Journal of proteome research

pubs.acs.org/jpr

Review

Protein Language Models: Applications and Perspectives

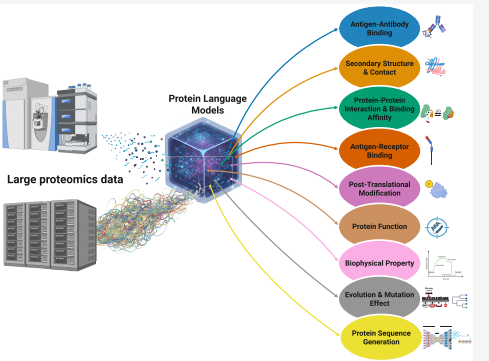
Mickael Leclercq and Arnaud Droit*

Cite This: *J. Proteome Res.* 2026, 25, 507−524

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Large language models (LLMs) originally developed for human text have been adapted to proteomics as protein language models (pLMs). These models treat amino acid sequences like sentences, and they learn patterns from millions of sequences. pLMs are used for several key tasks, including the prediction of protein structures, annotating protein functions, designing novel protein sequences with specific characteristics, and mapping the interactions between proteins and other molecules. Compared with traditional approaches, pLMs deliver insights more quickly but demand large computing resources and careful data management. Developers are focused on decreasing prediction inaccuracies and biases by exploring more efficient training techniques and smaller models to decrease the resources required. As sequence databases continue to grow, pLMs will improve to uncover links between proteins and disease pathways, speeding drug development and basic research while offering new proteome-scale insights that support experimental design and validation.



KEYWORDS: *protein language models (pLMs), transformer architectures, sequence embeddings, protein structure prediction, protein function annotation, de novo protein sequence generation, protein−protein interaction modeling, post-translational modification prediction, biophysical property prediction, computational scalability and efficiency*

INTRODUCTION

Proteomics is the large-scale study of proteins, including their sequences, structures, functions, and interactions, and plays a central role in modern biology. Since proteins drive cellular processes, their analysis is key to improving diagnostics, guiding drug development, and understanding biological mechanisms.[1] Despite its significance, proteomics faces substantial challenges, such as the inherent complexity of protein data, persistent batch effects, difficulty of integrating heterogeneous data across platforms and studies,[2] as well as the growing need to account for diverse protein forms arising from post-translational modifications and isoforms.[3] Traditional computational approaches for structure prediction and function analysis, often reliant on sequence alignment or homology modeling, frequently struggle to generalize across diverse protein families, accentuating the need for innovative methodologies.[4−7]

In this context, large language models (LLMs) have emerged as potentially transformative tools for proteomics.[8] By conceptualizing protein sequences as a "language", with amino acids analogous to words, LLMs use deep learning architectures to discern intricate patterns and relationships within extensive protein sequence data sets. These models, referred to as protein language models (pLMs), demonstrate strong potential in protein analysis.[8] Research highlights their growing adoption in proteomics for applications such as protein structure prediction, function annotation, and drug discovery.[9,10] Moreover, pLMs have shown the capability to generate functional protein sequences, which could significantly advance protein engineering efforts.[11]

Despite their promise, integration of LLMs into proteomics is not without challenges. High computational demands, adapting to protein-specific tasks, and balancing accuracy with generalizability represent notable hurdles that must be addressed to fully realize their potential.[12]

This review examines the applications of pLMs in key areas including protein structure prediction, function prediction, de novo sequence generation, design and optimization, post-translational modification prediction, evolutionary and mutation prediction, biophysical property prediction, protein−protein interactions, and other molecular interactions (e.g., protein−ligand, protein-nucleic acid), antigen−receptor binding, and antigen−antibody binding. The review evaluates the significant advantages of pLMs over conventional approaches, such as handling long-range dependencies, generalizability, scalability, and flexibility. It also critically assesses their limitations, including interpretability challenges, data bias, tokenization issues, generalizability across diverse protein families, computational demands, and fine-tuning complexities.

Received: May 28, 2025
Revised: October 29, 2025
Accepted: November 25, 2025
Published: December 26, 2025

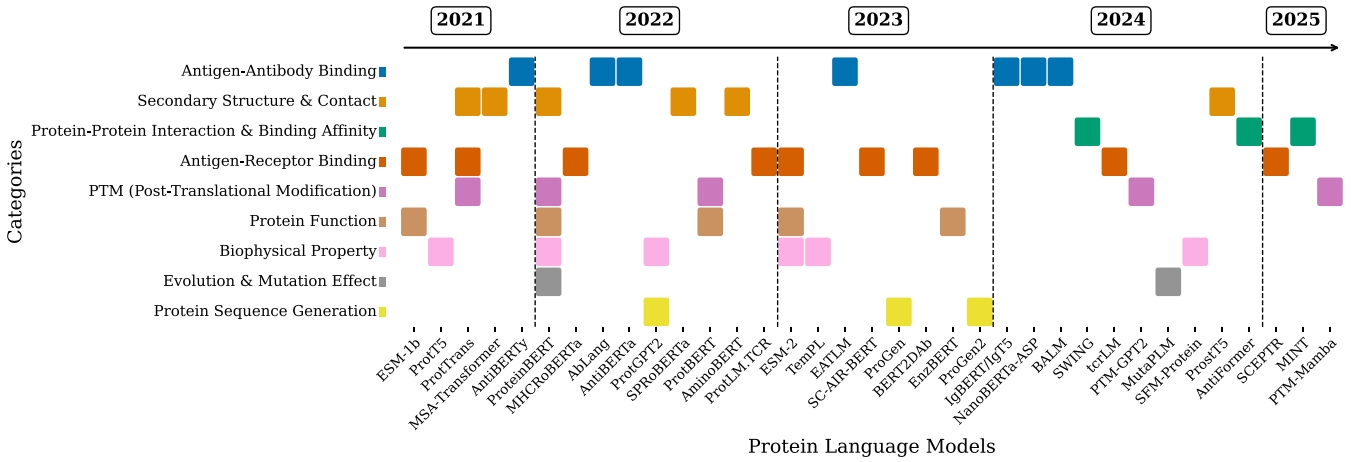


Figure 1. Timeline of standalone pLMs arranged by release date and color-coded by primary prediction category. Only standalone pLMs are shown; derivative methods that build on embeddings from existing pLMs are excluded.

Furthermore, it explores future research directions, highlighting the potential of LLMs to revolutionize drug development, personalized medicine, and fundamental biology. This article synthesizes recent developments and offers perspectives to help researchers apply LLMs effectively in advancing proteomic research.

LARGE LANGUAGE MODELS AND ADAPTATION TO PROTEINS

LLMs are advanced artificial intelligence systems designed to comprehend and generate human-like text by training on very large data sets.[13] These models predominantly rely on transformer architectures, which employ self-attention mechanisms to effectively capture contextual relationships within sequential data.[14] Transformer architecture is a type of neural network designed to handle sequences of words by focusing on how each word relates to every other word. Instead of reading text one token (i.e., unit of text: word, subword, or character) at a time, it applies a self-attention mechanism that assigns a weight to each word pair, allowing the model to capture both local and long-range connections in a single step. Key components include multihead attention layers, which let the network learn different types of relationships in parallel, and feed-forward layers, which transform those insights into richer representations. Positional encodings are added to give the model information about word order. Transformers process all tokens simultaneously using attention instead of recurrence, enabling faster training and strong performance on core natural language processing tasks.[14] The power of this design has made transformers the backbone of modern LLMs across text generation, translation, and classification.[15−17]

LLMs are typically trained on large general data sets, then adapted to specific domains using targeted methods. For example, SciBERT and BioBERT are trained on scientific and biomedical texts to improve accuracy on related tasks.[18,19] Parameter-efficient methods, such as adapters or low-rank adaptations, reduce compute costs by updating only a small subset of model parameters.[20,21] Prompt-based tuning can also adjust model behavior using a few example prompts, avoiding full model updates, and retrieval-augmented generation can link the model to external knowledge bases so outputs reflect up-to-date information.[22]

Originally designed for language tasks, LLMs have been adapted to proteomics by treating protein sequences as text with amino acids as tokens. This allows transformer models to scale the proteomic analysis. The adaptation of LLMs to proteomics begins with tokenization, where each amino acid in a protein sequence (e.g., "MVLSPADKT") is assigned a unique token ("MVLSPADKT" is tokenized into ["M", "V", "L", "S", "P", "A", "D", "K", "T"]), similar to words in a sentence. Amino acid tokens are converted to numerical vectors that encode sequence contexts and biochemically relevant patterns; vectors for residues with similar roles tend to be close to one another in this representation space. These embeddings are then processed by transformer layers with self-attention, allowing the model to capture long-range dependencies and to encode biochemical, evolutionary, and structural signals.[8] Training typically involves self-supervised learning on extensive proteomics data sets, such as UniRef[23−25] or The Big Fantastic Database,[26] which collectively include millions of protein sequences. But unlike text, protein sequences share ancestry, and random splits during training can inflate scores by placing close homologues in both train and test.[27,28] Thus, current practice is to use homology-aware evaluation, where sequences are clustered by identity and coverage (e.g., UniRef, MMseqs2, CD-HIT) and entire clusters are assigned to train, validation, and test.[24,29,30] Some benchmarks also include a time split in which test families postdate the pretraining freeze.[31] Common training objectives include masked-language modeling (predicting masked amino acids from context) and next amino acid prediction, similar to the approaches used in Bidirectional Encoder Representations from Transformers (BERT) and Generative Pretrained Transformer (GPT).[32] After pretraining, LLMs are fine-tuned on smaller labeled data sets for tasks like structure prediction, function classification, and interaction modeling. The resulting contextual embeddings, i.e., high-dimensional numerical vectors that encode functional signals, capture functional patterns and offer a stronger basis for interpreting protein sequences than traditional methods.[33] These capabilities have also enabled functional sequence generation,[11] supporting advances in protein engineering and drug discovery.

APPLICATIONS OF PROTEIN LANGUAGE MODELS

This section reviews how protein language models (pLMs) are used in proteomics. It covers de novo protein design, function

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

Table 1. Summary of pLMs[a]

pLM	Standalone pLM	Antigen-Antibody Binding	Secondary Structure & Contact	Protein-Protein Interaction & Binding Affinity	Antigen-Receptor Binding	PTM (Post-Translational Modifications)	Protein Function	Biophysical Property	Evolution & Mutation Effect	Protein Sequence Generation	Description	Training Data
ESM-1b [8]	X				X		X				General-purpose embedding model for structure, function, and interaction prediction	250 M UniParc sequences (86 B amino acids)
ProtT5 [34]	X							X			Sequence-to-sequence Transformer pretrained for protein sequences and structure for multiple tasks	UniRef and BFD containing 393 B amino acids
ProtTrans [35]	X		X		X	X					Suite of pretrained transformer models (BERT, T5, XLNet, etc.) on large-scale protein corpora using self-supervised objectives	UniRef and BFD containing 393 B amino acids
MSA-Transformer [36]	X		X								Transformer taking MSAs as input for unsupervised contact and structure prediction	Multiple sequence alignments of related protein families
NetSolP [37]								X			Predictor for E. coli expression solubility and usability using pLM embeddings from ESM variants	Curated solubility and usability datasets with strict sequence partitioning
AntiBERTy [38]	X	X									Antibody-specific masked language model fine-tuned for binding specificity	558 M natural antibody sequences
ProteinBERT [39]	X		X			X	X	X	X		Joint masked language and Gene Ontology annotation pretraining for versatile protein structure, function, other prediction tasks	~106 million UniRef90 sequences and Gene Ontology
MHCRoBERTa [40]	X				X						RoBERTa-based model for peptide–MHC class I binding prediction via transfer learning	Public peptide–MHC binding datasets (IC50 and SRCC benchmarks)
AbLang [41]	X	X									Antibody-specific model to restore missing residues in antibody sequences	Large antibody repertoires (OAS)
AntiBERTa [42]	X	X									Antibody-specific language model trained on large repertoires for engineering tasks	Hundreds of millions of antibody sequences
ProtGPT2 [43]	X							X		X	Autoregressive Transformer generating de novo protein sequences with natural-like amino acid and disorder profiles	~50 M sequences from UniRef50 for autoregressive training
SPRoBERTa [44]	X		X								Unsupervised protein tokenizer learns local fragment patterns for transformer-based representations	Unlabeled protein sequences from UniProt for fragment vocabulary learning
RGN2 [45]			X								Hybrid network using AminoBERT embeddings and a recurrent geometric module for single-sequence 3D structure prediction	AminoBERT pretrained on millions of unaligned proteins; trained on PDB structures for de novo prediction
ProtBERT [46]	X					X	X				Masked-language trasnformer model for protein sequences in the ProtTrans family	2.1 B sequences from the Big Fantastic Database (BFD)
AminoBERT [45]	X		X								Protein language model for single-sequence structure prediction without MSAs	Millions of unaligned protein sequences

