Allergology International 74 (2025) 534–548

Contents lists available at ScienceDirect

Allergology International

journal homepage: http://www.elsevier.com/locate/alit

Invited Review Article

Predictive and therapeutic applications of protein language models

Kairi Furui, Koh Sakano, Masahito Ohue*

Department of Computer Science, School of Computing, Institute of Science Tokyo, Tokyo, Japan

ARTICLE INFO

Article history:
Received 29 July 2025
Available online 18 September 2025

Keywords:
Artificial intelligence
Bioinformatics
Computational biology
Drug discovery
Protein language model

Abbreviations:
BCR, B Cell Receptor; BERT, Bidirectional Encoder Representations from Transformers; BFDB, Big Fantastic Database; CNN, Convolutional Neural Network; DMS, Deep Mutational Scanning; ESM, Evolutionary Scale Modeling; GPT, Generative Pretrained Transformer; GO, Gene Ontology; GVP, Geometric Vector Perceptron; IDP, Intrinsically Disordered Protein; LM, Language Model; LSTM, Long Short-Term Memory; MLM, Masked Language Model; MSA, Multiple Sequence Alignment; NLP, Natural Language Processing; NSP, Next Sentence Prediction; NTP, Next Token Prediction; PDB, Protein Data Bank; pLM, Protein Language Model; pMHC, Peptide Major Histocompatibility Complex; PSP, Protein Structure Prediction; RNN, Recurrent Neural Network; SP, Signal Peptide; TCR, T Cell Receptor

ABSTRACT

Protein language models (pLMs) are rapidly emerging as revolutionary artificial intelligence technologies that bring transformative changes to drug discovery and therapeutic research. pLMs acquire rich representational capabilities from large-scale sequence datasets, enabling the solution of various biological problems that were difficult with conventional methods. In this review, we provide a comprehensive overview of various pLMs and their implementations, exploring their potential utility in drug discovery and therapeutic research. First, we systematically classify pLMs based on their architectures and information sources while discussing their development to the present. We also explain recent trends in multimodal approaches that integrate co-evolutionary information, structural information, and functional information, as well as domain-specific models specialized for particular domains such as antibodies and T-cell receptors. We then provide a comprehensive overview of various therapeutic applications of pLMs, including mutation effect prediction, function prediction, and structure prediction. Finally, we discuss future prospects of pLMs toward therapeutic applications and challenges for transforming them into technologies that contribute to actual diseases.

Introduction

Proteins are the fundamental units of life, with their functions determined by amino acid sequences and three-dimensional structures. Elucidating the relationship between protein sequence, structure, and function is extremely important in fields such as biology, protein engineering, therapeutic research, and drug discovery.[1] Traditionally, various bioinformatics methods have been developed for protein research, including homologous sequence searches, prediction of disease-causing mutation effects,[2] structure prediction,[3–5] property prediction, and optimization.

In recent years, language models (LMs) that learn linguistic patterns from large amounts of text have made remarkable progress in natural language processing (NLP).[5–7] Against this backdrop of technological advancement, protein language models (pLMs) have emerged as powerful tools in protein science.[1,3,8] Protein sequences contain information acquired over the course of evolution and can be regarded as a kind of language. pLMs learn structural motifs, evolutionary patterns, and functional features through self-supervised learning[9] such as predicting masked residues or the next residue in a sequence, without explicit labeled

* Corresponding author. Yokohama, Kanagawa, 226-8501, Japan.
E-mail address: ohue@comp.isct.ac.jp (M. Ohue).

https://doi.org/10.1016/j.alit.2025.08.004

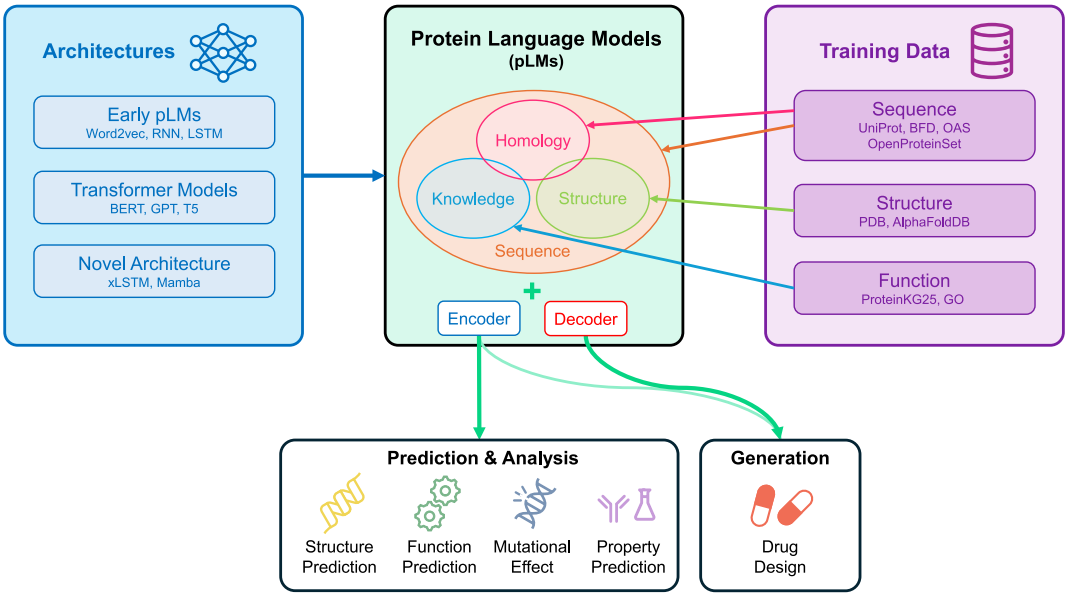

Fig. 1. An overview of protein language models.

training data. The latent space representations can achieve higher generalization performance in downstream tasks such as structure prediction and function prediction through transfer learning.[10–12] Thus, pLMs have been established as essential foundation models[13] for illuminating the relationship between sequence, structure, and function.

This article provides a comprehensive overview of the fundamental concepts and major architectures of pLMs, recent trends in drug discovery and therapeutic research, and a perspective on their potential and challenges (Fig. 1). While existing surveys comprehensively cover technical architectures and general applications,[3,14] this review specifically focuses on applications in therapeutic research. Furthermore, we emphasize the progress of domain-specific pLMs and the utilization of parameter-efficient fine-tuning (PEFT) methods in few-shot learning scenarios, which are particularly valuable topics for therapeutic research. First, we discuss the development of pLMs, focusing on the Transformer, which is the most important architecture of pLMs. Next, we describe recent trends in incorporating co-evolutionary information, structural information, and knowledge bases into pLMs to achieve richer representations. Then, we touch on domain-specific pLMs targeting specific protein domains such as antibodies and TCRs. We also outline the databases used for pretraining pLMs. Finally, we provide an overview of various applications of pLMs in drug discovery and therapeutic research.

Protein language models

In the development of pLMs, various architectures and approaches have been proposed. Table 1 summarizes the pLMs that have been proposed as general models based solely on amino acid sequences. In this section, we explain pLMs using major architectures such as LSTM and Transformer. For Transformer, we describe the purpose and representative pLMs for each of the three architectures: encoder, decoder, and encoder-decoder. We show representative models and their characteristics for each

Table 1
Sequence-based protein language model.

Name	Model	Dataset
SeqVec[15]	LSTM[16]	UniRef50
UniRep[17]	mLSTM[18]	UniRef50
UDSMProt[19]	AWD-LSTM[20]	Swiss-Prot[21,22]
TAPE[10]	Transformer	Pfam[23]
ESM-1b[24]	Transformer	UniRef50
ESM-1v[2]	Transformer	UniRef90
PRoBERTa[25]	RoBERTa	Swiss-Prot
ProtTrans[26]	Transformer-XL,[27] BERT, T5,[28] ALBERT,[29] XLNet,[30] ELECTRA[31]	UniRef50, BFD[32]
DistilProtBert[33]	BERT	UniRef50
RITA[34]	Autoregressive transformer	UniRef100
ProGPT2[35]	Autoregressive transformer	UniRef50
ESM-2[36]	BERT	UniRef50/UniRef90
ESM Cambrian[37]	BERT	UniRef, JGI,[38] MGnify[39]
ProtFlash[40]	Transformer + Mixed Chunk attention	UniRef50
ProGen[41]	Autoregressive transformer	UniParc,[42] UniProtKB,[43] Swiss-Prot, Pfam
ProGen2[44]	Autoregressive transformer	UniRef90, BFD30/BFD90, OAS[45]
CARP[46]	CNN	UniRef50
AMPLIFY[47]	BERT	UniRef100, OAS, SCOP[48]
Ankh[49]	T5	UniRef50

architecture. We also discuss pLMs using architectures proposed to address the computational cost of Transformers.

Early pLMs

Early approaches to treat protein sequences similarly to NLP included methods using word2vec[50] or doc2vec.[51,52] Word2vec is a method that converts words into fixed-length dense vectors based on co-occurrence relationships with surrounding words, using algorithms such as Skip-gram and CBOW (Continuous Bag-of-Words). ProtVec[52] was the first model to apply embedding methods to biological sequences.[53] ProtVec treats amino acid triplets as words and generates 100-dimensional protein vectors using word2vec. Furthermore, seq2vec[54] embedded entire protein sequences rather than amino acid k-mers using doc2vec,[51] an NLP method for embedding documents instead of words. However, embedding methods like word2vec had limited ability to capture context-dependent meanings and long−range interactions, and could not adequately represent the complex features of proteins.

Subsequently, LSTM[16] was widely adopted in the early stages of pLM.[15,17] LSTM is a type of RNN proposed to address the vanishing gradient problem. With memory cells equipped with input gates, forget gates, and output gates, it can selectively retain and update important information in long sequences. This architecture was used in models such as SeqVec[15] and UniRep.[17] SeqVec[15] demonstrated higher performance in protein family classification than the ProtVec. However, there were limitations in parallel processing efficiency and the ability to learn extremely long sequences or complex interaction patterns.

Transformer

Subsequently, the Transformer architecture,[6] which achieved success in NLP, was applied to protein modeling. Transformer[6] is the mainstream of current pLM research and has been adopted in popular models such as ESM-1b[24] and ESM-2.[36] The core of the Transformer is the self-attention mechanism, which allows direct modeling of relationships between arbitrary positions in a sequence. This characteristic is particularly effective for capturing interactions between distant amino acids, which are important for the formation of tertiary protein structures. Additionally, Transformers offer the advantages of parallel processing capability and high computational efficiency, making them suitable for learning from large datasets. In ESM-2[36] and ESM3,[55] Transformer models with tens of billions of parameters have been pretrained, achieving remarkable results in predicting protein structure and function.

Rao *et al.*[10] proposed early pLMs using Transformers and developed TAPE benchmark for evaluating protein sequence representation learning. TAPE includes five biology-related tasks: secondary structure prediction, contact prediction, remote homology detection, fluorescence intensity prediction, and stability prediction. TAPE provided a systematic comparison of pLMs.

Transformer encoder

Transformer models can be categorized into different types based on their architecture and pretraining approaches, broadly divided into encoder, decoder, and encoder-decoder. Transformer encoder models excel at representation learning of protein sequences and are suitable for transfer learning to downstream tasks.[56] In particular, BERT[9] is a representative encoder model. BERT is trained bidirectionally through self-supervised learning via MLM, which predicts randomly masked words from the corpus, and next sentence prediction (NSP) tasks. Alternatively, encoder models may be pretrained only with MLM. There are also pLMs based on RoBERTa,[57] an architecture that optimizes BERT's training step.[25] Many pLMs, including TAPE[10] and the ESM series,[24,36] have adopted BERT architecture.

