Protein Large Language Models: A Comprehensive Survey

Yijia Xiao[♣*], Wanjia Zhao[♠], Junkai Zhang[♣], Yiqiao Jin[✿], Han Zhang[♣], Zhicheng Ren[♣], Renliang Sun[♣], Haixin Wang[♣], Guancheng Wan[♣], Pan Lu[♠], Xiao Luo[✎], Yu Zhang[🏛], James Zou[♠], Yizhou Sun[♣], Wei Wang[♣]

[♣]UCLA [♠]Stanford [✿]Georgia Tech [✎]UW–Madison [🏛]Texas A&M

https://github.com/Yijia-Xiao/Protein-LLM-Survey

Abstract

Protein-specific large language models (Protein LLMs) are revolutionizing protein science by enabling more efficient protein structure prediction, function annotation, and design. While existing surveys focus on specific aspects or applications, this work provides the first comprehensive overview of Protein LLMs, covering their architectures, training datasets, evaluation metrics, and diverse applications. Through a systematic analysis of over 100 articles, we propose a structured taxonomy of state-of-the-art Protein LLMs, analyze how they leverage large-scale protein sequence data for improved accuracy, and explore their potential in advancing protein engineering and biomedical research. Additionally, we discuss key challenges and future directions, positioning Protein LLMs as essential tools for scientific discovery in protein science. Resources are maintained at https://github.com/Yijia-Xiao/Protein-LLM-Survey.

1 Introduction

"*Proteins are the machinery of life, and understanding their language unlocks the secrets of biology.*"
— David Baker (Nobel Prize laureate 2024)

Proteins are essential biological molecules, driving functions such as catalyzing biochemical reactions, maintaining cell structure, and enabling cellular communication. Understanding their sequence-structure-function relationships is central to biological research. However, traditional experimental methods, including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, are time-consuming and labor-intensive, posing bottlenecks for large-scale applications.

Recent advancements in language modeling have revolutionized computational biology, offering powerful tools for protein analysis. Protein large language models (**Protein LLMs**) share several foundational similarities with LLMs: 1) *Training objectives and learning paradigms*, both LLMs and Protein LLMs are trained in a self-supervised manner on large-scale datasets using objectives such as masked language modeling (Devlin et al., 2019), auto-regressive modeling (Luo et al., 2022), or sentence permutation (Lewis et al., 2020; Yuan et al., 2022), learning to predict missing or next elements in sequences from the vocabulary. While LLMs predict missing words or phrases within textual data (Reimers and Gurevych, 2019; Liu et al., 2019; Touvron et al., 2023), Protein LLMs predict amino acids or subsequences within protein sequences. 2) *Pretraining data.* Protein LLMs adopt a data-driven paradigm to learn directly from large-scale protein datasets (Liu et al., 2024b; Jones et al., 2024). The datasets for training Protein LLMs consist of vast collections of protein sequences, analogous to the textual corpora used for LLMs. This eliminates the need for explicit feature engineering, allowing Protein LLMs to learn intricate patterns, such as structural motifs, evolutionary relationships, and functional insights, similar to how LLMs capture semantic and syntactic structures in language.

This paradigm shift has led to the emergence of highly effective models that can predict protein folding, annotate biological functions, and even design novel proteins with desired characteristics. However, adapting language modeling techniques to protein modeling introduces unique challenges. Unlike natural language, protein sequences encode structural and biochemical constraints that are nonlinear and spatially dependent, making direct adaptation of NLP techniques suboptimal. Long-range dependencies in proteins arise from 3D folding rather than sentence-level syntax, and the scarcity of annotated biological data further limits the effectiveness of supervised fine-tuning. Additionally, interpretability and robustness are especially critical

*Contact Email: yijia.xiao@cs.ucla.edu

in biological domains, requiring models to incorporate domain-specific priors and rigorous evaluation standards. Beyond their predictive capabilities, Protein LLMs also provide interactive interfaces that allow users to upload protein sequences or structural files (e.g., PDB format), pose questions, and interact with the model in a conversational manner (Liu et al., 2024c; Xiao et al., 2024b), proving deeper insights into protein structure, function, and design. Recent surveys have covered the intersection of large language models and biological sciences. Some works have reviewed protein language models in terms of model design, pretraining, and downstream applications (Zhang et al., 2024b), while others have discussed scientific LLMs more broadly, spanning biology, chemistry, and bioinformatics (Zhang et al., 2024a; Ruan et al., 2025).

We present the first dedicated survey of Protein LLMs, analyzing their unique architectures, training methodologies, and practical applications in protein research. While previous studies have explored the applications of various computational methods for protein research (Chen et al., 2024c; Wu et al., 2022a) or discussed the role of language models in general scientific domains such as biomedicine (Wang et al., 2023a) and chemistry (Liao et al., 2024), this survey focuses specifically on Protein LLMs–a rapidly evolving area at the intersection of computational biology and NLP.

The key contributions are as follows:

- **Architectural Overview.** A structured taxonomy of state-of-the-art Protein LLMs (Figure 3) detailing their unique architectures for protein understanding (§2) and generation (§3), highlighting how these models surpass traditional experimental methods in both efficiency and accuracy (Appendix §A).
- **Data Insights.** A comprehensive summary of datasets for pretraining, fine-tuning, and benchmarking Protein LLMs, providing critical insights into data curation strategies and their impact on model performance (§4).
- **Evaluation Protocols.** A thorough discussion of methodologies for assessing the performance and impact of Protein LLMs, including comprehensive benchmarking strategies (§5).
- **Applications.** A detailed exploration of practical applications in protein prediction, annotation, and design, remarkably highlighting recent innovative advancements and showcasing the transformative potential of Protein LLMs in advancing biomedical research.

2 LLM Methods for Protein Understanding and Prediction

2.1 Problem Definition

A protein, composed of amino acids (residues), can be represented as a sequence $[x_1, \ldots, x_L]$ in the residue token space $\mathcal{P}$, where L denotes its length. According to Anfinsen's dogma, a protein's primary sequence determines its structure and function. General problems in protein understanding and prediction are as follows:
I. Sequence-to-Property Prediction: $f_\theta : \mathcal{P} \to \mathcal{R}^+$ mapping sequences to numerical properties, such as stability or fluorescence intensity.
II. Sequence-to-Label Prediction: $f_\theta : \mathcal{P} \to \mathcal{L}$ mapping sequences to categorical labels, including secondary structure types, contact maps, or functional annotations.
III. Sequence-to-Structure Prediction $f_\theta : \mathcal{P} \to \mathcal{S}$ mapping sequences to the 3D folding structures (i.e. tertiary structures).
IV. Sequence-to-Text Understanding: $f_\theta : \mathcal{P} \to \mathcal{T}$, where $\mathcal{T}$ represents generated textual descriptions of protein sequences.

2.2 Protein Sequence Models

Individual Protein Sequences Models. Protein language models process amino acid sequences into meaningful representations for downstream tasks including structure, function, and intermolecular interaction prediction, where they often serve as general-purpose sequence encoders. Like NLP models, they are usually first pretrained on large sequence datasets with masked language modeling (MLM) objective; and then the protein sequences' embeddings are adapted for downstream tasks. Initially, researchers leveraged long short-term memory (LSTM) architectures to learn representation of proteins (Alley et al., 2019; Bepler and Berger, 2019; Zhou et al., 2020). Following the breakthrough of transformer architectures (Vaswani et al., 2017) in NLP, transformer-based protein language models emerged as the new paradigm. Large-scale transformer models, scaling up to billions of parameters and trained on millions of protein sequences, have demonstrated remarkable effectiveness for protein understanding and prediction tasks (Rao et al., 2019; Elnaggar et al., 2021; Xiao et al., 2021; Hu et al., 2022), and 3D structure folding (Chowdhury et al., 2022; Fang

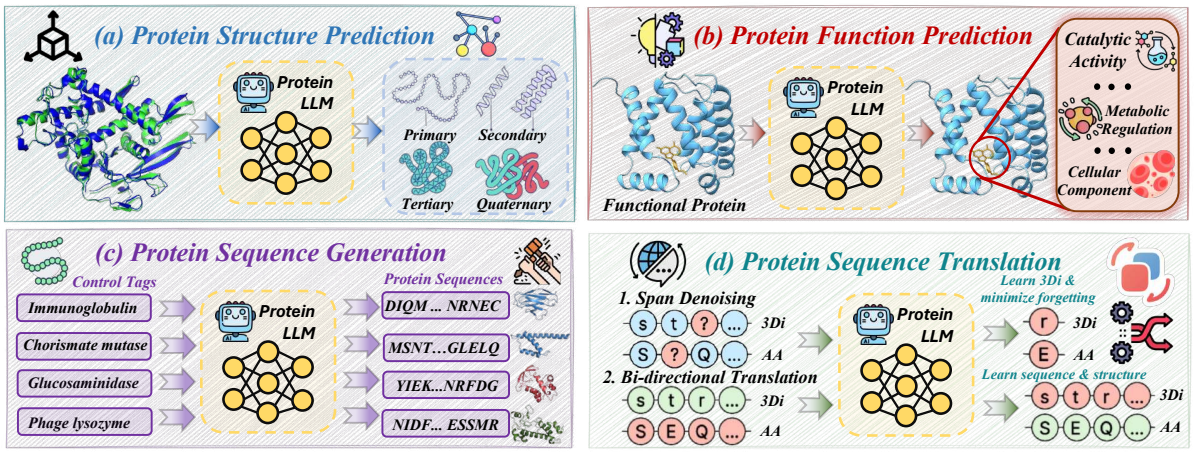

Figure 1: An Overview of Tasks in Protein Large Language Models.

et al., 2022; Chen et al., 2024a). The interpretability of these Protein LLMs has also been explored, with (Vig et al., 2021) analyzing learned representations through the lens of attention. Beyond general-purpose protein language models, several works have focused on domain-specific applications. For instance, Hie et al. (2021) applied BiLSTM to model viral escape patterns; TCR-BERT (Wu et al., 2024b) specialized in T-cell receptor (TCR) analysis for improved TCR-antigen binding prediction; PeptideBERT (Guntuboina et al., 2023) focused on predicting key properties of peptides; Kroll et al. (2023); Yu et al. (2023) adapted ESM-1b for enzymatic function prediction.