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

Table 1. continued

pLM	Standalone pLM	Antigen-Antibody Binding	Secondary Structure & Contact	Protein-Protein Interaction & Binding Affinity	Antigen-Receptor Binding	PTM (Post-Translational Modifications)	Protein Function	Biophysical Property	Evolution & Mutation Effect	Protein Sequence Generation	Description	Training Data
ProtLM.TCR [47]	X				X						Language model pretrained on T-cell receptor sequences for epitope binding prediction	~62 M TCR sequences from public repertoires
ESM-2 [48]	X				X		X	X			Next-generation embedding model for function and biophysical predictions	614 M protein sequences (UniRef)
TemPL [49]	X							X			Temperature-guided LM for zero-shot prediction of protein stability and activity without labels	Sequence and temperature annotations from ProTherm and enzyme assays
EATLM [50]	X	X									Evolution-aware antibody language model capturing germline-ancestral relationships and hypermutation using ATUE benchmark	Antibody sequences and ATUE benchmark tasks
SC-AIR-BERT [51]	X				X						Pretrained on paired single-cell TCR/BCR chain sequences, fine-tuned for antigen-binding specificity	Paired AIR chains from large single-cell immune repertoire datasets
ProGen [11]	X									X	Conditional transformer-based pLM for controlled artificial protein sequence generation	280 M sequences from UniParc, UniProtKB, Pfam, NCBI taxonomy
TAPIR [52]		X									CNN-based TCR language model encodes TCR and peptide sequences to predict common and novel target interactions	> 50 k paired/unpaired TCRs from VDJdb and Vcreate against 1,933 targets
BERT2DAb [53]	X				X						Pre-trained model incorporating antibody sequence and 2D structural annotations	Antibody sequences + structural annotations
EnzBERT [54]	X						X				Transformer models specialized for EC number prediction, boosting EC level-2 accuracy from 84% to 95%	EC40 benchmark and time-based level-4 test sets
DG-Affinity [55]		X		X							Sequence-based antigen–antibody affinity predictor using pLM embeddings and ConvNeXt	1,673 antibody–antigen pairs from sdAb-DB and Baidu PaddlePaddle competition
ProGen2 [56]	X									X	Scaled-up ProGen autoregressive model for sequence design and fitness prediction	> 1 B proteins from genomic, metagenomic, and immune repertoire databases
TCR-ESM [57]					X						Employs ESM1v embeddings of TCR, peptide, and MHC sequences via feedforward network for pMHC binding prediction	CDR3α+β and MHC embeddings from ESM1v with binding labels from netTCR and external datasets
HybridGCN [58]								X			Hybrid GCN combining deep ESM-1v embeddings and classical biophysical features with adaptive re-weighting for solubility	eSOL and S. cerevisiae solubility datasets
ProSTAGE [59]									X		Graph convolutional and pLM embedding integration for predicting stability changes upon single-point mutations	11,304 mutations across 318 proteins from curated datasets

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

Table 1. continued

pLM	Standalone pLM	Antigen-Antibody Binding	Secondary Structure & Contact	Protein-Protein Interaction & Binding Affinity	Antigen-Receptor Binding	PTM (Post-Translational Modifications)	Protein Function	Biophysical Property	Evolution & Mutation Effect	Protein Sequence Generation	Description	Training Data
DeepPTM [46]						X					ProtBERT-based predictor for PTM sites using embeddings and vision transformers	~1.2 M annotated PTM sites across species
ESMFold [48]			X								Structure predictor built on ESM-2	PDB structures and UR50 sequences
DiffPALM [60]				X							Method for pairing interacting protein sequences leveraging MSA Transformer to capture inter-chain coevolution	Shallow MSAs of prokaryotic protein families
IgBERT/IgT5 [61]	X	X									Antibody-specific models handling paired and unpaired variable regions for design and affinity prediction	Over 2 B unpaired and 2 M paired sequences from Observed Antibody Space
NanoBERTa-ASP [62]	X	X									Pretrained model for predicting nanobody paratope residues in CDR and non-CDR regions	Annotated paratope positions from crystal structures (PDB entries)
RXNAAMapper [63]							X				Unsupervised binding-site predictor using language modeling of reaction SMILES and protein sequences	~7 M biochemical and organic reactions from USPTO and ECREACT combined with UniProt sequences
BALM [64]	X	X									Bio-inspired antibody model trained to predict paratope and kinetics	300 M+ antibody sequence corpus
SWING [65]	X			X							Sliding Window INteraction Grammar generates an interaction vocabulary from amino acid property differences for peptide–protein interactions	pMHC binding assays from IEDB and curated interaction datasets
tcrLM [66]	X				X						Lightweight masked LM with virtual adversarial training pretrained for binding specificity prediction	> 100 M distinct TCR CDR3 sequences from public repertoires
ESM-therm [67]								X			Fine-tuned ESM-2 model for folding stability prediction on mega-scale dataset	528 k sequences from 461 protein domains with stability measurements
FEDKEA [68]							X				Integrates ESM-2 embeddings and distance-weighted KNN for hierarchical EC annotation	Enzyme-labeled sequences across four EC levels
PTM-GPT2 [69]	X					X					Interpretable GPT-2–based model using prompt-based fine-tuning to predict 19 types of post-translational modifications	Swiss-Prot PTM annotations from UniProt sequences used to fine-tune PROTGPT2 via custom prompts
MutaPLM [70]	X								X		Framework for explaining and engineering mutations with a protein delta network and chain-of-thought supervision	MutaDescribe dataset: large-scale mutation set with textual annotations and PLM features
SFM-Protein [71]	X							X			Co-evolutionary pretraining strategy emphasizing residue interactions for solubility	Large-scale protein sequence dataset optimized for co-evolutionary feature learning
MAMMAL [72]		X									Multi-modal transformer integrating small molecule, protein, and transcriptomic data for drug discovery	Small-molecule, protein, and transcriptomic datasets

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

Table 1. continued

pLM	Standalone pLM	Antigen-Antibody Binding	Secondary Structure & Contact	Protein-Protein Interaction & Binding Affinity	Antigen-Receptor Binding	PTM (Post-Translational Modifications)	Protein Function	Biophysical Property	Evolution & Mutation Effect	Protein Sequence Generation	Description	Training Data
ParaAntiProt [73]					X						Paratope predictor combining general and antibody-specific pLM embeddings with convolution and CDR encodings	Paratope datasets with 10-fold cross-validation splits
ProstT5 [74]	X		X								Bilingual pLM translating between protein sequences and structures	17 M proteins with high-quality 3D predictions from AlphaFoldDB
AntiFormer [75]	X			X							Graph-enhanced transformer for antibody binding affinity prediction	PPI networks + structural graphs
SCEPTR [76]	X				X						Contrastive TCR embedding model using autocontrastive pretraining for data-efficient transfer learning	Unlabeled TCR sequences from public repertoires for contrastive pretraining
UniPMT [77]					X						Unified multi-task GNN framework for peptide–MHC–TCR binding prediction across P-M-T, P-M, and P-T tasks	Curated P-M-T, P-M, and P-T interaction datasets from IEDB and VDJdb
MINT [78]	X			X							pLM for protein–protein interaction modeling capturing interaction patterns via language modeling	Datasets of experimentally verified PPIs and coevolutionary MSAs
PTM-Mamba [79]	X					X					PTM-aware LM integrating bidirectional Mamba blocks with ESM-2 embeddings for wild-type and PTM sequences	79 707 modified sequences from 311 350 Swiss-Prot PTM records
ProBASS [80]				X							Sequence and structure-based model predicting ΔΔGbind with high correlation on single and double mutations	12 M predicted structures from AlphaFold for mutation ΔΔGbind training

[a]This table lists all protein language models covered in this review, showing for each whether it was trained directly on proteomics data ("standalone pLM") or trained from existing pLM embeddings, a description of design or purpose, and the main annotation dataset used for training. A table with more information (Performance Highlights and release date) is provided in Supporting Information data Table S1.

prediction (enzymatic roles, binding sites), structure and contact prediction (secondary structure and backbone modeling), PTM site inference, evolutionary and mutation effect forecasting, biophysical property estimation (stability, solubility, aggregation), and protein−protein interaction and binding affinity prediction. Finally, it examines antigen−receptor and antibody−antigen-binding models for immunology. For each application, key pLMs are noted, compared with traditional methods, and their benefits and limitations are discussed. Several non-pLM models leveraging pLM embeddings are described to demonstrate the widespread application of pLMs in other learning architectures. The presented pLMs are summarized in Figure 1 and Table 1.

Protein Sequence Generation

Protein sequence generation aims to create new sequences with the desired functional or structural properties. Applications include designing therapeutic enzymes, antibodies for immunotherapy, and biocatalysts for industrial use.[81] Traditional methods like directed evolution or rational design are often slow and expensive, relying on trial-and-error or limited structural knowledge to guide changes.[82−84] Directed evolution mimics natural selection through repeated mutation and screening, which is a long and resource-intensive process. Rational design uses structural insights to guide changes, but a limited understanding of folding and function often leads to suboptimal outcomes.[85,86] In contrast, pLMs offer a scalable and efficient alternative to generate sequences tailored to desired attributes from vast data sets,[33] with models like ProGen showing their use in generating functional protein sequences[11] using model ability to interpret and extend sequence data through a structured workflow.[87] ProtGPT2, an autoregressive transformer model (738 M parameters) trained on ∼50 million UniRef50 protein sequences, is able to generate de novo protein sequences that recapitulate natural amino acid and disorder propensities while sampling unexplored regions of sequence space.[43] These models are trained via tokenization and self-supervised learning on very large data sets like UniRef or UniProt, which contain millions of sequences, with objectives to predict the next amino acid or

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

reconstruct masked sequence segments to learn the natural distribution of protein sequences.[11] To target specific properties, pLMs use conditional generation, where control tags (e.g., protein family or function) guide the output, such as prompting a language model to write in a chosen style. During inference, the model iteratively predicts subsequent tokens from an initial prompt, ensuring alignment with the specified attributes, as exemplified by ProGen's ability to produce sequences with lysozyme activity.[11] ProGen, trained on 280 million sequences from over 19,000 protein families, uses control tags to enable precise generation, producing artificial lysozymes with catalytic efficiencies comparable to natural ones, despite sequence identities as low as 31.4%. Its successor, ProGen2 has been scaled up to 6.4 billion parameters and contains over a billion sequences from genomic, metagenomic, and immune repertoire databases, achieving state-of-the-art performance in generating novel viable sequences and predicting protein fitness without additional fine-tuning.[56] ProGen-generated lysozymes were synthesized and tested,[11] the practical use of pLM-generated sequences in medical and industrial applications.

Protein Function Prediction

pLMs predict protein function by modeling sequences as language, inferring roles such as enzyme activity or binding sites from sequence alone. Unlike similarity-based tools such as BLAST, they capture deeper context and evolutionary signals, aiding annotation of novel or uncharacterized proteins. ESM-2,[48] trained on 614 million protein sequences, uses self-supervised learning to generate embeddings that are fine-tuned or fed into downstream classifiers to predict specific functions. ESM-2 embeddings have been shown to accurately predict Gene Ontology terms, such as enzymatic activity, outperforming BLAST by using evolutionary context (e.g., covariation and conserved-motif signals captured during large-scale unsupervised training) rather than direct homology.[88]

A key application is predicting enzymatic roles, where pLMs identify sequence motifs associated with the catalytic activity. Models such as ESM1b,[8] ESM2,[48] and ProtBERT[35] outperform traditional alignment methods in assigning EC numbers by capturing subtle sequence signals beyond homology-based approaches.[89] EnzBERT refines these predictions with a dedicated attention layer that assigns full EC codes to input sequences with high precision.[54] FEDKEA combines a fine-tuned ESM-2 backbone with a distance-weighted k-nearest neighbor classifier, achieving state-of-the-art performance in identifying catalytic proteins and classifying enzymes across all EC levels.[68]

Recent advancements highlight pLMs' ability to go beyond sequence data. By integrating genomic context with sequence information into a genomic Language Model, it is possible to predict enzymatic functions with up to 24.4% accuracy from context alone, demonstrating that surrounding genes can inform function.[90] This is particularly valuable for uncharacterized proteins in metagenomic data sets. Compared to traditional methods like BLAST, pLMs are faster, i.e., processing sequences in seconds rather than minutes, and handle remote homologues better, at the sacrifice of interpretability.[32,91] Another example is ProteinBERT, pretrained on ∼ 106 million UniRef90 sequences via joint masked-language modeling and multilabel Gene Ontology annotation prediction, incorporating both local and global representations to handle long proteins efficiently.[39] Its compact architecture yields embeddings that enable rapid fine-tuning across a wide range of downstream protein tasks, such as function prediction, with performance matching or exceeding larger models despite fewer parameters.

