A Survey on Efficient Protein Language Models

Anonymous authors
Paper under double-blind review

Abstract

Protein language models (pLMs) have become indispensable tools in computational biology, driving advances in variant effect prediction, functional annotation, structure prediction, and engineering. However, their rapid expansion from millions to tens of billions of parameters introduces significant computational, accessibility, and sustainability challenges that limit practical application in environments constrained by GPU memory, hardware availability, and energy budgets. This survey presents the first comprehensive review of efficient pLMs, synthesizing recent advancements across four key dimensions. We first examine *(1)* dataset efficiency through meta-learning-based few-shot and scaling-law-guided data allocation; and *(2)* architecture efficiency via lightweight alternatives including quantized transformers, embedding compression, and convolution-based designs. Furthermore, we review *(3)* training efficiency through scaling-law-informed pretraining, structure-integrated multimodal approaches, and low-rank adaptations with diverse distillation strategies; and *(4)* inference efficiency via quantization, dense-retrieval, and structure-search methods. By providing a structured taxonomy and practical guidance, this survey enables the development of high-performance, scalable, yet sustainable next-generation pLMs.

1 Introduction

Inspired by the success of large language models (LLMs) in natural language processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019), protein language models (pLMs) have emerged as a transformative force in computational biology (Bepler & Berger, 2019). By treating AA sequences as a "language of life," (Asgari & Mofrad, 2015) pLMs apply self-supervised learning to vast protein databases to learn the fundamental grammar of protein structure, function, and evolution (Rives et al., 2021; Elnaggar et al., 2021). A comparative overview between LLMs and pLMs is shown in Figure 3, providing intuitive context for their methodological parallels and distinctions. This paradigm shift enables protein analysis at an unprecedented scale of hundreds of millions of sequences (Lin et al., 2023), offering a powerful alternative to traditional experimental methods like X-ray crystallography, NMR spectroscopy, and Cryo-EM which, while foundational, are often slow and costly (Peng et al., 2023a; Xiao et al., 2025). The power of pLMs lies in their ability to bridge the enormous gap between the over 200 million known protein sequences and the fewer than 200 thousand experimentally determined structures (Burley et al., 2021; Consortium, 2021), thereby accelerating drug discovery and synthetic biology (Xiao et al., 2025). A historical overview of key milestones in pLM development is summarized in Figure 2.

The remarkable capabilities of pLMs stem from their deep neural architectures, predominantly transformer-based networks (Vaswani et al., 2017), which are trained extensively on massive protein sequence datasets (Suzek et al., 2015) to learn intrinsic biological patterns and representations. The entire pipeline of pLMs consists of dataset construction, model architecture design, training methodologies (pre-training, fine-tuning, and distillation), and inference strategies, each influencing both performance and efficiency (Ofer et al., 2021; Ferruz et al., 2022). Powered by large-scale computational resources, these models integrate evolutionary and structural information from massive-scale evolutionary protein corpora into parameter spaces reaching up to tens of billions (Lin et al., 2023), thereby achieving state-of-the-art accuracy on structure-prediction benchmarks such as CAMEO (Haas et al., 2013) and CASP14 (Kryshtafovych et al., 2021), a level not previously reached by single-sequence models.

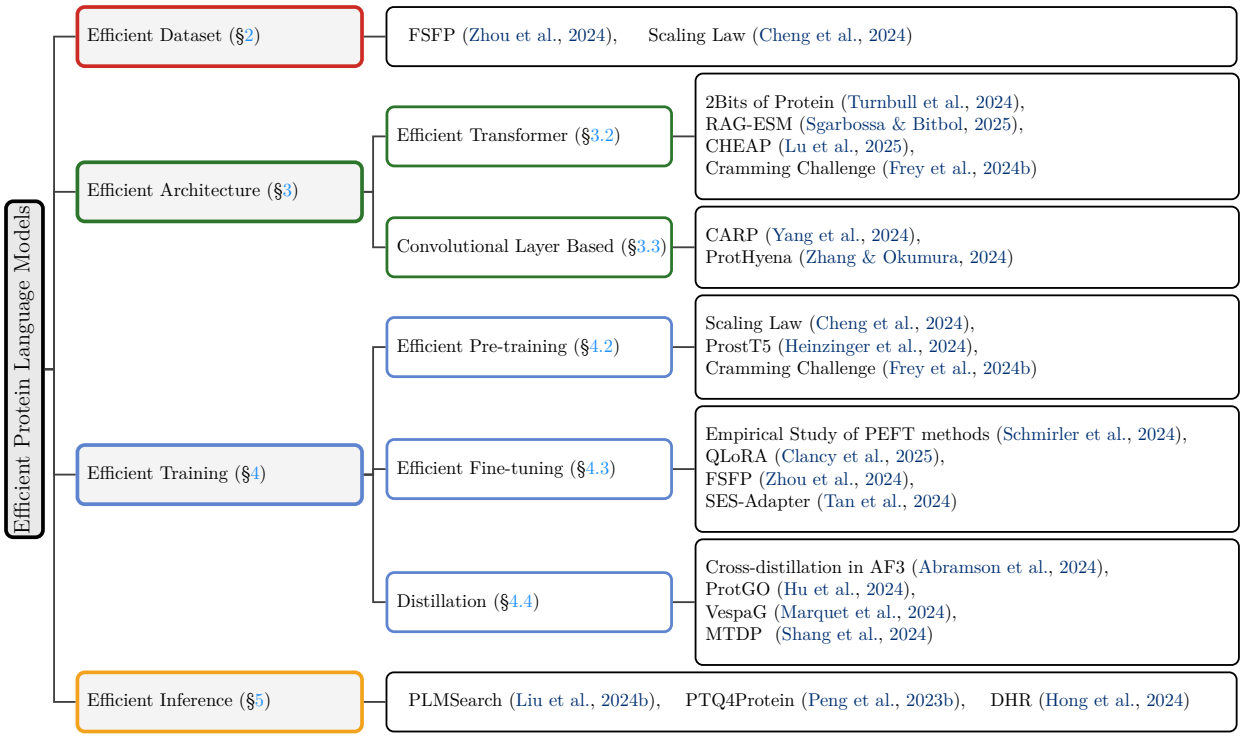

Figure 1: An overview and taxonomy of the efficient Protein Language Model methodologies discussed in this survey. We structure our review into four primary pillars—Dataset, Architecture, Training, and Inference—and highlight the key approaches and representative works within each category.

Despite their transformative potential, the effectiveness of large-scale pLMs comes at a substantial computational and memory cost, posing significant practical challenges. Typical large pLMs contain billions of parameters—such as the 15B parameter ESM-2 (Lin et al., 2023)—and require extensive resources; for instance, training models of this scale conventionally demands tens of thousands of GPU hours, translating to immense financial and environmental burdens (Lin et al., 2023; Strubell et al., 2019; 2020). This heavy reliance on compute limits accessibility for resource-constrained research groups (Kaplan et al., 2020b). Furthermore, deploying such extensive models in real-world scenarios, such as proteome-wide homology searches, becomes computationally formidable, severely restricting their broader applicability and practical utility (Hong et al., 2024).

Motivated by these challenges, a growing body of research has focused explicitly on improving the efficiency of pLMs across various stages of their lifecycle (Cheng et al., 2024; Zhou et al., 2024; Frey et al., 2024b). These *efficiency-focused* methods aim to achieve competitive or even superior performance while reducing resource consumption by orders of magnitude. Specifically, efficiency strategies in pLMs encompass several critical categories, the reviewed methods of which are listed in Figure 1:

- **Dataset Efficiency**: Enhancing pLM efficiency via optimizing the allocation and utilization of datasets (Cheng et al., 2024), to maximize learning from limited or expensive experimental data (Zhou et al., 2024).
- **Architecture Efficiency**: Developing compact or lightweight architectures, such as quantized transformer models (Turnbull et al., 2024) or convolutional architectures (Yang et al., 2024; Zhang & Okumura, 2024), reducing parameter counts by up to 10× while maintaining predictive performance.
- **Training Efficiency**: Optimizing the training procedures, including computational resource allocation guided by scaling laws (Cheng et al., 2024), efficient fine-tuning techniques (Schmirler et al., 2024; Clancy et al., 2025; Zhou et al., 2024), and knowledge distillation (Hinton et al., 2015) approaches that transfer knowledge from larger, teacher models to compact, efficient student models (Abramson et al., 2024; Hu et al., 2024; Marquet et al., 2024; Shang et al., 2024).

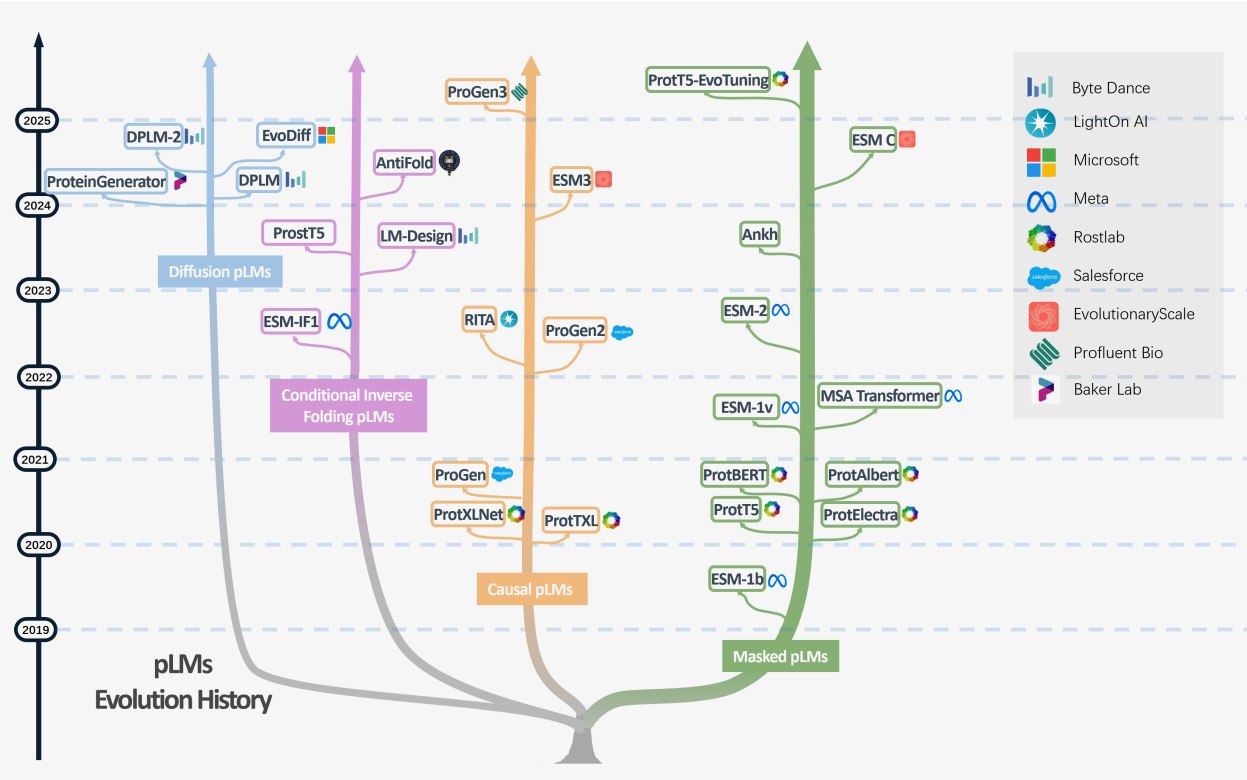

Figure 2: The evolution history of Protein Language Models (pLMs) from 2019 to 2025. The timeline illustrates the development and divergence of major pLM families, such as Masked pLMs and Conditional Inverse Folding pLMs, highlighting key models like the ESM series, ProtT5, and ProGen, along with their originating research labs.

- **Inference Efficiency**: Employing techniques such as post-training quantization (PTQ) (Peng et al., 2023b), dense embedding-based retrieval (Liu et al., 2024b), and alignment-free methods to accelerate search by over two orders of magnitude compared to traditional alignment methods (Hong et al., 2024).

While recent surveys have provided broad overviews of protein language models (pLMs) and their applications (Xiao et al., 2025; Wu et al., 2022), our work particularly focuses on the efficiency dimension of pLMs across dataset, architecture, training, and inference. We systematically review these efficiency-focused methodologies, highlighting their core principles, underlying mechanisms, and performance implications. While recent surveys have provided broad overviews of protein language models (pLMs) and their applications (Xiao et al., 2025; Wu et al., 2022), our work particularly focuses on the efficiency dimension of pLMs across dataset, architecture, training, and inference. We systematically review these efficiency-focused methodologies, highlighting their core principles, underlying mechanisms, and performance implications. The remainder of this survey is organized as follows: Section 2 details dataset strategies, Section 3 reviews architectural innovations, Section 4 discusses training methodologies, Section 5 covers efficient inference methods, and Section 6 outlines future directions.