The ESM (evolutionary scale modeling) series[24,36,55] is a group of large-scale pLMs based on BERT. ESM-1b[24] is a large Transformer model pretrained with MLM using the UniRef50 dataset. ESM-1v[2] is a model specialized for mutation prediction, demonstrating the ability to score missense mutations in a zero-shot manner. ESM-IF1,[58] though not a general pLM, incorporated 3D information through the Geometric Vector Perceptron (GVP) Transformer[59] to address the inverse folding problem. ESM-2,[36] released in 2022, was a significantly scaled-up pLM with 15 billion parameters, achieving state-of-the-art performance in structure prediction from a single sequence. Furthermore, ESM3,[55] the latest model as of 2025 developed by EvolutionaryScale, is a large-scale pLM trained with up to 98 billion parameters. ESM3 learns not only protein sequences but also structural information and knowledge from sources such as Gene Ontology (GO). ESM3 can directly process protein backbone coordinates through a geometric attention mechanism. Moreover, it features multi-track input and output, including sequence, structure, secondary structure, solvent-accessible surface area, functional keywords and residue annotations. By learning from these diverse data sources, ESM3 has high representational power, significantly surpassing previous methods in single-sequence structure prediction and sequence generation.

Additionally, Elnaggar *et al.*[26] proposed a series of models called ProtTrans, developed using multiple variants of Transformers including BERT and T5.[28] These have been utilized as foundational models for transfer learning in applications such as IDP-BERT[60] and PeptideBERT.[61] DistilProtBert[40] is a pLM that applied knowledge distillation to ProtBert, reducing the computational resources required for pretraining by 98 %.

Transformer decoder

Transformer decoder models are architectures specialized for protein sequence generation. Autoregressive models (also known as Causal Language Models, CLM) are decoder models trained through self-supervised learning using next token prediction (NTP). In autoregressive models, attention is masked to prevent attending to future positions, giving them the characteristic that the output of a token depends only on past tokens. The GPT series,[5,7,62] which became a breakthrough in the field of NLP, is a generative model based on this autoregressive transformer, and is adopted in pLM such as ProGPT2[35] and ProGen.[41]

ProGen adopts CTRL (Conditional Transformer Language Model),[63] which enables the generation of sequences with specific attributes or properties using control tags. This allows for the generation of protein sequences with specific functions based on keywords such as cellular component, biological process, and molecular function terms. ProGen2[44] improved protein sequence distribution capture, novel sequence generation, and protein fitness prediction by training a scaled-up pLM with 6.4 billion parameters on more than 1 billion sequences. Furthermore, the latest ProGen3[64] has scaled up to as many as 46 billion parameters. It achieves computational efficiency through sparse Mixture of Experts,[65] which selects the optimal network from multiple Expert networks according to input data.

Transformer encoder-decoder

In addition to encoder and decoder models, Transformer encoder-decoder models such as T5[28] are also used as architectures for pLMs.[26,66–69] T5 is a Transformer model with an encoder-decoder structure, proposed as a framework that uniformly treats various “text-to-text” tasks. According to Elnaggar *et al.*,[26] in a

comparison of multiple architectures in ProtTrans, ProtT5-XL-U50's latent variables showed the highest performance on downstream tasks.

Language models beyond transformer

Although Transformer has been a major breakthrough in both NLP and bioinformatics, it faces challenges that quadratically increasing computational cost and memory usage with sequence length. To overcome these issues, new architectures have been proposed.[40,46,70–73]

ProteinBERT[70] and CARP[46] use architectures with CNNs, while ProtFlash[40] employs a Transformer using a mechanism called Mixed Chunk Attention. Stärk *et al.*[71] proposed light attention for protein subcellular localization. These architectures have the advantage of linear cost increase with sequence length.

Furthermore, new architectures to replace Transformers such as xLSTM[72] and Mamba[73] have been proposed. xLSTM (Extended LSTM)[72] is an extension of traditional LSTM adapted to the era of large language models (LLMs). xLSTM achieves linear-time sequence processing through exponential gating and two new memory structures called sLSTM (scalar memory) and mLSTM (matrix memory), improving the ability to handle long contexts and track states, which were challenges in LSTM. Prot-xLSTM[74] is a new pLM that learned xLSTM on the OpenProteinSet,[75] demonstrating excellent performance in protein generation utilizing homology information.

Moreover, Mamba[73] is an extension of Structured State Space Models (SSM)[76] using a new structure called Selective State Space Model (Selective SSM), featuring a simple and GPU memory-efficient architecture. ProtMamba,[77] an alignment-free pLM using Mamba, has been proposed.

Prot-xLSTM and ProtMamba leverage their ability to handle much longer contexts than Transformer by concatenating multiple homologous sequences. These novel approaches have challenges such as insufficient optimization of architectures and learning, inadequate utilization of multimodal information, and unverified scaling laws for larger parameters. Additionally, pLMs based on diffusion models[78,79] such as DPLM[80] and DPLM-2[81] have been proposed in recent years. These models can generate plausible sequences and structures while maintaining the representational power of pLMs.

Understanding pLMs

Understanding the internal representations and learning processes of pLMs is important for improving model performance and developing new applications. For example, it has been revealed that the attention matrices of pLMs can reproduce contact maps,[82,83] and contact maps can be extracted by perturbing the input to pLMs.[84] This suggests that pLMs contain important information for structure prediction, which can be useful for identifying active sites in drug design and function prediction, and for providing evidence for pathogenicity prediction.

Efficient learning methods for pLMs are also noteworthy. Ankh[66] achieved the best performance on multiple tasks with less than 10 % of the parameters of ESM-2-15B by optimizing masking strategies, position encoding, architecture, and pretraining data. Furthermore, AMPLIFY[47] is a lightweight model by applying the latest improvements such as FlashAttention,[85] SwiGLU,[86] and RMSNorm,[87] challenging the trend of increasing model scale. As a result, AMPLIFY demonstrated performance equal to or better than ESM-2-15B, which is 43 times larger, while accelerating inference by 400–2000 times. Li *et al.*[88] systematically investigated the relationship between feature reuse and scaling in transfer learning of pLMs. The results showed that increasing pLM size or additional pretraining does not necessarily lead to performance improvements in downstream tasks. In particular, they found that many tasks, excluding the structure prediction task, rely on low-level features acquired early in pretraining. Additionally, an interesting study by Chen *et al.*[89] revealed that CLM and MLM have

Table 2
Protein language models with different information sources.

Name	Type	Model	Dataset
MSA transformer[91]	pLM + Alignment	Transformer	UniRef50, UniClust30
Tranception[92]	pLM + Alignment	Autoregressive transformer	UniRef100
MSAGPT[93]	pLM + Alignment	Autoregressive transformer	OpenProteinSet[75]
MSA-Generator[94]	pLM + Alignment	Transformer encoder-decoder	UniRef90
PoET[95]	pLM + Homology	Autoregressive transformer	UniRef50, UniRef100
ProGen3[64]	pLM + Alignment	Autoregressive transformer	PPA-1[64]
ProtMamba[77]	pLM + Homology	Mamba	OpenProteinSet
Prot-xLSTM[74]	pLM + Homology	xLSTM[72]	OpenProteinSet
LM-GVP[96]	pLM + Structure	ProtBERT	–
PromptProtein[97]	pLM + Structure	Transformer	UniRef50, PDB, STRING[98]
ESM-GearNet[99]	pLM + Structure	ESM-2	AlphaFoldDB
xTrimoPGLM[100]	pLM + Structure	GLM	UniRef90, ColabFoldDB[101]
SaProt[102]	pLM + Structure	ESM-2	AlphaFoldDB
ProstT5[67]	pLM + Structure	T5[28]	AlphaFoldDB, UniRef50
SI-pLM[103]	pLM + Structure	BERT	Pfam, PDB, AlphaFoldDB
S-PLM[104]	pLM + Structure	ESM-2 + Swin-transformer[105]	SwissProt, AlphaFoldDB
ProSST[106]	pLM + Structure	Transformer	AlphaFoldDB, CATH43-S40
ProteinBERT[70]	pLM + Function	BERT + CNN	UniRef90, GO
OntoProtein[107]	pLM + Function	ProtBert	ProteinKG25[107]
KeAP[69]	pLM + Function	Encoder-decoder	ProteinKG25
ProteinCLIP[108]	pLM + Function	Transformer	UniRef50, GO
ProLLaMA[49]	pLM + Function	LLaMA2[109]	UniRef, InterPro[110]
ESM3[55]	pLM + Structure +Function	Bidirectional transformer	Sequence: UniRef, JGI, MGnify Structure: AlphaFoldDB, PDB, ESMAtlas[36] Function: InterPro, GO

different optimal scaling strategies: MLM should prioritize expanding model size, whereas CLM should increase model size and training data equally. Furthermore, a recent study by Vieira *et al.*[90] shows that large-scale models do not necessarily outperform medium-scale models in transfer learning in practical situations where training data is limited. These findings provide important guidelines for optimizing model size and learning strategies to achieve high accuracy with limited computational resources in therapeutic applications.

Incorporating Co-evolutionary, structural, and functional information

Some pLMs incorporate information sources different from sequence, such as MSAs, 3D structural information, and functional knowledge from sources like Gene Ontology, to achieve more advanced representation learning.[3] Alignment information captures co-evolutionary patterns to identify functionally important residues, structural information enables learning structure-function relationships through three-dimensional features, and functional knowledge promotes biologically meaningful protein representations. This external knowledge can significantly improve prediction accuracy and reliability in therapeutic applications. Table 2 provides a list of pLMs that utilize these additional information sources. This additional information, either incorporated during model pretraining or utilized during inference, contributes to performance improvements in downstream tasks. This section provides an overview of the development of pLMs that utilize these diverse information sources.

Alignment/homology-based pLMs

In the evolutionary process of proteins, functionally important sites tend to be conserved, and interacting residues tend to co-evolve. Such evolutionary information can be extracted from MSAs, and pLMs leveraging this information have emerged.[3,91] MSA Transformer[91] is a pioneering method that takes MSA as input. MSA Transformer efficiently captures co-evolutionary patterns through alternating row and column attention mechanisms, achieving high accuracy in mutation effect and contact prediction. On the other hand, Tranception[92] is an autoregressive Transformer model for predicting protein sequence fitness. Unlike MSA Transformer, it uses only single sequences during training and utilizes MSA information during inference, making it effective for proteins with shallow alignments. Incidentally, Erckert and Rost[111] reported that when they explicitly incorporate MSA information into sequence-based pLMs to improve performance, recent pLMs such as ProtT5 did not benefit.

pLMs that handle homologous sequences as input and output without relying on alignments have also been proposed.[74,77,95] For example, PoET[95] is a hierarchical Transformer model that treats protein families as sequences-of-sequences. It is trained on homologous sequences extracted from UniRef50 and implements an attention mechanism that considers order within sequences but is order-independent between sequences. These methods enable transfer learning between families without MSA and demonstrate excellent mutation effect prediction ability even for proteins with shallow alignments.

Structure-Based pLMs

Some pLMs integrate structural information to enhance sequence representations. Since protein function is determined by both sequence and 3D structure, structural information is obviously important. One approach is to fuse structural encodings into sequence models.[67,96,99,102,112,113] LM-GVP[96] and ESM-GearNet[99] fused information from pretrained structural encoders like GVP Transformer[59] and GearNet[114] into pLMs. Also, ESM-GearNet-INR-MC[115] additionally fused surface encoding using their protein implicit neural representations (ProteinINR). Also, SaProt[102] and ProstT5[67] encode structural fragments from FoldSeek[116] as structural tokens. In the recent ProSST,[106] structural tokens from GVP transformer[59] are quantized into discrete tokens using a k-means model, creating effective protein representations that demonstrate excellent performance in zero-shot and supervised downstream tasks.