Similarly, pLMs excel at pinpointing binding sites, i.e., regions where proteins interact with ligands or other molecules, consequently elucidating their functional roles. RXNAAMapper, a model using biochemical reaction data, identifies binding sites with over 52% accuracy in unsupervised settings, surpassing other sequence-based methods.[63]

Despite their strengths, pLM-based function prediction performance depends on training data diversity, and predictions for out-of-distribution proteins can be unreliable.[33] Interpretability is limited, unlike BLAST's transparent alignment scores. Still, pLMs can annotate proteins with <30% identity to known sequences, offering an edge over homology-based methods.[92]

Secondary Structure and Contact Prediction

Secondary structure prediction, which identifies local folding patterns like helices, sheets, and loops, and contact prediction, which detects nearby residue pairs, are essential for understanding a protein's 3D structure and function. These predictions support drug design and the study of biological processes. Traditionally, they relied on sequence alignments or machine learning models trained on limited data, often using sequence features or structural inputs. Such methods struggle with scalability and accuracy, especially for novel proteins. pLMs improve this by using large data sets to capture complex patterns and context.[33]

In structure benchmarks, ProteinBERT achieves high accuracy on secondary structure assignments and contact-map inference, rivaling heavier models.[39] The Multiple Sequence Alignment (MSA) Transformer uses multiple sequence alignments to predict contact maps with state-of-the-art accuracy, incorporating evolutionary information to boost tertiary structure prediction.[93] ProtTrans, a family of models pretrained on large protein sequence corpora like UniProt, excels when fine-tuned for secondary structure prediction, achieving competitive accuracy against specialized tools.[94] SPRoBERTa adopts a local fragment-based approach, improving secondary structure prediction and offering insights into sequence-structure relationships, further demonstrating pLMs' versatility.[44] ProstT5 jointly learns from amino acid sequences and structural context encoded as 3D tokens. By training it to map sequences to these structure tokens (and back again), it builds a tighter link between sequence and fold, which improves tasks like fold classification.[74] These models improve prediction accuracy and enable the design of novel proteins with specific structures by capturing biochemical and evolutionary signals.[8] They produce context-aware embeddings for each residue, capturing structural and evolutionary signals. Secondary structure can be predicted with simple projection layers, while attention maps reveal residue−residue contacts by highlighting spatial interactions. Recent analyses show that pLMs can recover coevolutionary signals typically derived from MSA-based approaches such as direct coupling analysis and evolutionary couplings, yet do so without explicit alignments.[95−97] Alternatively, end-to-end models like RGN2 combine pLM embeddings from its AminoBERT module with a recurrent geometric network to predict backbone coordinates.[45] AminoBERT, pretrained on millions of

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

unaligned sequences, captures both local and global sequence patterns. It produces per-residue embeddings encoding structural signals, which the geometric network uses, via Frenet−Serret frames, to construct the Cα backbone with translational and rotational invariance. In this approach, secondary structures and contacts emerge from the predicted geometry rather than as explicit steps.

These techniques contrast sharply with Molecular Dynamics (MD), which simulates protein folding based on physical force fields.[98] While MD simulations offer detailed dynamic and energetic insights, de novo structure prediction via simulation is far more computationally intensive.[9,99] PLMs learn statistical links between sequence and structure, enabling fast and accurate predictions, particularly when using evolutionary data or advanced single-sequence models like RGN2 or ESMFold,[48] with less direct physical interpretability compared to MD.

Post-Translational Modifications Prediction

Post-translational modifications (PTMs) are chemical alterations to proteins after their synthesis, such as phosphorylation, acetylation, ubiquitination, and glycosylation, which play critical roles in regulating protein biochemical activity/functionality, localization, and stability.[100,101] These modifications are essential for cellular processes like signal transduction, protein degradation, and disease mechanisms.[102] Understanding PTMs is crucial for advancing drug discovery, personalized medicine, and disease research. Traditionally, PTM identification relied on experimental techniques like mass spectrometry or computational methods using sequence motifs and machine learning with handcrafted features.[100,103] However, prediction models perform worse when a PTM lacks a clear sequence motif or is strongly context-dependent, and by definition cannot exceed the accuracy of experimental methods such as mass spectrometry.[102] Practically, sequence-based PTM predictors estimate, for each residue and PTM type, the probability that a site is competent to carry a given modification from sequence alone;[8] this reflects site competence rather than context-dependent occupancy, which varies with cell type and cellular state.[104,105]

PTM prediction with pLMs typically uses models like ESM or ProtTrans, trained on large protein data sets (e.g., UniRef50, UniProt).[8,35] These generate residue-level embeddings, which feed into downstream models (e.g., neural networks and gradient boosting) to predict PTMs. Alternatively, pLMs can be fine-tuned directly for PTM tasks, as in DeepPTM.[46] DeepPTM used ProtBERT-based protein embeddings with attention-based vision transformers to predict ubiquitination, succinylation, crotonylation, and glycation sites with ROC AUCs of 0.776, 0.793, 0.764, and 0.734, respectively, in humans.[46] Prior to DeepPTM, reported ROC AUCs were 0.91 for ubiquitination,[106] ∼0.80−0.82 for succinylation,[107,108] ∼0.86−0.91 for crotonylation,[109,110] and ∼0.69 on a large human glycation set.[111] Similarly, ProtTrans embeddings improve ubiquitination site prediction, achieving a precision-recall AUC of 0.88.[35] Using pLM embeddings, it is also possible to improve PTM site prediction; for example, adding ProtT5 features boosted lysine glutarylation prediction, increasing recall and AUC over earlier methods.[112] Additionally, PTM-GPT2,[69] created by prompt-based fine-tuning of a protein language model, was trained on a compendium of ∼1.2 million experimentally annotated PTM sites covering 19 modification types. Benchmarks also demonstrate ProteinBERT's ability to predict PTM sites (e.g., phosphorylation, glycosylation) with performance comparable to specialized predictors.[39] Finally, PTM-Mamba extends a protein Transformer model by introducing special tokens representing PTMs and uses a gating mechanism to fuse these PTM tokens with sequence embeddings.[79]

These pLMs outperform traditional methods by capturing long-range dependencies and subtle patterns.[69] Their broad pretraining supports adaptation to various PTM types with little extra training. However, sequence-only models often fall short for PTMs that depend on structural or cellular context, like O-glycosylation.[113,114]

Evolution and Mutation Prediction

Prediction of protein evolutionary trajectories and mutation effects refers to estimating how amino acid substitutions influence protein stability, function, and fitness across different contexts and how such substitutions may accumulate over time within evolutionary lineages. This task underlies efforts to model disease variants and guide protein engineering. Traditional computational methods, including phylogenetic and multiple sequence alignments (MSAs) tools,[115−117] and mutation effect predictors relying on sequence conservation and structural features.[118,119] These have provided valuable frameworks but face limits related to running on very large data sets and long protein sequences without prohibitive compute or memory, dependence on high-quality MSAs or structures, and potential issues with accuracy, bias, and data circularity (i.e., evaluating a model on data that overlap with its training set).

pLMs like ESM variants, ProteinBERT, and others[39,120,121] propose a new paradigm to analyze proteomics data. Using Transformer-based architectures trained on large unlabeled data sets, they learn representations that capture evolutionary, structural, and functional context.[25] They often outperform traditional methods in predicting mutation effects, frequently in a zero-shot setting (i.e., without any task-specific examples seen during training), achieving strong performance without task-specific training data, and reducing reliance on labeled data sets and avoiding circularity issues.[11,122,123] They offer potential advantages in scalability and can operate independently of MSAs.[11,124] Specialized pLMs fine-tuned for tasks like stability prediction (e.g.,[125] ProSTAGE[59]) or enhanced with structural or textual inputs (e.g., ProtSSN,[126] MutapLM[70]) show improved performance for site- or function-level classification, better agreement with experimentally measured changes in protein stability, stronger generalization to unseen folds, and reduced labeled data requirements. They are also being used for phylogenetic inference[36] and in silico directed evolution.

Biophysical Properties Prediction

The prediction of protein biophysical properties, such as solubility, stability, aggregation propensity, and secondary structure, is fundamental to understanding protein function, the mechanism of disease, and guiding protein engineering and therapeutic development.[127] Previously dependent on experiments or tools using handcrafted features, pLMs offer new approaches to predict a range of biophysical properties directly from amino acid sequences.[45]

A key application is **solubility** prediction, which is essential for biological function and recombinant expression. pLM-based methods outperform older tools, particularly in generalization and accuracy. Two common strategies are used. First, feed pLM embeddings into a downstream classifier or

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

regressor; some systems add extra sequence or experimental features in a hybrid design, such as in hybrid GCN,[58] ESM-2,[128] and SFM-Protein,[71] on curated data sets.[129]

Stability, including thermodynamic stability (ΔG), and thermostability (e.g., melting temperature), is a property effectively predicted by pLMs. ESMtherm, for example, fine-tunes ESM-2 on a data set of 528 k domain stabilities and generalizes to unseen folds, predicting ΔG directly from sequence.[67] Also TemPL applies temperature-guided language modeling to millions of growth temperatures and ΔTm values, enabling zero-shot predictions of melting points and activity shifts without mutagenesis data.[49]

pLMs also show promise in predicting **solubility and aggregation propensity** and related phenomena like liquid−liquid phase separation (LLPS).[130] This ability often arises from the models' recognition of intrinsically disordered regions, which commonly underlie phase transitions. ProteinBERT embeddings have also been used to infer solvent accessibility, intrinsic disorder, and other biophysical attributes, showing robust performance across data sets.[39] Furthermore, NetSolP, a solubility model using pLM embeddings, reached state-of-the-art accuracy in predicting soluble expression in *E. coli*, surpassing earlier machine learning approaches.[37] Similarly, pLM-derived features also improve the prediction of aggregation and crystallization success. Recent benchmarks showed that embeddings from models like ESM2 and ProtT5[34] raised AUC/AUPR by 3−5% over older tools for crystallizability prediction.[131] Generative pLMs have even been used to design new proteins less prone to aggregation, e.g., fine-tuning ProtGPT2 to generate sequences predicted to be highly crystallizable.[131]

Beyond specific traits, pLMs can predict engineered properties such as **fluorescence intensity**.[132] This versatility stems from unsupervised pretraining, which lets pLMs capture biophysical signals and biochemical properties embedded in sequences.

Protein−Protein Interaction and Binding Affinity Prediction

Recent studies have explored pLMs to predict protein−protein interactions (PPIs) and binding affinities directly from sequences.[133] A common approach is to use high-dimensional embeddings from a pretrained model as input to a classifier or regressor to predict whether two proteins interact or to estimate their binding strength.[133] For example, ProtBert-BiGRU-Attention pairs ProtBert embeddings with a BiGRU-attention network to accurately predict binary protein−protein interactions using sequence data alone.[134] Fine-tuned pLMs can match or exceed state-of-the-art results on various PPI benchmarks. However, standard models trained on single sequences may miss interprotein context, as encoding partners separately, or just concatenating them, can overlook important interaction-specific features.[78] Another solution, SWING,[65] addresses this by modeling the language of protein−protein and protein−peptide interactions through recurrent local structural motifs. It assigns each residue an index based on its local structural environment, which summarizes recurrent interface motifs. These indices are then used by the model to score partner compatibility and predict interactions, improving generalization beyond a simple sequence concatenation.

To improve the PPI and binding affinity prediction from sequences, recent methods have adapted language models to better capture interactions. MINT (Multimeric Interaction Transformer) is one such example, extending ESM-2 to jointly process multiple protein sequences using cross-attention.[78] MINT was pretrained on a large corpus of known interactions (STRING database[135]) to learn contextual representations of interacting chains. This resulted in clear gains in PPI tasks: MINT reached an AUPRC of ∼0.69 on a standard PPI data set, well above earlier sequence-based models, and improved accuracy by 29% for predicting binding affinity changes on the SKEMPI benchmark.[136] In parallel, other work explored pairing strategies using masked-language modeling as an unsupervised score for interaction propensity. For example, DiffPALM, an MSA-based transformer, detects coevolutionary signals between sequences, allowing accurate pairing of interacting proteins without explicit training on known pairs.[60] Beyond binary interaction predictions, embeddings from models like AntiFormer[75] have been combined with graph neural networks or attention mechanisms to integrate context such as interaction networks or antibody lineage, improving PPI prediction accuracy.

The pLMs have also been applied to predict binding affinities and affinity changes from sequence.[137] One study introduced ProBASS, a sequence-based framework that combines embeddings from ESM-2 and ESM-IF1 to predict binding free energy changes ($\Delta\Delta G$) resulting from mutations.[138] Fine-tuned on a large $\Delta\Delta G$ data set, ProBASS reached correlations up to $r \approx 0.8$ with experimental binding changes, outperforming earlier methods for both single and multiple mutations. In the antibody−antigen domain, DG-Affinity[55] concatenates embeddings of an antibody and antigen (from two pretrained encoders) into a convolutional network, reaching a Pearson $r > 0.65$ on independent test sets and surpassing earlier structure-based tools. Similarly, AntiFormer combines transformer language models with a graph-based module to predict antibody binding affinities, showing higher accuracy than other antibody-specific predictors.[75] These findings suggest that sequence-only models can capture key factors influencing binding affinity even without direct structural input.