2 Datasets for Efficient pLMs

2.1 Background and Challenges

The training paradigm for modern language models is fundamentally data-driven, predicated on learning statistical and semantic regularities from vast corpora (Kaplan et al., 2020a). For Large Language Models (LLMs) like BERT (Devlin et al., 2019) and the GPT series (Radford et al., 2018; Brown et al., 2020),

Table 1: Summary of representative dataset-related methods for pLMs.

Category	Method	Brief Description
Dataset	scaling-law (Cheng et al., 2024)	Empirical study on optimal data/model allocation for pLM training.
	FSFP (Zhou et al., 2024)	few-shot with meta-learning and LoRA (Hu et al., 2022b) under severe data scarcity.

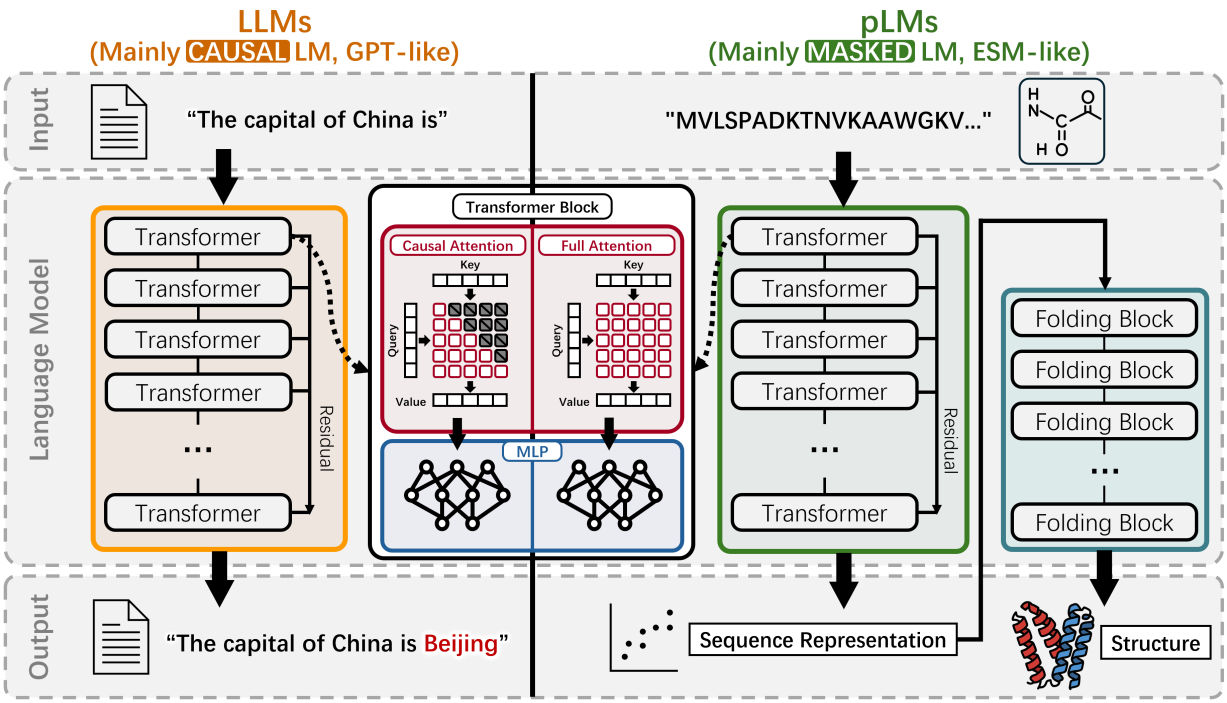

Figure 3: A comparative overview of general Large Language Models (LLMs) and Protein Language Models (pLMs). While both are based on the Transformer architecture, LLMs (left) typically use causal attention for autoregressive text generation. In contrast, pLMs (right) often use masked language modeling (with full attention) to learn sequence representations from amino acid inputs, which can then be used for downstream tasks like structure prediction. The folding blocks denote structure-prediction components, e.g., as in ESMFold (Lin et al., 2023); their internal architectural details are omitted here for brevity.

this involves processing immense textual datasets that are rigorously deduplicated, filtered, and ultimately discretized into a finite vocabulary via tokenization (Sennrich et al., 2015; Kudo & Richardson, 2018). Transposing this paradigm to computational biology, Protein Language Models (pLMs) operate on datasets of AA sequences drawn from large-scale biological repositories such as UniRef (Suzek et al., 2015) and expansive metagenomic collections (Mitchell et al., 2020). These biological datasets undergo analogous preprocessing, including stringent sequence identity filtering and task-specific curation, to construct a suitable training corpus (Cheng et al., 2024).

Within both NLP and computational biology, an efficiency-centric approach to dataset construction seeks to maximize model utility and generalization performance under a constrained computational-budget. For LLMs, this often involves curating data for optimal scale and diversity to avoid diminishing returns on compute (Hoffmann et al., 2022; Raffel et al., 2020). In the context of pLMs, efficiency-driven data strategies have emerged to address distinct biological challenges. Key approaches include meta-learning frameworks for few-shot adaptation, which are critical in scenarios where experimental labels are prohibitively expensive or scarce (Zhou et al., 2024). Additionally, researchers have formulated empirically-derived scaling laws that govern the co-scaling of model and data size to optimize the training trajectory within a fixed computational

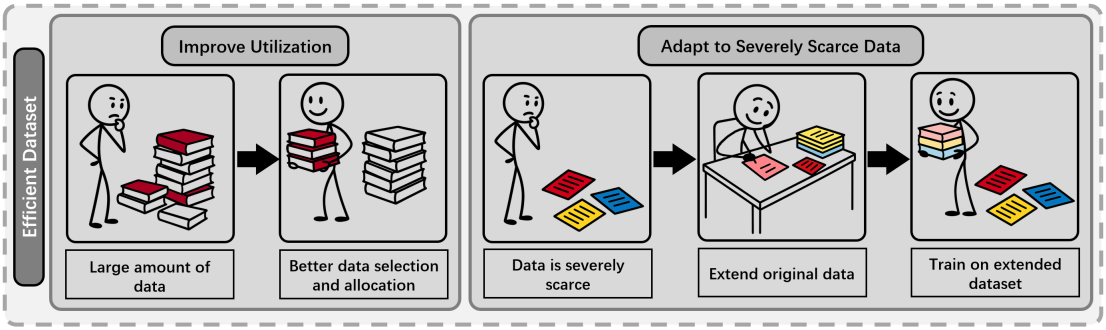

Figure 4: Conceptual illustration of efficient dataset strategies. The left subfigure, 'Improve Utilization,' represents using methods like scaling laws for better data selection and allocation from large-scale corpora. The right subfigure, 'Adapt to Severely Scarce Data,' represents few-shot learning approaches, for example using auxiliary tasks to extend limited datasets.

envelope (Cheng et al., 2024). Such methods directly confront data-related bottlenecks, aiming to enhance model performance within the practical constraints of bioinformatics research.

The development of these efficiency-focused strategies is motivated by several fundamental challenges inherent to biological data utilization:

- **Paucity of Labeled Data:** A critical bottleneck in many real-world applications, such as protein engineering and functional genomics, is the extreme scarcity of experimentally validated labels. Often, only a few dozen labeled protein variants are available, posing a fundamental impediment to traditional supervised fine-tuning and demanding highly data-efficient adaptation methods (Zhou et al., 2024).
- **Absence of Principled Scaling Guidelines:** Given a finite computational-budget, determining the optimal allocation between model parameter count and training data volume remains a significant challenge. An unprincipled approach, such as aggressively scaling model size while repeatedly cycling through a small dataset, can lead to suboptimal training dynamics, including premature convergence, overfitting, and a failure to capitalize on the model's architectural capacity (Cheng et al., 2024).

To surmount these obstacles, recent work has focused on developing sophisticated meta-learning protocols and establishing a principled, theoretical foundation for balancing model and data scaling. The subsequent section reviews these methodologies in detail, with an overview provided in Table 1. See Figure 4 for an illustration of these efficient dataset strategies.

2.2 Efficient Dataset Methods

Few-Shot Learning for Protein Fitness Prediction (FSFP) (Zhou et al., 2024) addresses the challenge of efficiently fine-tuning pLMs under severe data scarcity (Romero et al., 2013), enabling fitness-prediction with only tens of experimentally labeled mutants per target protein. The key innovation of FSFP lies not in increasing data volume, but in its auxiliary-task construction, evolutionary-informed pseudo-labeling, and meta-learning-driven adaptation, providing a robust solution for pLM fine-tuning in low-data scenarios.

FSFP frames fitness-prediction as a ranking task, employing a listwise-loss (Cao et al., 2007) throughout both meta-learning and fine-tuning, which better reflects practical needs in protein engineering research and applications (Buller et al., 2018). FSFP consists of the following major components:

- *Auxiliary Task Construction:* FSFP constructs three auxiliary tasks: it first retrieves two labeled DMS datasets (Fowler & Fields, 2014; Melamed et al., 2013) from proteins similar to the target, and additionally generates a third pseudo-labeled dataset using GEMME (Laine et al., 2019), which leverages evolutionary information from MSA (Thompson et al., 1994; Mirdita et al., 2022).
- *Meta-Learning:* These auxiliary tasks are used for meta-training via MAML (Finn et al., 2017), equipping the pLM for rapid adaptation.

- *Parameter-Efficient Adaptation:* To prevent overfitting under few-shot settings (Snell et al., 2017), FSFP freezes the backbone pLM and trains LoRA (Hu et al., 2022b) parameters during meta-training and when transferring to target task; meta-learning yields a better initialization of the LoRA parameters, and the subsequent adaptation continues to update these parameters.

Benchmarking on 87 protein datasets (Notin et al., 2023), FSFP demonstrates superior performance compared to strong baselines. It outperforms unsupervised zero-shot inference (including ESM-1v (Meier et al., 2021), ESM-2(Lin et al., 2023), and Saprot (Su et al., 2023)) and significantly surpasses the state-of-the-art supervised few-shot baseline, Ridge Regression (Hsu et al., 2022). FSFP boosts the average spearman correlation by up to 0.1 and raises positive hit rates in real-world engineering by 25% (Zhou et al., 2024).

Scaling laws for dataset allocation in pLM training (Cheng et al., 2024) have been empirically established to maximize training efficiency under fixed compute. Similar to LLMs in NLP tasks, simply increasing model size or compute is insufficient (Kaplan et al., 2020a; Hoffmann et al., 2022); the size *and* diversity of training data are critical. Empirical analysis shows that masked language models (MLMs) (Devlin et al., 2019) and causal language models (CLMs) (Radford et al., 2019; Brown et al., 2020) follow distinct power-law relationships between model size and required data, with sublinear data scaling (Hestness et al., 2017). Repeating small datasets across many epochs yields overfitting in MLMs (Roelofs et al., 2019) and diminishing returns in CLMs.

To address these issues, the study introduces the *UniMeta200B* dataset (Mitchell et al., 2020; Nayfach et al., 2021), aggregating diverse metagenomic protein sequences with strict deduplication. This improves out-of-domain perplexity and yields more stable independent and identically distributed validation curves with reduced variance across families.

The resulting scaling laws provide explicit formulas for allocating parameters and dataset size under a compute constraint, enabling principled, data-centric decisions in pLMs development. For fitted relations and resource-allocation details, see Section 4.2. These results highlight that enhancing dataset diversity—rather than simply expanding model size or repeating data—is essential for efficient protein language modeling.

Takeaway in §2 for Efficiency-focused pLMs Dataset Methods

Dataset efficiency-focused strategies—such as meta-learning-based few-shot and principled scaling laws—enable efficient pLMs training under label scarcity and compute constraints, offering concrete solutions to key bottlenecks in efficient dataset usage for pLMs.

3 Efficient Architecture

3.1 Background and Challenges

Model architecture fundamentally determines the representational capability and computational complexity of language models. LLMs in NLP predominantly employ transformer architectures (Vaswani et al., 2017), characterized by self-attention mechanisms that effectively capture long-range-dependencies. However, the quadratic scaling of transformers in computational complexity and memory consumption with respect to input sequence length poses significant efficiency challenges (Vaswani et al., 2017; Hoffmann et al., 2022). Alternatively, convolutional neural networks (CNNs) (Kalchbrenner et al., 2016; Gehring et al., 2017), leveraging local receptive fields, scale linearly with sequence length and inherently incorporate positional information through sliding window operations. These distinct architectural paradigms have informed similar developments in pLMs, where transformers remain dominant due to their robust performance on capturing long-range residue interactions (Rives et al., 2021), while CNN-based methods have emerged as efficient alternative (Yang et al., 2024).