Another approach is to learn structural representations by introducing additional training objectives related to structure during pretraining or fine-tuning.[97,103,117] PromptProtein[97] performed multi-level pretraining using three tasks: MLM, α carbon atom coordinate prediction, and protein–protein interaction prediction. SI-pLM[103] also learned structural representations by extending MLMs to predict structural properties of proteins such as secondary structure, relative solvent accessibility (RSA), and contact maps. ISM[118] also takes an approach of distilling knowledge about structure by predicting structural tokens during training. The recent S-PLM[104] integrates protein structural information into pLMs by contrastive learning between sequence and structure representations.

Knowledge-based pLMs

Several pLMs attempt to reflect biochemical knowledge by incorporating external information such as knowledge graphs and GO.[55,69,70,107,108] OntoProtein[107] constructed a large-scale knowledge graph called ProteinKG25, which was trained using contrastive learning to correctly distinguish between positive and negative samples. KeAP[69] further introduced a token-level knowledge graph to OntoProtein's approach. ProteinBERT[70] is pretrained to predict GO terms simultaneously with an MLM. ProteinCLIP[108] utilized the CLIP[119] (Contrastive Language-Image pretraining) approach. By learning a shared space from both protein sequence and text obtained from GO terms, it incorporated functional information from natural language into pLMs. ESM3[55] also incorporates functional annotations into pretraining in addition to sequence and structural data.

Use of sequences other than amino acids

While pLMs handle amino acid sequences as input, LMs utilizing sequences other than amino acid sequences can also be useful for predicting protein function, mutation effects, and sequence generation. CaLM[120] is codon language model trained on 9 million non-redundant protein-coding DNA sequences from ENA.[121,122] By capturing codon usage patterns, it demonstrated excellent predictive performance in species recognition, prediction of protein and transcript abundance, and melting point estimation.[120] Furthermore, genomic language models such as Evo[123] and Evo2,[124] trained on large-scale genomic sequences, have achieved performance competitive with pLMs in protein sequence generation and zero-shot mutation prediction.

ESM All-Atom[49] is a multi-scale model at both residue and atomic levels using the AlphaFold DB and Uni-Mol[125] datasets. ESM All-Atom can be used for protein-molecule tasks, protein tasks, and molecular tasks, achieving performance that surpasses or competes with existing methods for each.

Table 3
Domain-specific protein language model.

Name	Class	Model	Dataset
AbLSTM[126]	Antibody	LSTM	OAS[45] NGS
Sapiens[127]	Antibody	RoBERTa[57]	OAS
AntiBERTy[128]	Antibody	BERT	OAS
IgLM[128]	Antibody	Autoregressive transformer	OAS
AbLang[129]	Antibody	RoBERTa	OAS
AntiBERTa[130]	Antibody	RoBERTa	OAS
AbNatiV[131]	Antibody	VQ-VAE,[132] BERT	OAS, VHH, Vκ, Vλ
IgT5/IgBERT[68]	Antibody	T5, BERT	OAS
REALM[133]	Antibody	ESM-2	OAS
AbLang2[134]	Antibody	ESM-2	OAS
nanoBERT[135]	Nanobody	BERT	INDI[136]
TCR-BERT[137]	TCR	BERT	VDJdb[138] PIRD[139] TCRdb[140]
ProtLM.TCR[141]	TCR	RoBERTa	TCR CDR3β[142,143]
SC-AIR-BERT[144]	TCR, BCR	BERT	
TABR-BERT[145]	TCR-epitope	BERT	TCRdb IEDB[146]
TULIP[147]	TCR-epitope	Encoder-decoder	VDJdb[138] IEDB[146] McPAS-TCR[148] netMHC[149]

Domain-specific pLMs

The pLMs introduced in the previous section are trained using general protein sequence databases such as UniProt and can capture evolutionary features and patterns common to proteins. On the other hand, LMs trained on datasets specialized for specific protein domains such as antibodies and TCRs can learn the sequence distributions and conservation patterns unique to those domains more precisely. Table 3 summarizes such domain-specialized pLMs, and Figure 2 shows an overview of domain-specific pLMs. By training pLMs on domain-specific protein sequences such as antibodies and TCRs, foundation models tailored for specialized tasks can be developed. This section introduces domain-specialized pLMs such as antibody-specific pLMs and TCR-specific pLMs, and discusses their relevance to therapeutic drug development and the potential for clinical applications.

Antibody language models

Antibody language models are models specifically developed to learn the characteristics of antibody sequences.[126,128–130,134] In antibody drug development, multiple stages are involved: discovery of candidates that bind to antigens, affinity maturation, humanization to reduce immunogenicity, and ensuring physicochemical properties suitable for manufacturing. Representations learned from antibody language models can be used to enhance prediction and generation performance for these purposes.

BERT-based models include AntiBERTa[130] and AbLang,[129,134] while autoregressive decoder models include IgLM[128] and CloneLM.[150] These models are primarily trained using the large-scale antibody sequence data recorded in the OAS database.[45] In particular, IgLM[128] is a GPT-2-based antibody LM aimed at generating antibody sequences and infilling CDR loop regions. Olsen *et al.* point out that existing pLMs such as ESM-2 are biased, because the majority of natural antibody sequence data comes from germline-derived nucleotide sequences.[134] AbLang2[134] overcame this bias by focusing on learning non-germline residues that are important for antibody function[151] through focal loss[152] and incorporating multiple masking strategies. Ablang2 also utilized paired heavy and light chain sequence information, improving the prediction accuracy of non-germline mutations.

Furthermore, nanoBERT,[135] a pLM for nanobodies, was trained using a dataset of 10 million non-redundant NGS sequences from the INDI[136] database. AbNatiV[131] is a deep learning model for assessing the nativity of antibodies and nanobodies, adopting a vector quantized variational autoencoder (VQ-VAE) architecture.[132] Additionally, CloneLM trained an autoregressive transformer to generate antibody clonal families obtained from OAS and further proposed a Bayesian optimization method called CloneBO.[150] CloneBO demonstrated that sequences can be efficiently optimized from few samples by learning how the immune system optimizes antibodies.[150]

These antibody language models are expected to be utilized for improving antibody affinity and developability, as well as for repertoire analysis. However, insufficient training data for key properties like antigen specificity, affinity, and structure remains a major challenge for realizing the full potential of pLMs.[153]

TCR-specific pLMs

TCRs (T-cell receptors) recognize antigens presented as peptide-MHC complexes on cell surfaces and initiate immune responses. The specificity of this recognition plays a central role in cancer immunotherapy, autoimmune diseases, and infectious disease treatment. Several TCR-specific pLMs have been developed for this purpose.[137,141]

TCR-BERT[137] is an early TCR-specific pLM. TCR-BERT consists of two stages: representation learning utilizing large amounts of unlabeled TCR sequences, followed by fine-tuning with a small amount of labeled data on antigen specificity. ProtLM.TCR[141] fine-

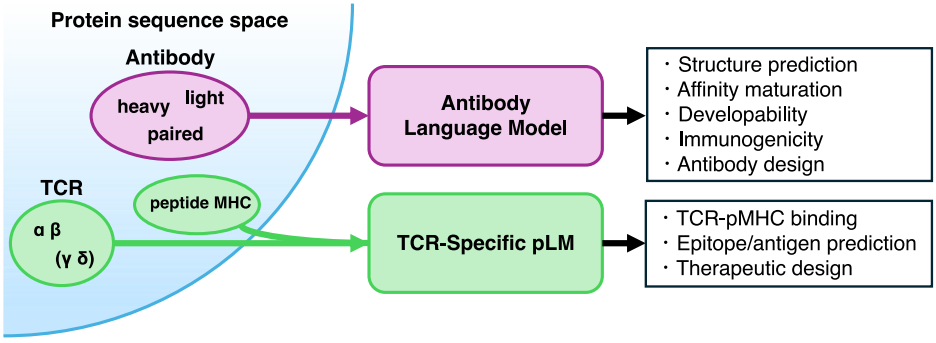

Fig. 2. Overview of domain-specific pLMs.

tuned a pLM pretrained on TCRβ sequences for predicting binding between TCRs and HLA class I epitopes. Furthermore, TCR-pMHC-specific pLMs that explicitly model the relationship between TCRs and epitopes during pretraining have also been proposed.[153,145,154–157] For example, TABR-BERT[145] is a BERT-based transfer learning model for predicting TCR-pMHC interactions. This model consists of three sub-models: a TCR embedding model (TCR-BERT), a pMHC embedding model (pMHC-BERT), and an MLP-based prediction model, demonstrating excellent performance particularly in zero-shot prediction for unknown epitopes. Additionally, TULIP[147] constructed a TCR-specific pLM based on a Transformer encoder-decoder to address incomplete information related to epitope, TCR α chain, and TCR β chain sequences in training data. SC-AIR-BERT[144] is a BERT-based model for predicting antigen binding specificity of adaptive immune receptors (AIRs), simultaneously learning paired immune receptor chains (TCR α and β chains, BCR light and heavy chains).

These TCR-specific pLMs play important roles in a wide range of therapeutic applications including cancer immunotherapy, vaccine design, and autoimmune disease treatment. TCR-specific pLMs can help identify high-affinity TCRs for specific cancer antigens in TCR-T therapy and predict immunogenic peptides for cancer vaccine design.[145] They are also useful for understanding the activation mechanisms of autoimmune diseases and developing new treatments targeting disease-causing cells.[137] However, challenges remain such as the lack of negative training data and the absence of pairing and structural information, though further performance improvements are expected by addressing these issues.

Training datasets

The selection of datasets is extremely important in pretraining pLMs. Large-scale sequence data forms the foundation in the learning process of pLMs, and is a crucial factor determining the quality and quantity of information acquired by the model. In this section, we provide an overview of the datasets used for pretraining pLMs. We also describe datasets related to previously mentioned pLMs that utilize co-evolutionary information, structural information, and functional knowledge.

Sequence databases

First, as a sequence database, there is UniProtKB, which contains comprehensive protein sequences and annotation information. UniRef is a dataset clustered from all protein sequences in the UniProt database, clustered at three similarity levels: 100 %, 90 %, and 50 %. Using clustered sequences allows for computationally efficient learning of the protein sequence space, so many pLMs use UniRef [158]. According to Elnaggar *et al.*,[66] using the non-redundant UniRef50 dataset was most efficient. Similarly, the UniClust database[159] clusters UniProtKB sequences at 90 %, 50 %, and 30 % pairwise sequence identity levels using MMseqs2.[160] It features higher consistency in functional annotations compared to UniRef database.

Also, some pLMs[10,41] use Pfam,[23] a database of 31 million protein domains widely used in bioinformatics, as their pretraining corpus. Pfam sequences are classified into evolutionary related groups, or protein families.

Besides regular protein sequence databases like UniRef, protein sequence databases collected for specific purposes such as BFD,[4,32] MGnify,[39] JGI,[38] and OAS[45] are also utilized. BFD (Big Fantastic Database)[4,32] is one of the largest public protein family collections, merging over 2.2 billion protein sequences from the UniProt database and multiple metagenomic sequencing projects. The sequences in BFD are clustered using Linclust/MMseqs2.[160] For example, ProtTrans[26] achieved higher performance by training ProtT5 on the BFD dataset and then fine-tuning on UniRef50, compared to using UniRef50 alone.

Additionally, ESM3[55] uses sequence data from sources other than UniRef [158], including the MGnify protein database,[39] JGI[38] database, and OAS database[45] for training. The MGnify protein database[39] is a microbiome-related database containing over 2.4 billion non-redundant protein sequences. Residue-level functional annotations from MGnify were also utilized in model training for ESM3. The JGI database[38] is a microbial genome database. As of August 2022, it contains over 451 million genome-derived genes and over 75.1 billion metagenome-derived genes from archaea, bacteria, eukarya, plasmids, and viruses, with standardized annotations. The OAS database[45] is a project to collect and annotate immune repertoires for large-scale analysis, containing over one billion sequences from more than 80 different studies. These repertoires cover diverse immune states, organisms (mainly human and mouse), and individuals, including both unpaired and paired antibody sequences.