Despite progress, sequence-only models may miss subtle conformational or context-specific effects better captured by structural data. They also depend on large labeled data sets for fine-tuning, though self-supervised interaction pretraining, as used in MINT,[78] helps mitigate this.

Antigen−Receptor Binding Prediction

Antigen−receptor binding is central to immunology. T-cell receptors (TCRs) recognize peptides presented by MHC molecules, while B-cell receptors (BCRs), mainly antibodies, bind free antigens to trigger immune responses essential for host defense.[139] The diversity and complexity of receptor−antigen sequences have made this task difficult, with experimental methods like surface plasmon resonance and yeast two-hybrid assays providing precise but slow and low-throughput results.[140] Recent advances in antigen−receptor binding prediction have been driven by pLM-derived contextual embeddings,[141] allowing sequence-only models to capture structural and functional features. This improves TCR and antibody interaction predictions by learning representations that reflect key biochemical and structural properties from large-scale sequence data.[66] For instance, ESM-1b, a 650-million-parameter transformer, has been applied to generate embeddings for peptide epitopes and TCR complementarity-determining regions (CDRs), achieving state-of-the-art affinity

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

prediction on benchmarks like the TDC challenge,[66] while ProtTrans-based approaches such as ParaAntiProt uses embeddings from pretrained antibody language models to identify paratope residues (i.e., the antibody's binding site that pairs with an antigenic epitope) without antigen input, showing that pLMs inherently capture contact-relevant information.[73] In the TCR domain, tcrLM uses a masked-segment objective and virtual adversarial training on over 100 million CDR3 sequences to capture biochemical motifs and positional preferences, achieving superior discrimination of binding pairs in various cohorts,[66] while TAPIR (T-cell receptor and Peptide Interaction Recognizer) employs separate convolutional encoders for TCR and peptide sequences, merging the embeddings by concatenation into a joint vector for prediction,[142] enabling zero-shot generalization and successful identification of cancer neoantigen-specific TCRs.[142] SC-AIR-BERT extends this paradigm by pretraining a six-layer BERT encoder on millions of paired TCR/BCR sequences from single-cell data sets using a k-mer masking strategy, then fine-tuning with a multilayer perceptron head to predict antigen-binding specificity with top performance across TCR and BCR benchmarks.[51] MHCRoBERTa, in turn, adapts the RoBERTa architecture to peptide−MHC class I binding by self-supervised pretraining on large peptide corpora and subsequent fine-tuning on affinity measurements, achieving higher Spearman correlation and AUC than dedicated tools such as NetMHCpan3.0 (a pan-specific peptide−HLA binding predictor) and MHCflurry (a neural-network ensemble for affinity estimation) on pan-allele benchmarks.[40] Similarly, BERT2DAb pretrains a BERT masked-language model on antibody sequences with 2D structural features, enabling richer residue embeddings that boost downstream tasks like paratope identification and antibody screening.[53] Complementing these, ProtLM.TCR, pretrained on millions of TCR sequences, yields high precision−recall performance.[47] TCR-ESM leverages embeddings from a general ESM model for TCR−pMHC binding prediction.[57] In parallel, contrastive models such as SCEPTR apply self-supervised learning to adapt efficiently to low-resource binding tasks.[76] Unified frameworks like UniPMT integrate peptide, MHC allele, and TCR β CDR3 sequences within a single transformer, showing improved precision-recall over separate models without needing per-allele retraining.[77] Benchmarking studies confirm that embeddings from pLMs like ESM-1b, ProtLM.TCR, and tcrLM consistently outperform traditional baselines, and this modular embedding paradigm facilitates rapid extension to specialized applications such as neoantigen recognition and cross-reactivity mapping, as demonstrated by TAPIR's discovery of novel cancer-specific TCRs.[52] Overall, pLM-based models offer a scalable, structure-agnostic approach for predicting antigen−receptor binding, enabling high-throughput in silico screening of therapeutic antibodies and TCRs. This supports faster vaccine design, immunotherapy, personalized medicine, and advances in understanding adaptive immunity.

Antigen−Antibody Binding Prediction

Transformer-based protein language models are now key tools for predicting antibody−antigen interactions. Recent approaches leverage both general pLMs and antibody-specialized LMs to encode sequences for binding prediction tasks. These models are typically pretrained on massive antibody sequence data sets,[61,143] using architectures from bidirectional masked-language models (BERT/RoBERTa variants) to sequence-to-sequence models. The rich sequence embeddings produced by these transformers capture immunoglobulin-specific features. For example, AntiBERTy's embedding space reflects antibody structural and evolutionary properties learned from unlabeled sequences.[38]

Several studies report that sequence-based language models can rival or outperform traditional structure-based methods on binding affinity and specificity benchmarks. For instance, DG-Affinity[55] uses two pretrained encoders, TAPE[144] for antigen sequences and AbLang[41] for antibody sequences, to generate representations that are fed into a ConvNeXt network,[145] a Transformer-inspired modern convolutional neural network, for affinity regression. This purely sequence-driven model achieved a Pearson correlation >0.65 on independent affinity test data, exceeding earlier structure-based predictors. Similarly, AntiBERTa, fine-tuned on large antibody repertoires, produces residue-level embeddings that align well with experimental paratope regions, matching structure-based methods.[42] And BALM, trained on over 300 million antibody sequences, learns representations that predict binding kinetics, mutational effects, and paratope or affinity across various antigens.[64] Moreover, fine-tuning antibody pLMs on known binding data has proven effective, as demonstrated by AntiBERTy,[38] AntiBERTa, and a fine-tuned ESM2[48] for binary classification of binding specificity to SARS-CoV-2 spike and influenza HA antigens.[146] The fine-tuned models showed improved accuracy over using static embeddings; in cross-validation, they reached high AUROC (∼0.86−0.88) on held-out test folds, significantly outperforming baseline SVM classifiers on frozen LM embeddings. On a similar theme, a recent large-scale model introduces IgBERT (encoder) and IgT5 (encoder−decoder) trained on >2 billion antibody chains and fine-tuned on 2 million paired heavy-light sequences.[61] These models achieve state-of-the-art results across antibody engineering tasks, outperforming earlier protein and antibody pLMs in binding affinity prediction. In one benchmark, embeddings from IgBERT/T5 led to more accurate affinity and developability predictions than those from older pLMs like AbLang[41] or AntiBERTy, highlighting the value of training on very large and paired antibody data sets.

Transformer models have been adapted to predict paratopes and integrate antigen context, where attention and embeddings can locate functional antibody sites. For example, ParaAntiProt[73] combines general (ESM-2/ProtTrans) and antibody-specific pLMs (AbLang, BALM, AntiBERTy) with convolutional layers and CDR encodings to predict paratopes from sequence alone, outperforming earlier structure- or alignment-based methods. Other techniques integrate antigen information via contrastive learning or multimodal models, such as the supervised contrastive fine-tuning of AbLang-PDB embeddings, which grouped antibodies by epitope in embedding space,[147] achieving high accuracy in distinguishing same- vs different-epitope pairs and improving generalization to unseen epitopes. Other hybrid strategies integrate the structural context: AntiFormer incorporates graph-based structural representations alongside sequence embeddings to enhance binding affinity prediction, outperforming prior sequence-only methods.[75] Additional antibody-tailored pLMs exist, such as EATLM,[148] an evolution-aware transformer incorporating germline lineage context, and even nanobody-specific models like NanoBERTa-ASP[62,149] for single-domain antibody paratopes.

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

Finally, foundation models have been extended to consider antigen features directly: MAMMAL multimodal framework,[150] predicts whether an antibody will bind and block a given influenza hemagglutinin using only sequences.[72] Their sequence-only model attained high AUROC (~0.90) on known antibody-HA interactions, though performance declined on novel, dissimilar antibodies, underscoring the need for diverse training data for robust generalization.

CHALLENGES AND LIMITATIONS

pLMs have achieved impressive results in learning protein sequence representations, but they still face significant challenges and limitations in proteomics.

Interpretability is a primary concern: these deep models act largely as black boxes, and although attention-based analyses have shown that some learned attention patterns correspond to structural contacts or binding sites, the models' internal representations remain difficult to interpret in biologically meaningful terms.[91] Understanding why large language models make certain predictions is challenging, which limits their trust and use in important biological applications.

Data bias in training corpora further limits these models. pLMs are trained on large protein databases (e.g., UniProt), which are skewed toward certain organisms and protein families; as a result, the models learn species- or lineage-specific biases.[151] For example, it has been observed that a pLM assigned systematically higher likelihoods (fitness proxies) to sequences from over-represented species regardless of function, a bias that can mislead protein design efforts.[151]

Train-test leakage, often due to sequence or metadata overlap in common protein benchmarks, inflates the apparent generalization of pLMs and poses a risk to the validity of results, especially when models are trained on broad corpora and evaluated with lenient splits.[152] Pretraining contamination is another failure mode: if the test proteins or close variants were seen during pretraining, downstream tasks like thermostability look easier than they are.[153] Mitigations include pretraining-aware splits, strict cluster or family splits with identity filtering, and deduplication across all stages.[154] Claims should include reported splits and decontamination to reflect true generalization, not memorized training properties.[155]

Tokenization of the protein sequences is another challenge. Unlike natural language, proteins lack clear word boundaries, delineated by whitespace and punctuation, to delineate "words" so most models resort to treating each amino acid as a token.[156] This one-letter tokenization misses higher-order motifs and contextual cues present in protein sequences. Some groups replace single-residue tokens with learned multiresidue units using byte-pair or unigram tokenization, which can capture common motifs but sometimes miss biologically critical single-residue changes.[157] These factors can limit the model's ability to learn grammar-like rules beyond local amino acid patterns.

Moreover, pLMs still have **limitations in downstream prediction tasks**. Sequence-only models can plateau, especially for functional predictions that depend on subtle biochemical or evolutionary contexts. Current pLMs largely overlook protein biophysics, relying on statistical patterns that hinder generalization to novel protein design or variant-effect prediction.[158] For instance, without structural context, language models may miss effects of distant mutations, making traditional methods or domain-specific features necessary for some tasks.

Generalizability across protein families remains an open issue: pLMs trained on millions of sequences can still struggle with rare or unseen families. They often fail on out-of-distribution proteins, such as orphans, showing that broad sequence coverage is key.[45,159] As a result, performance varies: strong on common families and weaker on remote or novel folds. Recent work such as PortalCG proposes an end-to-end sequence−structure−function framework with out-of-cluster meta-learning to address out-of-distribution settings in protein−ligand prediction, reporting improved generalization on understudied proteins.[160]

Scalability is another limitation, in terms of both sequence length and data set size. Most pLMs have quadratic complexity, limiting their ability to handle long sequences without truncation or optimizations.[48] Standard models typically cap input at a few thousand residues, and processing longer sequences requires efficient attention methods that can reduce accuracy. Likewise, training on the full range of protein sequences also demands significant computational resources due to large model sizes and data sets. This **trade-off between model size and performance** is a constant theme. Larger models generally improve protein task performance; for example, the ESM series showed that scaling up transformers leads to better representations and accuracy, following trends seen in natural language processing.[121] Larger models can yield diminishing returns without diverse training data and may overfit when data sets are redundant. Scaling up also increases the cost and complexity, so developers must balance size with practical gains in representation.

There are also important considerations in **fine-tuning and transfer learning** for biological problems. Fine-tuning large pLMs on specific proteomics tasks is difficult with limited labeled data, as they can overfit and lose the general protein features learned during pretraining. Early applications often avoided full fine-tuning using frozen pLM embeddings as features. Recently, however, studies show that fine-tuning pLMs, even with minimal added parameters or adapter layers, can significantly improve downstream predictions.[161] Task-specific fine-tuning consistently improves accuracy, and parameter-efficient methods can achieve similar results at lower cost. With proper regularization, fine-tuning helps to address data scarcity and extract task-relevant knowledge. Still, overfitting and the need for expert tuning remain challenges.

Finally, the **fast pace of new pLM architectures** and variants creates uncertainty. Many models appear briefly with limited validation in preprints, and there is no consensus on which are reliable for long-term use. Reported results can be biased toward the authors' own methods, reinforcing the need for objective, community-wide benchmarks. Establishing shared evaluation suites and updated model libraries will be essential for tracking progress and selecting suitable tools as new models emerge.

FUTURE PERSPECTIVES

Future pLM-based approaches in proteomics provide new opportunities for studying and engineering proteins. With over 240 million sequences known but <0.3% functionally characterized,[162] pLMs are becoming indispensable for extracting evolutionary knowledge from sequences.

In **structure prediction**, pLM-powered methods already achieve impressive results, eliminating the need for multiple sequence alignments and enabling rapid proteome-scale structure discovery.[163] However, combining pLM embeddings

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

with evolutionary profiles and physics-based refinement could balance speed and accuracy, cut costs, and improve predictions for large assemblies and flexible proteins.