Efficiency-focused model architectures in LLMs typically include strategies such as weight quantization, embedding dimension reduction, and structural sparsity to mitigate computational overheads (Tay et al., 2022; Frantar et al., 2022; Frankle & Carbin, 2018). Analogously, recent advancements in pLM architectures have pursued quantization of weights (Turnbull et al., 2024), embedding compression (Lu et al., 2025), and cross-attention (Vaswani et al., 2017) modules conditioned on homologous sequences (Sgarbossa & Bitbol,

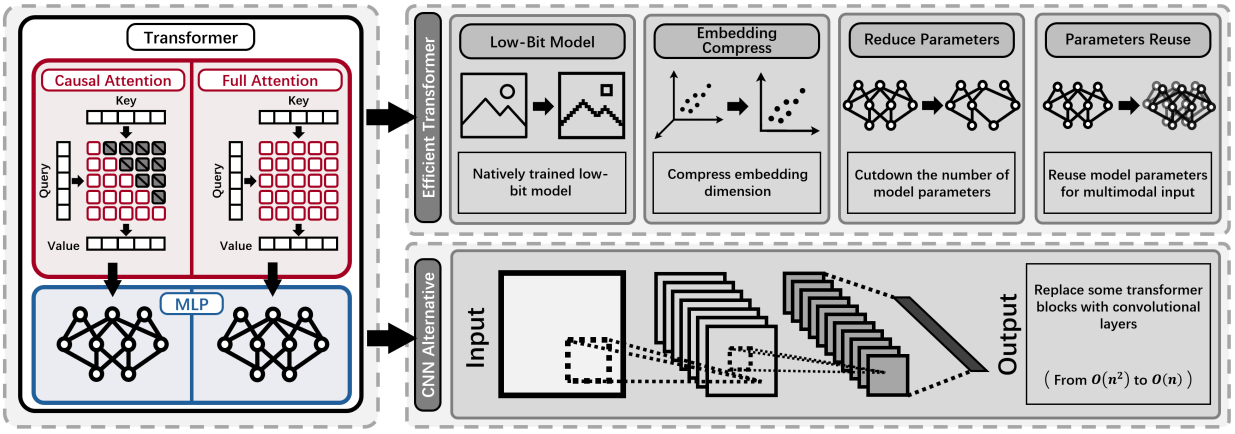

Figure 5: Overview of efficient architecture methodologies. These strategies diverge from the standard Transformer (left), branching into 'Efficient Transformer' approaches (top right) such as low-bit quantization, embedding compression, and parameter reduction/reuse, and 'CNN Alternatives' (bottom right) which replace attention blocks with convolutional layers to achieve linear $O(n)$ scaling complexity.

Table 2: Summary of representative efficient model architectures for pLMs.

Category	Method	Brief Description
Transformer-based	2Bits of Protein (Turnbull et al., 2024)	Training with ternary weights for memory and compute efficiency.
	RAG-ESM (Sgarbossa & Bitbol, 2025)	Lightweight retrieval-augmented model with parameter-sharing.
	CHEAP (Lu et al., 2025)	Compressing high-dimensional embeddings while preserving structure.
	Cramming Challenge (Frey et al., 2024b)	Train a performant pLM in 24 hours on a single GPU.
CNN-based	CARP (Yang et al., 2024)	Using dilated-convolutions (ByteNet-style) to replace attention layers.
	Hyena operator (Zhang & Okumura, 2024)	Implicit long convolutions for subquadratic scaling on long sequences.

2025) to integrate evolutionary information efficiently (Hopf et al., 2017; Jumper et al., 2021b). Convolution-based models, with inherent linear scaling, further enhance computational efficiency, presenting competitive alternatives for large-scale protein modeling tasks (Yang et al., 2024).

The emergence of efficiency-focused model architectures in pLMs is largely driven by several challenges in model architecture, including:

- **High Computational and Memory Cost:** Standard transformer-based pLMs require quadratic time and memory with respect to sequence length, leading to prohibitive resource demands for long proteins or large models. The use of full-precision weights and high-dimensional embeddings further increases GPU and energy costs, particularly during large-scale training and deployment (Vaswani et al., 2017; Rives et al., 2021; Turnbull et al., 2024; Lu et al., 2025; Yang et al., 2024).

- **Redundancy and Massive Activations in Embeddings:** Protein embeddings in large models are highly compressible, indicating over-parameterization, and also exhibit massive activations in certain channels, leading to inefficient memory use (Lu et al., 2025).

- **Limited Inductive Bias and Information Integration:** Existing architectures may not efficiently leverage important biological information, such as evolutionary relationships or sequence homology, constraining their performance and flexibility in specialized tasks (Sgarbossa & Bitbol, 2025).

The following section reviews recent architectural strategies developed to overcome these challenges. Table 2 provides an overview of these methods, and Figure 5 illustrates representative efficient architectural designs.

3.2 Efficient Transformer

2Bits of Protein (Turnbull et al., 2024) is a ternary-quantization architecture (based on (Courbariaux et al., 2016)) which offers an efficient solution to the computational and memory bottlenecks of large pLMs. In this approach, all linear layers except the embedding and output head are quantized to ternary precision, with weights constrained to $\{-1, 0, 1\}$ by scaling and rounding each weight matrix (Ma et al., 2024; Li et al., 2016; Zhu et al., 2016). Specifically, for a weight matrix W, the quantized matrix $\widetilde{W}$ is defined as:

$$\widetilde{W} = \text{RoundClip}\left(\frac{W}{\gamma + \epsilon}, -1, 1\right),$$

where γ is the mean absolute value of W's elements, and $\text{RoundClip}(x, a, b)$ rounds x to the nearest integer within $[a, b]$.

This design achieves substantial reductions in GPU memory and energy usage—prior NLP work reports up to 3.55× lower memory and 21.7× less energy consumption compared to full-precision models (Ma et al., 2024). The ternary pLM preserves the standard transformer encoder architecture (Vaswani et al., 2017; Lin et al., 2023) and remains trainable from scratch with stable performance (Banner et al., 2019). On ProteinGym (Notin et al., 2022), ternary ESM2-8M attains a mean spearman-correlation of 0.181 across 217 zero-shot tasks, competitive with efficient baselines such as ProtGPT2 (Ferruz et al., 2022) and UniRep (Alley et al., 2019). Under matched compute (24 GPU-hours, 'crammed' training (Frey et al., 2024a)), the performance gap becomes statistically insignificant, whereas under the standard training schedule, ESM2-8M is about 25% higher and statistically significant ($p < 0.05$). Thus, ternary-quantization provides a principled and highly efficient design for pLMs, substantially reducing computational cost while maintaining competitive predictive performance.

Compressed Hourglass Embedding Adaptations of Proteins (CHEAP) (Lu et al., 2025) provide an architectural solution for drastically reducing the dimensionality and memory footprint of pLMs embeddings, while retaining both sequence and structure information. CHEAP employs a transformer-based hourglass-autoencoder (Newell et al., 2016; Kingma & Welling, 2013), which applies linear downsampling and projection to the ESMFold (Lin et al., 2023) latent embedding. Its key innovations can be divided into:

- *Continuous Compression:* CHEAP enables up to 128× channel compression and 8× sequence length compression through a learnable bottleneck, allowing backbone structure reconstruction with less than 1.34 Å root-mean-squared distance (Kabsch, 1976) at 32× channel compression, and maintaining sequence recovery above 99% until fewer than 8 channels remain. To address abnormal massive activations in certain embedding channels, CHEAP introduces per-channel normalization (Ba et al., 2016; Ioffe & Szegedy, 2015), further improving compressibility.
- *Discrete (Tokenized) Compression:* CHEAP supports discrete embeddings via finite scalar quantization (Mentzer et al., 2023), which outperforms VQ-VAE (Van Gerven & Bohte, 2017; Van Den Oord et al., 2017) for large codebooks on latent and structure reconstruction, and it also provides a representation of protein structure obtainable from sequence alone.

These compressed embeddings enable efficient downstream applications—such as protein design, similarity search, and functional prediction—making advanced pLM representations accessible for resource-limited scenarios while preserving high-resolution information.

RAG-ESM (Sgarbossa & Bitbol, 2025) introduces a parameter-efficient encoder-decoder architecture that augments pretrained ESM2 (Lin et al., 2023) with a minimal number of cross-attention layers and extensive parameter-sharing. The encoder embeds a context (homologous) sequence using the original ESM2 model, while the decoder reuses pretrained ESM2 layers, inserting a few cross-attention layers to transfer information from the context embedding to the masked input sequence. Critically, self-attention and feed-forward layers in both modules share weights, so the number of new parameters is negligible compared to the backbone model. Formally, the total parameter count of RAG-ESM is

$$N_{\text{RAG-ESM}} = N + M \quad (M \ll N),$$

where N is the number of parameters in the pretrained ESM2 backbone, and M is the number of newly added cross-attention parameters. Conditioning on homologs during inference allows RAG-ESM to extract

evolutionary information (Altschul et al., 1990) without requiring MSAs (Jumper et al., 2021b; Remmert et al., 2012; Rao et al., 2021) or significantly increasing model size.

Empirically, RAG-ESM models with 12M and 165M parameters achieve 48% and 43% lower perplexity, respectively, than their ESM2-8M and ESM2-150M baselines on masked AA prediction. Both are trained under the discrete-diffusion (Austin et al., 2021; Hoogeboom et al., 2021) using the closest homolog as context (Steinegger & Söding, 2017), with modest compute budgets (50–120 GPU-hours). Cross-attention heads naturally learn alignment (Vig & Belinkov, 2019); several heads reach $\rho > 0.6$, and a logistic regression trained on all 60 heads achieves an average $\rho = 0.76$ against Needleman–Wunsch (Needleman & Wunsch, 1970) alignment matrices.

This design enables efficient and scalable homology-aware pLMs that achieve state-of-the-art performance among sequence-based models for homolog-conditioned generation and remain competitive on motif scaffolding (Dauparas et al., 2022), while retaining flexibility.

Cramming Protein Language Model Training in 24 GPU Hours (Frey et al., 2024b) introduces an efficient transformer-based architecture specifically optimized for rapid, resource-constrained pre-training. The corresponding pre-training setup is detailed in Section 4.2.

The core approach adapts the HuggingFace ESM2 (Lin et al., 2023) backbone by removing: (1) *all query, key, and value biases in attention blocks*, (2) *biases in intermediate linear layers*, thereby reducing computational overhead while retaining model capacity to maximize per-token throughput under a stringent compute budget: model weights are initialized from scratch, and training is limited within 24 hours on a single GPU.

These architectural and implementation choices collectively allow a 67M-parameter model to achieve competitive downstream performance—on tasks such as protein fitness landscape inference (Notin et al., 2023) and protein–protein interaction classification (Szklarczyk et al., 2021)—with models trained for over 15,000× greater GPU hours. It demonstrates that by simplifying architectural details, it is feasible to obtain expressive pLMs suitable for practical research and deployment under extreme computational constraints.

3.3 Convolutional Layer Based Methods

Convolutional Autoencoding Representations of Proteins (CARP) (Yang et al., 2024) demonstrates that convolutional architectures can match or exceed transformer-based pLMs (Rives et al., 2021; Lin et al., 2023) in both predictive power and computational efficiency. CARP replaces transformer self-attention with ByteNet-style (Kalchbrenner et al., 2016) dilated convolutional blocks (Yu & Koltun, 2015; Chen et al., 2017), exponentially expanding the receptive-field while ensuring linear scaling in computation and memory with sequence length. Formally, the computational complexity per layer is:

$$\begin{cases} \mathcal{O}(L^2) & \text{for transformer self-attention} \\ \boldsymbol{\mathcal{O}(L)} & \textbf{for convolution in CARP} \end{cases}$$

where L is the input sequence length.

CARP-640M (640M parameters) achieves a masked language modeling loss of 2.02, close to ESM-1b's (Rives et al., 2021) 1.96, and attains higher average zero-shot mutation effect prediction (Spearman 0.49 vs. 0.46) on 41 datasets. Unlike ESM-1b, CARP processes sequences up to 4,096 residues without memory overflow and maintains stable loss and runtime for longer sequences. These results support convolutional architectures as efficient, scalable, and competitive alternatives to transformers for large protein sequence modeling.

ProtHyena (Zhang & Okumura, 2024) introduces the Hyena operator (Poli et al., 2023) as an efficient alternative to self-attention and standard convolution in pLMs. The per-layer computational complexity is:

$$\begin{cases} \mathcal{O}(L^2) & \text{Transformer self-attention} \\ \mathcal{O}(L) & \text{Standard CNN} \\ \boldsymbol{\mathcal{O}(L \log_2 L)} & \textbf{The Hyena operator} \end{cases}$$

where L is the input sequence length.