Multiple Sequence Alignments (MSA) Models. MSA aligns homologous proteins within sequence space by mapping their residues to the coordinate framework of a designated seed sequence. MSA reveals evolutionary relationships between proteins and thus serves as a cornerstone of computational biology, particularly for mutation effects prediction (Ram and Bepler, 2022; Hawkins-Hooker et al., 2021). The MSA Transformer (Rao et al., 2021) processed MSAs instead of single sequences. It used a modified axial attention mechanism (Ho et al., 2019; Child et al., 2019) to model both intra- and inter-sequence relationships. In contrast, Tranception (Notin et al., 2022), was trained on individual non-aligned sequences but could leverage aligned sequences during inference. It extracted patterns from contiguous protein subsequences and improves fitness prediction by integrating MSAs retrieved at inference time. In specific subdomains, Lin et al. (2023a) developed a transfer learning framework that utilized ESM-MSA-1b for transmembrane protein complexes. Additionally, vcMSA (McWhite et al., 2023) and Poet (Truong Jr and Bepler, 2023) leveraged protein LLMs to identify MSAs or homologous sequences.

Evolutionary Scale Modeling (ESM) Series. ESM is a protein modeling transformer family. ESM-1b (Rives et al., 2021) was trained on 250 million sequences using masked language modeling, containing 669.2 million parameters. ESM-1v (Meier et al., 2021) focused on zero-shot mutation effect prediction, while integrating MSA Transformer (Rao et al., 2021) for few-shot predictions. Following AlphaFold2's success (Jumper et al., 2021), ESM-IF (Hsu et al., 2022) combined Geometric Vector Perceptron (Jing et al., 2021) with GNN or transformer for inverse folding. ESM-2 (Lin et al., 2023b) scaled to 15 billion parameters and added a folding head creating ESMFold for single-sequence structure prediction. The latest ESM-3 (Hayes et al., 2025), a 98-billion-parameter multimodal model, reasons across protein sequences, structures, and functions, successfully designing a novel fluorescent protein using chain-of-thought approaches.

2.3 Structure-Integrated and Knowledge-Enhanced Models

Beyond residue sequences, many models integrate additional information, such as structure data or external knowledge, to enhance protein understanding and prediction ability.

Structure-Integrated Models: Structural information plays an important role in protein understanding, as a protein's functions are determined by its structures. Therefore, many works have incorporated structural information to enhance protein modeling ability. Some works utilized struc-

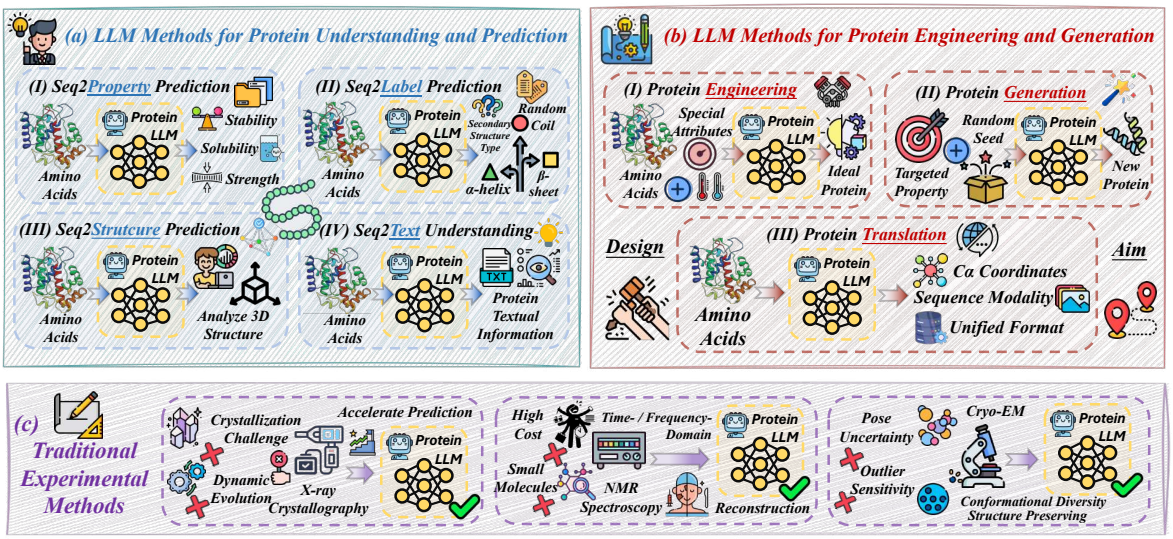

Figure 2: An Overview of Methods of Protein Large Language Models.

ture information as additional inputs (Chen et al., 2024b; Tan et al., 2024), a strategy often adopted in structure-based modeling of inter-molecular interactions. For instance, Zhang et al. (2023a) fused global structure information captured by structure encoder (GVP, GearNet (Zhang et al., 2023b), or CDConv (Fan et al., 2022)) into representations of ESM-2; SaProt (Su et al., 2024) incorporated local structural information for each amino acid, derived from Foldseek (Van Kempen et al., 2024), to generate structure-aware tokens. Alternatively, other works injected the structure information only in the training stage by either additional training tasks Wang et al. (2022); Sun and Shen (2024); Zhang et al. (2024c) or contrastive learning (Wang et al., 2025). Wu et al. (2022b) proposed OmegaFold, an end-to-end model that combines a protein language model with a geometry-aware transformer to directly predict 3D structures from single sequences, offering strong performance without MSAs. Some studies have also leveraged pretrained protein language models to improve structure models (Wu et al., 2023; Zheng and Li, 2024).

Knowledge-Enhanced Models: Beyond large protein sequence datasets, information in other formats can further enhance a model's understanding of proteins in the training stage. OntoProtein (Zhang et al., 2022) and KeAP (Zhou et al., 2023) incorporated knowledge graphs data during training by additional MLM objectives and/or contrastive learning to inject factual biological knowledge into the pre-trained Protein LLMs. ProteinBERT (Brandes et al., 2022) performed dual-task learning during pretraining to learn both protein sequence modeling and Gene Ontology (GO) annotation prediction. It utilized a specialized BERT architecture with parallel input pathways for sequences and annotations. To leverage the rich information in textual descriptions or other modalities, ProteinCLIP (Wu et al., 2024a) and MolBind (Xiao et al., 2024a) applied contrastive learning between protein sequences and textual descriptions and/or molecular representations. Such literature-based approaches mine semantic knowledge from existing biomedical texts to enhance interaction modeling.

2.4 Protein Description and Annotation Models

The previously mentioned models have primarily focused on learning protein representations and utilizing them for classification, regression, or 3D structure folding tasks. To enhance expressiveness and understanding, more recent models have been trained on both protein sequences and textual data, allowing them to integrate NLP capabilities with protein representation learning (Wang et al., 2023b; Liu et al., 2024c; Zhuo et al., 2024; Jin et al., 2024). Xu and Wang (2022) proposed ProTranslator, a bilingual translation framework between protein sequences and GO functions with textual descriptions. ProTranslator encoded and aligned the textual definitions of GO functions and protein sequences within the same low-dimensional space, facilitating the annotation of novel GO functions and the generation of textual descriptions for proteins. BioTranslator (Xu et al., 2023a) further improved ProTranslator by extending the bilingual framework to a mul-

tilingual translation framework, embedding text and multiple biomedical modalities into a shared space. ProtST (Xu et al., 2023b) was a framework designed to jointly learn from protein sequences and their associated biomedical text descriptions. It integrated protein language models (e.g., ESM or ProtBERT) with biomedical language models (e.g., PubMedBERT) to fuse sequence and text information through pre-training tasks. Prot2Text (Abdine et al., 2024) combined ESM-2 with a structure encoder (RGCN) and extended function prediction from categorical classification to free-text descriptions. BioT5 and BioT5+ (Pei et al., 2023, 2024) further unified molecular information within a more comprehensive training framework.

There have also been several interactive LLMs for protein understanding. These models enhanced pretrained LLMs with protein comprehension by integrating a protein processing module (Wu et al., 2024c; Wang et al., 2024a,b). For instance, ProteinChat (Guo et al., 2023) allowed users to input protein structures and query them using texts. ProteinGPT (Xiao et al., 2024b) extended this capability by supporting both protein sequences and structures as inputs. In these models, protein data were processed through Protein LLMs to generate embeddings, which were then projected to the natural language embedding space. The backbone LLMs integrated these adapted embeddings with user's queries to produce meaningful answers.

3 LLM Methods for Protein Engineering, Generation and Translation

Protein engineering and generation aims to design protein sequences with desired attributes (e.g. structures and properties). Given the desired attributes T and reference protein sequence $\mathcal{S}$ (optional), the model is expected to output a protein sequence $\mathcal{S}'$ with desired attributes. Key tasks include:

I. Protein Engineering: $f_\theta : (\mathcal{S}, T) \rightarrow \mathcal{S}'$ modifies protein $\mathcal{S}$ toward the desired attributes T, yielding the engineered protein $\mathcal{S}'$.

II. Protein Generation: $f_\theta : (T, R) \rightarrow \mathcal{P}$ generates proteins with attributes T by sampling from the protein space using random seeds R.

III. Protein Translation: $f_\theta : (\mathcal{P}, T) \rightarrow \mathcal{P}'$ translates a protein $\mathcal{P}$ into an alternative representation $\mathcal{P}'$ based on the target translation parameters T.

3.1 Protein Engineering Models

ProteinDT (Liu et al., 2023) is a multimodal protein design framework that robustly integrates textual protein knowledge with sequence-based generative modeling. ProteinDT employs contrastive alignment and a facilitator module, enabling zero-shot text-to-protein generation and editing. Meanwhile, PLMeAE (Zhang et al., 2025) is a closed-loop protein engineering framework that integrates protein language models with an automated biofoundry within a Design-Build-Test-Learn cycle. Furthermore, Toursynbio (Shen et al., 2024b) introduces an agent that is capable of facilitating the modification and engineering of wet lab proteins.