Another frontier is capturing protein **dynamics and context**, as present pLM approaches struggle with intrinsically disordered regions and conformation changes induced by partners or post-translational modifications.[164] Recent pLM-based efforts have begun to address this: DR-BERT annotates disordered regions,[165] LoRA-DR adapters improve intrinsic and soft disorder prediction from pLM embeddings,[166] and SeaMoon uses pLM embeddings to model motions from sequence,[167] suggesting that adding structural context or multistate training may enable prediction of ensembles and allosteric effects in vivo.

In **functional annotation**, pLM embeddings have already begun to outperform traditional sequence-homology methods across many tasks, leading to an irreversible trend of replacing handcrafted features with pLM-derived representations.[162] Fine-tuning these foundation models on specific tasks further boosts accuracy as demonstrated by recent benchmarks across diverse prediction challenges.[161] This means that a single pretrained pLM can be adapted to predict enzyme activities, subcellular localizations, or protein−protein interactions with high fidelity, vastly accelerating the annotation of the proteome.

In **evolution**, recent phylogenetic models like SiteRM[168] demonstrate that alternative probabilistic approaches can remain highly competitive, suggesting ongoing opportunities for innovation beyond standard pLM architectures.

Translational impacts are also on the horizon. For example, generative pLMs can design novel proteins in silico: ProGen model produced artificial enzymes that not only expressed in cells but also showed desired activity,[164] hinting at AI-driven directed evolution for enzyme engineering and therapeutic protein design. Such capabilities imply that pLM-based protein design will revolutionize drug development, enabling custom protein therapeutics or tailor-made targets for small molecules.[164]

In genomic medicine, pLM-informed tools are tackling **variants of unknown significance**. DeepMind's AlphaMissense leverages unsupervised pLM training to predict the pathogenicity of every possible missense mutation, showing enormous potential to aid rare disease diagnosis and treatment decisions.[169] Looking ahead, we anticipate pLM strategies to integrate with structural biology, systems biology, and clinical pipelines, ultimately yielding a holistic understanding of proteins.

In **PPIs**, ongoing work is likely to refine these models further (e.g., by incorporating partner context, coevolution signals, or lightweight structural cues) and to expand their applicability to more complex interaction scenarios, all while maintaining a fully sequence-based approach.

Finally, **Multimodal pLMs** are emerging, with models that integrate sequence with structure and function tokens and show gains on downstream tasks.[170−172] Extending this line toward true multiomics inputs alongside proteins is an active area, and we expect fast progress, but rigorous benchmarks are still limited

A last critical avenue for future work is the establishment of **evaluation frameworks** and model repositories to tame the ever-growing diversity of pLMs. Establishing shared benchmarks and model registries will be vital to managing the rapid proliferation of pLM architectures. Standardized evaluation suites and catalogs detailing training data, hyperparameters, and compute requirements allow a clear comparison of tools across structure, function, and interaction tasks. These will help researchers identify production-ready models, reduce redundant efforts, and maintain relevance as sequence landscapes evolve, ultimately speeding up the adoption of reliable pLMs in proteomic research and applications.

CONCLUSION

The applications of pLMs have shown that raw sequence data, when transformed into context-aware embeddings, can support a wide spectrum of proteomic tasks, from predicting three-dimensional folds and annotating enzymatic functions to designing novel sequences and modeling molecular interactions. By replacing manual feature engineering with self-supervised pretraining on millions of sequences, these models offer rapid, proteome-wide inference and, in many cases, match or exceed the accuracy of alignment- and structure-based methods. Their ability to generate realistic candidate proteins and guide experimental screens has already begun to shorten the cycle of design and validation in enzyme engineering, antibody discovery, and beyond.

Yet important hurdles must be overcome before protein language models become routine tools. The opacity of their internal representations makes it difficult to trace predictions back to biophysical principles, and biases in training databases can skew outputs toward over-represented species or folds. Large models demand substantial computation for both training and inference, posing practical limits on sequence length and throughput. At the same time, fine-tuning on specialized tasks risks overfitting when labeled data are scarce, and purely sequence-based approaches may miss crucial structural or cellular contexts for certain functions.

Moving forward, the field stands to gain by combining the strengths of language models with explicit structural, phylogenetic, and experimental data. Integrating attention analyses with graph- or physics-based modules could yield more interpretable and data-efficient predictors. Advances in efficient attention mechanisms and parameter-light tuning methods will help scale models to longer sequences and broader proteomes. Close collaboration between computational scientists and experimentalists will be essential to validate predictions, curate high-quality data sets, and iterate models in real-world settings. Additionally, the quick pace of new pLMs and changing protein data can make it hard to choose the right tool, so shared benchmarks and up-to-date model lists are essential. As these efforts mature, pLMs will not only accelerate basic discovery but also drive drug discovery and personalized medicine and support targeted interventions in biomedicine, industrial biotechnology, and synthetic biology, ultimately realizing the promise of decoding and engineering the language of life.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.5c00506.

Table S1: Summary of all protein language models discussed in this review, extending Table 1 with the columns "Performance Highlights" and "Release Date" (XLSX)

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524

AUTHOR INFORMATION

Corresponding Author

Arnaud Droit − *Axe Endo-Nephro, Centre de recherche du CHU de Québec-Université Laval, Québec, QC G1 V 4G2, Canada; Département de médecine moléculaire, Faculté de médecine, Université Laval, Québec, QC G1 V 0A6, Canada;* orcid.org/0000-0001-7922-790X; Email: arnaud.Droit@crchudequebec.ulaval.ca

Author

Mickael Leclercq − *Axe Endo-Nephro, Centre de recherche du CHU de Québec-Université Laval, Québec, QC G1 V 4G2, Canada;* orcid.org/0000-0001-6205-888X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.5c00506

Author Contributions

M.L. defined the review's scope, performed the literature search, organized and synthesized the selected studies, wrote the manuscript, and finalized it for submission. A.D. guided the topic, critically reviewed, and revised the manuscript.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Kavallaris, M.; Marshall, G. M. Proteomics and Disease: Opportunities and Challenges. *Med. J. Aust.* **2005**, *182* (11), 575−579.

(2) Yu, Y.; Mai, Y.; Zheng, Y.; Shi, L. Assessing and Mitigating Batch Effects in Large-Scale Omics Studies. *Genome Biol.* **2024**, *25*, No. 254.

(3) Po, A.; Eyers, C. E. Top-Down Proteomics and the Challenges of True Proteoform Characterization. *J. Proteome Res.* **2023**, *22*, 3663−3675.

(4) Yan, R.; Xu, D.; Yang, J.; Walker, S.; Zhang, Y. A Comparative Assessment and Analysis of 20 Representative Sequence Alignment Methods for Protein Structure Prediction. *Sci. Rep.* **2013**, *3* (1), 1−9.

(5) Chao, J.; Tang, F.; Xu, L. Developments in Algorithms for Sequence Alignment: A Review. *Biomolecules* **2022**, *12* (4), No. 546.

(6) Warnow, T. Revisiting Evaluation of Multiple Sequence Alignment Methods. *Methods Mol. Biol. (Clifton, N.J.)* **2021**, *2231*, 299−317.

(7) Chowdhury, B.; Gautam, G. A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, *109* (5−6), 419−431.

(8) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), No. e2016239118, DOI: 10.1073/pnas.2016239118.

(9) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(10) Kulikova, A. V.; Diaz, D. J.; Chen, T.; Cole, T. J.; Ellington, A. D.; Wilke, C. O. Two Sequence- and Two Structure-Based ML Models Have Learned Different Aspects of Protein Biochemistry. *Sci. Rep.* **2023**, *13* (1), No. 13280.

(11) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat. Biotechnol.* **2023**, *41* (8), 1099−1106.

(12) Valentini, G.; Malchiodi, D.; Gliozzo, J.; Mesiti, M.; Soto-Gomez, M.; Cabri, A.; Reese, J.; Casiraghi, E.; Robinson, P. N. The Promises of Large Language Models for Protein Design and Modeling. *Front. Bioinform.* **2023**, *3*, No. 1304099.

(13) Zhang, H.; Song, H.; Li, S.; Zhou, M.; Song, D. A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models. *ACM Comput. Surv.* **2024**, *56*, 1−37.

(14) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I.Attention Is All You Need *arXiv [cs.CL]* 2017, pp 5998−6008 https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

(15) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners, arXiv:2005.14165. arXiv.org e-Print archive. http://arxiv.org/abs/2005.14165 (accessed March 20, 2025). 2020.

(16) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805. arXiv.org e-Print archive. http://arxiv.org/abs/1810.04805 (accessed March 20, 2025). 2018.

(17) Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842−866.

(18) Beltagy, I.; Lo, K.; Cohan, A.*SciBERT: A Pretrained Language Model for Scientific Text*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019; pp 3615−3620 .

(19) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36* (4), 1234−1240.

(20) Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning*; PMLR, 2019; pp 2790−2799.

(21) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, arXiv:2106.09685. arXiv.org e-Print archive. http://arxiv.org/abs/2106.09685 (accessed May 17, 2025). 2021.

(22) Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv:2005.11401. arXiv.org e-Print archive. http://arxiv.org/abs/2005.11401 (accessed May 17, 2025). 2020

(23) UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43* (Database issue), D204−D212.

(24) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; UniProt Consortium. UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* **2015**, *31* (6), 926−932.

(25) Bepler, T.; Berger, B. Learning the Protein Language: Evolution, Structure and Function. *Cell Syst.* **2021**, *12* (6), 654−669.e3.

(26) Steinegger, M.; Söding, J. Clustering Huge Protein Sequence Sets in Linear Time. *Nat. Commun.* **2018**, *9* (1), No. 2542.

(27) Bernett, J.; Blumenthal, D. B.; List, M. Cracking the Black Box of Deep Sequence-Based Protein-Protein Interaction Prediction. *Brief. Bioinform.* **2024**, *25* (2), No. bbae076, DOI: 10.1093/bib/bbae076.

(28) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9689−9701.

