ProteinAligner: A Tri-Modal Contrastive Learning Framework for Protein Representation Learning

Li Zhang [*1] **Han Guo** [*1] **Leah Schaffer** [2] **Young Su Ko** [3] **Digvijay Singh** [4] **Danielle Grotjahn** [5] **Elizabeth Villa** [4 6] **Michael Gilson** [7] **Wei Wang** [3 8] **Trey Ideker** [2 9] **Eric Xing** [10 11] **Pengtao Xie** [1 2 11]

Abstract

Protein foundation models, particularly protein language models, have shown strong success in learning meaningful protein representations using transformer architectures pretrained on large-scale datasets through self-supervised learning. These representations have proven effective for downstream tasks such as predicting protein functions and properties. However, most existing models focus solely on amino acid sequences, overlooking other informative modalities such as 3D structures and literature text. While some recent efforts incorporate multiple modalities, they often suffer from limitations in modality coverage or training strategy. To address this gap, we propose a multimodal pretraining framework that integrates three complementary modalities — protein sequences, structures, and literature text. Our method uses the sequence modality as an anchor and aligns the other two modalities to it via contrastive learning, enabling the model to capture richer and more holistic protein representations. Across a diverse set of downstream tasks, ProteinAligner outperforms state-of-the-art foundation models in predicting protein functions and properties.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of California San Diego (UC San Diego), La Jolla, CA 92093, USA [2]Department of Medicine, UC San Diego, La Jolla, CA 92093, USA [3]Department of Chemistry and Biochemistry, UC San Diego, La Jolla, CA 92093, USA [4]School of Biological Sciences, UC San Diego, La Jolla, CA 92093, USA [5]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA [6]Howard Hughes Medical Institute, UC San Diego, La Jolla, CA 92093, USA [7]Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, CA 92093, USA [8]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA 92093, USA [9]Department of Bioengineering, UC San Diego, La Jolla, CA 92093, USA [10]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA [11]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, UAE. Correspondence to: Pengtao Xie <p1xie@ucsd.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

1. Introduction

Proteins play a fundamental role in virtually all biological processes. Understanding their functions and properties is central to advancing fields such as drug discovery (Wells & McClendon, 2007), diagnostics (Borrebaeck, 2017), and biotechnology (Nobeli et al., 2009). Recent advances in artificial intelligence, particularly in transformer-based models (Vaswani et al., 2017), have led to the development of protein foundation models capable of learning rich representations from large-scale protein datasets (Rives et al., 2021; Jumper et al., 2021; Elnaggar et al., 2021; Bepler & Berger, 2021; Jumper et al., 2021; Hsu et al., 2022; Brandes et al., 2022; Zhang et al., 2022; Xu et al., 2023; Chen et al., 2024; Shanker et al., 2024; Wu et al., 2024). These models, particularly protein language models (PLMs) (Rives et al., 2021; Elnaggar et al., 2021; Bepler & Berger, 2021; Brandes et al., 2022; Chen et al., 2024), have shown remarkable success in performing various downstream tasks such as protein function prediction (Unsal et al., 2022; Yu et al., 2023), property prediction (Flamholz et al., 2024; Teufel et al., 2022), structure prediction (Lin et al., 2023; Chowdhury et al., 2022), and protein design (Madani et al., 2023; Ferruz et al., 2022).

Despite these successes, current PLMs predominantly focus on amino acid sequences while overlooking the wealth of complementary information available in other modalities. Protein structures, for example, provide critical three-dimensional information that is essential for understanding how proteins fold and interact with other molecules, directly influencing their biological functions (Petsko & Ringe, 2004). The spatial arrangement of amino acids, which governs interactions such as binding affinities and functional sites, cannot be readily inferred from sequence data alone (Zhang et al., 2012; Mosca et al., 2013), making the integration of structural data crucial for a more comprehensive understanding of protein behavior. Similarly, the vast amount of biological literature contains experimentally validated insights into protein mechanisms, behavior, and

interactions that are often context-specific and difficult to infer from sequences or structures alone (Lee et al., 2020; Xu et al., 2023). Literature captures critical information about post-translational modifications, protein dynamics in various environments, and interaction networks - details accumulated from years of experimental studies. By incorporating these additional modalities - protein structures and related literature - protein foundation models can move beyond sequence prediction to a more robust, context-aware understanding of protein biology. This multimodal integration has the potential to greatly enhance the representational power of these models, enabling more accurate predictions of protein functions and behaviors in diverse biological scenarios. More detailed discussions about related works can be found in Appendix A.