3.2 Protein Generation Models

Protein generation models create novel protein sequences for engineering applications using large-scale datasets of existing proteins. These typically employ decoder-based architectures generating sequences conditioned on biological annotations. ProGen (Madani et al., 2023) treats protein engineering as unsupervised sequence generation, creating functional proteins conditioned on annotations like molecular function or taxonomy, using data from UniProt and Pfam. ProtGPT2 (Ferruz et al., 2022) generates de novo sequences with natural amino acid compositions, exploring previously uncharted areas of protein sequence space.

ProGen2 (Nijkamp et al., 2023) extends ProGen with larger model size and training dataset, enabling protein fitness prediction without additional fine-tuning. ProLLaMA (Lv et al., 2024) handles both generation and understanding tasks using a two-stage approach: continued pre-training on protein sequences followed by instruction tuning with a 13-million-sample multitask dataset.

Alternative architectures include Ankh (Elnaggar et al., 2023), employing an encoder-decoder approach that reduces parameters while maintaining generation quality, and PAAG (Yuan et al., 2024), which aligns textual annotations with protein sequences at multiple levels. Pinal (Dai et al., 2024) constrains the design space by generating structure tokens before predicting sequences to improve foldability.

Specialized models target specific applications like antibody design. IgLM (Shuai et al., 2023) performs conditional generation based on antibody chain type and species origin, while PALM-H3 (He et al., 2024) specifically targets SARS-CoV-2 anti-

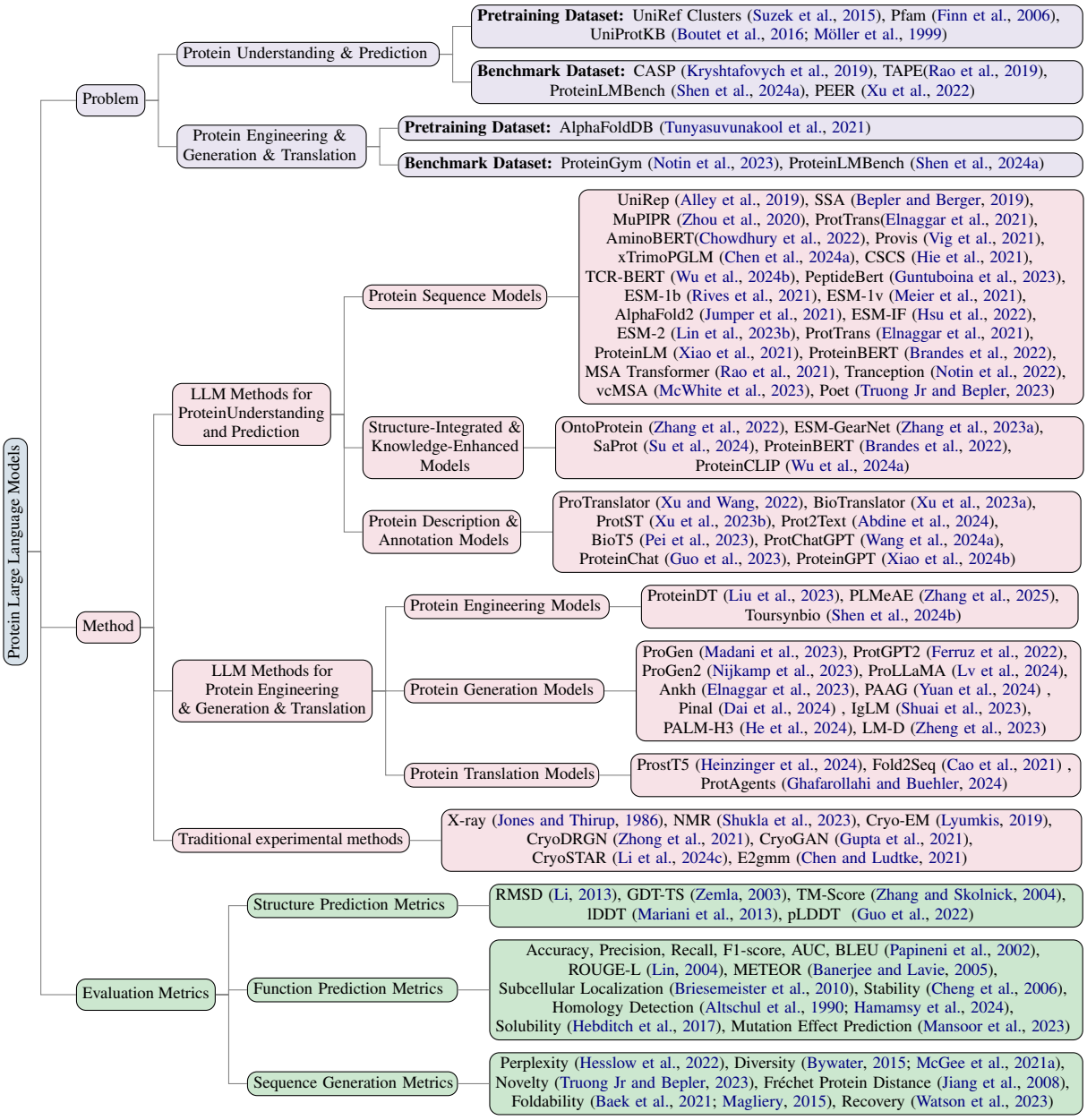

Figure 3: Taxonomy of Protein Large Language Models.

body generation, demonstrating how these models can be tailored for specific protein design tasks.

3.3 Protein Translation Models

Protein translation models are specifically developed to handle tasks that require translating between different protein representations, which could be helpful in protein design.

ProstT5 (Heinzinger et al., 2024) addresses the task of simultaneously modeling the dual nature of proteins — their linear one-dimensional (1D) sequences and three-dimensional (3D) structures — using a bilingual language model based on T5 (Raffel et al., 2020) and ProtT5 (Pokharel et al., 2022). It extracts features and patterns from both the sequence and the structure data Fold2Seq (Cao et al., 2021) is another model that learns structure-sequence relationships of proteins. The model could guide designs of protein sequences conditioned on desired structural folds. Recently, ProtAgents (Ghafarollahi and Buehler, 2024), a multiagent framework, has been proposed to handle 1D sequence generation and 3D fold generation simultaneously. LM-DESIGN (Zheng et al., 2023) is a method for reprogramming protein language models (pLMs) to design protein sequences for given structural folds.

4 Datasets

Datasets are crucial for training and evaluating Protein LLMs. They are categorized into pretraining datasets, comprising unlabeled protein sequences for self-supervised learning, and benchmark datasets, which contain labeled sequences for supervised fine-tuning and evaluation on specific biological tasks.

4.1 Pretraining Datasets

UniProtKB: A comprehensive protein sequence and annotation database composed of two main components: *Swiss-Prot* (Boutet et al., 2016), a manually curated, high-quality dataset with reliable annotations and *TrEMBL* (Möller et al., 1999), an automatically annotated dataset providing broader coverage.

UniRef Clusters (Suzek et al., 2015): A collection of clustered protein sequences designed to reduce data redundancy and improve computational efficiency. Provided by the UniProt database, UniRef is organized into three hierarchical levels: UniRef100, UniRef90, and UniRef50. UniRef100 contains a non-redundant set of all UniProt protein sequences where the latter two are created by clustering sequences with at least 90% and 50% sequence identity.

Pfam (Finn et al., 2006): A database of protein families and domains widely used for annotation and analysis of protein sequences. Each Pfam entry represents a group of related protein sequences defined by a multiple sequence alignment and a corresponding profile hidden Markov model (HMM). It provides insights into protein structure, function, and evolution, helping researchers identify conserved domains, predict functions, and classify proteins across organisms.

PDB (Bank, 1971): The Protein Data Bank is a repository for the 3D structural data of large biological molecules, such as proteins and nucleic acids. It provides valuable resources for understanding the structural aspects of proteins, which can be beneficial for training models that incorporate structural information.

AlphaFoldDB (Tunyasuvunakool et al., 2021): The AlphaFold Protein Structure Database offers predicted protein structures generated by the AlphaFold model containing over 200 million entries.

4.2 Benchmark Datasets

CASP (Kryshtafovych et al., 2019): Critical Assessment of Structure Prediction is a biennial competition that evaluates methods for protein structure prediction. Participants predict 3D structures of proteins from their sequences, compared against experimental results.

ProteinGym (Notin et al., 2023): A large-scale benchmark platform for protein design and fitness prediction. It includes over 250 Deep Mutational Scanning (DMS) assays, encompassing millions of mutated protein sequences, and curated clinical datasets with expert annotations. By integrating zero-shot and supervised evaluation frameworks, ProteinGym allows systematic comparison of over 70 machine learning models. It provides standardized metrics for tasks like mutation effect prediction and protein design, fostering innovation in computational biology and protein engineering.

TAPE (Rao et al., 2019): A benchmark designed to evaluate protein sequence embeddings in biologically relevant tasks using machine learning. It includes five tasks covering structure prediction, evolutionary understanding, and protein engineering. TAPE leverages self-supervised learning, enabling models to learn from unlabeled protein sequences, and offers standardized datasets and metrics for systematic comparisons. It aims to advance protein representation learning by addressing gaps in generalization and real-world applicability.

PEER (Xu et al., 2022): A comprehensive and multi-task benchmark designed to evaluate protein sequence understanding. It includes tasks such as protein function prediction, localization prediction, structure prediction, protein-protein interaction prediction, and protein-ligand interaction prediction.

ProteinLMBench (Shen et al., 2024a): A benchmark dataset comprising 944 manually verified multiple-choice questions aimed at assessing the protein understanding capabilities of LLMs. It incorporates protein-related details and sequences in multiple languages, setting a new standard for evaluating LLMs' abilities in protein comprehension.

5 Evaluation Metrics

Comprehensive and accurate evaluation is essential for understanding and applying Protein LLMs. Currently, these models are commonly assessed on tasks such as structure prediction, function prediction, and sequence generation.

5.1 Structure Prediction Metrics

Structure prediction evaluates how accurately a model predicts a protein's three-dimensional structure from its sequence (Kuhlman and Bradley, 2019). Common metrics include:

Root Mean Square Deviation (RMSD) measures the distance between the predicted and actual atomic coordinates. Lower RMSD indicates higher structural accuracy (Li, 2013).