(29) Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.
(30) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026−1028.
(31) Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N.; Davis, L.; Dogan, T.; Atalay, V.; Rifaioglu, A. S.; Dalkıran, A.; Cetin Atalay, R.; Zhang, C.; Hurto, R. L.; Freddolino, P. L.; Zhang, Y.; Bhat, P.; Supek, F.; Fernández, J. M.; Gemovic, B.; Perovic, V. R.; Davidović, R. S.; Sumonja, N.; Veljkovic, N.; Asgari, E.; Mofrad, M. R. K.; Profiti, G.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Boecker, F.; Schoof, H.; Kahanda, I.; Thurlby, N.; McHardy, A. C.; Renaux, A.; Saidi, R.; Gough, J.; Freitas, A. A.; Antczak, M.; Fabris, F.; Wass, M. N.; Hou, J.; Cheng, J.; Wang, Z.; Romero, A. E.; Paccanaro, A.; Yang, H.; Goldberg, T.; Zhao, C.; Holm, L.; Törönen, P.; Medlar, A. J.; Zosa, E.; Borukhov, I.; Novikov, I.; Wilkins, A.; Lichtarge, O.; Chi, P.-H.; Tseng, W.-C.; Linial, M.; Rose, P. W.; Dessimoz, C.; Vidulin, V.; Dzeroski, S.; Sillitoe, I.; Das, S.; Lees, J. G.; Jones, D. T.; Wan, C.; Cozzetto, D.; Fa, R.; Torres, M.; Warwick Vesztrocy, A.; Rodriguez, J. M.; Tress, M. L.; Frasca, M.; Notaro, M.; Grossi, G.; Petrini, A.; Re, M.; Valentini, G.; Mesiti, M.; Roche, D. B.; Reeb, J.; Ritchie, D. W.; Aridhi, S.; Alborzi, S. Z.; Devignes, M.-D.; Koo, D. C. E.; Bonneau, R.; Gligorijević, V.; Barot, M.; Fang, H.; Toppo, S.; Lavezzo, E.; Falda, M.; Berselli, M.; Tosatto, S. C. E.; Carraro, M.; Piovesan, D.; Ur Rehman, H.; Mao, Q.; Zhang, S.; Vucetic, S.; Black, G. S.; Jo, D.; Suh, E.; Dayton, J. B.; Larsen, D. J.; Omdahl, A. R.; McGuffin, L. J.; Brackenridge, D. A.; Babbitt, P. C.; Yunes, J. M.; Fontana, P.; Zhang, F.; Zhu, S.; You, R.; Zhang, Z.; Dai, S.; Yao, S.; Tian, W.; Cao, R.; Chandler, C.; Amezola, M.; Johnson, D.; Chang, J.-M.; Liao, W.-H.; Liu, Y.-W.; Pascarelli, S.; Frank, Y.; Hoehndorf, R.; Kulmanov, M.; Boudellioua, I.; Politano, G.; Di Carlo, S.; Benso, A.; Hakala, K.; Ginter, F.; Mehryary, F.; Kaewphan, S.; Björne, J.; Moen, H.; Tolvanen, M. E. E.; Salakoski, T.; Kihara, D.; Jain, A.; Šmuc, T.; Altenhoff, A.; Ben-Hur, A.; Rost, B.; Brenner, S. E.; Orengo, C. A.; Jeffery, C. J.; Bosco, G.; Hogan, D. A.; Martin, M. J.; O'Donovan, C.; Mooney, S. D.; Greene, C. S.; Radivojac, P.; Friedberg, I. The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens. *Genome Biol.* **2019**, *20* (1), No. 244.
(32) Rao, R.; Meier, J.; Sercu, T.; Ovchinnikov, S.; Rives, A. Transformer Protein Language Models Are Unsupervised Structure Learners. *bioRxiv* **2020**, DOI: 10.1101/2020.12.15.422761.
(33) Liu, J.; Yang, M.; Yu, Y.; Xu, H.; Li, K.; Zhou, X. Large Language Models in Bioinformatics: Applications and Perspectives *ArXiv* 2025.
(34) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing, arXiv:2007.06225. arXiv.org e-Print archive. http://arxiv.org/abs/2007.06225 (accessed May 17, 2025). 2020.
(35) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. *ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44; 2022; pp 7112−7127 .
(36) Tule, S.; Foley, G.; Bodén, M. Do Protein Language Models Learn Phylogeny? *bioRxiv* **2024**, No. 2024.09.23.614642, DOI: 10.1101/2024.09.23.614642.
(37) Thumuluri, V.; Martiny, H.-M.; Almagro Armenteros, J. J.; Salomon, J.; Nielsen, H.; Johansen, A. R. NetSolP: Predicting Protein Solubility in *Escherichia Coli* Using Language Models. *Bioinformatics* **2022**, *38* (4), 941−946.
(38) Ruffolo, J. A.; Gray, J. J.; Sulam, J. Deciphering Antibody Affinity Maturation with Language Models and Weakly Supervised Learning, arXiv:2112.07782. arXiv.org e-Print archive. https://arxiv.org/abs/2112.07782 (accessed May 23, 2025). 2021.
(39) Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics* **2022**, *38* (8), 2102−2110.
(40) Wang, F.; Wang, H.; Wang, L.; Lu, H.; Qiu, S.; Zang, T.; Zhang, X.; Hu, Y. MHCRoBERTa: Pan-Specific peptide−MHC Class I Binding Prediction through Transfer Learning with Label-Agnostic Protein Sequences. *Briefings Bioinf.* **2022**, *23* (3), No. bbab595.
(41) Olsen, T. H.; Moal, I. H.; Deane, C. M. AbLang: An Antibody Language Model for Completing Antibody Sequences. *Bioinform Adv.* **2022**, *2* (1), No. vbac046.
(42) Choi, Y. Artificial Intelligence for Antibody Reading Comprehension: AntiBERTa. *Patterns (N Y)* **2022**, *3* (7), No. 100535.
(43) Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. *Nat. Commun.* **2022**, *13* (1), No. 4348.
(44) Wu, L.; Yin, C.; Zhu, J.; Wu, Z.; He, L.; Xia, Y.; Xie, S.; Qin, T.; Liu, T.-Y. SPRoBERTa: Protein Embedding Learning with Local Fragment Modeling. *Briefings Bioinf.* **2022**, *23* (6), No. bbac401, DOI: 10.1093/bib/bbac401.
(45) Chowdhury, R.; Bouatta, N.; Biswas, S.; Floristean, C.; Kharkar, A.; Roy, K.; Rochereau, C.; Ahdritz, G.; Zhang, J.; Church, G. M.; Sorger, P. K.; AlQuraishi, M. Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning. *Nat. Biotechnol.* **2022**, *40* (11), 1617−1623.
(46) Soylu, N. N.; Sefer, E. DeepPTM: Protein Post-Translational Modification Prediction from Protein Sequences by Combining Deep Protein Language Model with Vision Transformers. *Curr. Bioinf.* **2024**, *19* (9), 810−824.
(47) Essaghir, A.; Sathiyamoorthy, N. K.; Smyth, P.; Ghiviriga, S.; Ghita, A.; Singh, A.; Kapil, S.; Phogat, S.; Singh, G. T-Cell Receptor Specific Protein Language Model for Prediction and Interpretation of Epitope Binding (ProtLM.TCR). *bioRxiv* **2022**, No. 2022.11.28.518167, DOI: 10.1101/2022.11.28.518167.
(48) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, No. 1123.
(49) Tan, P.; Li, M.; Zhang, L.; Hu, Z.; Hong, L. TemPL: A Novel Deep Learning Model for Zero-Shot Prediction of Protein Stability and Activity Based on Temperature-Guided Language Modeling, arXiv:2304.03780. arXiv.org e-Print archive. https://arxiv.org/abs/2304.03780. (accessed May 19, 2025). 2023.
(50) Wang, D.; Ye, F.; Zhou, H. On Pre-Trained Language Models for Antibody, arXiv:2301.12112. arXiv.org e-Print archive. https://arxiv.org/abs/2301.12112 (accessed May 26, 2025). 2023.
(51) Zhao, Y.; Su, X.; Zhang, W.; Mai, S.; Xu, Z.; Qin, C.; Yu, R.; He, B.; Yao, J. SC-AIR-BERT: A Pre-Trained Single-Cell Model for Predicting the Antigen-Binding Specificity of the Adaptive Immune Receptor. *Briefings Bioinf.* **2023**, *24* (4), No. bbad191.
(52) Fast, E.; Dhar, M.; Chen, B. TAPIR: A T-Cell Receptor Language Model for Predicting Rare and Novel Targets. *bioRxiv: the preprint server for biology* **2023**, DOI: 10.1101/2023.09.12.557285.
(53) Luo, X.; Tong, F.; Zhao, W.; Zheng, X.; Li, J.; Li, J.; Zhao, D. BERT2DAb: A Pre-Trained Model for Antibody Representation Based on Amino Acid Sequences and 2D-Structure. *MAbs* **2023**, *15* (1), No. 2285904.
(54) Buton, N.; Coste, F.; Le Cunff, Y. Predicting Enzymatic Function of Protein Sequences with Attention. *Bioinformatics (Oxford, England)* **2023**, *39* (10), No. btad620, DOI: 10.1093/bioinformatics/btad620.
(55) Yuan, Y.; Chen, Q.; Mao, J.; Li, G.; Pan, X. DG-Affinity: Predicting Antigen-Antibody Affinity with Language Models from Sequences. *BMC Bioinf.* **2023**, *24* (1), No. 430.
(56) Nijkamp, E.; Ruffolo, J. A.; Weinstein, E. N.; Naik, N.; Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. *Cell Syst.* **2023**, *14* (11), 968−978.e3.

(57) Yadav, S.; Vora, D. S.; Sundar, D.; Dhanjal, J. K. TCR-ESM: Employing Protein Language Embeddings to Predict TCR-Peptide-MHC Binding. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 165−173.
(58) Chen, L.; Wu, R.; Zhou, F.; Zhang, H.; Liu, J. K. HybridGCN for Protein Solubility Prediction with Adaptive Weighting of Multiple Features. *J. Cheminf.* **2023**, *15* (1), No. 118.
(59) Li, G.; Yao, S.; Fan, L. ProSTAGE: Predicting Effects of Mutations on Protein Stability by Using Protein Embeddings and Graph Convolutional Networks. *J. Chem. Inf. Model.* **2024**, *64*, 340−347.
(60) Lupo, U.; Sgarbossa, D.; Bitbol, A.-F. Pairing Interacting Protein Sequences Using Masked Language Modeling. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121* (27), No. e2311887121.
(61) Kenlay, H.; Dreyer, F. A.; Kovaltsuk, A.; Miketa, D.; Pires, D.; Deane, C. M. Large Scale Paired Antibody Language Models. *PLoS Comput. Biol.* **2024**, *20* (12), No. e1012646.
(62) Li, S.; Meng, X.; Li, R.; Huang, B.; Wang, X. NanoBERTa-ASP: Predicting Nanobody Paratope Based on a Pretrained RoBERTa Model. *BMC Bioinf.* **2024**, *25* (1), No. 122.
(63) Nana Teukam, Y. G.; Kwate Dassi, L.; Manica, M.; Probst, D.; Schwaller, P.; Laino, T. Language Models Can Identify Enzymatic Binding Sites in Protein Sequences. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 1929−1937.
(64) Jing, H.; Gao, Z.; Xu, S.; Shen, T.; Peng, Z.; He, S.; You, T.; Ye, S.; Lin, W.; Sun, S. Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model. *Briefings Bioinf.* **2024**, *25* (4), No. bbae245, DOI: 10.1093/bib/bbae245.
(65) Omelchenko, A. A.; Siwek, J. C.; Chhibbar, P.; Arshad, S.; Nazarali, I.; Nazarali, K.; Rosengart, A.; Rahimikollu, J.; Tilstra, J.; Shlomchik, M. J.; Koes, D. R.; Joglekar, A. V.; Das, J. Sliding Window INteraction Grammar (SWING): A Generalized Interaction Language Model for Peptide and Protein Interactions. *bioRxiv* **2024**, No. 2024.05.01.592062, DOI: 10.1101/2024.05.01.592062.
(66) Fang, X.; Yu, C.; Tian, S.; Liu, H. tcrLM: A Lightweight Protein Language Model for Predicting T Cell Receptor and Epitope Binding Specificity, arXiv:2406.16995. arXiv.org e-Print archive. https://arxiv.org/abs/2406.16995 (accessed April 22, 2025). 2024.
(67) Chu, S. K. S.; Narang, K.; Siegel, J. B. Protein Stability Prediction by Fine-Tuning a Protein Language Model on a Mega-Scale Dataset. *PLoS Comput. Biol.* **2024**, *20* (7), No. e1012248.
(68) Zheng, L.; Li, B.; Xu, S.; Chen, J.; Liang, G. FEDKEA: Enzyme Function Prediction with a Large Pretrained Protein Language Model and Distance-Weighted K-Nearest Neighbor. *bioRxiv* **2024**, No. 2024.08.12.604109, DOI: 10.1101/2024.08.12.604109.
(69) Shrestha, P.; Kandel, J.; Tayara, H.; Chong, K. T. Post-Translational Modification Prediction via Prompt-Based Fine-Tuning of a GPT-2 Model. *Nat. Commun.* **2024**, *15* (1), No. 6699.
(70) Luo, Y.; Nie, Z.; Hong, M.; Zhao, S.; Zhou, H.; Nie, Z. MutaPLM: Protein Language Modeling for Mutation Explanation and Engineering. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 79783−79818.
(71) He, L.; Jin, P.; Min, Y.; Xie, S.; Wu, L.; Qin, T.; Liang, X.; Gao, K.; Jiang, Y.; Liu, T.-Y. SFM-Protein: Integrative Co-Evolutionary Pre-Training for Advanced Protein Sequence Representation, arXiv:2410.24022. arXiv.org e-Print archive. https://arxiv.org/abs/2410.24022. (accessed April 21, 2025). 2024.
(72) Barkan, E.; Siddiqui, I.; Cheng, K. J.; Golts, A.; Shoshan, Y.; Weber, J. K.; Campos Mota, Y.; Ozery-Flato, M.; Sautto, G. A. Leveraging Large Language Models to Predict Antibody Biological Activity against Influenza A Hemagglutinin. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 1286−1295.
(73) Kalemati, M.; Noroozi, A.; Shahbakhsh, A.; Koohi, S. ParaAntiProt Provides Paratope Prediction Using Antibody and Protein Language Models. *Sci. Rep.* **2024**, *14* (1), No. 29141.
(74) Heinzinger, M.; Weissenow, K.; Sanchez, J. G.; Henkel, A.; Mirdita, M.; Steinegger, M.; Rost, B. Bilingual Language Model for Protein Sequence and Structure. *NAR:Genomics Bioinf.* **2024**, *6* (4), No. lqae150.
(75) Wang, Q.; Feng, Y.; Wang, Y.; Li, B.; Wen, J.; Zhou, X.; Song, Q. AntiFormer: Graph Enhanced Large Language Model for Binding Affinity Prediction. *Briefings Bioinf.* **2024**, *25* (5), No. bbae403, DOI: 10.1093/bib/bbae403.
(76) Nagano, Y.; Pyo, A. G. T.; Milighetti, M.; Henderson, J.; Shawe-Taylor, J.; Chain, B.; Tiffeau-Mayer, A. Contrastive Learning of T Cell Receptor Representations. *Cell Syst.* **2025**, *16* (1), No. 101165.
(77) Zhao, Y.; Yu, J.; Su, Y.; Shu, Y.; Ma, E.; Wang, J.; Jiang, S.; Wei, C.; Li, D.; Huang, Z.; Cheng, G.; Ren, H.; Feng, J. A Unified Deep Framework for Peptide−major Histocompatibility complex−T Cell Receptor Binding Prediction. *Nat. Mach. Intell.* **2025**, *7* (4), 650−660.
(78) Ullanat, V.; Jing, B.; Sledzieski, S.; Berger, B. Learning the Language of Protein-Protein Interactions. *bioRxiv* **2025**, DOI: 10.1101/2025.03.09.642188.
(79) Peng, F. Z.; Wang, C.; Chen, T.; Schussheim, B.; Vincoff, S.; Chatterjee, P. PTM-Mamba: A PTM-Aware Protein Language Model with Bidirectional Gated Mamba Blocks. *Nat. Methods* **2025**, *22*, No. 945.
(80) Gurusinghe, S. N. S.; Wu, Y.; DeGrado, W.; Shifman, J. M. ProBASS - a Language Model with Sequence and Structural Features for Predicting the Effect of Mutations on Binding Affinity. *bioRxiv* **2024**, DOI: 10.1101/2024.06.21.600041.
(81) Regan, L.; Caballero, D.; Hinrichsen, M. R.; Virrueta, A.; Williams, D. M.; O'Hern, C. S. Protein Design: Past, Present, and Future: Protein Design: Past, Present, and Future. *Biopolymers* **2015**, *104* (4), 334−350.
(82) Tobin, M. B.; Gustafsson, C.; Huisman, G. W. Directed Evolution: The "Rational" Basis for "Irrational" Design. *Curr. Opin. Struct. Biol.* **2000**, *10* (4), 421−427.
(83) Pongsupasa, V.; Anuwan, P.; Maenpuen, S.; Wongnate, T. Rational-Design Engineering to Improve Enzyme Thermostability. *Methods Mol. Biol. (Clifton, N.J.)* **2022**, *2397*, 159−178.
(84) Song, Z.; Zhang, Q.; Wu, W.; Pu, Z.; Yu, H. Rational Design of Enzyme Activity and Enantioselectivity. *Front. Bioeng. Biotechnol.* **2023**, *11*, No. 1129149.
(85) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (12), 866−876.
(86) Steiner, K.; Schwab, H. Recent Advances in Rational Approaches for Enzyme Engineering. *Comput. Struct. Biotechnol. J.* **2012**, *2*, No. e201209010, DOI: 10.5936/csbj.201209010.
(87) Sarumi, O. A.; Heider, D. Large Language Models and Their Applications in Bioinformatics. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 3498−3505.
(88) Kulmanov, M.; Guzmán-Vega, F. J.; Duek Roggli, P.; Lane, L.; Arold, S. T.; Hoehndorf, R. Protein Function Prediction as Approximate Semantic Entailment. *Nat. Mach. Intell.* **2024**, *6* (2), 220−228.
(89) Capela, J.; Zimmermann-Kogadeeva, M.; van Dijk, A. D. J.; de Ridder, D.; Dias, O.; Rocha, M. Comparative Assessment of Protein Large Language Models for Enzyme Commission Number Prediction. *BMC Bioinf.* **2025**, *26* (1), 1−21.
(90) Hwang, Y.; Cornman, A. L.; Kellogg, E. H.; Ovchinnikov, S.; Girguis, P. R. Genomic Language Model Predicts Protein Co-Regulation and Function. *Nat. Commun.* **2024**, *15* (1), No. 2880.
(91) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. BERTology Meets Biology: Interpreting Attention in Protein Language Models, arXiv:2006.15222. arXiv.org e-Print archive. https://arxiv.org/abs/2006.15222 (accessed April 28, 2025). 2020.
(92) Littmann, M.; Heinzinger, M.; Dallago, C.; Weissenow, K.; Rost, B. Protein Embeddings and Deep Learning Predict Binding Residues for Various Ligand Classes. *Sci. Rep.* **2021**, *11* (1), No. 23916.
(93) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. In *International Conference on Machine Learning*; PMLR, 2021; pp 8844−8856.
(94) Le, V.-T.; Zhan, Z.-J.; Vu, T.-T.-P.; Malik, M.-S.; Ou, Y.-Y. ProtTrans and Multi-Window Scanning Convolutional Neural