To address these limitations, we introduce ProteinAligner, a multimodal pretraining framework that combines protein sequences, structures, and literature text. Our framework aligns these modalities with the protein sequence as the anchor, enabling the model to learn richer and more comprehensive representations of proteins. By integrating diverse protein-related data, ProteinAligner improves the model's ability to capture intricate biological phenomena, paving the way for more accurate predictions of protein functions and properties. ProteinAligner utilizes three specialized encoders - a sequence encoder, a structure encoder, and a text encoder - to learn representations for each modality. These distinct representations are projected into a shared latent space, enabling direct comparison across modalities. By employing a contrastive alignment strategy (Oord et al., 2018), ProteinAligner uses protein sequences as the anchor to align corresponding structures and textual descriptions, encouraging similar representations for the same protein and dissimilar representations for different proteins. This approach not only maximizes data utilization by allowing pretraining on incomplete modality data but also captures the biological insights provided by each modality.

ProteinAligner offers several advantages over existing multimodal protein foundation models. While models such as ProtST, ESM-IF1, ESM-S, ProteinCLIP, and ProtCLIP are limited to two modalities, ProteinAligner simultaneously integrates three: sequence, structure, and functional text. In contrast to ESM-3, which relies on masked token prediction that emphasizes local reconstruction without explicit cross-modal alignment, and ProTrek, which employs a multi-task learning framework prone to task interference, ProteinAligner uses a unified contrastive learning objective that promotes global semantic alignment across modalities and avoids task conflict. Additionally, unlike ESM-3 and ProTrek, which discretize protein structures into tokens and incur quantization errors, ProteinAligner retains continuous structural representations, enabling more precise encoding of geometric features.

ProteinAligner demonstrated superior performance compared to state-of-the-art baselines across various downstream prediction tasks, including predicting pathogenic missense variants, predicting protein thermostability, detecting type I anti-CRISPR activities, identifying potent bioactive peptides, estimating the minimum inhibitory concentration of antimicrobial peptides, and protein fitness prediction. An detailed overview of ProteinAligner can be found in Appendix B.

2. Results

ProteinAligner predicts pathogenic missense variants. Pathogenic missense variants refer to specific types of genetic mutations where a single nucleotide change in a DNA sequence results in the substitution of one amino acid for another in the corresponding protein (Cheng et al., 2023). This change can disrupt the protein's normal function, potentially leading to diseases or disorders. In the context of pathogenicity, these variants are considered harmful because they alter the protein's structure or function in a way that impairs biological processes. Depending on the protein's role, this can lead to a variety of outcomes, from minor effects to severe genetic disorders, such as cystic fibrosis, sickle cell disease, or certain forms of cancer. Identifying and characterizing pathogenic missense variants is crucial in genetic research and clinical diagnostics for understanding inherited diseases and developing targeted treatments.

The inputs for this task are two protein sequences: the wild-type sequence and the mutant sequence. We employed the sequence encoder in ProteinAligner to extract representation vectors for both proteins. These vectors were then concatenated and passed through a multi-layer perceptron to predict whether the mutant protein is pathogenic (Fig. 2a). We used 200 labeled examples from the VariPred (Lin et al., 2024) dataset, with 100 examples allocated for training and the remaining 100 for testing.

We benchmarked ProteinAligner against five state-of-the-art protein foundation models: (1) ESM-2 (650M) (Lin et al., 2023), a protein language model pretrained solely on amino acid sequences; (2) ProtST (Xu et al., 2023), which uses contrastive learning to align protein sequences with functional texts; (3) ESM-3 (1.4B) (Hayes et al., 2025), pretrained jointly on sequences, structures, and functional annotations using a masked language modeling objective across all three modalities; (4) ProTrek (Su et al., 2024), trained on sequences, structures, and texts via a multi-task framework combining masked modeling and contrastive learning; and (5) ESM-S (Zhang et al., 2024), which incorporates 3D structural priors into the ESM-2 model via remote homology supervision. We did not include direct comparisons with ProteinCLIP (Wu et al., 2024) and ProtCLIP (Zhou et al., 2025), as both follow a similar two-modality con-

trastive framework as ProtST. Model performance was assessed using F1-score, precision, and recall. For this and all other downstream tasks, we retrained every model five times using different random initializations of the task-specific prediction head. We then reported the mean and standard deviation of the evaluation metrics across the five runs.