Global Distance Test (GDT-TS) calculates the percentage of alpha-carbon atoms within thresholds (1, 2, 4, and 8 Å) of the reference structure after iterative superimposition (Zemla, 2003).

GDT-TS usually uses thresholds of 1, 2, 4, and 8 Å. The higher the GDT-TS score, the closer the predicted structure is to the reference structure.

Template Modeling (TM) Score evaluates the global structural similarity of proteins with values ranging from 0 to 1 (Zhang and Skolnick, 2004).

$$\mathrm{TM} = \max \left[\frac{1}{L_{\mathrm{tgt}}} \sum_{i}^{L_{\mathrm{com}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\mathrm{tgt}})} \right)^2} \right], \quad (1)$$

$$d_0(L_{\mathrm{tgt}}) = 1.24 \sqrt[3]{L_{\mathrm{tgt}} - 15} - 1.8. \quad (2)$$

Here, L_{tgt} is the length of the target protein amino acid sequence. L_{com} is the number of residues in the template and target structures. d_i represents the distance between the i-th residue pair in the template structure and the target structure. Higher scores indicate closer similarity.

lDDT, Local Distance Difference Test, evaluates the local accuracy of protein structure prediction by comparing distances between atom pairs in the predicted structures and those in the reference structures (Mariani et al., 2013).

A distance is considered preserved if it falls within a specified threshold. lDDT is calculated as the proportion of preserved distances, with higher values indicating better local accuracy.

pLDDT, Predicted Local Distance Difference Test, is a per-residue measure of local confidence (Guo et al., 2022). pLDDT evaluates the local quality of the predicted structure without a reference structure. Its computation usually relies on models such as AlphaFold (Jumper et al., 2021), which learns patterns from large-scale protein data. Scores range from 0 to 100, with higher scores indicating greater confidence and more accurate predictions.

5.2 Function Prediction Metrics

Protein function prediction aims to determine biological roles, including interactions with other biomolecules (Radivojac et al., 2013). The evaluation methods involve machine learning performance metrics and biomedical relevance validation.

Machine learning evaluation metrics include classification metrics (precision, recall, F-1 scores, accuracy, AUC) for protein classification and interaction prediction, and generative metrics (BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005)) for tasks like question answering.

In addition to machine learning metrics, there are also biometric-related evaluation metrics:

Subcellular Localization refers to the specific location of proteins within a cell (Briesemeister et al., 2010). The location of a protein is closely related to the function it performs, so by predicting the subcellular localization of a protein, it is possible to speculate on the biological function it may have (Holm, 2020).

Homology Detection aims to identify proteins that share an evolutionary relationship (homologous) with the target protein, usually reflected in similarities in sequences, structure, and functions. Traditional methods such as BLAST (Altschul et al., 1990) perform sequence alignment to identify homologs by comparing the query sequence against a database.

Recent deep learning approaches such as TM-vec (Hamamsy et al., 2024) focus on structural similarity and generate vector representations of proteins.

Stability of the protein is critical for many applications, such as drug development. Predicting the stability of a protein can help determine whether the protein can perform its function efficiently in the cellular environment (Cheng et al., 2006).

Solubility reflects the solubility characteristics of a protein in a particular solvent. Predictions of solubility can help to understand whether a protein can exist and function properly within a cell (Hebditch et al., 2017).

Mutation Effect Prediction of proteins refers to the assessment of the impact on various properties, structures, and functions of proteins when their amino acid sequences are changed (Mansoor et al., 2023). Commonly used methods include molecu-

lar dynamics-based methods, deep learning-based prediction models, and structural comparison methods.

5.3 Sequence Generation Metrics

Protein sequence generation is the process of creating new protein sequences using specific methods, models, or algorithms (Anand and Achim, 2022). Common evaluation methods include:

Perplexity (PPL) can be used to measure how accurately a model predicts amino acids (Hesslow et al., 2022). The lower the perplexity, the more accurate the prediction.

Novelty refers to the degree of uniqueness of the generated protein sequence compared to a database of known protein sequences (Truong Jr and Bepler, 2023).

Fréchet Protein Distance (FPD) is used to measure the similarity between the distribution represented by the generated protein sequence and the distribution of the real protein sequence (Jiang et al., 2008), denoted as:

$$\delta_{\mathcal{F}}(f, g) = \inf_{\alpha,\beta} \max_{s\in[0,1]} \text{dist}(f(\alpha(s)), g(\beta(s))) \quad (3)$$

where α and β are continuous non-decreasing functions. The sequence distribution can be denoted by f and g.

Diversity evaluates the degree of difference between protein sequences generated by a model. Common methods include Shannon Entropy (Bywater, 2015) and Hamming Distance (McGee et al., 2021b).

Foldability focuses on whether the generated protein sequence can be folded into a stable three-dimensional structure. Measuring foldability is usually performed with tools such as RoseTTAFold (Baek et al., 2021) or computational methods based on physicochemical principles (Magliery, 2015) to predict the likelihood that the generated sequence will form a stable structure.

Recovery is focused on the ability of a model to predict the corresponding sequence for a given structure accurately (Watson et al., 2023). Evaluating recovery includes methods sequence comparison, structure comparison, functionality comparison, etc.

6 Conclusion and Future Work

This survey provides an overview of Protein Large Language Models, highlighting architectures, datasets, evaluation, and applications. These works represent significant advancements in protein science and offer innovative approaches to protein analysis and design. Several challenges remain.

Protein Dynamics. AlphaFold (Jumper et al., 2021) has been shown to provide accurate static 3D structures. However, proteins are naturally dynamic molecules with various conformations (Ohnuki and Okazaki, 2024). Although several works incorporate 3D structures into LLMs, the conformational dynamics of proteins have not yet been considered. Since conformational dynamics are highly related to the transporter functions of proteins, it would benefit the model to include protein dynamics.

Combination with Single-cell Data. Recently, single-cell proteomics sequencing technology (Li et al., 2024b; Liu et al., 2024a; Bennett et al., 2023) has attracted extensive attention in biology, which can help us understand the pathways in specific cells. Since LLMs have shown effectiveness in understanding both proteins and single-cell data, they can be extended to learn from single-cell proteomics data in the future.

Interpretability. Model interpretability is crucial for building trustworthy Protein LLMs (Huang et al., 2024). Unlike traditional models (e.g., decision trees, SVMs), Protein LLMs are often opaque, making it difficult to identify which amino acids or structural features drive predictions—hindering adoption in biology and chemistry. Progress has been made through attention-based attribution (Gu et al., 2023), latent space analysis (Vecchietti et al., 2024), and recent methods like InterPLM (Simon and Zou, 2024), which use sparse autoencoders to uncover biologically meaningful representations.

Towards Biological Applications. Although several biological applications have been studied in recent works, a range of detailed and complex problems remain unsolved, including protein-ligand interaction learning (Koh et al., 2024), cryptic pocket identification (Ge et al., 2024), and rational ligand generation (Li et al., 2024a). These applications require extensive and diverse domain knowledge of proteins and their related fields. We believe LLMs have the potential to incorporate and utilize more domain knowledge to solve these problems.

In summary, Protein LLMs represent a transformative direction in computational biology, and addressing these challenges will be crucial for unlocking their potential in biological applications.

Limitations

This survey primarily focuses on Protein LLMs. We acknowledge that the study of protein interactions with other molecules (e.g., DNA, RNA) in the inter-molecular domain is a broad and valuable field worth reviewing. Given its vast scope, we do not extensively cover it in this survey, and instead focus on Protein LLMs centered on proteins themselves. In the future, we may either expand our review to include these areas or write a separate survey specifically dedicated to this domain, providing more comprehensive coverage for researchers.

Acknowledgments

This research work has been supported by NIH U54OD036472, U54HG012517, U24DK097771, NSF 2106859, 2312501, Optum AI, NEC, and Amazon.

References

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *AAAI*, pages 10757–10765.

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

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

Namrata Anand and Tudor Achim. 2022. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv:2205.15019*.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.

Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. 2005. The universal protein resource (uniprot). *Nucleic Acids Research*, 33(suppl_1):D154–D159.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Protein Data Bank. 1971. Crystallography: Protein data bank. *Nature New Biology*, 233(42):223–223.

Irène Barbarin-Bocahu and Marc Graille. 2021. Artificial intelligence to solve the x-ray crystallography phase problem: a case study report. *bioRxiv*, pages 2021–12.

Winona C Barker, John S Garavelli, Zhenglin Hou, Hongzhan Huang, Robert S Ledley, Peter B McGarvey, Hans-Werner Mewes, Bruce C Orcutt, Friedhelm Pfeiffer, Akira Tsugita, et al. 2001. Protein information resource: a community resource for expert annotation of protein data. *Nucleic Acids Research*, 29(1):29–32.

Hayley M Bennett, William Stephenson, Christopher M Rose, and Spyros Darmanis. 2023. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods*, 20(3):363–374.

Tristan Bepler and Bonnie Berger. 2019. Learning protein sequence embeddings using information from structure. In *ICLR*.

Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. 2016. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. *Plant Bioinformatics: Methods and Protocols*, pages 23–54.

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

Sebastian Briesemeister, Jï¿ 1/2rg Rahnenfï¿ 1/2hrer, and Oliver Kohlbacher. 2010. Yloc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, 38(suppl_2):W497–W502.

Robert Paul Bywater. 2015. Prediction of protein structural features from sequence data based on shannon entropy and kolmogorov complexity. *PloS One*, 10(4):e0119306.

Yue Cao, Payel Das, Vijil Chenthamarakshan, Pin-Yu Chen, Igor Melnyk, and Yang Shen. 2021. Fold2seq: A joint sequence (1d)-fold (3d) embedding-based generative model for protein design. In *ICML*, pages 1261–1271.

Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. 2024a. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*.

Dexiong Chen, Philip Hartout, Paolo Pellizzoni, Carlos Oliver, and Karsten Borgwardt. 2024b. Endowing protein language models with structural knowledge. *arXiv preprint arXiv:2401.14819*.