Networks for the Prediction of Protein-Peptide Interaction Sites. *J. Mol. Graphics Modell.* **2024**, *130*, No. 108777.
(95) Zhang, Z.; Wayment-Steele, H. K.; Brixi, G.; Wang, H.; Kern, D.; Ovchinnikov, S. Protein Language Models Learn Evolutionary Statistics of Interacting Sequence Motifs. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121* (45), No. e2406285121, DOI: 10.1073/pnas.2406285121.
(96) Cocco, S.; Monasson, R.; Weigt, M. From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes Are Needed for Structure Prediction. *PLoS Comput. Biol.* **2013**, *9* (8), No. e1003176.
(97) de Juan, D.; Pazos, F.; Valencia, A. Emerging Methods in Protein Co-Evolution. *Nat. Rev. Genet.* **2013**, *14* (4), 249−261.
(98) Huang, B.; Kong, L.; Wang, C.; Ju, F.; Zhang, Q.; Zhu, J.; Gong, T.; Zhang, H.; Yu, C.; Zheng, W.-M.; Bu, D. Protein Structure Prediction: Challenges, Advances, and the Shift of Research Paradigms. *Genomics, Proteomics Bioinf.* **2023**, *21* (5), 913−925.
(99) Qi, G.; Tollefson, M. R.; Gogal, R. A.; Smith, R. J. H.; AlQuraishi, M.; Schnieders, M. J. Protein Structure Prediction Using a Maximum Likelihood Formulation of a Recurrent Geometric Network. *bioRxiv* **2021**, No. 2021.09.03.458873, DOI: 10.1101/2021.09.03.458873.
(100) Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and Structure-Based Prediction of Eukaryotic Protein Phosphorylation Sites. *J. Mol. Biol.* **1999**, *294* (5), 1351−1362.
(101) Hornbeck, P. V.; Kornhauser, J. M.; Latham, V.; Murray, B.; Nandhikonda, V.; Nord, A.; Skrzypek, E.; Wheeler, T.; Zhang, B.; Gnad, F. 15 Years of PhosphoSitePlus: Integrating Post-Translationally Modified Sites, Disease Variants and Isoforms. *Nucleic Acids Res.* **2019**, *47* (D1), D433−D441, DOI: 10.1093/nar/gky1159.
(102) Lee, T. Y.; Lin, Z. Q.; Hsieh, S. J.; Bretaña, N. A.; Lu, C. T. Exploiting Maximal Dependence Decomposition to Identify Conserved Motifs from a Group of Aligned Signal Sequences. *Bioinformatics (Oxford, England)* **2011**, *27* (13), 1780−1787.
(103) Kim, J. H.; Lee, J.; Oh, B.; Kimm, K.; Koh, I. Prediction of Phosphorylation Sites Using SVMs. *Bioinformatics (Oxford, England)* **2004**, *20* (17), 3179−3184.
(104) Wandall, H. H.; Nielsen, M. A. I.; King-Smith, S.; de Haan, N.; Bagdonaite, I. Global Functions of O-Glycosylation: Promises and Challenges in O-Glycobiology. *FEBS J.* **2021**, *288* (24), 7183−7212.
(105) Wilkinson, H.; Saldova, R. Current Methods for the Characterization of O-Glycans. *J. Proteome Res.* **2020**, *19* (10), 3890−3905.
(106) Fu, H.; Yang, Y.; Wang, X.; Wang, H.; Xu, Y. DeepUbi: A Deep Learning Framework for Prediction of Ubiquitination Sites in Proteins. *BMC Bioinf.* **2019**, *20* (1), 1−10.
(107) Thapa, N.; Chaudhari, M.; McManus, S.; Roy, K.; Newman, R. H.; Saigo, H.; Kc, D. B. DeepSuccinylSite: A Deep Learning Based Approach for Protein Succinylation Site Prediction. *BMC Bioinf.* **2020**, *21* (3), 1−10.
(108) Jia, J.; Wu, G.; Li, M.; Qiu, W. pSuc-EDBAM: Predicting Lysine Succinylation Sites in Proteins Based on Ensemble Dense Blocks and an Attention Module. *BMC Bioinf.* **2022**, *23*, No. 450.
(109) Chen, Y.-Z.; Wang, Z.-Z.; Wang, Y.; Ying, G.; Chen, Z.; Song, J. nhKcr: A New Bioinformatics Tool for Predicting Crotonylation Sites on Human Nonhistone Proteins Based on Deep Learning. *Briefings Bioinf.* **2021**, *22* (6), No. bbab146.
(110) Khanal, J.; Kandel, J.; Tayara, H.; Chong, K. T. CapsNh-Kcr: Capsule Network-Based Prediction of Lysine Crotonylation Sites in Human Non-Histone Proteins. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 120−127.
(111) Liu, Y.; Liu, Y.; Wang, G.-A.; Cheng, Y.; Bi, S.; Zhu, X. BERT-Kgly: A Bidirectional Encoder Representations From Transformers (BERT)-Based Model for Predicting Lysine Glycation Site for Homo Sapiens. *Front. Bioinform.* **2022**, *2*, No. 834153.
(112) Indriani, F.; Mahmudah, K. R.; Purnama, B.; Satou, K. ProtTrans-Glutar: Incorporating Features From Pre-Trained Transformer-Based Models for Predicting Glutarylation Sites. *Front. Genet.* **2022**, *13*, No. 885929.
(113) Bennett, E. P.; Mandel, U.; Clausen, H.; Gerken, T. A.; Fritz, T. A.; Tabak, L. A. Control of Mucin-Type O-Glycosylation: A Classification of the Polypeptide GalNAc-Transferase Gene Family. *Glycobiology* **2012**, *22* (6), 736−756.
(114) Li, G. X. H.; Vogel, C.; Choi, H. PTMscape: An Open Source Tool to Predict Generic Post-Translational Modifications and Map Modification Crosstalk in Protein Domains and Biological Processes. *Mol. Omics* **2018**, *14* (3), 197−209.
(115) Jacques, F.; Bolivar, P.; Pietras, K.; Hammarlund, E. U. Roadmap to the Study of Gene and Protein Phylogeny and evolution—A Practical Guide. *PLoS One* **2023**, *18* (2), No. e0279597.
(116) Stamatakis, A.; Ludwig, T.; Meier, H. RAxML-III: A Fast Program for Maximum Likelihood-Based Inference of Large Phylogenetic Trees. *Bioinformatics* **2005**, *21* (4), 456−463.
(117) Spirin, S.; Sigorskikh, A.; Efremov, A.; Penzar, D.; Karyagina, A. PhyloBench: A Benchmark for Evaluating Phylogenetic Programs. *Mol. Biol. Evol.* **2024**, *41* (6), No. msae084, DOI: 10.1093/molbev/msae084.
(118) Adzhubei, I.; Jordan, D. M.; Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*, No. Unit7.20.
(119) Ng, P. C.; Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **2001**, *11* (5), 863−874.
(120) Liu, X.; Yang, X.; Ouyang, L.; Guo, G.; Su, J.; Xi, R.; Yuan, K.; Yuan, F. Protein Language Model Predicts Mutation Pathogenicity and Clinical Prognosis. *bioRxiv* **2022**, No. 2022.09.30.510294, DOI: 10.1101/2022.09.30.510294.
(121) Wang, L.; Li, X.; Zhang, H.; Wang, J.; Jiang, D.; Xue, Z.; Wang, Y. A Comprehensive Review of Protein Language Models, arXiv:2502.06881. arXiv.org e-Print archive. https://arxiv.org/abs/2502.06881, 2025.
(122) Lin, W.; Wells, J.; Wang, Z.; Orengo, C.; Martin, A. C. R. VariPred: Enhancing Pathogenicity Prediction of Missense Variants Using Protein Language Models. *bioRxiv* **2023**, No. 2023.03.16.532942, DOI: 10.1101/2023.03.16.532942.
(123) James, J. K.; Norland, K.; Johar, A. S.; Kullo, I. J. Deep Generative Models of LDLR Protein Structure to Predict Variant Pathogenicity. *J. Lipid Res.* **2023**, *64* (12), No. 100455.
(124) Sun, Y.; Shen, Y. Structure-Informed Protein Language Models Are Robust Predictors for Variant Effects. *Hum. Genet.* **2025**, *144* (2−3), 209−225.
(125) Dieckhaus, H.; Brocidiacono, M.; Randolph, N. Z.; Kuhlman, B. Transfer Learning to Leverage Larger Datasets for Improved Prediction of Protein Stability Changes. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121* (6), No. e2314853121.
(126) Tan, Y.; Zhou, B.; Zheng, L.; Fan, G.; Hong, L. Semantical and Geometrical Protein Encoding Toward Enhanced Bioactivity and Thermostability. *bioRxiv* **2025**, No. 2023.12.01.569522, DOI: 10.1101/2023.12.01.569522.
(127) Tcherkasskaya, O.; Eugene, D.; Vladimir, U. Biophysical Constraints for Protein Structure Prediction. *J. Proteome Res.* **2003**, *2* (1), 37−42, DOI: 10.1021/pr025552q.
(128) Zhang, Z.; Lu, J.; Chenthamarakshan, V.; Lozano, A.; Das, P.; Tang, J. Structure-Informed Protein Language Model, arXiv:2402.05856. arXiv.org e-Print archive. https://arxiv.org/abs/2402.05856 (accessed April 21, 2025). 2024.
(129) Paranou, D.; Chatzigoulas, A.; Cournia, Z. Using Deep Learning and Large Protein Language Models to Predict Protein-Membrane Interfaces of Peripheral Membrane Proteins. *Bioinf. Adv.* **2024**, *4* (1), No. vbae078.
(130) Wasim, A.; Pramanik, U.; Das, A.; Latua, P.; Rudra, J. S.; Mondal, J. Harnessing Transformers to Generate Protein Sequences Prone to Liquid Liquid Phase Separation. *bioRxiv* **2024**, No. 2024.03.02.583105, DOI: 10.1101/2024.03.02.583105.
(131) Mall, R.; Kaushik, R.; Martinez, Z. A.; Thomson, M. W.; Castiglione, F. Benchmarking Protein Language Models for Protein Crystallization. *Sci. Rep.* **2025**, *15* (1), No. 2381.
(132) Gelman, S.; Johnson, B.; Freschlin, C. R.; D'Costa, S.; Gitter, A.; Romero, P. Green Fluorescent Protein Engineering with a