Finally, OpenProteinSet[75] is a large-scale open dataset for protein structure prediction and design research. This dataset consists of over 16 million MSAs, structural homologs from PDB, and AlphaFold2 protein structure predictions, has been used for training recent alignment-free pLMs.[74,77]

Structure databases

Structure-Based pLMs utilize crystal structures available in the Protein Data Bank (PDB)[161] or predicted structures from AlphaFoldDB.[162] The CATH (Class, Architecture, Topology, Homology) dataset[163] is a system for hierarchically classifying protein domain structure and function; ProSST[106] utilizes CATH for training quantized structure encoders. Additionally, ESM3 uses ESM-Atlas,[36] a large-scale ESM-2 predicted structure database, in addition to AlphaFoldDB.

Functional information databases

GO terms provide a standardized terminology system for protein functions and are used as functional annotations in some pLMs. ProteinKG25 is a large-scale knowledge graph dataset constructed in OntoProtein.[107] From protein sequence information from Swiss-Prot and GO, it contains 612,483 entities (565,254 proteins and 47,229 GO terms) and approximately 5 million triples, with 31 types of relationships defined. ProGen[41] uses GO and NCBI taxonomic information[164] as control tags for conditional sequence generation.

Additionally, ProLLaMA[49] and ESM3[55] also use a dataset on functional annotations called InterPro.[110] It integrates information on protein sequence patterns, domains, and families from multiple databases.

Applications of pLMs

Pretrained pLMs have been shown to improve performance on downstream tasks through transfer learning with additional labeled data.[10–12] Therefore, this section discusses how pLMs are utilized in various downstream task areas including sequence analysis, function prediction, mutation effect prediction, protein optimization, and structure prediction.

Table 4 shows a systematic classification of pLM applications by domain and approach. The classification covers diverse applications from basic biology to clinical applications, categorized from the perspective of three approaches: prediction, design, and

Table 4
Classification of pLM applications by domain and approach.

	Prediction	Design	Analysis
Basic biology	Structure prediction Function prediction Subcellular localization Protein–protein interaction	Biological circuit	Sequence analysis Evolutionary analysis Functional annotation
Protein engineering	Enzyme activity Stability prediction	Enzyme design *De novo* design Directed evolution	
Drug discovery	Drug-target interaction Binding affinity ADMET prediction Off-target prediction	Lead optimization	Binding site analysis Allosteric site analysis target discovery Interaction mechanism
Immunology	TCR-pMHC binding Epitope prediction Antigen specificity Antibody developability	Antibody design Affinity maturation TCR design Vaccine design	Repertoire analysis
Clinical applications	Mutation effect prediction Pathogenicity Disease risk evaluation	Personalized design	Disease mechanism Viral evolution

analysis. The following subsections focus on representative examples of these applications and provide detailed descriptions.

Sequence analysis

pLMs pretrained on large-scale sequence databases implicitly acquire co-evolutionary information similar to MSAs.[36,165,166] Therefore, some attempts have been made to evaluate sequence similarity using pLMs and to incorporate sequence alignments. In sequence analysis, pLMs enable therapeutic applications by detecting distant homologs missed by traditional methods, facilitating function inference, drug target discovery, and off-target prediction. In a pioneering case, Bepler and Berger learned structural similarity of sequence pairs with an LSTM through supervised learning.[167] Lupo *et al.*[168] demonstrated that MSA Transformer can reproduce inter-species relationships captured by traditional phylogenetic tree methods. Furthermore, research incorporating pLMs into MSA has been proposed.[169–172] Becker *et al.*[171] proposed a method to improve MSA quality by incorporating pLM embedding representations into hidden Markov models (HMMs), which performed particularly well in regions of low sequence similarity. Additionally, for protein structure prediction methods like AlphaFold that take MSAs as input, the quality of predicted structures is limited when the quality and quantity of MSAs are low.[93,94] Therefore, MSAGPT[93] and MSA-Generator[94] have been proposed as pLMs for generating MSAs from minimal MSAs. Using generated MSAs has been reported to significantly improve the performance of AlphaFold2, even for proteins with limited MSA evolutionary information.[93,94]

Function prediction

Function prediction using pLMs is useful for applications such as identifying drug targets and elucidating disease processes. Enzyme Commission (EC) number prediction, which predicts EC numbers assigned to classified enzymes, is sometimes used to evaluate pLM performance.[99,115,173] EnzBert[174] is a model that fine-tunes ProtBERT to predict EC numbers. Additionally, CLEAN[175] using contrastive learning[176] and ProtDETR,[177] which provides residue-level interpretation for EC number prediction, have been proposed.

The application of pLMs to predicting[173] and improving[178] turnover numbers, an indicator of catalytic efficiency in enzyme-catalyzed reactions, is also noteworthy. Eom *et al.*[178] integrated pLMs and homology search to narrow down candidate enzymes, experimentally validated them, and discovered promising enzymes (kynureninases) that improved turnover numbers.

Protein subcellular localization[71,179] and identification of functional sites[180] are similarly important for understanding function. DeepLoc2.0[179] is a pioneering method using pLMs to predict subcellular localization with multi-labels. Additionally, DeepLoc[181] is used as a benchmark for subcellular localization prediction, and has been evaluated in many pLMs[66,106,182,183] as a downstream task.

The utilization of pLMs for functional annotation of unknown proteins[184–188] is also attracting attention. In the CAFA (Critical Assessment of Functional Annotation) challenge,[189] LM-based methods outperformed most protein function prediction methods.[190]

Mutation effect prediction

Quantifying the pathogenicity of protein variants in human disease-related genes is crucial for clinical decision-making.[191,192] pLMs can predict mutation effects in zero-shot or few-shot settings by implicitly acquiring preferences for mutations through pretraining. ESM-1v[2] is a pLM specialized for mutation effect prediction. ESM-1v demonstrated extremely high performance on pathogenic variants from ClinVar[191] and HGMD,[193] along with putatively benign missense variants from gnomAD[194] in the assessment by Livesey *et al.*[195]

Notin*et al.* proposed TranceptEVE,[196] a hybrid approach that weights the mutation scores from Tranception,[92] a pLM, according to the number of sequences included in the MSA of EVE,[192] a variational autoencoder (VAE) trained on MSAs. Furthermore, CPT-1[197] is a supervised model that learns a logistic regression model on deep mutational scanning (DMS) data, incorporating mutation scores from EVE and ESM-1v and other features. In ECNet,[198] features extracted via direct coupling analysis[199] from MSAs and embeddings from TAPE were combined using an LSTM-based supervised approach. Luo *et al.*[122,198] successfully created actual variants with up to approximately 8-fold higher activity than the wild type by designing candidate sequences with a prediction model trained on point and double mutants of TEM-1 β-lactamase using ECNet.

Conventionally, MSA-based mutation effect prediction models like GEMME[200] achieved high accuracy but required enormous computational costs for MSA computation. Therefore, VESPA[201]

was proposed as an ensemble logistic regression model for predicting mutations based on information such as predicted conservation scores, substitution scores from BLOSUM62,[202] and substitution probabilities from ProtT5. Furthermore, Marquet *et al.*[203] constructed VespaG, a neural network model that takes ESM-2 embeddings as input and learns with GEMME as ground truth, proposing a model that can predict mutations quickly and accurately without MSA.

Notably, ProteinGym[11] is a benchmark set of 1.5 million missense mutations collected from 87 DMS assays, enabling systematic comparison of zero-shot and supervised mutation prediction models.

Predicting viral evolution is important for anticipating the emergence of new variants and developing effective therapeutics and vaccines. As a pioneering method, CSCS[204] was proposed, which is a BiLSTM model for predicting mutations that escape the immune system for influenza, Human Immunodeficiency Virus 1 (HIV-1), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Additionally, SARS-CoV-2, the cause of the recent pandemic, has received particular attention, and pLM applications specialized in predicting SARS-CoV-2 mutations have been proposed.[205–208] Recently, PLANT[209] was also introduced to map influenza H3N2 sequences into antigenic space for improved vaccine strain selection.

Parameter-efficient fine-tuning

In mutation effect prediction, supervised models showed excellent performance, but these results heavily depend on vast data obtained from high-throughput mutagenesis experiments. Therefore, few-shot learning, i.e., efficiently learning appropriate protein fitness landscapes from a realistic small number of experimental data points, is extremely important.[183,210,211]

Hsu *et al.*[210] reported that for fitness prediction from dozens of data points, Ridge regression using existing evolutionary probability densities such as VAE[212] and Potts models[213] and one-hot encoded amino acid features performed well. While updating all parameters of pLMs for small amounts of data led to overfitting.

PEFT (parameter-efficient fine-tuning)[183,211] has recently gained attention in transfer learning of LLMs as a promising approach for enhancing adaptability to downstream tasks. Updating all parameters of a large-scale model is computationally inefficient and carries the aforementioned risk of overfitting. Therefore, PEFT adjusts only part of the model's parameters, enabling efficient construction of specialized prediction models for small supervised datasets while preserving pLMs' broad biological knowledge. This approach is powerful in therapeutic applications where available experimental data is insufficient, such as rare diseases or early stages of drug development.

LoRA (Low-Rank Adaptation),[214] a representative PEFT technique, significantly reduces the number of learnable parameters in downstream tasks based on the assumption that weight changes in model's tuning are sufficiently low-rank. This method efficiently fine-tunes pre-trained models by expressing the updated weight matrix $W \in \mathbb{R}^{d\times k}$ given a pre-trained model's weight matrix $W_0 \in \mathbb{R}^{d\times k}$ as:

$$W = W_0 + BA \quad (1)$$

Where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are rank r matrices, and $r \ll \min(d,k)$. By freezing W_0 during training and updating only A and B, the number of learnable parameters is significantly reduced.

Schmirler *et al.* investigated multiple PEFT methods and found that LoRA was superior in terms of performance and computational efficiency.[183] Zhou *et al.* proposed Pro-FSFP,[211] a training strategy combining PEFT,[214–216] meta-transfer learning,[217] and learning-to-rank.[218,219] They demonstrated that the performance of various pLMs could be significantly improved using just a few dozen labeled single-site mutants from proteins. Furui & Ohue[220] also proposed a multi-objective active learning approach using pLMs and LoRA to efficiently utilize limited experimental data in antibody CDR-H3 optimization. Furthermore, Sledzieski *et al.*[221] showed that fine-tuning with PEFT demonstrated excellent performance in tasks such as protein–protein interactions and homooligomer symmetry prediction. Additionally, SI-Tuning[222] is a structure-aware PEFT method that significantly improved downstream task performance by efficiently injecting structural information into embeddings and attention through LoRA. Furthermore, Gorantla *et al.*[223] reported that advanced LoRA-based PEFT such as LoKR (Low-Rank Kronecker Product) and LoHA (Low-Rank Hadamard Product)[224] showed higher prediction accuracy than LoRA in protein-ligand binding affinity prediction. This suggests that more efficient and expressive techniques such as LoKR and LoHA should be used when sufficient fine-tuning data is available. Thus, PEFT deserves attention as a powerful method that can be applied regardless of pLM type and can be used in various transfer learning scenarios. In the context of viral proteins, Sawhney *et al.*[225] significantly improved the analysis accuracy of viral sequences by fine-tuning foundation pLMs such as ESM-2 on viral protein sequences using LoRA. These PEFT techniques are particularly effective in constructing domain-specific models where limited training data is available.