Compared to transformer models whose self-attention layers require quadratic complexity, the Hyena operator enables substantially improved scalability; compared to standard CNNs, whose long-range modeling is limited by kernel size, Hyena leverages implicit long convolutions (Gu et al., 2021) and element-wise gating (Dauphin et al., 2017; Hochreiter & Schmidhuber, 1997) to efficiently capture both local and global dependencies. The core block applies:

$$y = x_N \cdot (h_N * (x_{N-1} \cdot (h_{N-1} * \cdots x_1 \cdot (h_1 * v))))$$

where N is block depth, h_i are long convolution filters, v is the input projection, and $\cdot$ denotes gating.

This enables ProtHyena to process sequences up to a million residues and achieve accuracy competitive with or better than larger transformer-based pLMs, such as ESM-1b (Rives et al., 2021), ESM-2 (Lin et al., 2023), and ProteinBERT (Brandes et al., 2022), using about 10% of their parameters and up to 60× speedup.

Takeaway in §3 for Efficiency-focused pLMs Model Architectures

Recent architectural advances—including quantization, embedding compression, cross-attention with homology conditioning, convolutional alternatives, and the design of transformer architectures optimized for rapid "cramming" pre-training—directly address key computational and representational bottlenecks in pLMs, enabling efficient and scalable model design without sacrificing predictive power.

4 Efficient Training

4.1 Background and Challenges

Training LLMs generally follows two stages: *pre-training* and *post-training*. During pre-training, models are trained on large unlabeled corpora to learn general representations (Radford et al., 2018; Devlin et al., 2019). post-training then adapts these models to downstream objectives (including, for example, supervised fine-tuning (SFT) (Raffel et al., 2020) and reinforcement learning (RL)-based tuning such as RLHF (Ouyang et al., 2022) or GRPO (Shao et al., 2024)). In this paper, we mainly focus on *fine-tuning* (Howard & Ruder, 2018) within post-training. In parallel, *distillation* transfers knowledge from high-capacity teachers to compact students for efficiency (Hinton et al., 2015; Gou et al., 2021; Sanh et al., 2019).

Training pLMs adopt a conceptually similar pipeline to LLMs but with several domain-specific characteristics: (1) *Pre-training:* pLMs are trained on large corpora of unlabeled protein sequences at the amino-acid level to learn general biological representations (Elnaggar et al., 2021; Rives et al., 2021; Bepler & Berger, 2019; Alley et al., 2019). (2) *Post-training:* pLMs are further adapted on task-specific datasets for specialized bioinformatics objectives (Min et al., 2017), for example, function prediction (Schmirler et al., 2024; Gligorijević et al., 2021). (3) *Distillation:* knowledge is transferred from large teacher models to smaller students for efficient deployment, which closely mirrors the approach used in LLMs (Shang et al., 2024; Hu et al., 2024; Sanh et al., 2019; Sun et al., 2020).

Given these characteristics, improving computational and data efficiency has become a key objective in pLMs training. Efficiency-focused pLMs training therefore focuses on the following bottlenecks:

- *Challenges in Pre-training:*
  - **Lack of Resource Allocation Guidelines:** It is challenging to optimally balance model size and training dataset size under fixed compute, with suboptimal allocation leading to overfitting (Srivastava et al., 2014) or diminished returns (Cheng et al., 2024).
  - **Limited Structural Integration:** Most pre-training pipelines rely solely on sequence data, whereas structural information remains underutilized, which may ultimately limit both the efficiency and generalization capabilities of the models (Heinzinger et al., 2024).
- *Challenges in Fine-tuning:*
  - **Lack of Comprehensive Empirical Studies:** Compared to NLP (Ding et al., 2022), systematic evaluations of parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Hu et al., 2022b; Lester

Table 3: Summary of representative training-stage efficiency methods for pLMs.

Category	Method	Brief Description
Pre-training	Scaling Law (Cheng et al., 2024)	Empirical study on pLMs optimal data/model allocation.
	Structure Integration (Heinzinger et al., 2024)	Joint sequence-structure pre-training (e.g., ProstT5).
	Cramming Challenge (Frey et al., 2024b)	Train a performant pLM in 24 hours on a single GPU.
Fine-tuning	A study of PEFT Methods (Schmirler et al., 2024)	LoRA for parameter-efficient tuning.
	QLoRA (Clancy et al., 2025)	4-bit quantization + LoRA for efficient fine-tuning.
	SES-Adapter (Tan et al., 2024)	Structure-aware adapters for cross-modal representation.
	FSFP (Zhou et al., 2024)	few-shot with meta-learning and LoRA.
Distillation	Cross-distillation in AlphaFold3 (Abramson et al., 2024)	Distillation from a single large teacher-model.
	ProtGO (Hu et al., 2024)	Transferring functional and structural knowledge.
	VespaG (Marquet et al., 2024)	Distillation via evolutionary expert supervision.
	MTDP (Shang et al., 2024)	Aggregating knowledge from multiple teacher models.

et al., 2021; Li & Liang, 2021a)methods for pLMs across diverse tasks has been limited, leaving practical guidelines unclear (Schmirler et al., 2024).
  - **High Computational Demands:** Full-model fine-tuning of large pLMs requires substantial memory and compute, limiting accessibility and scalability, especially under limited hardware situations (Schmirler et al., 2024; Clancy et al., 2025).
  - **Low Data Efficiency:** Adapting models to new tasks with limited labeled data can be inefficient, as conventional fine-tuning is prone to overfitting and resource waste (Zhou et al., 2024).
- *Challenges in Distillation:*
  - **Teacher-student Discrepancy:** Differences in size, architecture, or modality (e.g., sequence, structure, function) between teacher and student models can hinder efficient knowledge transfer and limit student performance (Abramson et al., 2024; Hu et al., 2024; Mirzadeh et al., 2020).
  - **Knowledge Transfer Bottlenecks:** Distilling large teacher models or integrating multi-modal knowledge (e.g., structure, function) into compact students can reduce predictive power or require extensive tuning (Marquet et al., 2024; Hu et al., 2024; Shang et al., 2024).

Following sections review methods addressing challenges through efficient pre-training, fine-tuning, and knowledge distillation methods.

4.2 Efficient Pre-training Methods

To orient the discussion, Figure 6 summarizes the efficient pre-training methods reviewed in this section.

Scaling law-guided pre-training for pLMs (Cheng et al., 2024) empirically establishes optimal resource allocation between model size and data for transformer-based pLMs. On a large dataset of 939M protein sequences, over 300 models were trained to reveal that both masked (MLM) and causal (CLM) objectives follow distinct power-law scaling similar to Kaplan et al. (2020a) and Hestness et al. (2017). For a 10× increase in compute, optimal MLM size increases 6× with 1.7× more data, while CLM size grows 4× with 3× more data. The fitted scaling relations for MLMs are:

$$\begin{cases} N_{\text{MLM}}^{\text{opt}} = 6.19 \times 10^{-8} C^{0.776} & \text{MLMs' model parameters} \\ D_{\text{MLM}}^{\text{opt}} = 2.02 \times 10^{6} C^{0.230} & \text{MLMs' training tokens} \end{cases}$$

and for CLMs are:

$$\begin{cases} N_{\text{CLM}}^{\text{opt}} = 1.26 \times 10^{-3} C^{0.578} & \text{CLMs' model parameters} \\ D_{\text{CLM}}^{\text{opt}} = 1.23 \times 10^{2} C^{0.422} & \text{CLMs' training tokens} \end{cases}$$

where C represents the total pre-training FLOPs, N is the number of forward-activated non-embedding parameters, and D represents the number of training tokens.

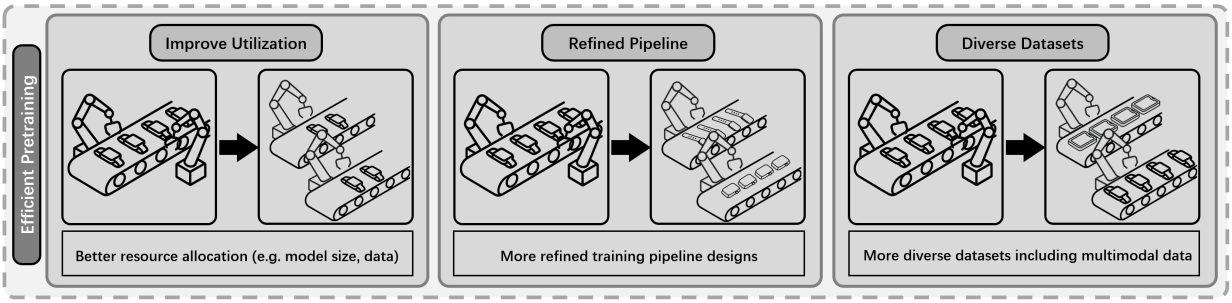

Figure 6: Conceptual illustration of efficient pretraining strategies: Improve Utilization — optimizing resource allocation such as model size and data scaling; Refined Pipeline — developing more streamlined and effective training pipeline designs; Diverse Datasets — incorporating more varied datasets, including multimodal data, to enhance generalization and robustness.

Repeating data leads to overfitting for MLMs and diminishing returns for CLMs (Hernandez et al., 2022), which is mitigated by introducing a diverse, deduplicated dataset (UniMeta200B). Transfer learning—pretraining with CLM and then fine-tuning with MLM—reduces the compute needed for optimal MLMs by up to 7.7×. Large-scale validations confirm that these scaling laws yield models with better generalization and downstream performance under fixed compute budgets, particularly in resource-constrained environments.

ProstT5 (Heinzinger et al., 2024) is a T5-based model (Raffel et al., 2020) that integrates structure information (Jing et al., 2020) into pLM pretraining to boost efficiency and versatility. ProstT5 unifies protein sequence (Amino Acid, AA) and structural (3Di token) modalities within a single encoder-decoder architecture. Its key innovations in the pretraining stage can be divided into:

- *Multimodal Pre-training:* By encoding 3D structures from AlphaFoldDB (Varadi et al., 2024) as 3Di-token sequences, and expanding the vocabulary with 3Di tokens, ProstT5 applies span-based denoising objectives to both AA and 3Di sequences, teaching the model new structural tokens while avoiding catastrophic forgetting of sequence information.
- *Bi-directional Translation Objectives:* ProstT5 is trained to translate between AA and 3Di representations using direction tags (`<fold2AA>`, `<AA2fold>`), enabling both "folding" (AA→3Di) and "inverse folding" (3Di→AA), thus robustly linking sequence and structure information.

This bilingual setup allows ProstT5-predicted 3Di strings to be used for structure-level similarity search (e.g., with Foldseek (Van Kempen et al., 2024)), achieving remote homology detection (Rost, 1999) sensitivity nearly matching experimental structures, but with a speedup of over three orders of magnitude (e.g., 43 seconds vs. 48 hours for full proteome annotation) (Van Kempen et al., 2024; Mirdita et al., 2022). Embedding-based annotation transfer further shows improved CATH fold classification (Sillitoe et al., 2021) accuracy over ProtT5 (Elnaggar et al., 2021) and ESM-1b (Rives et al., 2021), and secondary structure prediction with ProstT5 matches or exceeds prior state-of-the-art. In inverse folding (Dauparas et al., 2022; Hsu et al., 2022), ProstT5 generates diverse sequences with predicted structures closely matching targets, demonstrating both enhanced efficiency and versatility.

Overall, integrating structure into pLM pretraining via bilingual modeling delivers substantial gains in inference speed and structural generalization, advancing protein design and large-scale annotation capabilities.

Cramming Protein Language Model Training in 24 GPU Hours (Frey et al., 2024b) proposes a pre-training paradigm specifically tailored for rapid and resource-constrained scenarios. In this framework, a transformer-based pLM is trained from scratch under a strict 24 GPU-hour budget, with no use of pre-trained models at any stage. The pre-training pipeline consists of several key components:

- *Random initialization:* All model weights are initialized randomly.
- *On-the-fly masked language modeling:* Masked language modeling is performed directly on UniRef50 (Suzek et al., 2015) splits, with all data processing (e.g., tokenization, filtering, sorting) occurring on-the-fly during training, and no offline pre-processing.
- *Large effective batch size:* Training employs a batch size of 128, sequence length of 512, and gradients are accumulated over 16 steps for an effective batch of 2048 sequences.
- *Critical hyperparameter tuning:* The optimizer is AdamW (Loshchilov & Hutter, 2017) with carefully tuned parameters. The learning rate schedule is central: the optimal regime employs a fast warmup (Goyal et al., 2017; Liu et al., 2020) to a peak learning rate of 1×10^{-3} over 1,000 steps, followed by a slow linear decay (Loshchilov & Hutter, 2016) to near zero over the remaining steps, with a training process capped at 50,000 updates.

These pre-training choices allow the resulting 67M-parameter model to match or approach the downstream performance of much larger models such as ESM2-3B (Lin et al., 2023)—trained for over 15,000× more GPU hours—on tasks including protein fitness landscape inference and PPI classification. The architectural design supporting these results is detailed in Section 3.2. This work demonstrates that, through systematic pre-training design and hyperparameter tuning (Bergstra & Bengio, 2012; Snoek et al., 2012), performant pLMs can be obtained at a fraction of the conventional computational cost, enabling broader accessibility for rapid biological modeling.

Takeaway in §4.2 for Efficiency-focused pLM Pre-Training

Scaling law-guided resource allocation, the integration of structural information, and the development of rapid "cramming" pre-training protocols have substantially improved the efficiency and versatility of pLMs pre-training, providing practical solutions for balancing compute, data, and biological knowledge—even under stringent resource constraints—in large-scale protein modeling.

4.3 Efficient Fine-tuning Methods

Fine-tuning in pLMs involves adapting a pre-trained model to specific downstream biological tasks using relatively small amounts of labeled protein data (Schmirler et al., 2024). This step is inherently efficiency-focused, enabling rapid model adaptation without extensive retraining. Advanced PEFT techniques, such as LoRA (Hu et al., 2022a) and structure-aware adapters (Tan et al., 2024), further minimize computational and memory demands while maintaining high performance. Subsequent subsections review these PEFT methods and associated efficiency gains.

As a summary, Figure 7 provides an overview of the key techniques covered for efficient fine-tuning.

A study of Parameter-efficient fine-tuning (PEFT) methods systematically compared full-model fine-tuning with several PEFT approaches—including LoRA (Hu et al., 2022b), DoRA (Liu et al., 2024a), IA3 (Liu et al., 2022), and prefix tuning (Li & Liang, 2021b)—across diverse pLMs and protein prediction tasks (Schmirler et al., 2024). Its main findings include:

- *Efficiency and accuracy:* By freezing most model weights and only updating a small fraction (e.g., 0.25% for LoRA), PEFT methods, such as LoRA, achieved nearly equivalent accuracy to full fine-tuning.
- *Training speedup and memory savings:* For larger models, LoRA offered up to a 4.5× training speedup with comparable GPU memory requirements, and saving only adapter weights, further improving efficiency.
- *Compatibility and resource efficiency:* LoRA, DoRA, and similar methods are compatible with mixed precision, gradient accumulation, and CPU-offloading, enabling fine-tuning even on 8GB or 16GB GPUs.
- *Method comparison:* Among PEFT methods, LoRA was generally most compute-efficient for large pLMs, with DoRA sometimes up to 30% slower, and all PEFT methods yielded average prediction gains of 61.3% compared to using static embeddings.

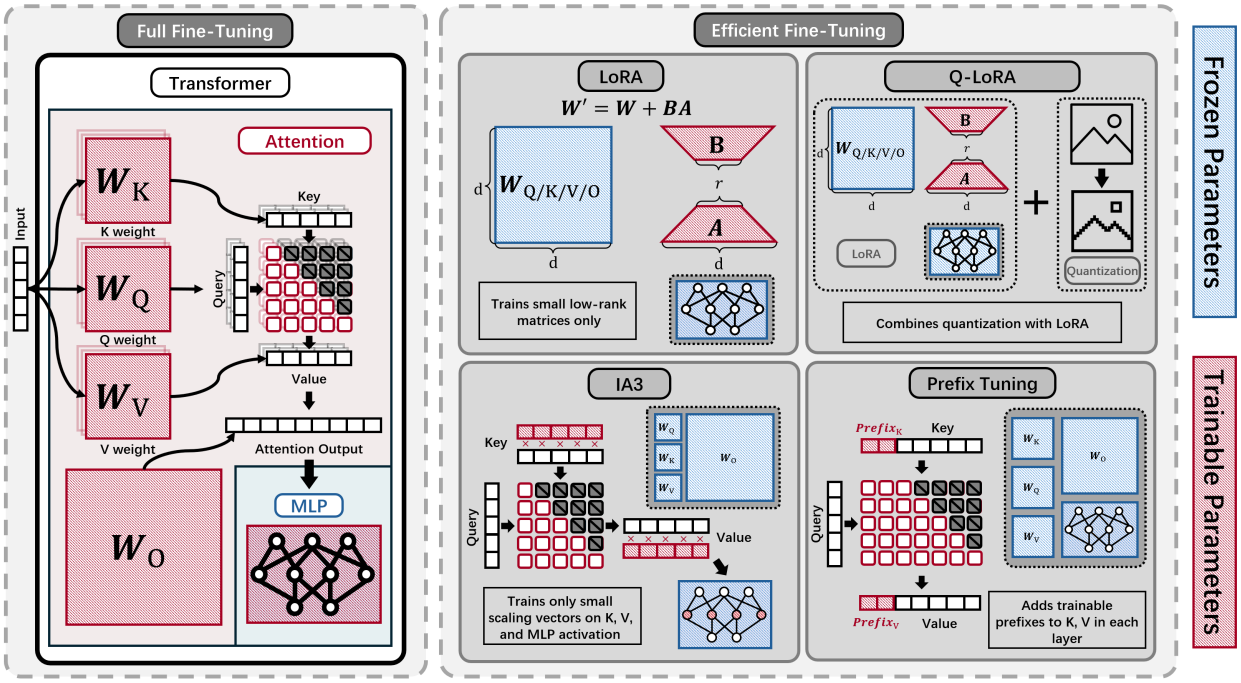

Figure 7: A comparison of Full Fine-Tuning (left) with several mainstream parameter Parameter-Efficient Fine-Tuning (PEFT) methods (right) for pLMs. In full fine-tuning, all model parameters are trainable. PEFT methods keep the backbone parameters frozen (blue) and introduce a small number of new trainable parameters (pink). Examples shown include LoRA, which adds low-rank adaptation matrices; Q-LoRA, which combines LoRA with quantization; IA3, which trains scaling vectors; and Prefix Tuning, which adds trainable prefixes to key and value vectors.

PEFT—especially LoRA—provides a practical solution for adapting large pLMs in resource-constrained environments, with all tested PEFT methods demonstrating substantial improvements over static, pre-trained embeddings.

Quantization with LoRA (QLoRA) (Clancy et al., 2025) combines 4-bit quantization with LoRA to enable highly efficient fine-tuning of pLMs while maintaining strong performance (Dettmers et al., 2023). By quantizing most model weights to 4 bits and updating only a small set of adapter parameters, QLoRA substantially reduces memory and computation costs.

Clancy et al. (Clancy et al., 2025) systematically evaluated QLoRA across multiple pLMs, including ESM-2 (Lin et al., 2023), ESM C (Team, 2024), ProtBERT (Brandes et al., 2022), ProtT5-half (Elnaggar et al., 2021), and Ankh-base (Elnaggar et al., 2023), and diverse tasks, including GFP fluorescence (Sarkisyan et al., 2016), protease stability (Rocklin et al., 2017), and protein secondary structure prediction (Berman et al., 2000; Kryshtafovych et al., 2021; Klausen et al., 2019). Experiments show that QLoRA achieves an average GPU memory reduction of 46.7%, with generative models such as ProtGPT2 and ProLLaMA reaching up to 76.4% reduction.

Performance remains largely preserved: for ESM C 600M, fluorescence prediction SpearmanR is 0.850 for QLoRA versus 0.863 for full precision, and secondary structure accuracy remains identical at 0.870. In generative tasks, 4-bit quantization yields negligible differences in predicted protein quality, with metrics such as Foldseek pass rate, local distance difference test, and predicted local distance difference test showing no significant decline.

Overall, QLoRA lowers hardware barriers for fine-tuning and inference of large pLMs, making efficient protein modeling feasible on modest computing infrastructure.

Few-Shot Learning for Protein Fitness Prediction (FSFP) (Zhou et al., 2024) builds on the dataset construction and meta-learning strategies described in Section 2.2, and focuses on PEFT for robust fitness prediction with minimal labeled data.

After meta-training on auxiliary tasks constructed from related proteins and MSA-based pseudo-labels, FSFP fine-tunes only a small subset of parameters (e.g., via LoRA with rank $r = 16$) to mitigate overfitting. Fitness prediction is formalized as a ranking problem, with ListMLE loss (Xia et al., 2008) used to optimize the ordering of variants. On 87 deep mutational scanning datasets, FSFP outperforms both zero-shot and regression-based baselines, improves spearman-correlation by up to 0.1 with just 20 labeled mutants, and demonstrates strong extrapolation to unseen mutations. Wet-lab validation further highlights its efficiency and real-world utility.

SES-Adapter (Tan et al., 2024) is a structure-aware, PEFT framework for pLMs, integrating sequence and structural information via cross-modal multi-head attention. In SES-Adapter, the pLM backbone is frozen, and only the adapter module is trained. Structural features representations are serialized via, such as, FoldSeek (Van Kempen et al., 2024) or DSSP (Kabsch & Sander, 1983), into embeddings and serve as queries (Q), while pLM sequence embeddings serve as keys and values (K, V). Uniquely, both Q and K are first embedded with rotary position embedding (RoPE). For an embedding vector $x \in \mathbb{R}^{d_k}$ at position p, RoPE performs a rotation on each coordinate pair $(2i, 2i+1)$:

$$\begin{bmatrix} (\mathrm{RoPE}(x,p))_{2i} \\ (\mathrm{RoPE}(x,p))_{2i+1} \end{bmatrix} = \begin{bmatrix} \cos\theta_{p,i} & -\sin\theta_{p,i} \\ \sin\theta_{p,i} & \cos\theta_{p,i} \end{bmatrix} \begin{bmatrix} x_{2i} \\ x_{2i+1} \end{bmatrix}, \quad i = 0, \ldots, \tfrac{d_k}{2} - 1,$$

where $\theta_{p,i}$ is a position-dependent rotation angle varying with p and the dimension index i.

This fusion enables SES-Adapter to consistently outperform vanilla pLMs and prior tuning strategies: on nine diverse benchmarks, it yields up to 11% (avg. 3%) accuracy improvement, convergence acceleration up to 1034% (avg. 362%), and maintains robustness to structural noise (performance variation <0.6% across different structure predictors). SES-Adapter thus offers an efficient, scalable, and model-agnostic solution for fine-tuning pLMs in protein research.

Takeaway in §4.3 for Efficiency-focused pLMs Fine-Tuning

Recent advances in PEFT—such as LoRA, QLoRA, meta-learning, and structure-aware adapters—enable large pLMs to be effectively adapted with limited computational resources and small labeled datasets. These approaches address key bottlenecks in scalability and accessibility, and provide practical solutions for efficient, robust downstream adaptation in diverse biological tasks.

4.4 Efficient Training via Distillation

knowledge distillation (Hinton et al., 2015) in pLMs is a widely adopted model compression technique that transfers knowledge from larger, high-capacity teacher models to smaller, efficient student models. It typically involves guiding student models to mimic teacher outputs, intermediate representations, or other informative signals(Zagoruyko & Komodakis, 2016). Distillation inherently emphasizes efficiency, enabling deployment of lightweight models that substantially reduce computational and memory requirements (Cheng et al., 2017; Han et al., 2015) while retaining predictive capabilities comparable to larger models (Wang & Yoon, 2021; Geffen et al., 2022). Recent literature highlights diverse distillation strategies—including single-teacher (Abramson et al., 2024), expert-guided (Marquet et al., 2024), cross-modal (Hu et al., 2024), and multi-teacher adaptive methods (Shang et al., 2024), demonstrating significant efficiency improvements.

Cross-distillation of AlphaFold 3 (Abramson et al., 2024) is a strategy for mitigating hallucinations and improving data efficiency in diffusion-based protein structure prediction (Watson et al., 2023). Its core methodological steps are:

- *Teacher-augmented training data:* The training set is augmented by including high-confidence structures predicted by AlphaFold-Multimer v2.3 (Evans et al., 2021; Židek, 2022), in addition to experimental structures. Teacher-model predictions act as additional supervisory targets during training (Lee et al., 2013; Xie et al., 2020).
- *Joint supervision with ground-truth and teacher outputs:* AlphaFold 3 is explicitly optimized to align its outputs not only with experimental ground-truth structures, but also with reliable teacher-generated structures. This enables the model to learn more physically realistic disorder regions, as teacher-generated structures tend to represent unstructured segments as extended loops, reducing hallucinated compact order in intrinsically disordered regions.

On the CAID 2 benchmark (Kryshtafovych et al., 2021), cross-distillation led to better alignment between predicted confidence (predicted local distance difference test) and actual disorder, effectively suppressing generative artifacts. By increasing both the diversity and quantity of usable training data, intra-chain metrics reach 97% of their maximum within the first 20,000 training steps, reflecting rapid and efficient learning. Overall, cross-distillation exemplifies how leveraging both ground-truth and high-quality model-generated supervision can simultaneously reduce hallucinations and accelerate robust protein structure learning in next-generation diffusion-based models.

ProtGO (Hu et al., 2024) is a multimodal protein representation learning framework that distills functional knowledge from a teacher to a student network via distribution alignment in latent space. The teacher leverages sequence, structure, and GO (GO) (Ashburner et al., 2000; Aleksander et al., 2023) function annotations, while the student uses only sequence and structure, making it practical for annotation-scarce scenarios. Instead of aligning individual embeddings, ProtGO minimizes the KL-divergence (Kullback & Leibler, 1951) between the batch-level Gaussian-distributed graph embeddings of teacher and student:

$$L_{\text{kd}} = \text{KL}\left[p_S(z_S) \,\|\, p_T(z_T)\right]$$

where z_S and z_T are graph-level embeddings of the teacher and student models, and $p_S(\cdot)$, $p_T(\cdot)$ denote their estimated batch-wise distributions (Gaussian with mean and variance computed over each batch). The overall loss for training the student is:

$$L = L_{\text{student}} + \beta L_{\text{kd}}$$

where L_{student} is the cross entropy loss for downstream classification and β is a trade-off hyperparameter.

This strategy enables the student to inherit rich functional information, achieving state-of-the-art results: e.g., 60.5% accuracy for fold classification (3.8 points higher than prior methods), 89.4% for enzyme reaction prediction, and top F_{max} values on GO term and Enzyme Commission (EC) number prediction (Bairoch, 2000). Ablation studies confirm the distillation module is critical for performance (fold accuracy drops from 60.5% to 57.8% when removed). ProtGO thus demonstrates the efficiency and effectiveness of distribution-based multimodal distillation for robust protein representation learning.

Expert-guided distillation in VespaG (Marquet et al., 2024) offers an efficient strategy for protein variant effect prediction by training a lightweight neural network (student) to imitate an evolutionary expert model (teacher). Here, the teacher is GEMME (Laine et al., 2019)—a MSA (MSA)-based method that generates mutational effect scores—while the student is VespaG, a feed-forward neural network with a single hidden layer. VespaG takes embeddings from a pretrained pLM (ESM-2 (Lin et al., 2023)) as input and learns to predict GEMME's variant effect scores for each possible mutation. The student is trained by minimizing the mean squared error between its predictions $\hat{y}$ and the GEMME scores y^*:

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i^*)^2$$

This distillation approach removes the need for log-odds calculations and MSAs at inference time, enabling VespaG to rapidly predict mutational landscapes across entire proteomes on standard hardware. On the ProteinGym (Notin et al., 2022) benchmark (over 3 million variants), VespaG achieves a mean spearman-correlation of 0.48 ± 0.02, matching state-of-the-art methods and sometimes surpassing its expert teacher

in certain protein families. Importantly, VespaG is over three orders of magnitude faster than MSA-based methods. This demonstrates that expert-guided distillation with pLM embeddings yields robust, scalable, and practical variant effect predictors.

Multi-teacher distillation in MTDP (Shang et al., 2024) enables efficient and high-fidelity embedding by aggregating knowledge from multiple large pre-trained models, specifically ESM2-33 and ProtT5-XL-UniRef50.MTDP student, a T5-based model with ∼20M parameters, is trained via an adaptive teacher selection mechanism that dynamically assigns the most suitable guidance to each sample using RL strategy (Yuan et al., 2021). The training objective jointly minimizes a masked language modeling loss (L_{MLM}) and a distillation loss (L_{Distill}) defined by the KL divergence between student and teacher output distributions:

$$L = \alpha L_{\text{MLM}} + (1-\alpha) L_{\text{Distill}}, \quad \alpha = 0.2$$

where,

$$L_{\text{MLM}} = -\sum_{i=1}^{N} \sum_{j=1}^{V} I(y_i = j) \log p(x_i = j)$$

$$L_{\text{Distill}} = \sum_{i=1}^{N} \text{KL}\left(p_T(x_i) \parallel p_S(x_i)\right)$$

where N is sequence length, V is vocabulary size, y_i/x_i are the true/predicted amino acids at position i, $I(\cdot)$ is the indicator function, $p_T(x_i)/p_S(x_i)$ are the teacher/student output probabilities, and KL denotes KL divergence. At each step, selected teacher T for each sample is chosen via a RL-based scheduling mechanism (Yuan et al., 2021).

Compared to its teachers (650M and 120M parameters), the MTDP student is dramatically more efficient, achieving ±1.5% accuracy of large models on tasks like function and subcellular localization prediction, and requiring ∼70% less encoding time for 10,000 protein sequences. Ablation studies confirm multi-teacher distillation outperforms single-teacher setups. MTDP can be deployed on modest GPU hardware, providing a scalable, resource-efficient solution for large-scale, real-time protein informatics applications.

Takeaway in §4.4 for pLMs Distillation

Recent progress in distillation—including expert-guided, cross-modal, and multi-teacher strategies—enables efficient pLMs with reduced memory and computational requirements. These approaches address knowledge transfer and model compression challenges, making large-scale protein analysis and deployment more practical while preserving strong predictive performance.

5 Efficient Inference

5.1 Background and Challenges

Inference in language models refers to the process of generating outputs, such as predictions or embeddings, based on trained models given new input data. Autoregressive models (e.g. GPT series (Brown et al., 2020)) generate text by predicting subsequent tokens, while masked language models (e.g. BERT (Devlin et al., 2019)) infer masked tokens or produce contextual embeddings for various tasks. This inference typically involves passing inputs through multiple transformer layers to capture contextual relationships and generate meaningful, task-relevant representations.

Analogously, pLMs infer structural, functional, or evolutionary properties from protein sequences. Given a protein sequence, pLM inference generates embeddings that encode rich biological information, facilitating downstream tasks like homolog detection, structure prediction, and functional annotation (Rives et al., 2021).

Efficiency-focused inference methods in LLMs encompass techniques such as quantization (Jacob et al., 2018; Nagel et al., 2021) and embedding-based dense retrieval, aiming to reduce computational costs and memory

Table 4: Summary of representative inference-stage efficiency methods for pLMs.

Category	Method	Brief Description
Inference	PLMSearch (Liu et al., 2024b)	Efficient remote homology search using learned embeddings.
	PTQ4Protein (Peng et al., 2023b)	PTQ for reducing inference memory and compute.
	DHR (Hong et al., 2024)	Transforms protein homology detection from pairwise alignment into vector retrieval.

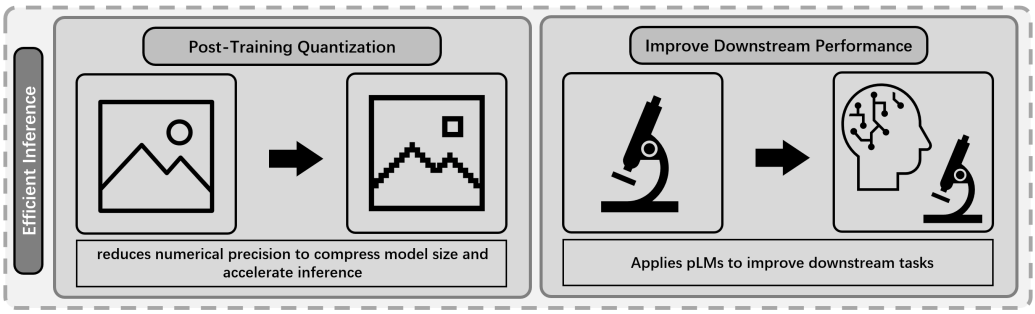

Figure 8: Strategies for improving inference efficiency: Post-Training Quantization (PTQ) — reduces numerical precision to compress model size and accelerate inference; Improve Downstream Performance — applies pLMs to enhance downstream task performance and practical applicability.

usage while preserving model performance (Bondarenko et al., 2021; Lewis et al., 2020). Similarly, efficient inference in pLMs integrates analogous strategies, including PTQ (Peng et al., 2023b), embedding-based retrieval via dual-encoder architectures (Hong et al., 2024), and structure-informed similarity scoring (Liu et al., 2024b). An overview of these efficient inference strategies is provided in Figure 8.

The emergence of efficiency-focused inference methods in pLMs is largely motivated by several challenges related to inference process, including:

- **Resource Constraints:** The inference process of large pLMs require substantial computation and memory, hindering deployment in resource-limited settings (Peng et al., 2023b).
- **Inference Latency:** Traditional alignment-based and MSA methods are inherently slow and cannot efficiently scale to large databases, which is critical for many practical applications, such as proteome-wide homology search (Liu et al., 2024b; Hong et al., 2024).
- **Limited Sensitivity:** Sequence-based retrieval may miss remote homologs and deeper structural relationships (Liu et al., 2024b; Hong et al., 2024).

The following section reviews methods that address these issues, with an overview provided in Table 4.

5.2 Efficient Inference

PTQ4Protein (Peng et al., 2023b) enables efficient inference of large pLMs such as ESMFold (Lin et al., 2023) by applying PTQ (Banner et al., 2019) to both weights and activations. Baseline experiments show that standard 8-bit uniform quantization (Jacob et al., 2018) substantially reduces memory usage (4×) but leads to significant accuracy loss—mainly due to the asymmetric and wide activation distributions before LayerNorm layers. To address this, PTQ4Protein employs piecewise linear quantization for activations, splitting the value range $[r_l, r_u]$ at breakpoints p_l and p_u, and quantizing each region separately:

$$\hat{r} = \begin{cases} \text{uni}(r_1; b-1, p_l, p_u, p_l) & r_1 \in [p_l, p_u] \\ \text{uni}(r_2; b-1, r_l, p_l, r_l) & r_2 \in [r_l, p_l) \\ \text{uni}(r_3; b-1, p_u, r_u, p_u) & r_3 \in (p_u, r_u] \end{cases}$$

where $[r_l, r_u]$ denotes the full activation range, p_l and p_u are the central breakpoints, b is the total bit width, and uni$(\cdot)$ denotes uniform quantization within each segment. Each r_1, r_2, r_3 represents the activation value in the central and two tail regions, respectively.

Experiments demonstrate that PTQ4Protein achieves nearly lossless accuracy: for 8-bit quantization, TM-score (Zhang & Skolnick, 2004) drops by only 0.36% (CASP14) (Kryshtafovych et al., 2021) and 0.08% (CAMEO) (Haas et al., 2018), with memory usage reduced to 25% of baseline. Even at 6 bits, accuracy loss remains below 0.5%. Ablation studies show ESM-2 (Lin et al., 2023) is particularly robust, and the visual fidelity of predicted structures is preserved. Overall, PTQ4Protein provides a practical solution for deploying large pLMs on resource-limited hardware with minimal accuracy degradation.

PLMSearch (Liu et al., 2024b) is an efficient inference framework for large-scale remote homology search that leverages pLM embeddings and structure-informed similarity prediction. Its pipeline consists of:

1. *PfamClan-based pre-filtering:* PfamClan (Mistry et al., 2021) assigns proteins to evolutionarily related superfamilies (clans); this stage retains only protein pairs sharing the same clan, ensuring high recall while greatly reducing the candidate search space.
2. *SS-predictor:* A bilinear neural network is trained to predict structural similarity (TM-score (Zhang & Skolnick, 2005)) between protein pairs based on pLM embeddings; it integrates predicted TM-scores with cosine similarity, capturing global and local sequence features for remote homolog detection.
3. *PLMAlign:* For top-scoring pairs, per-residue alignment is performed using pLM embeddings to refine hit precision and enhance biological relevance.

A key efficiency innovation is the precomputation and caching of all target protein embeddings, enabling scalable, million-level searches via a single forward pass through the SS-predictor. By restricting downstream alignment to high-confidence pre-filtered pairs, PLMSearch dramatically reduces computation while maintaining high sensitivity. On standard benchmarks, PLMSearch achieves orders-of-magnitude faster search and higher sensitivity than conventional structure or sequence-based methods. These results establish PLMSearch as a scalable, structure-aware solution for accurate and ultra-fast remote protein homology inference.

Dense Homolog Retriever (DHR) (Hong et al., 2024) is an structure-aware inference framework for large-scale protein homolog detection, integrating pLMs with deep dense retrieval. Its workflow consists of:

1. *Dual-encoder architecture:* Both query and candidate sequences are embedded into dense vectors using a pair of ESM-initialized encoders, allowing alignment-free similarity calculation.
2. *Contrastive learning:* The dual-encoders are trained to embed homologous pairs close together and nonhomologous pairs far apart, enabling effective discrimination via dot-product scoring.
3. *Scalable dense retrieval:* All database embeddings are precomputed and cached, allowing rapid, large-scale retrieval of homologs through efficient vector similarity search (Johnson et al., 2019; Malkov & Yashunin, 2018) on standard hardware.

As an MSA prefilter, DHR accelerates MSA (with JackHMMER (Eddy, 2011)) by up to 93-fold, increases MSA diversity (log Meff (Morcos et al., 2011)), and improves AlphaFold2 (Jumper et al., 2021a) structure prediction—boosting TM-score (Zhang & Skolnick, 2005) by up to 0.03 and lowering root-mean-squared distance (Kabsch, 1976) by 0.15 Å on hard targets. DHR also maintains robust performance on massive datasets such as BFD (Steinegger & Söding, 2018) or (Mitchell et al., 2020) MGnify (over 500 million sequences) (Mirdita et al., 2022), demonstrating its practicality for structure-aware, scalable, and accurate protein inference at unprecedented speed.

Takeaway in §5 for pLMs Efficiency-focused Infernece

Recent methods—including advanced quantization, embedding-based retrieval, and structure-informed similarity prediction—enable pLMs to achieve efficient, scalable inference with lower computational cost and memory usage, while maintaining or improving predictive performance for large-scale biological applications.

6 Future Directions

Quantum Algorithms for Protein Structure Prediction A central challenge in classical, physics-driven approaches to protein folding (like molecular dynamics (MD)) is their immense computational complexity. These methods simulate the intricate energy landscape of a protein, a task that scales poorly on classical hardware. This has motivated the exploration of quantum algorithms, which are naturally suited for simulating complex systems. An overview of quantum–classical learning pipeline is provided in Figure 9.

The core physical principle of protein folding is that a protein chain will configure itself into its most stable, lowest-energy state. In physics, the total energy of a system is described by a mathematical operator known as a *Hamiltonian*. Therefore, a protein's most stable 3D structure can be *reframed* as finding the *lowest possible energy value* of this Hamiltonian. This lowest-energy solution is known as the *ground state*.

Quantum Protein Structure Prediction (QPSP) attempts to solve this problem by first simplifying it (Robert et al., 2021; Chandarana et al., 2023; Doga et al., 2024; Li et al., 2025). The vast, continuous space of all possible protein folds is too complex to model directly. Instead, the protein is "discretized" by placing its residues onto the points of a 3D grid, or *lattice*. Common choices include cubic lattice (Wong & Chang, 2021; Babej et al., 2018; Fingerhuth et al., 2018) and tetrahedral lattice, which build upon earlier 2D models (Perdomo et al., 2008; Perdomo-Ortiz et al., 2012; Babbush et al., 2014).

With the problem now on a lattice, the Hamiltonian is assembled from terms that encode the structural energy: terms that capture the interactions between nearby residue and terms that act as constraints to enforce a structurally valid chain. Finding the ground state of this lattice-based Hamiltonian becomes an optimization task for a quantum computer. Research in this emerging field is focused on improving the realism of these models—progressing from 2D to 3D lattices—and developing more efficient encoding (e.g., turn-based rather than coordinate-based) to reduce the number of qubits required.

A fundamental limitation remains: native protein structures are not confined to a discrete lattice. This mismatch limits the fidelity of such models. Consequently, the accuracy of any lattice-based prediction is ultimately constrained by the resolution of the lattice, thereby necessitating models with higher degrees of freedom to yield more realistic fits (Godzik et al., 1993).

Hybrid quantum–classical Transformers as a path toward quantum PLMs. A more immediate and pragmatic future direction lies in creating *hybrid quantum-classical models*. Instead of attempting to replace the entire classical pipeline, this approach seeks to identify computationally expensive sub-components of existing models (like PLMs) and "enhance" them with a quantum subroutine.

The self-attention mechanism is a prime target. At its core, the attention score is a similarity measure—typically a simple dot product—between a *Query (Q)* vector and a *Key (K)* vector. This score determines how much influence one token has on another. Recent work on hybrid transformers (Smaldone et al., 2025) proposes the replacement of this classical dot product.

In this hybrid model, the Q and K vectors are used to prepare quantum states. A quantum circuit then measures the *fidelity* (or *overlap*) between these two states, and this measurement becomes the new attention score. This is not just a faster way to compute a dot product; it is a *fundamentally different, quantum-native similarity metric*.

This pathway is promising for several reasons. As *a new similarity metric*, this quantum-derived similarity is a specialized nonlinear function that may be able to capture complex relationships between token embeddings that a simple dot product misses, potentially improving model generalization in the low-data regimes common to protein engineering. It also offers *theoretical efficiency*: while the primary benefit may be representative power, this quantum subroutine also offers a theoretical cost reduction for the *scoring* portion, from $O(n^2 d)$ to $O(n^2 \log d)$, where n is the sequence length and d the embedding dimension. Furthermore, it has *NISQ-era (Noisy Intermediate-Scale Quantum) feasibility*, as the model remains "mostly classical," retaining standard tokenization, embedding, and feed-forward blocks. Crucially, the quantum subroutine only requires a qubit count that scales as $\log d$, making its parameter-efficiency a viable candidate for NISQ hardware. Finally, it represents *a practical prototyping path* that can be prototyped *today*. Using frameworks like CUDA-

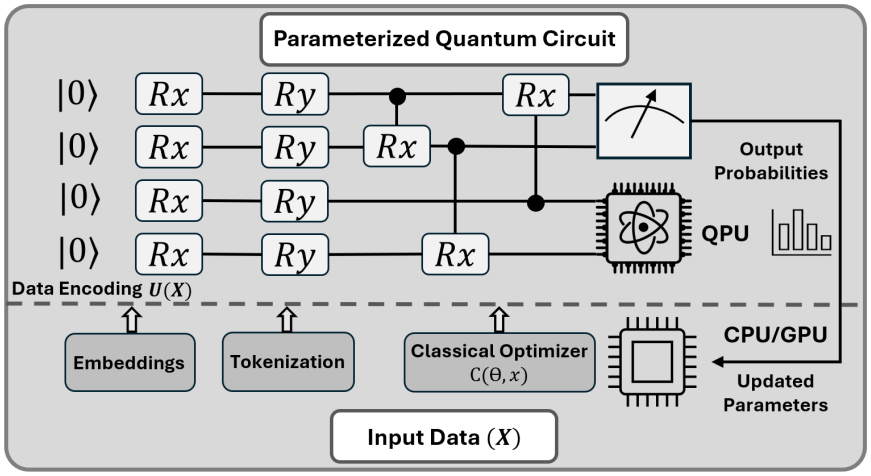

Figure 9: Conceptual illustration of a hybrid quantum–classical learning framework. Input data X are first processed through classical preprocessing steps such as tokenization and embedding, and subsequently encoded into a parameterized quantum circuit via a data-dependent unitary $U(X)$. The quantum circuit consists of trainable single-qubit rotation gates and entangling operations, and is executed on a quantum processing unit (QPU), where measurements produce output probability distributions. These measurement outcomes are used to evaluate a classical objective function $C(\theta, x)$, whose feedback is computed on classical hardware (CPU/GPU) to iteratively update the circuit parameters θ. This variational optimization loop illustrates a potential direction for integrating quantum computing with machine learning in future work.

Q (NVIDIA Corporation, 2025b;a), the quantum circuit can be simulated on classical CPU/GPU hardware and integrated directly into standard ML training loops, allowing rapid, end-to-end validation of the concept.

Positive results from such simulations would position hybrid attention as a credible and practical first step toward future, quantum-enhanced PLMs for protein modeling.

7 Conclusion

In this survey, we systematically reviewed advancements in efficient pLMs across four dimensions: dataset optimization, model architecture, training strategies, and inference. We highlighted a spectrum of techniques—from meta-learning and scaling law-guided design to PEFT, knowledge distillation, and quantization—that collectively address the formidable computational demands of modern pLMs. These approaches enable substantial cost reductions, democratize accessibility, and lay the groundwork for scalable, practical deployment in computational biology.

Despite this substantial progress, significant challenges remain. While efficient pLMs achieve competitive results on standard benchmarks, a performance gap often emerges in complex tasks such as de novo structure prediction or out-of-distribution generalization. The optimal strategies for downstream adaptation, such as fine-tuning protocols and representation pooling in extreme resource-limited settings, also remain underexplored. Future research must not only extend these efficiency paradigms to more diverse biological tasks but also systematically benchmark adaptation strategies and develop even more scalable solutions through advances in optimization, quantization, and architectural innovation.

We anticipate that continued innovation in these areas, coupled with broader community engagement, will be essential for unlocking the full potential of efficient pLMs and accelerating biological discovery.

Glossary

3Di Three-Dimensional Interaction (token). 12

AA Amino Acid. 1, 4, 9, 12

Cryo-EM Cryo-Electron Microscopy. 1

cubic lattice A simple three-dimensional lattice in which residues occupy vertices of a cube and can move to six nearest neighbors, used as a basic discretization model in coarse-grained protein folding. 20, *see also* lattice

energy landscape A multidimensional surface representing the potential energy of a molecular system as a function of its atomic coordinates; valleys correspond to stable conformations and barriers to transition states. 20

Enzyme Commission (EC) number A numerical classification code for enzymes, based on the chemical reactions they catalyze.. 16

functional genomics A field of molecular biology that aims to understand the function of genes and proteins, often at a genome-wide scale.. 5

GFP Green Fluorescent Protein. 14

GO Gene Ontology. 16

Hamiltonian The energy operator of a system, representing the total energy (kinetic plus potential) and governing its evolution in classical or quantum mechanics; used to encode energy terms in protein folding and quantum models. 20

homolog detection The computational task of identifying proteins that share a common evolutionary ancestor (homologs), which is often challenging when sequence similarity is low (remote homologs).. 17, 19

homology search A bioinformatics method used to identify sequences with a shared evolutionary ancestry, typically by comparing a query sequence against a database.. 18, 19

lattice A discrete grid (2D or 3D) used in coarse-grained protein folding models where residues occupy vertices and moves are restricted to adjacent lattice points. Common choices include square, cubic, and face-centered cubic (FCC) lattices, which approximate steric constraints and allow efficient enumeration of conformations. 20

metagenomic Relating to the study of genetic material recovered directly from environmental samples, which often contain a vast collection of uncultured and unknown organisms.. 4, 6

molecular dynamics (MD) A computational simulation method that models the physical movements of atoms and molecules over time by numerically integrating Newton's equations of motion, often used to explore protein folding and conformational changes. 20

MSA Multiple Sequence Alignment. 5, 15–19

NISQ An acronym for *Noisy Intermediate-Scale Quantum*, referring to the current generation of quantum devices that contain tens to a few hundred qubits. These systems are powerful enough to perform small-scale quantum computations but still suffer from noise, decoherence, and limited qubit connectivity.. 20

NMR Nuclear Magnetic Resonance Spectroscopy. 1

proteome The complete set of proteins expressed by an organism, cell, or tissue at a given time.. 12, 18

qubit The fundamental unit of quantum information, represented by a two-level quantum system that can exist in superposition of basis states $|0\rangle$ and $|1\rangle$; used to encode residue interactions or Hamiltonian terms in quantum protein models. 20

remote homology detection The task of identifying distant evolutionary relationships between proteins, typically when their sequence similarity is very low but their 3D structures are still related.. 12

residue An individual amino acid unit within a protein chain; in lattice and coarse-grained models, each residue is mapped to a vertex or bead. 20

secondary structure prediction The task of predicting the local secondary structures of a protein (e.g., alpha-helices, beta-sheets) from its primary amino acid sequence.. 12, 14

structural energy The potential energy associated with a given molecular or protein conformation, determined by intra- and inter-residue interactions, often minimized during structure prediction or folding simulations. 20

subcellular localization The specific location or compartment within a cell where a protein resides and performs its function (e.g., nucleus, mitochondria, cytoplasm).. 17

tetrahedral lattice A three-dimensional lattice where each vertex connects to four nearest neighbors forming tetrahedral geometry, offering improved angular representation of protein backbones compared to cubic lattices. 20, *see also* lattice

variant effect prediction The task of predicting the functional impact (e.g., on fitness, stability) of a mutation or a set of mutations on a protein.. 16

X-ray crystallography A foundational experimental technique used to determine the precise three-dimensional atomic structure of a protein by analyzing the diffraction pattern of X-rays passing through a protein crystal.. 1

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024. 2, 11, 15

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023. 16

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12): 1315–1322, 2019. 8, 10

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. 9

Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015. 1

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. 16

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021. 9

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 8

Ryan Babbush, Alejandro Perdomo-Ortiz, Bryan O'Gorman, William Macready, and Alán Áspuru-Guzik. Construction of energy functions for lattice heteropolymer models: Efficient encodings for constraint satisfaction programming and quantum annealing. In *Advances in Chemical Physics*, volume 155, pp. 201–244. John Wiley & Sons, Inc., 2014. doi: 10.1002/9781118755815.ch05. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118755815.ch05. 20

Tomáš Babej, Christopher Ing, and Mark Fingerhuth. Coarse-grained lattice protein folding on a quantum annealer, November 2018. URL https://arxiv.org/abs/1811.00713. arXiv:1811.00713. 20

Amos Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000. 16

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32, 2019. 8, 18

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019. 1, 10

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012. 13

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000. 14

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. *arXiv preprint arXiv:2109.12948*, 2021. 18

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022. 10, 14

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3, 6, 17

Andrew R Buller, Paul Van Roye, Jackson KB Cahn, Remkes A Scheele, Michael Herger, and Frances H Arnold. Directed evolution mimics allosteric activation by stepwise tuning of the conformational ensemble. *Journal of the American Chemical Society*, 140(23):7256–7266, 2018. 5

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021. 1

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007. 5

Pranav Chandarana, Narendra N. Hegade, Iraitz Montalban, Enrique Solano, and Xi Chen. Digitized counterdiabatic quantum algorithm for protein folding. *Physical Review Applied*, 20(1), July 2023. ISSN 2331-7019. doi: 10.1103/physrevapplied.20.014024. URL http://dx.doi.org/10.1103/PhysRevApplied.20.014024. 20

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 9

Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *Advances in Neural Information Processing Systems*, 37:69386–69418, 2024. 2, 4, 5, 6, 10, 11

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017. 15

Sebastian Clancy, Ilan Yaniv Zeisler, Pouriya Bayat, Matthew Xie, Vivian White, Spencer Perkins, Sepehr Bayat, and Keith Pardee. Assessing quantization and efficient fine-tuning for protein language models. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025. 2, 11, 14

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49 (D1):D480–D489, 2021. 1

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016. 8

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. 9, 12

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017. 10

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023. 14

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019. 1, 3, 6, 10, 17

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022. 10

Hakan Doga, Bryan Raubenolt, Fabio Cumbo, Jayadev Joshi, Frank P. DiFilippo, Jun Qin, Daniel Blankenberg, and Omar Shehab. A perspective on protein structure prediction using quantum computers. *Journal of Chemical Theory and Computation*, 20(9):3359–3378, May 2024. ISSN 1549-9626. doi: 10.1021/acs.jctc.4c00067. URL http://dx.doi.org/10.1021/acs.jctc.4c00067. 20

Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011. 19

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021. 1, 10, 12, 14

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023. 14

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021. 16

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022. 1, 8

Mark Fingerhuth, Tomáš Babej, and Christopher Ing. A quantum alternating operator ansatz with hard and soft constraints for lattice protein folding, October 2018. URL https://arxiv.org/abs/1810.13411. arXiv:1810.13411. 20

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017. 5

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014. 5

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 6

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022. 6

Nathan C Frey, Taylor Joren, Aya Abdelsalam Ismail, Allen Goodman, Richard Bonneau, Kyunghyun Cho, and Vladimir Gligorijević. Cramming protein language model training in 24 gpu hours. *bioRxiv*, pp. 2024–05, 2024a. 8

Nathan C Frey, Taylor Joren, Aya Abdelsalam Ismail, Allen Goodman, Richard Bonneau, Kyunghyun Cho, and Vladimir Gligorijević. Cramming protein language model training in 24 gpu hours. *bioRxiv*, pp. 2024–05, 2024b. 2, 7, 9, 11, 12

Yaron Geffen, Yanay Ofran, and Ron Unger. Distilprotbert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, 38(Supplement_2): ii95–ii98, 2022. 15

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pp. 1243–1252. PMLR, 2017. 6

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021. 10

Adam Godzik, Andrzej Kolinski, and Jeffrey Skolnick. Lattice representations of globular proteins: How good are they? *Journal of Computational Chemistry*, 14(10):1194–1202, October 1993. ISSN 1096-987X. doi: 10.1002/jcc.540141009. URL http://dx.doi.org/10.1002/jcc.540141009. 20

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021. 10

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 13

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 10

Juergen Haas, Steven Roth, Konstantin Arnold, Florian Kiefer, Tobias Schmidt, Lorenza Bordoli, and Torsten Schwede. The protein model portal—a comprehensive resource for protein structure and model information. *Database*, 2013:bat031, 2013. 1

Jürgen Haas, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018. 19

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 15

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024. 2, 10, 11, 12

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022. 12

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 6, 11

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 10, 15

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 10

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 4, 6

Liang Hong, Zhihang Hu, Siqi Sun, Xiangru Tang, Jiuming Wang, Qingxiong Tan, Liangzhen Zheng, Sheng Wang, Sheng Xu, Irwin King, et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nature Biotechnology*, pp. 1–13, 2024. 2, 3, 18, 19

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021. 9

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2): 128–135, 2017. 7

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019. 10

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 10

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022. 6, 12

Bozhen Hu, Cheng Tan, Yongjie Xu, Zhangyang Gao, Jun Xia, Lirong Wu, and Stan Z Li. Protgo: Function-guided protein modeling for unified representation learning. *Advances in Neural Information Processing Systems*, 37:88581–88604, 2024. 2, 10, 11, 15, 16

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022a. 13

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022b. 4, 6, 10, 13

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015. 8

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018. 17, 18

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020. 12

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 19

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021a. 19

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021b. 7, 9

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976. 8, 19

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12): 2577–2637, 1983. 15

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016. 6, 9

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020a. 3, 6, 11

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020b. 2

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 8

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019. 14

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021. 1, 14, 16, 19

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 4

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 16

Elodie Laine, Yasaman Karami, and Alessandra Carbone. Gemme: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution*, 36(11):2604–2619, 2019. 5, 16

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013. 16

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 10

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 18

Fengfu Li, Bin Liu, Xiaoxing Wang, Bo Zhang, and Junchi Yan. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016. 8

Rui-Hao Li, Hakan Doga, Bryan Raubenolt, Sarah Mostame, Nicholas DiSanto, Fabio Cumbo, Jayadev Joshi, Hanna Linn, Maeve Gaffney, Alexander Holden, Vinooth Kulkarni, Vipin Chaudhary, Kenneth M. Merz Jr, Abdullah Ash Saki, Tomas Radivoyevitch, Frank DiFilippo, Jun Qin, Omar Shehab, and Daniel Blankenberg. Quantum algorithm for protein structure prediction using the face-centered cubic lattice, 2025. URL https://arxiv.org/abs/2507.08955. 20

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021a. 11

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021b. 13

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 1, 2, 4, 6, 8, 9, 10, 13, 14, 16, 18, 19

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 13

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020. 13

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024a. 13

Wei Liu, Ziye Wang, Ronghui You, Chenghan Xie, Hong Wei, Yi Xiong, Jianyi Yang, and Shanfeng Zhu. Plmsearch: Protein language model powers accurate and fast sequence search for remote homology. *Nature communications*, 15(1):2775, 2024b. 2, 3, 18, 19

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 13

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 13

Amy X Lu, Wilson Yan, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, and Nathan C Frey. Tokenized and continuous embedding compressions of protein sequence and structure. *Patterns*, 6(6), 2025. 2, 6, 7, 8

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Lifeng Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 1(4), 2024. 8

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 19

Céline Marquet, Julius Schlensok, Marina Abakarova, Burkhard Rost, and Elodie Laine. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics*, 40(11): btae621, 2024. 2, 11, 15, 16

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021. 6

Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *Rna*, 19(11):1537–1551, 2013. 5

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 8

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017. 10

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022. 5, 12, 19

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020. 11

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021. 19

Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020. 4, 6, 19

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011. 19

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021. 17

Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of earth's microbiomes. *Nature biotechnology*, 39(4):499–509, 2021. 6

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 9

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pp. 483–499. Springer, 2016. 8

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022. 8, 16

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023. 6, 9

NVIDIA Corporation. CUDA-Q documentation (latest). https://nvidia.github.io/cuda-quantum/latest/, 2025a. Accessed: 2025-10-31. 21

NVIDIA Corporation. CUDA-Q: A platform for hybrid quantum–classical computing. https://github.com/NVIDIA/cuda-quantum, 2025b. Version 0.12.0. Accessed: 2025-10-31. 21

Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021. 1

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 10

Chun-Xiang Peng, Fang Liang, Yu-Hao Xia, Kai-Long Zhao, Ming-Hua Hou, and Gui-Jun Zhang. Recent advances and challenges in protein structure prediction. *Journal of Chemical Information and Modeling*, 64(1):76–95, 2023a. 1

Shuang Peng, Fei Yang, Ning Sun, Sheng Chen, Yanfeng Jiang, and Aimin Pan. Exploring post-training quantization of protein language models. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 602–608. IEEE, 2023b. 2, 3, 18

Alejandro Perdomo, Colin Truncik, Ivan Tubert-Brohman, Geordie Rose, and Alán Áspuru-Guzik. Construction of model hamiltonians for adiabatic quantum computation and its application to finding low-energy conformations of lattice protein models. *Physical Review A*, 78(1):012320, July 2008. doi: 10.1103/PhysRevA.78.012320. URL https://link.aps.org/doi/10.1103/PhysRevA.78.012320. 20

Alejandro Perdomo-Ortiz, Neil Dickson, Marshall Drew-Brook, Geordie Rose, and Alán Áspuru-Guzik. Finding low-energy conformations of lattice protein models by quantum annealing. *Scientific Reports*, 2:571, 2012. doi: 10.1038/srep00571. URL https://doi.org/10.1038/srep00571. 20

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023. 9

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *Advances in neural information processing systems*, 2018. 3, 10

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4, 10, 12

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021. 9

Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012. 9

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. 1, 6, 7, 9, 10, 12, 17

Anton Robert, Panagiotis Kl. Barkoutsos, Stefan Woerner, and Ivano Tavernelli. Resource-efficient quantum algorithm for protein folding. *npj Quantum Information*, 7(1), February 2021. ISSN 2056-6387. doi: 10.1038/s41534-021-00368-4. URL http://dx.doi.org/10.1038/s41534-021-00368-4. 20

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017. 14

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. *Advances in neural information processing systems*, 32, 2019. 6

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013. 5

Burkhard Rost. Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94, 1999. 12

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 10

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016. 14

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024. 2, 10, 11, 13

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 4

Damiano Sgarbossa and Anne-Florence Bitbol. Rag-esm: Improving pretrained protein language models via sequence retrieval. *bioRxiv*, pp. 2025–04, 2025. 2, 6, 7, 8

Jiayu Shang, Cheng Peng, Yongxin Ji, Jiaojiao Guan, Dehan Cai, Xubo Tang, and Yanni Sun. Accurate and efficient protein embedding using multi-teacher distillation learning. *Bioinformatics*, 40(9):btae567, 2024. 2, 10, 11, 15, 17

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 10

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021. 12

Anthony M. Smaldone, Yu Shee, Gregory W. Kyro, Marwa H. Farag, Zohim Chandani, Elica Kyoseva, and Victor S. Batista. A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation, 2025. URL https://arxiv.org/abs/2502.19214. 20

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 6

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012. 13

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014. 10

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017. 9

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):2542, 2018. 19

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019. URL https://arxiv.org/abs/1906.02243. 2

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13693–13696, 2020. 2

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pp. 2023–10, 2023. 6

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020. 10

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. 1, 4, 13

Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021. 9

Yang Tan, Mingchen Li, Bingxin Zhou, Bozitao Zhong, Lirong Zheng, Pan Tan, Ziyi Zhou, Huiqun Yu, Guisheng Fan, and Liang Hong. Simple, efficient, and scalable structure-aware adapter boosts protein language models. *Journal of Chemical Information and Modeling*, 64(16):6338–6349, 2024. 2, 11, 13, 15

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2022. URL https://arxiv.org/abs/2009.06732. 6

ESM Team. "esm cambrian: Revealing the mysteries of proteins with unsupervised learning". https://www.evolutionaryscale.ai/blog/esm-cambrian, December 2024. Accessed: YYYY-MM-DD. 14

Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994. 5

Oliver M Turnbull, Mohamed Baioumy, and Charlotte Deane. 2bits of protein: Efficient protein language models at the scale of 2-bits. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024. 2, 6, 7, 8

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 8

Marcel Van Gerven and Sander Bohte. Artificial neural networks as models of neural information processing, 2017. 8

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024. 12, 15

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024. 12

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 6, 7, 8

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019. 9

Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021. 15

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. 15

Renata Wong and Weng-Long Chang. Quantum speedup for protein structure prediction. *IEEE Transactions on NanoBioscience*, 20(3):323–330, July 2021. ISSN 1536-1241. doi: 10.1109/TNB.2021.3065051. URL https://doi.org/10.1109/TNB.2021.3065051. 20

Lirong Wu, Yufei Huang, Haitao Lin, and Stan Z Li. A survey on protein representation learning: Retrospect and prospect. *arXiv preprint arXiv:2301.00813*, 2022. 3

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008. 15

Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025. 1, 3

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020. 16

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024. 2, 6, 7, 9

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 9

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14284–14291, 2021. 17

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 15

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. 19

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005. 19

Yiming Zhang and Manabu Okumura. Prothyena: A fast and efficient foundation protein language model at single amino acid resolution. *bioRxiv*, pp. 2024–01, 2024. 2, 7, 9

Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15(1):5566, 2024. 2, 4, 5, 6, 11, 15

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016. 8

A. Židek. Alphafold v.2.3.0 technical note. https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md, 2022. Accessed: 2025-07-14. 16