Biophysics-Based Protein Language Model *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design* 2024.
(133) Hallee, L.; Gleghorn, J. P. Protein-Protein Interaction Prediction Is Achievable with Large Language Models. *bioRxiv* **2023**, No. 2023.06.07.544109, DOI: 10.1101/2023.06.07.544109.
(134) Gao, Q.; Zhang, C.; Li, M.; Yu, T. Protein-Protein Interaction Prediction Model Based on ProtBert-BiGRU-Attention. *J. Comput. Biol.* **2024**, *31* (9), 797−814.
(135) Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A. L.; Fang, T.; Doncheva, N. T.; Pyysalo, S.; Bork, P.; Jensen, L. J.; von Mering, C. The STRING Database in 2023: Protein-Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. *Nucleic Acids Res.* **2023**, *51* (D1), D638−D646, DOI: 10.1093/nar/gkac1000.
(136) Jankauskaitė, J.; Jiménez-García, B.; Dapkunas, J.; Fernández-Recio, J.; Moal, I. H. SKEMPI 2.0: An Updated Benchmark of Changes in Protein-Protein Binding Energy, Kinetics and Thermodynamics upon Mutation. *Bioinformatics* **2019**, *35* (3), 462−469.
(137) Wang, B.; Li, W. Advances in the Application of Protein Language Modeling for Nucleic Acid Protein Binding Site Prediction. *Genes* **2024**, *15* (8), No. 1090.
(138) Gurusinghe, S. N. S.; Wu, Y.; DeGrado, W.; Shifman, J. M. ProBASS—a Language Model with Sequence and Structural Features for Predicting the Effect of Mutations on Binding Affinity. *Bioinformatics* **2025**, *41* (5), No. btaf270.
(139) Wu, L. C.; Tuot, D. S.; Lyons, D. S.; Garcia, K. C.; Davis, M. M. Two-Step Binding Mechanism for T-Cell Receptor Recognition of Peptide MHC. *Nature* **2002**, *418* (6897), 552−556.
(140) Paiano, A.; Margiotta, A.; De Luca, M.; Bucci, C. Yeast Two-Hybrid Assay to Identify Interacting Proteins. *Curr. Protoc. Protein Sci.* **2019**, *95* (1), No. e70.
(141) Meng, F.; Zhou, N.; Hu, G.; Liu, R.; Zhang, Y.; Jing, M.; Hou, Q. A Comprehensive Overview of Recent Advances in Generative Models for Antibodies. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2648−2660.
(142) Nielsen, M.; Eugster, A.; Jensen, M. F.; Goel, M.; Tiffeau-Mayer, A.; Pelissier, A.; Valkiers, S.; Rodríguez Martínez, M.; Meynard-Piganeau, B.; Greiff, V.; Mora, T.; Walczak, A. M.; Croce, G.; Moreno, D. L.; Gfeller, D.; Meysman, P.; Barton, J. Lessons Learned from the IMMREP23 TCR-Epitope Prediction Challenge. *ImmunoInformatics* **2024**, *16*, No. 100045.
(143) Ruffolo, J. A.; Chu, L.-S.; Mahajan, S. P.; Gray, J. J. Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies. *Nat. Commun.* **2023**, *14* (1), No. 2389.
(144) Chen, Y.; Wang, Y.; Chen, Y.; Cheng, Y.; Wei, Y.; Li, Y.; Wang, J.; Wei, Y.; Chan, T.-F.; Li, Y. Deep Autoencoder for Interpretable Tissue-Adaptive Deconvolution and Cell-Type-Specific Gene Analysis. *Nat. Commun.* **2022**, *13* (1), No. 6735.
(145) Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S.*A ConvNet for the 2020s*, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 DOI: 10.1109/CVPR52688.2022.01167.
(146) Wang, M.; Patsenker, J.; Li, H.; Kluger, Y.; Kleinstein, S. H. Supervised Fine-Tuning of Pre-Trained Antibody Language Models Improves Antigen Specificity Prediction. *PLoS Comput. Biol.* **2025**, *21* (3), No. e1012153.
(147) Holt, C. M.; Janke, A. K.; Amlashi, P. B.; Marinov, T. M.; Georgiev, I. S. Contrastive Learning Enables Epitope Overlap Predictions for Targeted Antibody Discovery. *bioRxiv* **2025**, DOI: 10.1101/2025.02.25.640114.
(148) Wang, D.; Fei, Y. E.; Zhou, H. *On Pre-Training Language Model for Antibody*, Eleventh International Conference on Learning Representations; 2023.
(149) Li, S.; Meng, X.; Li, R.; Huang, B.; Wang, X. Correction: NanoBERTa-ASP: Predicting Nanobody Paratope Based on a Pretrained RoBERTa Model. *BMC Bioinf.* **2024**, *25* (1), No. 190.
(150) Shoshan, Y.; Raboh, M.; Ozery-Flato, M.; Ratner, V.; Golts, A.; Weber, J. K.; Barkan, E.; Rabinovici-Cohen, S.; Polaczek, S.; Amos, I.; Shapira, B.; Hazan, L.; Ninio, M.; Ravid, S.; Danziger, M. M.; Shamay, Y.; Kurant, S.; Morrone, J. A.; Suryanarayanan, P.; Rosen-Zvi, M.; Hexter, E. MAMMAL -- Molecular Aligned Multi-Modal Architecture and Language, arXiv:2410.22367. arXiv.org e-Print archive. https://arxiv.org/abs/2410.22367 (accessed Sept 19, 2025). 2024.
(151) Ding, F.; Steinhardt, J. Protein Language Models Are Biased by Unequal Sequence Sampling across the Tree of Life *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design* 2024.
(152) Bushuiev, A.; Bushuiev, R.; Sedlar, J.; Pluskal, T.; Damborsky, J.; Mazurenko, S.; Sivic, J. Revealing Data Leakage in Protein Interaction Benchmarks, arXiv:2404.10457. arXiv.org e-Print archive. https://arxiv.org/abs/2404.10457 (accessed Sept 25, 2025). 2024.
(153) Hermann, L.; Fiedler, T.; Nguyen, H. A.; Nowicka, M.; Bartoszewicz, J. M. Beware of Data Leakage from Protein LLM Pretraining. In *Machine Learning in Computational Biology*; PMLR, 2024; pp 106−116.
(154) Joeres, R.; Blumenthal, D. B.; Kalinina, O. V. Data Splitting to Avoid Information Leakage with DataSAIL. *Nat. Commun.* **2025**, *16*, No. 3337.
(155) Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4* (9), No. 100804.
(156) Dotan, E.; Jaschek, G.; Pupko, T.; Belinkov, Y. Effect of Tokenization on Transformers for Biological Sequences. *Bioinformatics (Oxford, England)* **2024**, *40* (4), No. btae196, DOI: 10.1093/bioinformatics/btae196.
(157) Zhang, Y.; Okumura, M. ProtHyena: A Fast and Efficient Foundation Protein Language Model at Single Amino Acid Resolution. *bioRxiv* **2024**, No. 2024.01.18.576206, DOI: 10.1101/2024.01.18.576206.
(158) Gelman, S.; Johnson, B.; Freschlin, C.; Sharma, A.; D'Costa, S.; Peters, J.; Gitter, A.; Romero, P. A. Biophysics-Based Protein Language Models for Protein Engineering. *bioRxiv* **2025**, No. 2024.03.15.585128, DOI: 10.1101/2024.03.15.585128.
(159) Wu, F.; Wu, L.; Radev, D.; Xu, J.; Li, S. Z. Integration of Pre-Trained Protein Language Models into Geometric Deep Learning Networks. *Commun. Biol.* **2023**, *6* (1), No. 876.
(160) Cai, T.; Xie, L.; Zhang, S.; Chen, M.; He, D.; Badkul, A.; Liu, Y.; Namballa, H. K.; Dorogan, M.; Harding, W. W.; Mura, C.; Bourne, P. E.; Xie, L. End-to-End Sequence-Structure-Function Meta-Learning Predicts Genome-Wide Chemical-Protein Interactions for Dark Proteins. *PLoS Comput. Biol.* **2023**, *19* (1), No. e1010851, DOI: 10.1371/journal.pcbi.1010851.
(161) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-Tuning Protein Language Models Boosts Predictions across Diverse Tasks. *Nat. Commun.* **2024**, *15* (1), No. 7407.
(162) Chen, J.-Y.; Wang, J.-F.; Hu, Y.; Li, X.-H.; Qian, Y.-R.; Song, C.-L. Evaluating the Advancements in Protein Language Models for Encoding Strategies in Protein Function Prediction: A Comprehensive Review. *Front. Bioeng. Biotechnol.* **2025**, *13*, No. 1506508.
(163) Le, N. Q. K. Leveraging Transformers-Based Language Models in Proteome Bioinformatics. *Proteomics* **2023**, *23* (23−24), No. 2300011, DOI: 10.1002/pmic.202300011.
(164) Huang, T.; Li, Y. Current Progress, Challenges, and Future Perspectives of Language Models for Protein Representation and Protein Design. *Innovation* **2023**, *4* (4), No. 100446.
(165) Nambiar, A.; Forsyth, J. M.; Liu, S.; Maslov, S. DR-BERT: A Protein Language Model to Annotate Disordered Regions. *Structure* **2024**, *32* (8), 1260−1268.e3.
(166) Lombardi, G.; Seoane, B.; Carbone, A. LoRA-DR-Suite: Adapted Embeddings Predict Intrinsic and Soft Disorder from Protein Sequences. *Bioinformatics* **2025**, *41* (Supplement_1), i439−i448.
(167) Lombard, V.; Timsit, D.; Grudinin, S.; Laine, E. SeaMoon: From Protein Language Models to Continuous Structural Heterogeneity. *Structure* **2025**, *33* (9), 1577−1590.e8.

(168) Prillo, S.; Wu, W.; Song, Y. S. Ultrafast Classical Phylogenetic Method Beats Large Protein Language Models on Variant Effect Prediction. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 130265−130290.
(169) Minton, K. Predicting Variant Pathogenicity with AlphaMissense. *Nat. Rev. Genet.* **2023**, *24* (12), 804.
(170) Xiao, H.; Lin, W.; Chen, X.; Wang, H.; Chen, K.; Li, J.; Sun, Y.; Dai, S.; Wu, B.; Ye, Q. STELLA: Towards Protein Function Prediction with Multimodal LLMs Integrating Sequence-Structure Representations, arXiv:2506.03800. arXiv.org e-Print archive. https://arxiv.org/abs/2506.03800 (accessed Sept 26, 2025). 2025.
(171) Wang, Z.; Ma, Z.; Cao, Z.; Zhou, C.; Zhang, J.; Gao, Y. Q. Prot2Chat: Protein Large Language Model with Early Fusion of Text, Sequence, and Structure. *Bioinformatics* **2025**, *41* (8), No. btaf396, DOI: 10.1093/bioinformatics/btaf396.
(172) Fei, X.; Chatzianastasis, M.; Carneiro, S. A.; Abdine, H.; Petalidis, L. P.; Vazirgiannis, M. Prot2Text-V2: Protein Function Prediction with Multimodal Contrastive Alignment, arXiv:2505.11194. arXiv.org e-Print archive. https://arxiv.org/abs/2505.11194 (accessed Sept 26, 2025). 2025.

https://doi.org/10.1021/acs.jproteome.5c00506
J. Proteome Res. 2026, 25, 507−524