Protein structure prediction

In the field of protein structure prediction (PSP) from sequence, the emergence of AlphaFold2[4] achieved a qualitative leap in prediction accuracy.[226] The AlphaFold series[4,227,228] successfully predicted three-dimensional structures with amazing accuracy using MSAs searched from single sequences and residue pair features as input. However, these methods rely on MSAs, and computing MSAs is time-consuming, creating a significant bottleneck in inference time.[36,166,229] Therefore, PSP methods utilizing pLMs such as ESMFold[36] and OmegaFold[166] have emerged.[165,230–233] These models can perform fast and accurate PSP from a single sequence without using MSA, accelerating prediction up to 60 times faster.[36,165] Furthermore, RGN2,[165] trRosettaX-Single,[231] and OmegaFold[166] report outperforming AlphaFold2[4] and RoseTTAFold[234] in predictions for orphan proteins, designed proteins, and antibodies.[165,166] Additionally, Weissenow *et al.*[233] mention that EMBER2 captures the correlation between changes in predicted structures and mutation effects from DMS[235] more than ColabFold.[101]

Recent models such as ESM3[55] and xT-Fold,[100] a PSP model using xTrimoPGLM, have significantly outperformed existing methods like ESMFold. ESM3 generates higher quality structures than ESMFold even with smaller models, suggesting that multimodal representations are important for PSP. However, PSP methods using MSA like AlphaFold series still remain more powerful than those using only pLMs, and further research in this area is anticipated.

In antibody structure prediction, accurately predicting CDR loop structures is a major challenge, and several PSP methods using antibody LMs have been proposed.[236–238] Nevertheless, predicting the CDR H3 loop, which is most important for antigen recognition, remains a significant challenge.[238]

Protein design

Generating protein sequences with desired functions is an important challenge in protein engineering. In therapeutic

research, it can be useful for rational design of antibody drugs[128] and development of universal vaccines that suppress the generation of immune escape mutations.[239] Gasser *et al.*[239] combined autoregressive Transformers and VAEs for vaccine and therapeutic protein design. Verkuil*et al.*[240] used ESM-2 to generate sequences in two scenarios, fixed backbone design and free generation, successfully generating novel protein sequences with low similarity to known sequences that were also soluble. pLMs are beginning to succeed in designing proteins with desired functions that differ from known sequences. For example, sequences designed by ProGen fine-tuned on the lysozyme family were confirmed by biochemical experiments to have enzyme activity comparable to known natural proteins.[41] More recently, ESM3 successfully designed a novel protein called esmGFP without model fine-tuning, which has only 58 % sequence identity with known natural green fluorescent proteins (GFP) but is equally bright.[55] This was achieved by starting with prompts for the minimal set of sequence and structure information near the chromophore formation site from a template protein, and performing iterative joint optimization of sequence and structure tokens called chain of thought.

Drug–target interaction prediction

In small molecule drug discovery, predicting drug–target interactions is important for improving the efficiency and success rate. This enables the identification of targets for drugs, screening of candidate compounds that could bind to specific targets,[241] and drug repositioning.[242] pLMs help by predicting drug–protein interactions from sequence information alone, useful when structural information cannot be utilized or when exploring vast combinations of drugs and proteins. TransformerCPI[243] used protein representations through word2vec. Then, TransformerCPI2.0[244] performed compound–protein interaction prediction using protein representations from TAPE-BERT. DLM-DTI[241] developed a drug–target interaction prediction method utilizing two modality language models: a chemical language model[245] and a pLM. Additionally, sequence-based binding site prediction methods[246–248] have lower computational cost and wider applicability, but face challenges in accuracy and interpretability due to the lack of structural information.[249] Other interesting pLMs applications for small molecule include detection of potential binding sites[250] and prediction of allosteric sites.[251]

Applications in various domains

Finally, we overview models for predicting various physical properties and functions of proteins. Flamholz *et al.* demonstrated the effectiveness of pLMs for annotating viral protein families.[252] They trained classifiers using embeddings generated by pLMs for distant viral proteins that were difficult to annotate using traditional homology search-based methods.

Signal peptides (SPs) are short amino acid sequences essential for protein subcellular localization and secretion, and their accurate prediction is important for protein function analysis. SignalP 6.0[253] is an SP prediction model combining ProtTrans[26] with conditional random field (CRF),[254] enabling prediction of all five SP types, whereas previous research could only predict specific SP types. Furthermore, Zeng *et al.*[255] proposed an SP prediction framework based on LoRA, and it outperformed both SignalP 6.0 and full fine-tuning of ESM-2.

Various physical properties of proteins are being predicted using pLMs. IDP-BERT[60] is a model for predicting the physical properties of intrinsically disordered proteins (IDPs). Using ProtBERT (the BERT model of ProtTrans) as a backbone, it was trained on IDP sequences from the DisProt[256,257] database to predict IDP properties such as structural, dynamic, and thermodynamic characteristics. PeptideBERT[61] is a BERT-based LM specialized for peptide property prediction. Using ProtBERT as a backbone, it was fine-tuned for predicting important peptide properties such as hemolytic activity, solubility, and non-adhesiveness. PeptideBERT showed excellent performance particularly in predicting hemolytic activity and non-adhesiveness. Kim *et al.*[258] proposed GPCR-BERT, in which ProtBERT was fine-tuned with 254 class A G-protein coupled receptor (GPCR) sequences collected from the GPCRdb database[259] to analyze motif sequences of GPCRs.

Other applications exist, including repertoire analysis,[153,260–263] antimicrobial peptides,[264,265] solubility of proteins expressed in *E. coli*,[266] thermal stability,[267] allergens,[268–270] and protein–protein interaction prediction,[271,272] but these are omitted due to space constraints.

Challenges of pLMs in therapeutic applications

Finally, to transform pLMs from basic research tools into technologies that contribute to the actual prevention, diagnosis, and treatment of diseases, several important requirements must be met. First, biological validity and interpretability are essential. For researchers to trust model predictions, it must be possible to provide biological explanations of how the model's predictions relate to disease mechanisms. Second, showing high performance on general benchmarks alone is insufficient; rigorous validation with disease-specific data is required. It is necessary to verify that models remain robust and accurate when applied to real-world data, such as disease-specific datasets or patient data, despite potential biases in training data. Third, the lack of supervised data in therapeutic research poses another significant challenge. PEFT approaches have only recently gained recognition in pLM research and represent a key technology for successful clinical applications. Fourth, the development of multimodal pLMs that integrate diverse information sources is important. Currently, research utilizing individual information sources such as sequence, structure, and co-evolutionary information is progressing, but multimodal approaches that comprehensively utilize these are still in their early stages. More comprehensive protein understanding derived from diverse information sources could solve the fundamental data shortage in therapeutic research. By addressing these challenges, pLMs could become more practically useful for drug discovery and therapeutic research.

Conclusion

This article provided an overview of pLMs from fundamentals to the latest applications, explaining how representations of evolutionary, structural, and functional information acquired through large-scale pretraining of sequence data contribute to performance improvements in various protein-related tasks. Looking at the overview, there seem to be mainly two directions in recent research approaches for pLMs as pretrained models. One involves efforts to expand model parameters and datasets, following the success in NLP[5] by pretraining larger pLMs according to scaling laws[55,124,100] or incorporating information from different data sources beyond amino acid sequence information. On the other hand, learning LMs with domain-specific datasets,[153] exploring alternative architectures to Transformers,[46,70,71,74,77] and parameter-efficient pretraining[66,47] or fine-tuning[183,211] approaches aimed at specialized domains or optimizing computational efficiency are also flourishing. The development through these approaches in pLM research is expected to continue in the future. In either case, for pLMs to perform sufficiently well in

downstream tasks, high-quality labeled training data related to the task and utilization of domain knowledge are important.

pLMs have emerged as powerful tools that unlock the language of life - proteins - enabling structure prediction, novel protein design, and prediction of mutational effects and functions. In the future, as pLM research addresses challenges such as computational efficiency, data shortage, and interpretability, significant advances in drug discovery processes and personalized medicine are expected.

Acknowledgments

This work was partly supported by JSPS KAKENHI (JP23H04880, JP23H04887, JP24KJ1091), AMED BINDS (JP25ama121026), and JST FOREST (JPMJFR216J). The authors are grateful to Mr. Apakorn Kengkanna for insightful comments of the manuscript.

Conflict of interest

The authors have no conflict of interest to declare.

References

1. Bepler T, Berger B. Learning the protein language: evolution, structure, and function. *Cell Syst* 2021;**12**:654–69.
2. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv Neural Inf Process Syst* 2021;**34**:29287–303.
3. Xiao Y, Zhao W, Zhang J, Jin Y, Zhang H, Ren Z, et al. Protein Large Language Models: A Comprehensive Survey. arXiv; 2025. https://doi.org/10.48550/arXiv.2502.17504. arXiv:2502.17504.
4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
5. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;**33**:1877–901.
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**.
7. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. *Language Models are Unsupervised Multitask Learners*. 2019.
8. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 2021;**19**:1750–8.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North*. Association for Computational Linguistics; 2019. p. 4171–86.
10. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst* 2019;**32**:9689–701.
11. Notin P, Kollasch A, Ritter D, van Niekerk L, Paul S, Spinner H, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv Neural Inf Process Syst* 2023;**36**:64331–79.
12. Dallago C, Mou J, Johnston KE, Wittmann B, Bhattacharya N, Goldman S, et al. FLIP: benchmark tasks in fitness landscape inference for proteins. In: *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*; 2021.
13. Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *Int J Mach Learn Cybern* 2024:1–65.
14. Wang L, Li X, Zhang H, Wang J, Jiang D, Xue Z, et al. A Comprehensive Review of Protein Language Models. arXiv; 2025. https://doi.org/10.48550/arXiv.2502.06881. arXiv:2502.06881.
15. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 2019;**20**:723.
16. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
17. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
18. Krause B, Lu L, Murray I, Renals S. Multiplicative LSTM for sequence modelling. In: *International conference on learning representations workshop track*; 2017.
19. Strodthoff N, Wagner P, Wenzel M, Samek W. UDSMProt: universal deep sequence models for protein classification. *Bioinformatics* 2020;**36**:2401–9.
20. Merity S, Keskar NS, Socher R. Regularizing and optimizing LSTM language models. In: *International conference on learning representations*; 2018.
21. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-prot: juggling between evolution and stability. *Brief Bioinform* 2004;**5**:39–55.
22. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The "SWISS-PROT" protein knowledgebase and its supplement "TrEMBL" in 2003. *Nucleic Acids Res* 2003;**31**:365–70.
23. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res* 2014;**42**:D222–30.
24. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
25. Nambiar A, Heflin M, Liu S, Maslov S, Hopkins M, Ritz A. Transforming the language of life: transformer neural networks for protein prediction tasks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics***5**; 2020. p. 1–8.
26. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27.
27. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*; 2019. p. 2978–88.
28. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;**21**:1–67.
29. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. In: *International conference on learning representations*; 2019.
30. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. In: *Proceedings of the 33rd international conference on neural information processing systems*; 2019. p. 5753–63.
31. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. In: *International conference on learning representations*; 2019.
32. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;**9**:2542.
33. Geffen Y, Ofran Y, Unger R. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics* 2022;**38**(Suppl 2):ii95–8.
34. Hesslow D, Zanichelli N, Notin P, Poli I, Marks D. RITA: a study on scaling up generative protein sequence models. In: *The 2022 ICML workshop on computational biology*; 2022.
35. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 2022;**13**:4348.
36. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30.
37. ESM team. *ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning*. 2024. Available at: https://www.evolutionaryscale.ai/blog/esm-cambrian.
38. Chen IMA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* 2023;**51**:D723–32.
39. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;**51**:D753–9.
40. Wang L, Zhang H, Xu W, Xue Z, Wang Y. Deciphering the protein landscape with ProtFlash, a lightweight language model. *Cell Rep Phys Sci* 2023;**4**:101600.
41. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023;**41**:1099–106.
42. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. UniProt archive. *Bioinformatics* 2004;**20**:3236–7.
43. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). *Nucleic Acids Res* 2005;**33**:D154–9.
44. Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A. ProGen2: exploring the boundaries of protein language models. *Cell Syst* 2023;**14**:968–978.e3.
45. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci* 2022;**31**:141–6.
46. Yang KK, Fusi N, Lu AX. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst* 2024;**15**:286–94.e2.
47. Fournier Q, Vernon RM, van der Sloot A, Schulz B, Chandar S, Langmead C. *Protein Language Models: Is Scaling Necessary?*. bioRxiv; 2024. https://doi.org/10.1101/2024.09.23.614603. bioRxiv:2024.09.23.614603.
48. Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. SCOPe: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res* 2022;**50**:D553–9.
49. Zheng K, Long S, Lu T, Yang J, Dai X, Zhang M, et al. *ESM All-Atom: Multi-Scale Protein Language Model for Unified Molecular Modeling*. arXiv; 2024. https://doi.org/10.48550/arXiv.2403.12995. arXiv:2403.12995.
50. Mikolov T, Chen K, Corrado G, Dean J. *Efficient Estimation of Word Representations in Vector Space*. arXiv; 2013. https://doi.org/10.48550/arXiv.1301.3781. arXiv:1301.3781.

51. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *Proceedings of the 31st international conference on machine learning*vol. 32. PMLR; 2014. p. 1188–96. 2.
52. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 2015;**10**:e0141287.
53. Iuchi H, Matsutani T, Yamada K, Iwano N, Sumi S, Hosoda S, et al. Representation learning applications in biological sequence analysis. *Comput Struct Biotechnol J* 2021;**19**:3198–208.
54. Kimothi D, Soni A, Biyani P, Hogan JM. *Distributed Representations For Biological Sequence Analysis*. arXiv; 2016. https://doi.org/10.48550/arXiv.1608.05949. arXiv:1608.05949.
55. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. *Science* 2025;**387**:850–8.
56. Chen R, Palpant N, Foley G, Boden M. *Multilingual Model Improves Zero-Shot Prediction of Disease Effects on Proteins*. bioRxiv; 2025. https://doi.org/10.1101/2025.03.12.642937. bioRxiv:2025.03.12.642937.
57. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv; 2019. https://doi.org/10.48550/arXiv.1907.11692. arXiv:1907.11692.
58. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, et al. Learning inverse folding from millions of predicted structures. In: *Proceedings of the 39th international conference on machine learning*vol. 162. PMLR; 2022. p. 8946–70.
59. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R. Learning from protein structure with geometric vector perceptrons. In: *International conference on learning representations*; 2021.
60. Mollaei P, Sadasivam D, Guntuboina C, Barati Farimani A. IDP-BERT: predicting properties of intrinsically disordered proteins using large language models. *J Phys Chem B* 2024;**128**:12030–7.
61. Guntuboina C, Das A, Mollaei P, Kim S, Barati Farimani A. PeptideBERT: a language model based on transformers for peptide property prediction. *J Phys Chem Lett* 2023;**14**:10427–34.
62. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Preprint]. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
63. Keskar NS, McCann B, Varshney LR, Xiong C, Socher R. *CTRL: A Conditional Transformer Language Model For Controllable Generation*. arXiv; 2019. https://doi.org/10.48550/arXiv.1909.05858/. arXiv:1909.05858.
64. Bhatnagar A, Jain S, Beazer J, Curran SC, Hoffnagle AM, Ching K, et al. *Scaling Unlocks Broader Generation and Deeper Functional Understanding of Proteins*. bioRxiv; 2025. https://doi.org/10.1101/2025.04.15.649055. bioRxiv: 2025.04.15.649055.
65. Shazeer N, Mirhoseini a, Maziarz k, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: *International conference on learning representations*; 2017.
66. Elnaggar A, Essam H, Salah-Eldin W, Moustafa W, Elkerdawy M, Rochereau C, et al. *Ankh: Optimized Protein Language Model Unlocks General-purpose Modelling*. arXiv; 2023. https://doi.org/10.48550/arXiv.2301.06568. arXiv: 2301.06568.
67. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M, et al. Bilingual language model for protein sequence and structure. *NAR Genom Bioinform* 2024;**6**:lqae150.
68. Kenlay H, Dreyer FA, Kovaltsuk A, Miketa D, Pires D, Deane CM. Large scale paired antibody language models. *Plos Comput Biol* 2024;**20**:e1012646.
69. Zhou HY, Fu Y, Zhang Z, Cheng B, Yu Y. Protein representation learning via knowledge enhanced primary structure reasoning. In: *The eleventh international conference on learning representations*; 2022.
70. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10.
71. Stärk H, Dallago C, Heinzinger M, Rost B. Light attention predicts protein location from the language of life. *Bioinform Adv* 2021;**1**:vbab035.
72. Pöppel K, Beck M, Spanring M, Auer A, Prudnikova O, Kopp MK, et al. xLSTM: extended long short-term memory. In: *First workshop on long-context foundation models at ICML 2024*; 2024.
73. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. In: *First conference on language modeling*; 2024.
74. Schmidinger N, Schneckenreiter L, Seidl P, Schimunek J, Hoedt PJ, Brandstetter J, et al. Bio-xLSTM: generative modeling, representation and in-context learning of biological and chemical sequences. In: *The thirteenth international conference on learning representations*; 2024.
75. Ahdritz G, Bouatta N, Kadyan S, Jarosch L, Berenberg D, Fisk I, et al. OpenProteinSet: training data for structural biology at scale. *Adv Neural Inf Process Syst* 2023;**36**:4597–609.
76. Gu A, Johnson I, Goel K, Saab K, Dao T, Rudra A, et al. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv Neural Inf Process Syst* 2021;**34**:572–85.
77. Sgarbossa D, Malbranke C, Bitbol AF. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics* 2025;**41**:btaf348.
78. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: *International conference on machine learning*vol. 37. PMLR; 2015. p. 2256–65.
79. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 2020;**33**:6840–51.
80. Wang X, Zheng Z, Ye F, Xue D, Huang S, Gu Q. Diffusion language models are versatile protein learners. In: *Proceedings of the 41st international conference on machine learning. ICML'24*vol. 235. PMLR; 2024. p. 52309–33.
81. Wang X, Zheng Z, Fei YE, Xue D, Huang S, Gu Q. DPLM-2: a multimodal diffusion protein language model. In: *The thirteenth international conference on learning representations*; 2024.
82. Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani N. BERTology meets biology: interpreting attention in protein language models. In: *International conference on learning representations*; 2020.
83. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. In: *International conference on learning representations*; 2020.
84. Zhang Z, Wayment-Steele HK, Brixi G, Wang H, Kern D, Ovchinnikov S. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc Natl Acad Sci* 2024;**121**:e2406285121.
85. Dao T, Fu DY, Ermon S, Rudra A, Re C. FlashAttention: fast and memory-efficient exact attention with IO-Awareness. In: *Adv. Neural inf. Process. Syst.*; 2022.
86. Shazeer N. *GLU Variants Improve Transformer*. arXiv; 2020. https://doi.org/10.48550/arXiv.2002.05202. arXiv:200205202.
87. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International conference on learning representations*; 2018.
88. Li FZ, Amini AP, Yue Y, Yang KK, Lu AX. Feature reuse and scaling: understanding transfer learning with protein language models. In: *Proceedings of the 41st international conference on machine learning. ICML'24*vol. 235. PMLR; 2024. p. 27351–75.
89. Cheng X, Chen B, Li P, Gong J, Tang J, Song L. Training compute-optimal protein language models. In: *ICML 2024 workshop on efficient and accessible foundation models for biological discovery*; 2024.
90. Vieira LC, Handojo ML, Wilke CO. Medium-sized protein language models perform well at transfer learning on realistic datasets. *Sci Rep* 2025;**15**: 21400.
91. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, et al. MSA transformer. In: Meila M, Zhang T, editors. *International conference on machine learning*, vol. 139. PMLR; 2021. p. 8844–56.
92. Notin P, Dias M, Frazer J, Marchena-Hurtado J, Gomez AN, Marks D, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: *International conference on machine learning*vol. 162. PMLR; 2022. p. 16990–7017.
93. Chen B, Bei Z, Cheng X, Li P, Tang J, Song L. MSAGPT: neural prompting protein structure prediction via msa generative pre-training. *Adv Neural Inf Process Syst* 2025;**37**:37504–34.
94. Zhang L, Chen J, Shen T, Li Y, Sun S. MSA generation with Seqs2Seqs pre-training: advancing protein structure predictions. *Adv Neural Inf Process Syst* 2024;**37**:57324–48.
95. Truong Jr T, Bepler T. PoET: a generative model of protein families as sequences-of-sequences. *Adv Neural Inf Process Syst* 2023;**36**:77379–415.
96. Wang Z, Combs SA, Brand R, Calvo MR, Xu P, Price G, et al. LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Sci Rep* 2022;**12**:6832.
97. Wang Z, Zhang Q, Shuang-Wei HU, Yu H, Jin X, Gong Z, et al. Multi-level protein structure pre-training via prompt learning. In: *The eleventh international conference on learning representations*; 2022.
98. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13.
99. Zhang Z, Wang C, Xu M, Chenthamarakshan V, Lozano A, Das P, et al. *A Systematic Study of Joint Representation Learning on Protein Sequences and Structures*. arXiv; 2023. https://doi.org/10.48550/arXiv.2303.06275. arXiv: 2303.06275.
100. Chen B, Cheng X, Li P, Geng YA, Gong J, Li S, et al. xTrimoPGLM: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nat Methods* 2025;**22**:1028–39.
101. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**: 679–82.
102. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. SaProt: protein language modeling with structure-aware vocabulary. In: *The twelfth international conference on learning representations*; 2023.
103. Sun Y, Shen Y. Structure-informed protein language models are robust predictors for variant effects. *Hum Genet* 2025;**144**:209–25.
104. Wang D, Pourmirzaei M, Abbas UL, Zeng S, Manshour N, Esmaili F, et al. S-PLM: structure-aware protein language model via contrastive learning between sequence and structure. *Adv Sci* 2025;**12**:e2404212.
105. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF international conference on computer vision (ICCV)*; 2021. p. 9992–10002.
106. Li M, Tan Y, Ma X, Zhong B, Yu H, Zhou Z, et al. ProSST: protein language modeling with quantized structure and disentangled attention. *Adv Neural Inf Process Syst* 2025;**37**:35700–26.
107. Zhang N, Bi Z, Liang X, Cheng S, Hong H, Deng S, et al. OntoProtein: protein pretraining with gene ontology embedding. In: *International conference on learning representations*; 2021.