Muyuan Chen and Steven J Ludtke. 2021. Deep learning-based mixed-dimensional gaussian mixture model for characterizing variability in cryo-em. *Nature Methods*, 18(8):930–936.

Xinhui Chen, Yiwen Yuan, Joseph Liu, Chak Tou Leong, Xiaoye Zhu, and Jiaqi Chen. 2024c. Generative models in protein engineering: A comprehensive survey. In *NeurIPS 2024 FM4Science Workshop*.

Jianlin Cheng, Arlo Randall, and Pierre Baldi. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1125–1132.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv:1904.10509*.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M Church, et al. 2022. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623.

Fengyuan Dai, Yuliang Fan, Jin Su, Chentong Wang, Chenchen Han, Xibin Zhou, Jianming Liu, Hui Qian, Shunzhi Wang, Anping Zeng, et al. 2024. Toward de novo protein design from natural language. *bioRxiv*, pages 2024–08.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

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

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

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. 2022. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *ICLR*.

Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Xiaonan Zhang, Hua Wu, Hui Li, and Le Song. 2022. Helixfold-single: Msa-free protein structure prediction by using protein language model as an alternative. *arXiv preprint arXiv:2207.13921*.

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

Robert D Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl_1):D247–D251.

Yunhui Ge, Vineet Pande, Mark J Seierstad, and Kelly L Damm-Ganamet. 2024. Exploring the application of sitemap and site finder for focused cryptic pocket identification. *The Journal of Physical Chemistry B*, 128(26):6233–6245.

Alireza Ghafarollahi and Markus J Buehler. 2024. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409.

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. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168.

Zhonghui Gu, Xiao Luo, Jiaxiao Chen, Minghua Deng, and Luhua Lai. 2023. Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics*, 39(7):btad410.

Chakradhar Guntuboina, Adrita Das, Parisa Mollaei, Seongwon Kim, and Amir Barati Farimani. 2023. Peptidebert: A language model based on transformers for peptide property prediction. *The Journal of Physical Chemistry Letters*, 14(46):10427–10434.

Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. 2023. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. *Authorea Preprints*.

Hao-Bo Guo, Alexander Perminov, Selemon Bekele, Gary Kedziora, Sanaz Farajollahi, Vanessa Varaljay, Kevin Hinkle, Valeria Molinero, Konrad Meister, Chia Hung, et al. 2022. Alphafold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, 12(1):10696.

Harshit Gupta, Michael T McCann, Laurene Donati, and Michael Unser. 2021. Cryogan: A new reconstruction paradigm for single-particle cryo-em via deep adversarial learning. *IEEE Transactions on Computational Imaging*, 7:759–774.

Tymor Hamamsy, James T Morton, Robert Blackwell, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Charlie EM Strauss, Julia Koehler Leman,

Kyunghyun Cho, and Richard Bonneau. 2024. Protein remote homology detection and structural alignment using deep learning. *Nature Biotechnology*, 42(6):975–985.

D Flemming Hansen. 2019. Using deep neural networks to reconstruct non-uniformly sampled nmr spectra. *Journal of Biomolecular NMR*, 73(10):577–585.

Alex Hawkins-Hooker, David T Jones, and Brooks Paige. 2021. Msa-conditioned generative protein language models for fitness landscape modelling and design. In *Machine Learning for Structural Biology Workshop, NeurIPS*.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. 2025. Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Haohuai He, Bing He, Lei Guan, Yu Zhao, Feng Jiang, Guanxing Chen, Qingge Zhu, Calvin Yu-Chian Chen, Ting Li, and Jianhua Yao. 2024. De novo generation of sars-cov-2 antibody cdrh3 with a pre-trained generative large language model. *Nature Communications*, 15(1):6867.

Max Hebditch, M Alejandro Carballo-Amador, Spyros Charonis, Robin Curtis, and Jim Warwicker. 2017. Protein–sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19):3098–3100.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. 2024. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, page lqae150.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. 2022. Rita: a study on scaling up generative protein sequence models. *arXiv:2205.05789*.

Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. 2021. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2019. Axial attention in multidimensional transformers. *arXiv:1912.12180*.

Liisa Holm. 2020. Dali and the persistence of protein shape. *Protein Science*, 29(1):128–140.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *ICML*, pages 8946–8970.

Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. 2022. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. In *ICML*.

Minghui Jiang, Ying Xu, and Binhai Zhu. 2008. Protein structure–structure alignment with discrete fréchet distance. *Journal of Bioinformatics and Computational Biology*, 6(01):51–64.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. Prollm: protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *bioRxiv*, pages 2024–04.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. 2021. Learning from protein structure with geometric vector perceptrons. In *ICLR*.

Precious Jones, Weisi Liu, I-Chan Huang, and Xiaolei Huang. 2024. Examining imbalance effects on performance and demographic fairness of clinical language models. *arXiv:2412.17803*.

T Alwyn Jones and Soren Thirup. 1986. Using known substructures in protein model building and crystallography. *The EMBO journal*, 5(4):819–822.

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

Gogulan Karunanithy and D Flemming Hansen. 2021. Fid-net: A versatile deep neural network architecture for nmr spectral reconstruction and virtual decoupling. *Journal of Biomolecular NMR*, 75(4):179–191.

Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. 2024. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 6(6):673—687.

Alexander Kroll, Yvan Rousset, Xiao-Pan Hu, Nina A Liebrand, and Martin J Lercher. 2023. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature communications*, 14(1):4139.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2019. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020.

Brian Kuhlman and Philip Bradley. 2019. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Pengyong Li, Kaihao Zhang, Tianxiao Liu, Ruiqiang Lu, Yangyang Chen, Xiaojun Yao, Lin Gao, and Xiangxiang Zeng. 2024a. A deep learning approach for rational ligand generation with toxicity control via reactive building blocks. *Nature Computational Science*, 4(11):851—-864.

Shuai Cheng Li. 2013. The difficulty of protein structure alignment under the rmsd. *Algorithms for Molecular Biology*, 8:1–9.

Wei Li, Fan Yang, Fang Wang, Yu Rong, Linjing Liu, Bingzhe Wu, Han Zhang, and Jianhua Yao. 2024b. scprotein: a versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nature Methods*, 21(4):623–634.

Yilai Li, Yi Zhou, Jing Yuan, Fei Ye, and Quanquan Gu. 2024c. Cryostar: leveraging structural priors and constraints for cryo-em heterogeneous reconstruction. *Nature Methods*, pages 1–9.

Chang Liao, Yemin Yu, Yu Mei, and Ying Wei. 2024. From words to molecules: A survey of large language models in chemistry. *arXiv:2402.01439*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Peicong Lin, Yumeng Yan, Huanyu Tao, and Sheng-You Huang. 2023a. Deep transfer learning for inter-chain contact predictions of transmembrane protein complexes. *Nature Communications*, 14(1):4935.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023b. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. 2023. A text-guided protein design framework. *arXiv:2302.04611*.

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, W Jim Zheng, and Hongyu Zhao. 2024a. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of EMNLP*, pages 4819—4836.

Weisi Liu, Zhe He, and Xiaolei Huang. 2024b. Time matters: Examine temporal effects on biomedical language models. *arXiv:2407.17638*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024c. Prott3: Protein-to-text generation for text-based protein understanding. In *ACL*, pages 5949–5966.

Jie Luo, Qing Zeng, Ke Wu, and Yanqin Lin. 2020. Fast reconstruction of non-uniform sampling multidimensional nmr spectroscopy via a deep neural network. *Journal of Magnetic Resonance*, 317:106772.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv e-prints*, pages arXiv–2402.

Dmitry Lyumkis. 2019. Challenges and opportunities in cryo-em single-particle analysis. *Journal of Biological Chemistry*, 294(13):5181–5197.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106.

Thomas J Magliery. 2015. Protein stability: computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, 33:161–168.

Sanaa Mansoor, Minkyung Baek, David Juergens, Joseph L Watson, and David Baker. 2023. Zero-shot mutation effect prediction on protein stability and function using rosettafold. *Protein Science*, 32(11):e4780.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728.

Andrew G McDonald, Sinéad Boyce, and Keith F Tipton. 2009. Explorenz: the primary source of the iubmb enzyme list. *Nucleic Acids Research*, 37(suppl_1):D593–D597.

Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M Levy, Vincenzo Carnevale, and Allan Haldane. 2021a. The generative capacity of probabilistic protein sequence models. *Nature Communications*, 12(1):6302.

Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M. Levy, Vincenzo Carnevale, and Allan Haldane. 2021b. The generative capacity of probabilistic protein sequence models. *Nature Communications*, 12(1).

Claire D McWhite, Isabel Armour-Garb, and Mona Singh. 2023. Leveraging protein language models for accurate multiple sequence alignments. *Genome Research*, 33(7):1145–1153.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. In *NeurIPS*, pages 29287–29303.

S Möller, Ulf Leser, Wolfgang Fleischmann, and Rolf Apweiler. 1999. Edittotrembl: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, 15(3):219–227.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. 2023. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. 2022. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *ICML*, pages 16990–17017.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. 2023. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *NeurIPS*.

Jun Ohnuki and Kei-ichi Okazaki. 2024. Integration of alphafold with molecular dynamics for efficient conformational sampling of transporter protein nark. *The Journal of Physical Chemistry B*, 128(31):7530–7537.

Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. 1997. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *EMNLP*, pages 1102–1123.

Suresh Pokharel, Pawel Pratyush, Michael Heinzinger, Robert H Newman, and Dukka B Kc. 2022. Improving protein succinylation sites prediction using embeddings from protein language model. *Scientific Reports*, 12(1):16933.

Ali Punjani and David J Fleet. 2023. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870.

Xiaobo Qu, Yihui Huang, Hengfa Lu, Tianyu Qiu, Di Guo, Tatiana Agback, Vladislav Orekhov, and Zhong Chen. 2020. Accelerated nuclear magnetic resonance spectroscopy with deep learning. *Angewandte Chemie*, 132(26):10383–10386.

Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227.

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

Soumya Ram and Tristan Bepler. 2022. Few shot protein generation. *arXiv preprint arXiv:2204.01168*.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. 2019. Evaluating protein transfer learning with tape. In *NeurIPS*, pages 9686–9698.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. Msa transformer. In *ICML*, pages 8844–8856.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, pages 3980–3990.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15):e2016239118.