ProteinAligner outperformed all baseline methods across F1 score, precision, and recall (Fig. 2b). The corresponding p-values from two-sided *t*-tests — computed based on five repeated runs under identical settings with variation only in random seed initialization — comparing ProteinAligner to ESM-S, ESM-3, and ProTrek on F1 score are 0.045, < 0.01, and 0.037, respectively. All p-values fall below the 0.05 threshold, indicating that the F1 score improvements achieved by ProteinAligner are statistically significant. Despite ESM-3's substantially larger scale — with 1.4 billion parameters and a pretraining corpus comprising 2.78 billion natural protein sequences, which are augmented to over 771 billion sequence tokens, in addition to 236 million structure tokens and 539 million function annotation tokens — its performance remains markedly below that of ProteinAligner. ESM-3 achieves only 0.53 precision, 0.58 recall, and 0.47 F1 score, compared to ProteinAligner's scores of 0.72, 0.72, and 0.72, respectively, despite ProteinAligner's considerably smaller model size and pretraining dataset.

ProteinAligner predicts protein thermostability. Protein thermostability refers to a protein's ability to maintain its structure and function when exposed to elevated temperatures (Modarres et al., 2016). This characteristic is critical because proteins typically lose their functional shape, or denature, at high temperatures, rendering them ineffective. Thermostability is an important factor in various biological processes and industrial applications. For instance, enzymes with high thermostability are essential in industries such as biotechnology and pharmaceuticals, where reactions often require high temperatures for optimal efficiency. Predicting protein thermostability allows researchers to design or engineer proteins that can withstand challenging conditions, improving their functionality and longevity. Additionally, thermostable proteins are valuable in drug design, as they tend to have better shelf lives and performance under physiological conditions. Accurate predictions of thermostability are crucial for advancing protein engineering and enhancing the reliability of proteins in various applications.

Unlike the previous task, this task takes the 3D structures of proteins, specifically their atomic coordinates, as input. The 3D structure of each protein was processed through ProteinAligner's structure encoder, generating a representation vector. This vector was then passed through a multilayer perceptron to predict the protein's thermostability class (Fig. 3a). We employed the HP-S^{2}C5 dataset (Chen et al., 2023b), which comprises 1,040 proteins spanning five thermostability classes: Hyperthermophilic (above 75°C), Thermophilic (45-75°C), Mesophilic (25-45°C), Psychrophilic (5–25°C), and Cryophilic (−20–5°C). 936 proteins were used for training and 104 for testing. We compared ProteinAligner with ESM-IF1 (Hsu et al., 2022), a protein structure encoder pretrained on both protein structures and sequences. We used accuracy, F1 score, and area under ROC curve as evaluation metrics. ProteinAligner remarkably outperformed ESM-IF1 (Fig. 3b), achieving an F1 score of 0.608 compared to 0.559, and an accuracy of 0.577 compared to 0.542.

ProteinAligner detects type I anti-CRISPR activities. We evaluated the effectiveness of ProteinAligner in detecting type I anti-CRISPR (Acr) activities. Acr proteins are produced by certain viruses, such as bacteriophages, or mobile genetic elements to inhibit the type I CRISPR-Cas immune system in bacteria and archaea (Hasani et al., 2023). The CRISPR-Cas system functions as an adaptive immune mechanism in these microorganisms, recognizing and cleaving foreign DNA from viral invaders. In type I systems that involve multi-subunit Cas proteins, Acr proteins disrupt this defense by preventing Cas proteins from binding to target DNA or carrying out their cleavage functions. Understanding and detecting these Acr activities is crucial for controlling CRISPR-Cas systems in genetic engineering and applying bacteriophages to combat antimicrobial resistance.

