https://doi.org/10.1038/s43588-026-00993-z

Protein language models for structural biology

Chenxiao Xiang, Bin Cheng, Zhenling Peng & Jianyi Yang

Protein language models can effectively decode evolution's grammar, making structure prediction and design scalable. This transformative capability is accelerating biological discovery and engineering across all scales.

Structural biology seeks to elucidate the three-dimensional (3D) architectures of biomolecules to understand function, interactions, and mechanisms at the atomic level. Experimental techniques such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy (cryo-EM) have yielded over 200 thousand entries in the Protein Data Bank as of early 2026. Yet these structures represent only a tiny fraction of known proteomes. Meanwhile, genomic and metagenomic sequencing have created a vast 'sequence universe' that far outpaces experimental structure determination.

Bridging this enormous gap demands accurate, scalable methods for structure prediction. Protein language models (PLMs) have emerged as a powerful solution, by learning to decode the rich evolutionary, structural, and functional information embedded within protein sequences. Originally developed for natural language processing, large language models (LLMs) have been repurposed as PLMs that treat amino acid sequences as a biological language. The central insight is that a protein sequence — composed of a 20-amino-acid alphabet — exhibits contextual dependencies and semantics analogous to human language[1] (Fig. 1a).

Figure 1b presents a visual timeline of key milestones, spanning from the introduction of the transformer architecture in 2017[2] to the latest advances in agentic artificial intelligence (AI) models in 2026. By pretraining on millions to billions of unlabeled protein sequences, PLMs such as the ESM series[1,3,4], learn evolutionary patterns, co-evolutionary signals, and structural constraints through self-supervised objectives like masked token prediction. One of the most profound findings is that structural and functional properties emerge as a byproduct of training on sequence data alone[1]. This emergent capability makes PLMs powerful as highly compressed, differentiable databases of evolutionary information.

Single-sequence PLM methods for efficient structure prediction

Concepts from language models have increasingly influenced AI-based structure prediction methods. Two representative examples — AlphaFold2[5] and RoseTTAFold[6] — employ transformer-inspired components. However, a key limitation of these early methods is their heavy reliance on multiple sequence alignments (MSAs) derived from evolutionarily related proteins. For so-called 'orphan' proteins, homologous sequences may be scarce or difficult to identify — a common challenge for many viral proteins.

PLMs offer a complementary — and in some cases superior — alternative, as they can learn and encode evolutionary constraints from large-scale sequence data during pretraining. For example, single-sequence structure prediction methods such as RGN2[7] and trRosettaX-Single[8] leverage contextual embeddings from PLMs: each amino acid is mapped to a context-aware vector that encodes sequence patterns and structural propensities, enabling accurate structure prediction from a single sequence. These methods achieve higher accuracy than AlphaFold2 for orphan proteins. In parallel, ESMFold[3] — built on the ESM2 language model — demonstrates that high-accuracy structures can be predicted directly from a single sequence, reaching inference speeds orders of magnitude faster than MSA-based approaches. As a result, ESMFold has enabled high-throughput structure prediction at the proteome and metagenome scales.

Together, these tools have transformed structural biology by dramatically expanding structural coverage and enriching protein structure databases. As of early 2026, the AlphaFold Database contains over 200 million predicted structures, and the ESM Metagenomic Atlas provides more than 700 million predictions by ESMFold — in comparison to ~200 thousand experimentally determined structures before AlphaFold. These resources are facilitating discoveries in drug design, enzyme engineering, and disease variant interpretation.

PLMs for inverse folding and protein design

Inverse folding — which generates amino acid sequences likely to fold into a desired 3D backbone — is central to de novo protein engineering and optimization. Early structure-based models, such as ESM-IF[9], demonstrated strong sequence recovery by learning geometric and physical constraints directly from backbone coordinates. Pure sequence-based PLMs such as ProtGPT2[10] excel at unconditional de novo generation by learning the underlying evolutionary distribution of protein sequences, enabling the sampling of novel yet natural-like proteins that can fold into stable structures.

A major step forward came with ESM3[4], a frontier multimodal generative language model that simultaneously reasons over three fundamental tracks: sequence, structure, and function. Unlike previous protein design models that operate on a single modality — either generating sequences from structures or designing sequences alone — ESM3 uses a multi-track transformer with discrete tokenization for all three aspects. ESM3 also supports controllable generation at atomic precision, such as coordinating distant residues in 3D space, and has demonstrated the ability to simulate vast evolutionary distances. Notably, it generated esmGFP, a bright fluorescent protein sharing only 58% sequence identity with natural counterparts — an evolutionary leap estimated as equivalent to 500 million years.

PLMs for integrative modeling with experiments

PLMs have extended their impact into integrative structural modeling. By providing rich, high-dimensional sequence embeddings that encode evolutionary, biophysical, and functional information, PLMs serve as powerful feature extractors that can be seamlessly fused with experimental data. For instance, in model building methods such as ModelAngelo[11] and CryoAtom[12,13], ESM embeddings are shown to

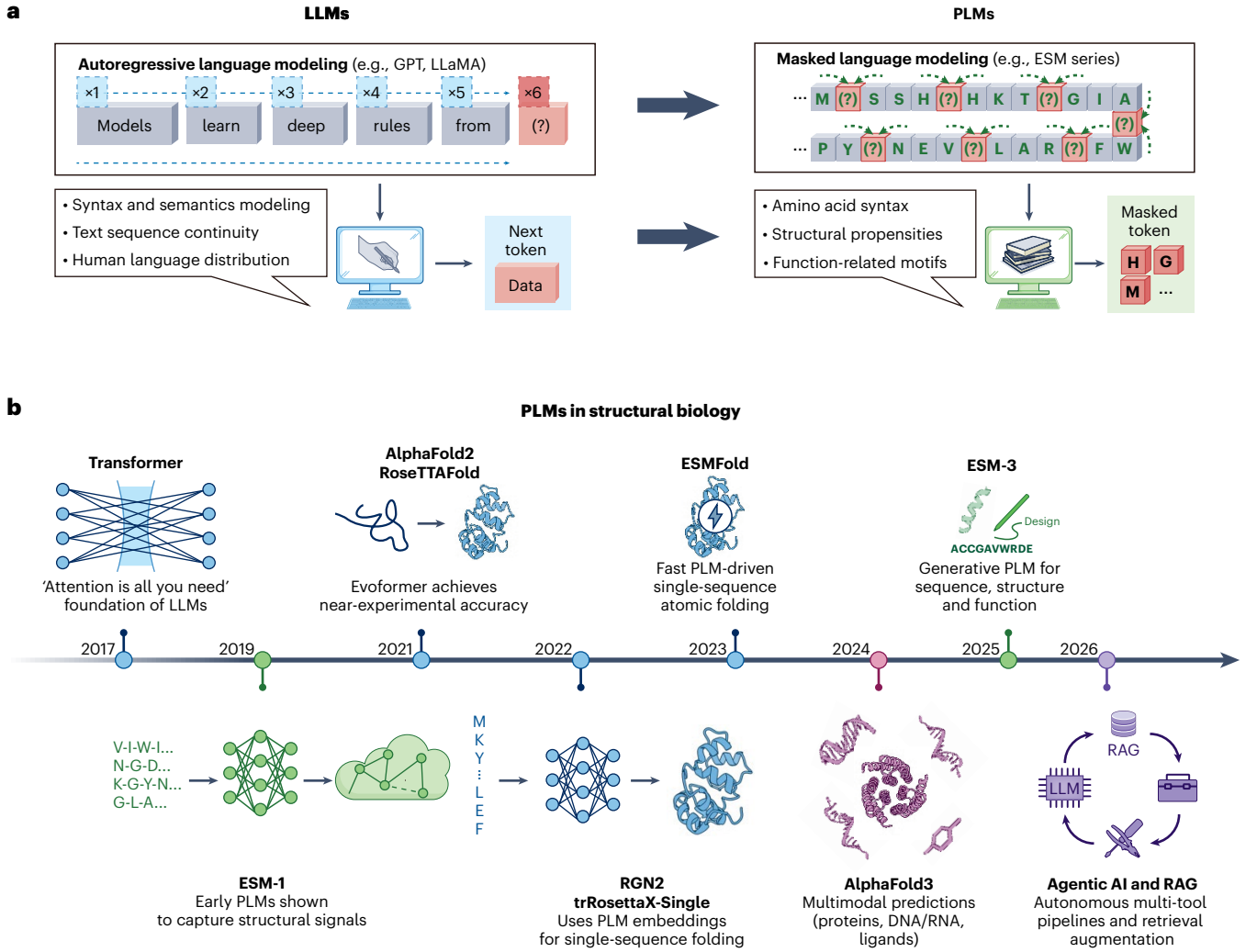

Fig. 1 | Protein language models: conceptual framework and evolution in structural biology. a, Schematic illustration of how protein sequences are modeled using language modeling frameworks, where amino acid sequences are treated as a biological language. **b**, Timeline of developments in protein language models in structural biology, from the introduction of the transformer architecture in 2017 to recent advances in multimodal and agentic AI models.

be important for correct assignment of amino acid types, implemented through cross-attention mechanisms between sequence and cryo-EM density maps. This integration improves the accuracy and completeness of atomic models derived from experimental data.

Challenges and opportunities in PLMs

Generalization to non-protein biomolecules and biomolecular interactions. Most language models in structural biology are designed exclusively for proteins. However, the functional units of the cell often involve nucleic acids (DNA, RNA), sugars, ligands, ions, and post-translational modifications. Predicting the structure and interactions of these hybrid complexes remains a frontier challenge. Models like AlphaFold3[14] are pioneering this area by incorporating separate tracks for different molecular modalities, yet data scarcity for such complexes poses a substantial hurdle. Developing a unified model capable of handling all classes of biomolecules represents a promising path forward.

Hallucination. Like their natural language processing counterparts, PLMs can 'hallucinate' — generating sequences or predicting structures that appear plausible but are non-functional or physically impossible. While generative models can produce a vast array of designs, experimental validation often reveals that only a fraction are soluble, stable, or active. Distinguishing functionally correct sequences from merely plausible ones remains an open problem, and there is a growing need for uncertainty quantification to prevent overconfidence in in silico predictions.

Interpretability. The 'black box' nature of PLMs poses a fundamental interpretability challenge. Recent studies have begun to address this by applying sparse autoencoders to decompose ESM embeddings into interpretable features, revealing that specific latent dimensions correspond to structural motifs, binding sites, or evolutionary constraints[15]. Nevertheless, much work remains to fully understand the internal representations learned by PLMs and to translate these insights into mechanistic biological understanding.

Chenxiao Xiang[1,2], Bin Cheng[1,2], Zhenling Peng[1] ✉ & Jianyi Yang[1] ✉

[1]MOE Frontiers Science Center for Nonlinear Expectations, State Key Laboratory of Cryptography and Digital Economy Security, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China. [2]These authors contributed equally: Chenxiao Xiang, Bin Cheng.

✉e-mail: zhenling@email.sdu.edu.cn; yangjy@sdu.edu.cn

Published online: 28 May 2026

References

1. Rives, A. et al. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
2. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* (NIPS, 2017).
3. Lin, Z. et al. *Science* **379**, 1123–1130 (2023).
4. Hayes, T. et al. *Science* **387**, 850–858 (2025).
5. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
6. Baek, M. et al. *Science* **373**, 871–876 (2021).
7. Chowdhury, R. et al. *Nat. Biotechnol.* **40**, 1617–1623 (2022).
8. Wang, W., Peng, Z. & Yang, J. *Nat. Comput. Sci.* **2**, 804–814 (2022).
9. Hsu, C. et al. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning* 8946–8970 (PMLR, 2022).
10. Ferruz, N., Schmidt, S. & Hocker, B. *Nat. Commun.* **13**, 4348 (2022).
11. Jamali, K. et al. *Nature* **628**, 450–457 (2024).
12. Su, B. et al. *Nat. Struct. Mol. Biol.* **33**, 351–361 (2026).
13. Su, B. et al. Preprint at LangTaoSha Preprint Server https://doi.org/10.65215/ltspreprints.2026.04.14.000185 (2026).
14. Abramson, J. et al. *Nature* **630**, 493–500 (2024).
15. Simon, E. & Zou, J. *Nat. Methods* **22**, 2107–2117 (2025).

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2024YFA0916901), National Natural Science Foundation of China (T2225007 and 32430063), the Fundamental Research Funds for the Central Universities.

Competing interests

The authors declare no competing interests.