Wei Ruan, Yanjun Lyu, Jing Zhang, Jiazhang Cai, Peng Shu, Yang Ge, Yao Lu, Shang Gao, Yue Wang, Peilong Wang, Lin Zhao, Tao Wang, Yufang Liu, Luyang Fang, Ziyu Liu, Zhengliang Liu, Yiwei Li, Zihao Wu, Junhao Chen, Hanqi Jiang, Yi Pan, Zhenyuan Yang, Jingyuan Chen, Shizhe Liang, Wei Zhang, Terry Ma, Yuan Dou, Jianli Zhang, Xinyu Gong, Qi Gan, Yusong Zou, Zebang Chen, Yuanxin Qian, Shuo Yu, Jin Lu, Kenan Song, Xianqiao Wang, Andrea Sikora, Gang Li, Xiang Li, Quanzheng Li, Yingfeng Wang,

Lu Zhang, Yohannes Abate, Lifang He, Wenxuan Zhong, Rongjie Liu, Chao Huang, Wei Liu, Ye Shen, Ping Ma, Hongtu Zhu, Yajun Yan, Dajiang Zhu, and Tianming Liu. 2025. Large language models for bioinformatics. *Preprint*, arXiv:2501.06271.

Martin H Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E Wanker, and Miguel A Andrade-Navarro. 2012. Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS One*, 7(2):e31826.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024a. A fine-tuning dataset and benchmark for large language models for protein understanding. In *BIBM*, pages 2390–2395.

Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. 2024b. Toursynbio: A multimodal large model and agent framework to bridge text and protein sequences for protein engineering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2382–2389. IEEE.

Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. 2023. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.

Vaibhav Kumar Shukla, Gabriella T Heller, and D Flemming Hansen. 2023. Biomolecular nmr spectroscopy in the era of artificial intelligence. *Structure*, 31(11):1360–1374.

Elana Simon and James Zou. 2024. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pages 2024–11.

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

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2024. Saprot: Protein language modeling with structure-aware vocabulary. In *ICLR*.

Yuanfei Sun and Yang Shen. 2024. Structure-informed protein language models are robust predictors for variant effects. *Human Genetics*, pages 1–17.

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Timothy Truong Jr and Tristan Bepler. 2023. Poet: A generative model of protein families as sequences-of-sequences. pages 77379–77415.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Luiz Felipe Vecchietti, Minji Lee, Begench Hangeldiyev, Hyunkyu Jung, Hahnbeom Park, Tae-Kyun Kim, Meeyoung Cha, and Ho Min Kim. 2024. Recent advances in interpretable machine learning using structure-based protein representations. *arXiv:2409.17726*.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. Bertology meets biology: Interpreting attention in protein language models. In *ICLR*.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pretrained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. 2024a. Protchatgpt: Towards understanding proteins with large language models. *arXiv:2402.09649*.

Duolin Wang, Mahdi Pourmirzaei, Usman L Abbas, Shuai Zeng, Negin Manshour, Farzaneh Esmaili, Biplab Poudel, Yuexu Jiang, Qing Shao, Jin Chen, et al. 2025. S-plm: Structure-aware protein language model via contrastive learning between sequence and structure. *Advanced Science*, 12(5):2404212.

Yingheng Wang, Zichen Wang, Gil Sadeh, Luca Zancato, Alessandro Achille, George Karypis, and Huzefa Rangwala. 2024b. Long-context protein language model. *arXiv preprint arXiv:2411.08909*.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023b.

Instructprotein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*.

Zeyuan Wang, Qiang Zhang, HU Shuang-Wei, Haoran Yu, Xurui Jin, Zhichen Gong, and Huajun Chen. 2022. Multi-level protein structure pre-training via prompt learning. In *The Eleventh International Conference on Learning Representations*.

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. 2023. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100.

Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Z Li. 2023. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6(1):876.

Kevin E Wu, Howard Chang, and James Zou. 2024a. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pages 2024–05.

Kevin E Wu, Kathryn Yost, Bence Daniel, Julia Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Chang, and James Zou. 2024b. Tcr-bert: learning the grammar of t-cell receptors for flexible antigen-binding analyses. In *Machine Learning in Computational Biology*, pages 194–229.

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

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. 2022b. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07.

Wei Wu, Chao Wang, Liyi Chen, Mingze Yin, Yiheng Zhu, Kun Fu, Jieping Ye, Hui Xiong, and Zheng Wang. 2024c. Structure-enhanced protein instruction tuning: Towards general-purpose protein understanding. *arXiv preprint arXiv:2410.03553*.

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G Honavar. 2024a. Molbind: Multimodal alignment of language, molecules, and proteins. *arXiv preprint arXiv:2403.08167*.

Yijia Xiao, Jiezhong Qiu, Ziang Li, Chang-Yu Hsieh, and Jie Tang. 2021. Modeling protein using large-scale pretrain language model. *arXiv:2108.07435*.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. 2024b. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv:2408.11363*.

Hanwen Xu and Sheng Wang. 2022. Protranslator: zero-shot protein function prediction using textual description. In *RECOMB*, pages 279–294.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B Altman, and Sheng Wang. 2023a. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1):738.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023b. Protst: Multi-modality learning of protein sequences and biomedical texts. In *ICML*, pages 38749–38767.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. In *NeurIPS*, pages 35156–35173.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. 2023. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363.

Chaohao Yuan, Songyou Li, Geyan Ye, Yikun Zhang, Long-Kai Huang, Wenbing Huang, Wei Liu, Jianhua Yao, and Yu Rong. 2024. Annotation-guided protein design with multi-level domain alignment. *arXiv preprint arXiv:2404.16866*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In *ACL 2022 BioNLP Workshop*, pages 97–109.

Adam Zemla. 2003. Lga: a method for finding 3d similarities in protein structures. *Nucleic Acids research*, 31(13):3370–3374.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Qiang Zhang, Jiazhang Lian, and Huajun Chen. 2022. Ontoprotein: Protein pretraining with gene ontology embedding. In *ICLR*.

Qiang Zhang, Wanyi Chen, Ming Qin, Yuhao Wang, Zhongji Pu, Keyan Ding, Yuyue Liu, Qunfeng Zhang, Dongfang Li, Xinjia Li, et al. 2025. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications*, 16(1):1553.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Kehua Feng, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Tao Huang, Pengju Yan, Renjun Xu, Hongyang Chen, Xiaolin Li, Xiaohui Fan, Huabin Xing, and Huajun Chen. 2024a. Scientific large language models: A survey on biological & chemical domains. *Preprint*, arXiv:2401.14656.

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

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *EMNLP'24*, pages 8783–8817.

Zuobai Zhang, Jiarui Lu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. 2024c. Structure-informed protein language model. *arXiv preprint arXiv:2402.05856*.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. 2023a. A systematic study of joint representation learning on protein sequences and structures. *arXiv:2303.06275*.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. 2023b. Protein representation learning by geometric structure pretraining. In *ICLR*.

Jiangbin Zheng and Stan Z Li. 2024. Ccpl: Cross-modal contrastive protein learning. In *International Conference on Pattern Recognition*, pages 22–38. Springer.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. 2023. Structure-informed language models are protein designers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. 2021. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 18(2):176–185.

Guangyu Zhou, Muhao Chen, Chelsea JT Ju, Zheng Wang, Jyun-Yu Jiang, and Wei Wang. 2020. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genomics and Bioinformatics*, 2(2):lqaa015.

Hong-Yu Zhou, Yunxiang Fu, Zhicheng Zhang, Cheng Bian, and Yizhou Yu. 2023. Protein representation learning via knowledge enhanced primary structure modeling. *arXiv preprint arXiv:2301.13154*.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. 2024. Protllm: An interleaved protein-language llm with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920*.

A Experimental Methods in Proteomics and Their Limitations

Traditional experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) in protein science have laid the foundation for studying protein structure and functions. However, computational approaches and also embrace the progress of AI development. This section briefly covers methods, which are essential for determining protein structures and functions.

X-ray Crystallography is a widely utilized method for determining the 3D structures of proteins (Jones and Thirup, 1986). In this method, X-rays are directed at a crystallized sample, and the resulting diffraction patterns are analyzed to reveal the arrangement of atoms within the crystal. This process provides detailed insights into the protein's electron density and overall structure. However, crystallization can be challenging, especially for large, flexible, or membrane-associated proteins. The technique typically offers a static snapshot of the protein, which may not fully capture its dynamic nature in solution. Advancements in AI have led to the development of structure prediction tools like AlphaFold (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021). For instance, the crystal structure of the KlNmd4 protein is predicted to consist of a single PIN domain (Barbarin-Bocahu and Graille, 2021). The study demonstrates that the high-quality models significantly accelerate the determination of KlNmd4's structure, while existing models fail to achieve similar results.

Nuclear Magnetic Resonance (NMR) Spectroscopy is a non-destructive technique for determining the structure, dynamics, and interactions of molecules at the atomic level under near-physiological conditions (Shukla et al., 2023). It provides 3D structural data of proteins in solution and captures real-time dynamics, making it highly effective for studying protein flexibility and weak protein-ligand interactions. NMR exploits the magnetic properties of atomic nuclei (e.g., hydrogen nuclei in proteins) to provide detailed information about the local chemical environment.

With the development of AI, deep learning methods are more and more promising to advance the reconstruction of sparsely sampled data in NMR spectroscopy, particularly in the context of non-uniform sampling. The input data typically consists of sparsely sampled NMR spectra, while the output is

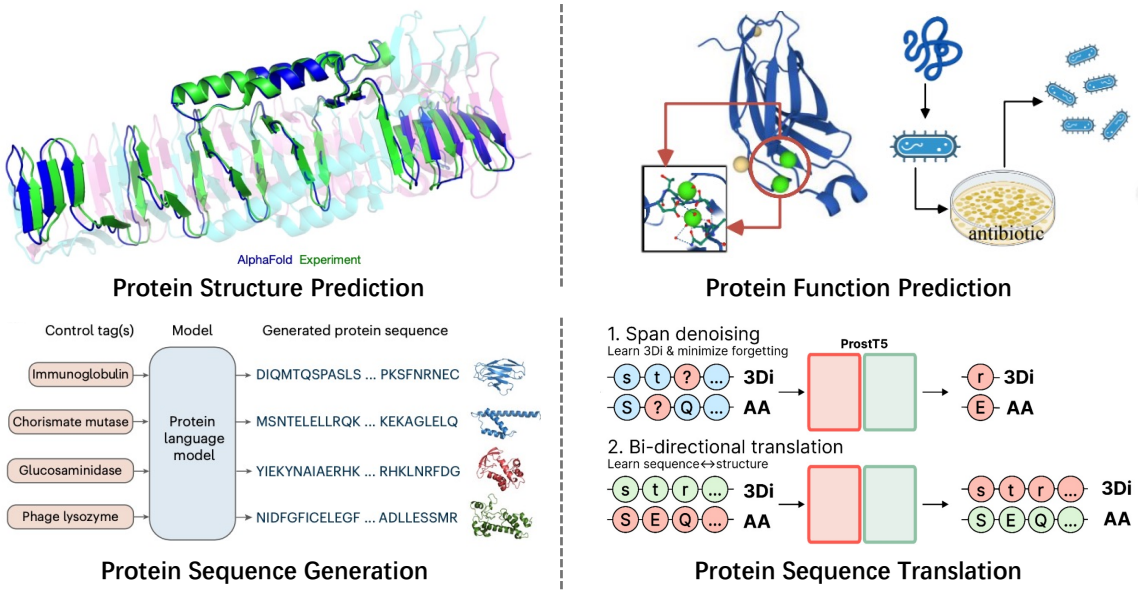

Figure 4: Illustrations on General Tasks of Protein Language Models.

the fully sampled spectrum, reconstructed either in the time (Hansen, 2019; Karunanithy and Hansen, 2021) or frequency domain (Qu et al., 2020; Luo et al., 2020). For time-domain reconstructions, neural networks effectively predict the missing data points. In frequency-domain reconstructions, they excel at removing artifacts caused by sparse and non-Nyquist sampling. Studies across various research groups have consistently demonstrated the high accuracy of DNN-based reconstructions, even under conditions of extremely sparse sampling, highlighting the potential of deep learning to enhance data acquisition and analysis in NMR.

However, NMR has limited size range: NMR is mostly suitable for proteins smaller than 30–50 kDa (larger proteins become challenging due to signal overlap). Protein sample preparation and data collection can also be expensive and take weeks to months.

Cryo-EM is a structural biology technique that enables the direct observation of conformational heterogeneity in individual dynamic macromolecules (Lyumkis, 2019). Researchers aim to reconstruct high-resolution 3D structural landscapes from numerous 2D observed projections, which may represent different conformational states. However, the cryo-EM reconstruction task is challenging because each particle's pose is unknown during imaging. Recently, deep learning methods have demonstrated powerful capabilities in representing heterogeneity within datasets by mapping them onto nonlinear manifold embeddings. On the one hand, CryoDRGN (Zhong et al., 2021) is a pioneering work that captures this heterogeneity by employing variational autoencoders (VAEs) to map the data into a low-dimensional latent space. A generative decoder then reconstructs a 3D volume from a sampled point in this latent space. CryoGAN (Gupta et al., 2021) introduces an entirely new possibility to learn to reconstruct in a distributional sense with a generative adversarial framework. Because of its likelihood-free nature, CryoGAN does not require any additional processing steps such as pose estimation and can be directly applied to cryo-EM measurements. This greatly simplifies the reconstruction procedure. On the other hand, E2gmm (Chen and Ludtke, 2021) models the 3D structure using a set of Gaussians to automatically resolve the structural heterogeneity, whereas 3DFlex (Punjani and Fleet, 2023) employs a neural network to fit the 3D displacement field of each particle by concurrently exploring its deformation field and refining a canonical density. More recently, CryoSTAR (Li et al., 2024c) resolves continuous conformational heterogeneity by constructing reasonable coarse-grained models, meanwhile, density maps are also estimated for different conformations. It meticulously preserves local structures, minimizes erroneous solutions, and ultimately achieves enhanced, accelerated convergence. Overall, the current trend is to incorporate atomic information to better activate deep models, aiming for more precise 3D structures that better comply with natural laws.

Table 1: LLM Methods for Protein Understanding and Prediction: Protein Sequence Models

Model	Time	Base Model	Dataset	Keywords
SSA (Bepler and Berger, 2019)	2019	BiLSTM	UniRef30	sequence embedding, sequence alignment, contextual embeddings, transmembrane prediction
UniRep (Alley et al., 2019)	2019	BiLSTM	UniRef50	Representation learning, Stability prediction, Functional effects of mutations
Bepler and Berger (2019)	2019	BiLSTM	SCOPe ASTRAL, Pfam, PDB, TOPCONS, CASP12	Structural property prediction, Soft symmetric alignment, Transmembrane
MuPIPR (Zhou et al., 2020)	2020	BiLSTM	STRING, PDB, SKP1402m, SKP1102s	Protein–Protein Interactions (PPI), binding affinity, buried surface area
CSCS (Hie et al., 2021)	2020	BiLSTM	IRD,LANL HIV database,ViPR,NCBI Virus,GISAID	Viral escape patterns, Constrained Semantic Change Search
ProtTrans (Elnaggar et al., 2021)	2021	Transformer-XL, XLNet, BERT, Albert, Electra, T5	UniRef, BFD	Protein secondary structure, sub-cellular localization, membrane vs. water-soluble
ESM-1b (Rives et al., 2021)	2021	Transformer	Uniparc	Large-scale pretraining, protein structure, functional effects of mutations
ESM-1v (Meier et al., 2021)	2021	ESM-1b	Uniref90	Functional effects of mutations, zero-shot prediction
ESM-2, ESMFold (Lin et al., 2023b)	2023	Transformer	UniRef, PDB, CAMEO, CASP14, MGnify, trRosetta Dataset	Atom-level resolution structure prediction
AminoBERT (Chowdhury et al., 2022)	2022	BERT	ProteinNet12, SCOPe ASTRAL	Single-sequence protein structure prediction
TCR-BERT (Wu et al., 2024b)	2021	BERT	VDJdb, PIRD, LCMV dataset	TCR–antigen binding
MSA Transformer (Rao et al., 2021)	2021	Transformer	UniRef50, UniClust30, CASP13, CAMEO	Multiple sequence alignment, evolutionary relationships
Tranception (Notin et al., 2022)	2022	Transformer	UniRef	Homologous sequences retrieval, fitness prediction
XTrimoPGLM (Chen et al., 2024a)	2024	Transformer	UniRef90, ColabFoldDB, UniProt, AlphaFold Database, PDB	100B parameters, Unified Protein Language Model
TurNuP (Kroll et al., 2023)	2022	ESM-1b	BRENDA, UniProt, Sabio-RK	Turnover number predictions, Differential Reaction Fingerprints
CLEAN (Yu et al., 2023)	2023	ESM-1b	UniProt,SwissProt	Contrastive Learning, Enzymatic function prediction
DeepTMP (Lin et al., 2023a)	2023	ESM-MSA-1b	PDB, PDBTM, UniRef30, BFD	Transfer learning, Transmembrane protein complexes,Inter-chain Contact Prediction
vcMSA (McWhite et al., 2023)	2023	ProtT5-XL-UniRef50	Quantest2, HOMSTRAD, UniRef50	MSA identification, Reciprocal Best Hits
Poet (Truong Jr and Bepler, 2023)	2023	Transformer	UniRef50, UniRef100, ProteinGym	Homologous Sequences, Retrieval-augmented LM

Table 2: LLM Methods for Protein Understanding and Prediction: Structure-Integrated and Knowledge-Enhanced Models

Model	Time	Base Model	Dataset	Keywords
ProteinBERT (Brandes et al., 2022)	2022	BERT	UniRef90, TAPE	GO annotations, protein structures, post-translational modifications, biophysical properties
OntoProtein (Zhang et al., 2022)	2022	ProtBert, Bert	ProteinKG25, UniRef100, TAPE, STRING, SHS27k, SHS148k	Knowledge graphs, gene ontology, PPI, structure prediction
ProteinCLIP (Wu et al., 2024a)	2024	ESM2, ProtT5, Text-Embedding-3-Large	UniProt	Contrastive learning, PPI, homology identification
SaProt (Su et al., 2024)	2023	ESM2	AlphaFoldDB, UniProt, ProteinGym, ClinVar, thermostability, metal ion binding, DeepLoc, TAPE, PEER, FLIP, PDB	Structure-aware vocabulary, Foldseek
ESM-GearNet (Zhang et al., 2023a)	2023	GVP, GearNet, CDConv	AlphaFold Database, GO (Gligorijević et al., 2021), Atom3D	Structural encoders for protein modeling
SES-Adapter(Tan et al., 2024)	2024	ESM2, ProtBert, ProtT5, Ankh	GO (Gligorijević et al., 2021)	Parameter-Efficient Fine-Tuning, Structure Representation
PromptProtein (Wang et al., 2022),	2023	Transformer	UniRef50, PDB, STRING, GO (Gligorijević et al., 2021)	Prompt Learning, Multi-level of structures
SI-pLMs (Sun and Shen, 2024)	2024	BERT	Pfam, PDB, AlphaFold Database	Variant Effect Prediction, Structural Information
Zhang et al. (2024c)	2024	ESM-2	SCOPe, GO and EC (Gligorijević et al., 2021), Swiss-Prot	Remote Homology Detection, Structural Information
S-plm (Wang et al., 2025)	2025	ESM3, discrete diffusion	BPTI, RMSD, Apo/holo, Fold-switch, ATLAS	Contrastive Learning, Structural Information
Wu et al. (2023)	2023	ESM-2, MSA-Transformer, GVP-GNN, EGNN, SE(3)-Transformer, Schnet, DimeNet	CASP, DB5.5, DIPS, PDBbind	Geometric Deep Learning
CCPL (Zheng and Li, 2024)	2023-2024	GVP-GNN, ESM-2	PDB, AlphaFoldDB, ProteinGym, trRosetta, CASP14, CATH, Ts 50&Ts500	Contrastive Learning, Structure-Sequence Pairing
KeAP (Zhou et al., 2023)	2023	ProteinKG25	ProteinNet,TAPE	Knowledge Graph, Contrastive Learning
MolBind (Xiao et al., 2024a)	2024	SciBERT, GIN, Uni-Mol	MolBind-M4, CASF-2016	Contrastive Learning, Protein-text-molecule Alignment