108. Wu KE, Chang H, Zou J. *ProteinCLIP: Enhancing Protein Language Models With Natural Language*. bioRxiv; 2024. https://doi.org/10.1101/2024.05.14.594226. bioRxiv:2024.05.14.594226.
109. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*, 2023, arXiv, https://doi.org/10.48550/arXiv.2307.09288, arXiv230709288.
110. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. *Nucleic Acids Res* 2023;**51**:D418–27.
111. Erckert K, Rost B. Assessing the role of evolutionary information for enhancing protein language model embeddings. *Sci Rep* 2024;**14**:20692.
112. Tan Y, Li M, Zhou B, Zhong B, Zheng L, Tan P, et al. Simple, efficient, and scalable structure-aware adapter boosts protein language models. *J Chem Inf Model* 2024;**64**:6338–49.
113. Zhang Z, Lu J, Chenthamarakshan V, Lozano A, Das P, Tang J. *Structure-informed Protein Language Model*. arXiv; 2024. https://doi.org/10.48550/arXiv.2402.05856. arXiv: 2402.05856.
114. Zhang Z, Xu M, Jamasb AR, Chenthamarakshan V, Lozano A, Das P, et al. Protein representation learning by geometric structure pretraining. In: *The eleventh international conference on learning representations*; 2022.
115. Lee Y, Yu H, Lee J, Kim J. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In: *The twelfth international conference on learning representations*; 2023.
116. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with foldseek. *Nat Biotechnol* 2024;**42**:243–6.
117. Zheng J, Li SZ. CCPL: cross-modal contrastive protein learning. In: *The 27th international conference on pattern recognition (ICPR 2024), LNCS***15327**; 2025. p. 22–38.
118. Ouyang-Zhang J, Gong C, Zhao Y, Kraehenbuehl P, Klivans A, Diaz DJ. Distilling structural representations into protein sequence models. In: *The thirteenth international conference on learning representations*; 2024.
119. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*vol. 139. PMLR; 2021. p. 8748–63.
120. Outeiral C, Deane CM. Codon language embeddings provide strong signals for use in protein engineering. *Nat Mach Intell* 2024;**6**:170–9.
121. Cummins C, Ahamed A, Aslam R, Burgin J, Devraj R, Edbali O, et al. The European nucleotide archive in 2021. *Nucleic Acids Res* 2022;**50**:D106–10.
122. Yamaguchi H, Saito Y. Protein language models. *JSBi Bioinform Rev* 2023;**4**:52–67.
123. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, et al. Sequence modeling and design from molecular to genome scale with evo. *Science* 2024;**386**:eado9336.
124. Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, et al. *Genome Modeling and Design Across All Domains of Life With Evo 2*. bioRxiv; 2025. https://doi.org/10.1101/2025.02.18.638918. bioRxiv:2025.02.18.638918.
125. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-mol: a universal 3D molecular representation learning framework. In: *The eleventh international conference on learning representations*; 2022.
126. Wollacott AM, Xue C, Qin Q, Hua J, Bohnuud T, Viswanathan K, et al. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Eng Des Sel* 2019;**32**:347–54.
127. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, et al. BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* 2022;**14**:2020203.
128. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. In: *Machine Learning for Structural Biology Workshop at NeurIPS 2021*; 2021.
129. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. *Bioinform Adv* 2022;**2**:vbac046.
130. Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. *Patterns* 2022;**3**:100513.
131. Ramon A, Ali M, Atkinson M, Saturnino A, Didi K, Visentin C, et al. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nat Mach Intell* 2024;**6**:74–91.
132. van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. *Adv Neural Inf Process Syst* 2017;**30**.
133. Nishino T, Kato N, Tsutaoka T, Li Y, Ohue M. Ohue, REALM: Region-Empowered Antibody Language Model for Antibody Property Prediction. In: *2024 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2024, 7104–7106.
134. Olsen TH, Moal IH, Deane CM. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics* 2024;**40**:btae618.
135. Hadsund JT, Satława T, Janusz B, Shan L, Zhou L, Röttger R, et al. nanoBERT: a deep learning model for gene agnostic navigation of the nanobody mutational space. *Bioinform Adv* 2024;**4**:vbae033.
136. Deszyński P, Młokosiewicz J, Volanakis A, Jaszczyszyn I, Castellana N, Bonissone S, et al. INDI—Integrated nanobody database for immunoinformatics. *Nucleic Acids Res* 2022;**50**:D1273–81.
137. Wu KE, Yost K, Daniel B, Belk J, Xia Y, Egawa T, et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. In: *Machine learning in computational biology*vol. 240. PMLR; 2024. p. 194–229.
138. Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res* 2020;**48**:D1057–62.
139. Zhang W, Wang L, Liu K, Wei X, Yang K, Du W, et al. PIRD: Pan immune repertoire database. *Bioinformatics* 2020;**36**:897–903.
140. Chen SY, Yue T, Lei Q, Guo AY. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res* 2021;**49**:D468–74.
141. Essaghir A, Sathiyamoorthy NK, Smyth P, Postelnicu A, Ghiviriga S, Ghita A, et al. *T-cell Receptor Specific Protein Language Model for Prediction and Interpretation of Epitope Binsd42ding (ProtLM.TCR)*. bioRxiv; 2022. https://doi.org/10.1101/2022.11.28.518167. bioRxiv:2022.11.28.518167.
142. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, Holzleitner M, et al. Modern hopfield networks and attention for immune repertoire classification. *Adv Neural Inf Process Syst* 2020;**33**:18832–45.
143. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017;**49**:659–65.
144. Zhao Y, Su X, Zhang W, Mai S, Xu Z, Qin C, et al. SC-AIR-BERT: a pre-trained single-cell model for predicting the antigen-binding specificity of the adaptive immune receptor. *Brief Bioinform* 2023;**24**:bbad191.
145. Zhang J, Ma W, Yao H. Accurate TCR-pMHC interaction prediction using a BERT-based transfer learning method. *Brief Bioinform* 2023;**25**:bbad436.
146. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43.
147. Meynard-Piganeau B, Feinauer C, Weigt M, Walczak AM, Mora T. TULIP: a transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc Natl Acad Sci U S A* 2024;**121**:e2316401121.
148. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* 2017;**33**:2924–9.
149. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**:W449–54.
150. Amin AN, Gruver N, Kuang Y, Li YL, Elliott H, McCarter C, et al. Bayesian optimization of antibodies informed by a generative model of evolving sequences. In: *The thirteenth international conference on learning representations*; 2024.
151. Kitaura K, Yamashita H, Ayabe H, Shini T, Matsutani T, Suzuki R. Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. *Front Immunol* 2017;**8**:389.
152. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:318–27.
153. Dounas A, Cotet TS, Yermanos A. *Learning Immune Receptor Representations With Protein Language Models*. arXiv; 2024. https://doi.org/10.48550/arXiv.2402.03823. arXiv:2402.03823.
154. Myronov A, Mazzocco G, Krol P, Plewczynski D. BERTrand—peptide:TCR binding prediction using bidirectional encoder representations from transformers augmented with random TCR pairing. *Bioinformatics* 2023;**39**:btad468.
155. Kwee BPY, Messemaker M, Marcus E, Oliveira G, Scheper W, Wu C, et al. *STAPLER: Efficient Learning of TCR-peptide Specificity Prediction From Full-length TCR-peptide Data*. bioRxiv; 2023. https://doi.org/10.1101/2023.04.25.538237. bioRxiv:2023.04.25.538237.
156. Zhang P, Bang S, Lee H. PiTE: TCR-epitope binding affinity prediction pipeline using Transformer-based sequence encoder. *Pac Symp Biocomput* 2023;**28**:347–58.
157. Zhang P, Bang S, Cai M, Lee H. *Context-aware Amino Acid Embedding Advances Analysis of TCR-epitope Interactions*. bioRxiv; 2023. https://doi.org/10.1101/2023.04.12.536635. bioRxiv:2023.04.12.536635.
158. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium UniProt. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;**31**:926–32.
159. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 2017;**45**:D170–6.
160. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8.
161. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 2019;**47**:D464–74.
162. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;**52**:D368–75.
163. Yang KK, Zanichelli N, Yeh H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Eng Des Sel* 2023;**36**:gzad015.
164. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res* 2012;**40**:D136–43.

165. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol* 2022;**40**:1617–23.
166. Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. *High-resolution de novo Structure Prediction From Primary Sequence*. bioRxiv; 2022. https://doi.org/10.1101/2022.07.21.500999. bioRxiv:2022.07.21.500999.
167. Bepler T, Berger B. *Learning Protein Sequence Embeddings Using Information From Structure*. arXiv; 2019. https://doi.org/10.48550/arXiv.1902.08661. arXiv:1902.08661.
168. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat Commun* 2022;**13**:6298.
169. McWhite CD, Armour-Garb I, Singh M. Leveraging protein language models for accurate multiple sequence alignments. *Genome Res* 2023;**33**:1145–53.
170. Kaminski K, Ludwiczak J, Pawlicki K, Alva V, Dunin-Horkawicz S. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* 2023;**39**:btad579.
171. Becker F, Stanke M. learnMSA2: deep protein multiple alignments with large language and hidden Markov models. *Bioinformatics* 2024;**40**(Suppl 2): ii79–86.
172. Liu W, Wang Z, You R, Xie C, Wei H, Xiong Y, et al. PLMSearch: protein language model powers accurate and fast sequence search for remote homology. *Nat Commun* 2024;**15**:2775.
173. Kroll A, Rousset Y, Hu XP, Liebrand NA, Lercher MJ. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nat Commun* 2023;**14**:4139.
174. Buton N, Coste F, Le Cunff Y. Predicting enzymatic function of protein sequences with attention. *Bioinformatics* 2023;**39**:btad620.
175. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. *Science* 2023;**379**:1358–63.
176. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. *Adv Neural Inf Process Syst* 2020;**33**:18661–73.
177. Yang Z, Su B, Chen J, Wen JR. *Interpretable Enzyme Function Prediction Via Residue-level Detection*. arXiv; 2025. https://doi.org/10.48550/arXiv.2501.05644.
178. Eom H, Park S, Cho KS, Lee J, Kim H, Kim S, et al. Discovery of highly active kynureninases for cancer immunotherapy through protein language model. *Nucleic Acids Res* 2025;**53**:gkae1245.
179. Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res* 2022;**50**:W228–34.
180. Yeung W, Zhou Z, Li S, Kannan N. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. *Brief Bioinform* 2023;**24**:bbac599.
181. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 2017;**33**:3387–95.
182. Xu H, Wang S. ProTranslator: zero-shot protein function prediction using textual description. In: *RECOMB2022, lecture notes in computer science***13278**; 2022. p. 279–94.
183. Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat Commun* 2024;**15**:7407.
184. Zhu YH, Zhang C, Yu DJ, Zhang Y. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. *Plos Comput Biol* 2022;**18**:e1010793.
185. Yuan Q, Xie J, Xie J, Zhao H, Yang Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief Bioinform* 2023;**24**:bbad117.
186. Kulmanov M, Guzmań-Vega FJ, Duek Roggli P, Lane L, Arold ST, Hoehndorf R. Protein function prediction as approximate semantic entailment. *Nat Mach Intell* 2024;**6**:220–8.
187. Wang S, You R, Liu Y, Xiong Y, Zhu S. NetGO 3.0: protein language model improves large-scale functional annotations. *Genom Proteom Bioinf* 2023;**21**: 349–58.
188. Chua ZM, Rajesh A, Sinha S, Adams PD. *PROTGOAT : Improved Automated Protein Function Predictions Using Protein Language Models*. bioRxiv; 2024. https://doi.org/10.1101/2024.04.01.587572. bioRxiv:2024.04.01.587572.
189. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 2021;**36**:5516–8.
190. Chen JY, Wang JF, Hu Y, Li XH, Qian YR, Song CL. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Front Bioeng Biotechnol* 2025;**13**: 1506508.
191. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. *Hum Mutat* 2018;**39**:1623–30.
192. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5.
193. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human gene mutation database (HGMD): 2003 update: hgmd 2003 update. *Hum Mutat* 2003;**21**:577–81.
194. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.
195. Livesey BJ, Marsh JA. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol Syst Biol* 2023;**19**:e11474.
196. Notin P, Van Niekerk L, Kollasch AW, Ritter D, Gal Y, Marks DS. TranceptEVE: combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In: *NeurIPS 2022 workshop on learning meaningful representations of life*; 2022.
197. Jagota M, Ye C, Albors C, Rastogi R, Koehl A, Ioannidis N, et al. Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biol* 2023;**24**:182.
198. Luo Y, Jiang G, Yu T, Liu Y, Vo L, Ding H, et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat Commun* 2021;**12**:5743.
199. Seemayer S, Gruber M, Söding J. CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 2014;**30**:3128–30.
200. Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol* 2019;**36**:2604–19.
201. Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. *Hum Genet* 2022;**141**:1629–47.
202. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9.
203. Marquet C, Schlensok J, Abakarova M, Rost B, Laine E. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics* 2024;**40**:btae621.
204. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. *Science* 2021;**371**:284–8.
205. Zhou B, Zhou H, Zhang X, Xu X, Chai Y, Zheng Z, et al. TEMPO: a transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Comput Biol Med* 2023;**152**:106264. 106264.
206. Han W, Chen N, Xu X, Sahil A, Zhou J, Li Z, et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. *Nat Commun* 2023;**14**:3478.
207. Ito J, Strange A, Liu W, Joas G, Lytras S, Genotype to Phenotype Japan (G2P-Japan) Consortium, et al. A protein language model for exploring viral fitness landscapes, *Nat Commun*, **16**, 2025, 4236.
208. Ma E, Guo X, Hu M, Wang P, Wang X, Wei C, et al. A predictive language model for SARS-CoV-2 evolution. *Signal Transduct Target Ther* 2024;**9**:353.
209. Ito J, Kawakubo S, Unno H, Strange A, Lytras S, Okumura K, et al. *Integrative modeling of seasonal influenza evolution via AI-powered antigenic cartography*. *bioRxiv* 2025. https://doi.org/10.1101/2025.08.04.668423. bioRxiv.2025.08.04.668423.
210. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* 2022;**40**:1114–22.
211. Zhou Z, Zhang L, Yu Y, Wu B, Li M, Hong L, et al. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nat Commun* 2024;**15**:5566.
212. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 2018;**15**:816–22.
213. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 2017;**35**:128–35.
214. Hu JE, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. In: *International conference on learning representations*; 2022.
215. Liu H, Tam D, Muqeeth M, Mohta J, Huang T, Bansal M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Adv Neural Inf Process Syst* 2022:1950–65.
216. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell* 2023;**5**: 220–35.
217. Sun Q, Liu Y, Chua TS, Schiele B. Meta-transfer learning for few-shot learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 2019. p. 403–12.
218. Chen W, Liu TY, Lan Y, Ma ZM, Li H. Ranking measures and loss functions in learning to rank. *Adv Neural Inf Process Syst* 2009;**22**.
219. Xia F, Liu TY, Wang J, Zhang W, Li H. Listwise approach to learning to rank: theory and algorithm. In: *Proceedings of the 25th international conference on machine learning - ICML'08*; 2008. p. 1192–9.
220. K. Furui K, Ohue M, ALLM-Ab: Active Learning-Driven Antibody Optimization Using Fine-Tuned Protein Language Models, 2025, bioRxiv, doi:10.1101/2025.08.05.668775, bioRxiv:10.1101/2025.08.05.668775
221. Sledzieski S, Kshirsagar M, Baek M, Dodhia R, Lavista Ferres J, Berger B. Democratizing protein language models with parameter-efficient fine-tuning. *Proc Natl Acad Sci U S A* 2024;**121**:e2405840121.
222. Zhang Z, Zhou Y, Zheng J, Feng C, Cui S, Wang S, et al. Boost protein language model with injected structure information through parameter efficient fine-tuning. *Comput Biol Med* 2025;**195**:110607.
223. Gorantla R, Gema AP, Yang IX, Serrano-Morrás Á, Suutari B, Jiménez JJ, et al. *Learning Binding Affinities Via Fine-tuning of Protein and Ligand Language Models*. bioRxiv; 2024. https://doi.org/10.1101/2024.11.01.621495. bioRxiv: 2024.11.01.621495.
224. Yeh SY, Hsieh YG, Gao Z, Yang BBW, Oh G, Gong Y. Navigating text-to-image customization: from LyCORIS fine-tuning to model evaluation. In: *The twelfth international conference on learning representations*; 2023.

225. Sawhney R, Ferrell BD, Dejean T, Schreiber ZD, Harrigan W, Polson SW, et al. *Fine-tuning Protein Language Models Unlocks the Potential of Underrepresented Viral Proteomes*. bioRxiv; 2025. https://doi.org/10.1101/2025.04.17.649224. 2025.04.17.649224.
226. AlQuraishi M. Machine learning in protein structure prediction. *Curr Opin Chem Biol* 2021;**65**:1–8.
227. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. *Protein complex prediction with AlphaFold-Multimer. bioRxiv;* 2021. https://doi.org/10.1101/2021.10.04.463034. bioRxiv:2021.10.04.463034.
228. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500.
229. Hu B, Xia J, Zheng J, Tan C, Huang Y, Xu Y, et al. *Protein Language Models and Structure Prediction: Connection and Progression*, 2022, arXiv, https://doi.org/10.48550/arXiv.2211.16742, arXiv:2211.16742.
230. Wang W, Peng Z, Yang J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nat Comput Sci* 2022;**2**:804–14.
231. Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat Mach Intell* 2023;**5**:1087–96.
232. Jing X, Wu F, Luo X, Xu J. Single-sequence protein structure prediction by integrating protein language models. *Proc Natl Acad Sci U S A* 2024;**121**:e2308788121.
233. Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 2022;**30**:1169–77. e4.
234. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
235. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods* 2014;**11**:801–7.
236. Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat Commun* 2023;**14**:2389.
237. Wang Y, Gong X, Li S, Yang B, Sun Y, Shi C, et al. *xTrimoABFold: De novo Antibody Structure Prediction without MSA*, 2022, arXiv, https://doi.org/10.48550/arXiv.2212.00735. arXiv.2212.00735.
238. Jing H, Gao Z, Xu S, Shen T, Peng Z, He S, et al. Accurate prediction of antibody function and structure using bio-inspired antibody language model. *Brief Bioinform* 2024;**25**:bbae245.
239. Gasser HC, Oyarzún DA, Rajan A, Alfaro JA. Guiding a language-model based protein design method towards MHC Class-I immune-visibility targets in vaccines and therapeutics. *ImmunoInformatics* 2024;**14**:100035.
240. Verkuil R, Kabeli O, Du Y, Wicky BIM, Milles LF, Dauparas J, et al. *Language Models Generalize Beyond Natural Proteins*. bioRxiv; 2022. https://doi.org/10.1101/2022.12.21.521521. bioRxiv:2022.12.21.521521.
241. Lee J, Jun DW, Song I, Kim Y. DLM-DTI: a dual language model for the prediction of drug-target interaction with hint-based learning. *J Cheminform* 2024;**16**:14.
242. Wei J, Zhuo L, Fu X, Zeng X, Wang L, Zou Q, et al. DrugReAlign: a multisource prompt framework for drug repurposing based on large language models. *BMC Biol* 2024;**22**:226.
243. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 2020;**36**:4406–14.
244. Chen L, Fan Z, Chang J, Yang R, Hou H, Guo H, et al. Sequence-based drug design as a concept in computational drug design. *Nat Commun* 2023;**14**:4217.
245. S. Chithrananda, G. Grand and B. Ramsundar, ChemBERTa: Large-scale Self-supervised Pretraining for Molecular Property Prediction, 2020, arXiv, doi: 10.48550/arXiv.2010.09885, arXiv:2010.09885.
246. Zhang S, Xie L. Protein language model-powered 3D ligand binding site prediction from protein sequence. In: *NeurIPS 2023 AI for science workshop*; 2023.
247. Hosseini S, Golding GB, Ilie L. Seq-InSite: sequence supersedes structure for protein interaction site prediction. *Bioinformatics* 2024;**40**:btad738.
248. Seo S, Choi J, Choi S, Lee J, Park C, Park S. Pseq2Sites: enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Eng Appl Artif Intell* 2024;**127**:107257.
249. Vural O, Jololian L. Machine learning approaches for predicting protein-ligand binding sites from sequence data. *Front Bioinform* 2025;**5**:1520382.
250. Škrhák V, Novotný M, Riedlova K, Hoksza D. Cryptic binding site prediction with protein language models. In: *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*; 2023. p. 2883–8.
251. Nerín-Fonz F, Cournia Z. Machine learning approaches in predicting allosteric sites. *Curr Opin Struct Biol* 2024;**85**:102774.
252. Flamholz ZN, Biller SJ, Kelly L. Large language models improve annotation of prokaryotic viral proteins. *Nat Microbiol* 2024;**9**:537–49.
253. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* 2022;**40**:1023–5.
254. Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the eighteenth international conference on machine learning*. ICML '01; 2001. p. 282–9.
255. Zeng S, Wang D, Jiang L, Xu D. Parameter-efficient fine-tuning on large protein language models improves signal peptide prediction. *Genome Res* 2024;**34**:1445–54.
256. Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res* 2019;**48**:D269–76.
257. Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* 2016;**45**:D219–27.
258. Kim S, Mollaei P, Antony A, Magar R, Barati Farimani A. GPCR-BERT: interpreting sequential design of G protein-coupled receptors using protein language models. *J Chem Inf Model* 2024;**64**:1134–44.
259. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* 2003;**31**:294–7.
260. Lee Y, Lee H, Shin K, Kwon S. GRIP: graph representation of immune repertoire using graph neural network and transformer. *Proc Conf AAAI Artif Intell* 2023;**37**:5160–8.
261. Zaslavsky ME, Craig E, Michuda JK, Sehgal N, Ram-Mohan N, Lee JY, et al. Disease diagnostics using machine learning of B cell and T cell receptor sequences. *Science* 2025;**387**:eadp2407.
262. Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* 2020;**36**:3594–6.
263. Zhao Y, He B, Xu F, Li C, Xu Z, Su X, et al. DeepAIR: a deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Sci Adv* 2023;**9**:eabo5128.
264. Dee W. LMPred: predicting antimicrobial peptides using pre-trained language models and deep learning. *Bioinform Adv* 2022;**2**:vbac021.
265. Han J, Kong T, Liu J. PepNet: an interpretable neural network for anti-inflammatory and antimicrobial peptides prediction using a pre-trained protein language model. *Commun Biol* 2024;**7**:1198.
266. Thumuluri V, Martiny HM, Armenteros JJA, Salomon J, Nielsen H, Johansen AR. NetSolP: predicting protein solubility in E. coli using language models. *Bioinformatics* 2021;**38**:941–6.
267. Chen T, Gong C, Diaz DJ, Chen X, Wells JT, Liu Q, et al. HotProtein: a novel framework for protein thermostability prediction and editing. In: *The eleventh international conference on learning representations*; 2022.
268. Basith S, Pham NT, Manavalan B, Lee G. SEP-AlgPro: an efficient allergen prediction tool utilizing traditional machine learning and deep learning techniques with protein language model features. *Int J Biol Macromol* 2024;**273**(Pt 2):133085.
269. Hu X, Li J, Liu T. Alg-MFDL: a multi-feature deep learning framework for allergenic proteins prediction. *Anal Biochem* 2025;**697**:115701.
270. Zhang L, Liu T. PreAlgPro: prediction of allergenic proteins with pre-trained protein language model and efficient neutral network. *Int J Biol Macromol* 2024;**280**(Pt 3):135762.
271. Jha K, Karmakar S, Saha S. Graph-BERT and language model-based framework for protein-protein interaction identification. *Sci Rep* 2023;**13**:5663.
272. Lupo U, Sgarbossa D, Bitbol AF. Pairing interacting protein sequences using masked language modeling. *Proc Natl Acad Sci U S A* 2024;**121**:e2311887121.