Given the amino acid sequences of an Arc protein and a set of Cas proteins from a CRISPR-Cas system, we employed ProteinAligner's pretrained sequence encoder to extract representation vectors for each protein. These vectors were then input into a convolutional neural network (CNN) based classification module to predict whether the Arc protein could inhibit the CRISPR-Cas system (Fig. 4a). We utilized the Acr-CRISPR-Cas inhibition dataset (Hasani et al., 2023) for experiments, which comprises 227 pairs of Acr proteins and CRISPR-Cas systems, including 132 experimentally verified positive pairs (Acr inhibits CRISPR-Cas) and 95 negative pairs (Acr does not inhibit CRISPR-Cas). It was randomly split into training and test sets in an 8:2 ratio.

ProteinAligner outperformed all baselines in terms of accuracy. For area under the ROC curve (AUC), it achieved the second-highest performance, slightly trailing ESM-3 (1.4B). For F1 score, the p-values for comparisons between ProteinAligner and ESM-S, ESM-3, and ProTrek are 0.018, 0.011, and 0.033. All values fall below the 0.05 threshold, indicating that the improvements achieved by ProteinAligner over these baselines are statistically significant.

ProteinAligner identifies potent bioactive peptides. Bioactive peptides are short chains of amino acids with specific biological activities (Bahar & Ren, 2013). They play critical roles in regulating physiological processes, in-

cluding immune function, metabolism, and cardiovascular health. Identifying bioactive peptides is important because they offer significant potential for developing new therapeutic agents and functional foods. These peptides can serve as natural, targeted treatments with fewer side effects compared to traditional drugs, and their discovery can lead to advancements in both medical applications and nutrition, benefiting public health and disease prevention efforts.

Given the amino acid sequence of a peptide, we employed ProteinAligner's protein sequence encoder to extract a representation vector, which was subsequently input into a convolutional neural network (CNN)-based classification head to predict whether the peptide has a specific bioactivity. We examined seven distinct bioactivities, including inhibition of dipeptidyl peptidase IV (DPP-IV) (Rasmussen et al., 2003), modulation of brain activity (Bin et al., 2020), antiviral properties (Vilas Boas et al., 2019), antioxidant activity (Zou et al., 2016), umami taste induction (Zhang et al., 2017), blood-brain barrier penetration (Dai et al., 2021), and T-cell immune response induction (Charoenkwan et al., 2020b). Given that a peptide can exhibit multiple bioactivities concurrently, we approached each bioactivity prediction as a binary classification task, avoiding the use of a multi-class model that would assign the peptide to a single category. Separate datasets were used for each bioactivity (Methods). The evaluation metrics for this task included accuracy (ACC), balanced accuracy (BACC) (He & Garcia, 2009), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC) (Matthews, 1975), and the area under the ROC curve (AUC). ProteinAligner outperformed the baselines in most cases (Figs. 5 and 6). For example, in terms of average performance, ProteinAligner outperformed ESM-2, ProtST, and ESM-S in all seven tasks, outperformed ProTrek in six of the seven tasks, and outperformed ESM-3 in five out of the seven tasks.

ProteinAligner predicts the minimum inhibitory concentration (MIC) of antimicrobial peptides. Antimicrobial peptides (AMPs) are short chains of amino acids that serve as a crucial part of the innate immune response in many organisms, exhibiting broad-spectrum activity against bacteria, viruses, fungi, and even cancer cells (Pandi et al., 2023). They function by disrupting microbial membranes, leading to cell death, and are considered potential alternatives to conventional antibiotics, especially in the face of rising antibiotic resistance. The minimum inhibitory concentration (MIC) is the lowest concentration of an antimicrobial agent, such as an AMP, that prevents visible microbial growth. Accurately predicting the MIC values of AMPs is essential as it allows for the optimization of peptide design for therapeutic use, minimizes potential toxicity, and helps in the early-stage screening of effective peptides before in vitro or in vivo testing. This predictive capability is vital for accelerating the development of AMPs as a novel class of antimicrobial agents in clinical applications.

Given the amino acid sequence of a peptide, we applied ProteinAligner's protein sequence encoder to extract a representation vector, which was then fed into a multi-layer perceptron-based regression module to predict the MIC of the peptide against a specific pathogen (Fig. 7a). We focused on Escherichia coli (E. coli), a gram-negative bacterium. We utilized the dataset from (Ledesma-Fernandez et al., 2023), comprising 3,695 training and 924 testing examples. Mean squared error was used as the evaluation metric. ProteinAligner achieved lower prediction error compared to ESM-2, ProtST, ESM-S, ESM-3, and ProTrek (Fig. 7b). ESM-3 had the highest error among all methods, with a value of 1.1 — substantially higher than ProteinAligner's error of 0.449. This observation aligns with the findings reported in (Zhao et al., 2025).

3. Discussions

ProteinAligner introduces a comprehensive approach to protein representation learning by integrating sequences, structures, and literature texts into a unified framework. This multimodal design allows the model to capture complementary information from each modality, providing a richer and more holistic understanding of proteins. By employing contrastive alignment, ProteinAligner learns representations that incorporate structural and functional attributes alongside contextual knowledge from experimental literature, enabling superior performance across a range of challenging protein-related tasks. Its ability to bridge critical gaps in existing models demonstrates its potential for advancing research in protein biology, drug development, and biotechnology, highlighting the importance of multimodal frameworks in addressing complex biological challenges. More detailed discussions can be found in Appendix E.

Software and Data

The FASTA and PDB datasets are publicly available in UniProtKB Swiss-Prot and RCSB PDB, respectively. The FASTA and PDB entries for protein sequences and structures in the pretraining data, along with their textual descriptions, are available at repository All data used in downstream tasks is also publicly available. The data for predicting the pathogenicity of missense variants is available at VariPred. The data used in thermostability prediction is available at HotProtein. The data used in type I anti-CRISPR activity detection is available at AcrTransAct. The data used in peptide bioactivity prediction is available at UniDL4BioPep. The data for predicting the minimum inhibitory concentration (MIC) of antimicrobial peptides is available at DeepAMP. See Appendix G for code availability.

Impact Statement

Proteins are vital to nearly all biological functions, and understanding their roles is critical for advancements in medicine and biotechnology. ProteinAligner introduces a multimodal framework integrating protein sequences, 3D structures, and scientific literature, enabling comprehensive protein representation learning. This approach enhances the ability to predict protein functions, thermostability, pathogenic mutations, and so on.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bahar, A. A. and Ren, D. Antimicrobial peptides. *Pharmaceuticals*, 6(12):1543–1575, 2013.

Bepler, T. and Berger, B. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6): 654–669, 2021.

Bin, Y., Zhang, W., Tang, W., Dai, R., Li, M., Zhu, Q., and Xia, J. Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. *Journal of proteome research*, 19(9):3732–3740, 2020.

Borrebaeck, C. A. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nature Reviews Cancer*, 17(3):199–204, 2017.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., Craig, P. A., Crichlow, G. V., Dalenberg, K., Duarte, J. M., et al. Rcsb protein data bank (rcsb. org): delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic acids research*, 51(D1):D488–D508, 2023.

Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, M. M., and Shoombuatong, W. idppiv-scm: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase iv (dpp-iv) inhibitory peptides using a scoring card method. *Journal of proteome research*, 19 (10):4125–4136, 2020a.

Charoenkwan, P., Nantasenamat, C., Hasan, M. M., and Shoombuatong, W. ittca-hybrid: Improved and robust identification of tumor t cell antigens by utilizing hybrid feature representation. *Analytical biochemistry*, 599: 113747, 2020b.

Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, M. M., and Shoombuatong, W. iumami-scm: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. *Journal of Chemical Information and Modeling*, 60(12):6666–6678, 2020c.

Chen, B., Cheng, X., Li, P., Geng, Y.-a., Gong, J., Li, S., Bei, Z., Tan, X., Wang, B., Zeng, X., et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint*, arXiv:2401.06199, 2024.

Chen, L., Zhang, Z., Li, Z., Li, R., Huo, R., Chen, L., Wang, D., Luo, X., Chen, K., Liao, C., et al. Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Systems*, 14(8):706–721, 2023a.

Chen, T., Gong, C., Diaz, D. J., Chen, X., Wells, J. T., Liu, Q., Wang, Z., Ellington, A., Dimakis, A., and Klivans, A. Hotprotein: A novel framework for protein thermostability prediction and editing. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://github.com/VITA-Group/HotProtein.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

Consortium, U. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., De Spiegeleer, B., and Xia, J. Bbppred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. *Journal of Chemical Information and Modeling*, 61(1): 525–534, 2021.

Du, