for Enzyme k_{cat} and K_M Prediction **2025**.
116. 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**.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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.
122. 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.

123. 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.
124. 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.
125. 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.
126. 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.
127. 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.
128. 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.
129. 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.
130. Jendrusch, M.; Korbel, J.O.; Jendrusch, M.; Korbel, J.O.; Jendrusch, M.; Korbel, J.O. Efficient Protein Structure Generation with Sparse Denoising Models **2025**.
131. Vieira, L.C.; Handojo, M.L.; Wilke, C.O. Scaling Down for Efficiency: Medium-Sized Transformer Models for Protein Sequence Transfer Learning.
132. 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.
133. 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.
134. Mahabadi, R.K.; Henderson, J.; Ruder, S. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers **2021**.
135. 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.
136. Ren, J.; Merced, U.; Rajbhandari, S.; Aminabadi, R.Y.; Ruwase, O.; Yang, S.; Merced, U.; Zhang, M.; Li, D.; Merced, U.; et al. ZeRO-Offload: Democratizing Billion-Scale Model Training.
137. 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.
138. 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.
139. 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.
140. 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.
141. 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.
142. 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.
143. 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.

144. 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.