Table 3: LLM Methods for Protein Understanding and Prediction: Protein Description and Annotation Models

Model	Time	Base Model	Dataset	Keywords
ProtST (Xu et al., 2023b)	2023	ProtBert, PubMedBERT, etc.	ProtDescribe	Multimodal learning, protein function annotation, zero-shot text-to-protein retrieval
ProtChatGPT (Wang et al., 2024a)	2024	ESM-1b, Transformer	PDB-QA, ProteinKG25	Protein Q&A, cross-modal protein retrieval, qualitative dialogs
ProteinChat (Guo et al., 2023)	2023	ESM-IF1, Vicuna-13B	RCSB-PDB Protein Description	Interactive protein inquiries, automated protein understanding
Prot2Text (Abdine et al., 2024)	2024	RGCN, ESM2, GPT2	SwissProt	Multimodality, textual function prediction
ProTranslator (Xu and Wang, 2022)	2022	DeepGOCNN, Transformer	CAFA3, SwissProt, GOA, Reactome, KEGG, MSigDB	Function annotation based on text description, text description generation
BioTranslator (Xu et al., 2023a)	2023	PubMedBERT	GOA, Swiss-Prot, CAFA3, STRING, GeneCards, Tabula Muris, Tabula Sapiens, Tabula Microcebus, GDSC, STITCH, Monarch Initiative, Reactome	Multimodality, text-to-bio-identity translation
BioT5 (Pei et al., 2023)	2023	T5	ZINC20, UniRef, C4, PubMed articles, PubChem, ChEBI20, SwissProt, MoleculeNet, PEER, BindingDB, BioSNAP, HPRD, Yeast PPI dataset	SELFIES-based molecular representation, wrapped text for bio-entities
BioT5+ (Pei et al., 2024)	2024	T5	MoleculeNet, ChEBI-20, PEER, BioSNAP, BindingDB	Multi-task instruction tuning, Molecular
ProLLaMA (Lv et al., 2024)	2024	LLaMA2	UniRef, InterPro	Instruction understanding, protein understanding and generation
ProteinGPT (Xiao et al., 2024b)	2024	ESM-2, ESM-IF1, Vicuna, LLaMA-2, LLaMA-3	ProteinQA	Multimodal, interactive protein Q&A
ProLLM(Jin et al., 2024)	2024	Flan-T5-large	Human, STRING, Mol-Instructions	Chain-of-Thought, PPI

Table 4: LLM Methods for Protein Engineering, Generation and Translation

Model	Time	Base Model	Dataset	Keywords
ProGen (Madani et al., 2023)	2020	Transformer	UniParc, UniProtKB, Swiss-Prot, TrEMBL	Controllable protein generation, de novo protein design
ProGen2 (Nijkamp et al., 2023)	2022	Autoregressive	UniRef50	Protein generation, de novo protein design
ProtGPT2 (Ferruz et al., 2022)	2022	Autoregressive	UniRef50	Autoregressive transformer, BPE tokenization, zero-shot protein generation
ProLLaMA (Lv et al., 2024)	2024	LLaMA2	UniRef50, InterPro	Multi-task, instruction tuning
IgLM (Shuai et al., 2023)	2023	GPT-style Transformer	OAS Training Data, Thera-SAbDab	Infilling, conditioned generation, controllable diversity
PALM-H3 (He et al., 2024)	2024	ESM2, RoFormer	Observed Antibody Space, CoV-AbDab, BioMap	Strong generalization to novel proteins, interpretability, antibody
ProstT5 (Heinzinger et al., 2024)	2023	T5, ProtT5	3Di from AlphaFoldDB, CASP12/14, NetSurfP2.0	Bilingual LM, Foldseek, inverse folding
Fold2Seq (Cao et al., 2021)	2021	Transformer	CATH 4.2	Inverse protein design, fold-level representation
Ankh (Elnaggar et al., 2023)	2023	T5	UniRef50, CASP12/14, NetSurfP-2.0, DeepSF, etc	Contact prediction, secondary structure, fold classification, efficiency
ProteinDT (Liu et al., 2023)	2023	ProtBert, SciBERT, ProteinDiff, T5	SwissProtCLAP	Multimodal learning, text-to-protein generation, autoregressive
PLMeAE (Zhang et al., 2025)	2025	ESM-2	GB1, UBC9 dataset, Ubiquitin	Protein engineering, automatic biofoundry
ESM-IF (Hsu et al., 2022)	2022	GVP, GNN, Transformer	UniRef50, CATH	Inverse folding, AlphaFold2 augmented dataset
ESM-3 (Hayes et al., 2025)	2024	Transformer	UniProt, PDB, AlphaFoldDB, Pfam, InterPro, MGnify, JGI, GO Consortium	Multimodal Learning, Evolutionary Simulation
PAAG (Yuan et al., 2024)	2024	ProtBERT, SciBERT	ProtAnnotation	Text alignment, annotation
Pinal (Dai et al., 2024)	2024	T2struct, SaProt-T	SwissProt, UniRef50-ProTrek	Multi-step, functional labels
ProtAgents (Ghafarollahi and Buehler, 2024)	2024	GPT-4, Chroma, OmegaFold	GPTProteinPretrained	Multi-agent, de novo protein design, protein folding
Toursynbio (Shen et al., 2024b)	2024	InternLM2-7B	ProteinLMDataset	Multi-modal, agent, interactive
LM-DESIGN (Zheng et al., 2023)	2024	ESM-1b, ESM-2, ProteinMPNN	CATH 4.2, CATH 4.3, TS50, TS500	De novo protein design, protein folding

Table 5: Summary of Datasets for Protein Language Model

	Dataset	Last Update	Scale	Keywords
Pretraining	UniProtKB/Swiss-Prot (Boutet et al., 2016)	2025	573K	Manually curated, high-quality annotations, reviewed
	UniProtKB/TrEMBL (Möller et al., 1999)	2025	253M	Computationally annotated, unreviewed, automated predictions
	UniRef Clusters (Suzek et al., 2015)	2025	>250M	Clustered sequences, reduced redundancy, hierarchical organization
	Pfam (Finn et al., 2006)	2024	22k	Protein families, HMMs, functional domains
	PDB (Bank, 1971)	2025	231K	Protein structures, crystallography, molecular modeling
	BFD (Steinegger and Söding, 2018)	2021	2.5B	Massive protein database, sequence clustering, structure prediction
	UniParc (Bairoch et al., 2005)	2025	>250M	Non-redundant, protein sequence archive, database cross-referencing
	PIR (Barker et al., 2001)	2025	513M	Protein sequence database, functional annotation, evolutionary classification
	AlphaFoldDB (Tunyasuvunakool et al., 2021)	2025	>200M	Predicted protein structures, deep learning, proteome coverage
Benchmark	CASP (Kryshtafovych et al., 2019)	2024	N/A	Protein structure prediction, modeling competitions
	ProteinGym (Notin et al., 2023)	2024	2.7M	Protein mutations, deep mutational scanning
	TAPE (Rao et al., 2019)	2021	∼120K	Protein embeddings, sequence modeling
	CATH (Orengo et al., 1997)	2024	>150M	Structure classification, evolutionary relationships, domain hierarchy
	PEER (Xu et al., 2022)	2022	>60K	Protein understanding, multi-task benchmark, sequence evaluation
	ExplorEnz (McDonald et al., 2009)	2025	8K	Enzyme classification, EC numbering, catalytic reactions
	HIPPIE (Schaefer et al., 2012)	2022	39K	Human protein interactions, network analysis
	ProteinLMBench (Shen et al., 2024a)	2024	893K	Protein language understanding, multiple-choice QA, model evaluation

Table 6: **Comparison of Protein LLM Model Categories**

Model Category	Input	Output	Purpose	Typical Use Case
Protein Engineering	Existing protein sequence and desired attributes	Modified sequence	Refine or optimize an existing protein	Enhancing stability or introducing new functions
Protein Generation	Desired attributes or functions	Novel protein sequence	Create entirely new proteins with desired attributes	Designing de novo enzymes or antibodies
Protein Translation	One protein representation and translation target	Alternative protein representation	Convert between modalities (sequence - structure)	Predicting structure from sequence or vice versa

Table 7: **Comparison of Pretraining Datasets for Protein LLMs**

Dataset	Scale	Key Features	Advantages	Limitations
UniProtKB/Swiss-Prot	573K	Manually curated, high-quality annotations	Highly reliable, ideal for function prediction	Small in size
UniProtKB/TrEMBL	253M	Automatically annotated, unreviewed	Broad coverage, supports large-scale training	Lower annotation accuracy
UniRef Clusters (100/90/50)	>250M	Clustered to reduce redundancy	Efficient for training, scalable	May lose fine-grained details
Pfam	22K fams	Protein families, HMM-based domains	Captures conserved functional/structural domains	Limited to well-known domains
PDB	231K	Experimentally resolved 3D structures	High-quality structural data	Expensive, limited coverage
AlphaFoldDB	>200M	Predicted protein structures	High coverage, useful for structure-aware models	Prediction noise, not experimental

Table 8: **Comparison of Benchmark Datasets for Protein LLM Evaluation**

Dataset	Scale	Focus Tasks	Advantages	Limitations
CASP	N/A	Protein structure prediction	Gold standard with experimental validation	Biennial updates, slow cycle
ProteinGym	2.7M	Mutation effect prediction, protein design	Comprehensive DMS-based evaluation	Task-specific, focused on mutations
TAPE	$\sim$120K	Structure prediction, evolutionary tasks	Well-established, enables fair comparisons	Small dataset, limited scope
PEER	>60K	Multi-task understanding	Diverse tasks: function, interaction, structure	More complex benchmarking setup
ProteinLMBench	893K	Protein QA, LLM-based understanding	LLM-oriented, supports multilingual evaluation	Biased toward language understanding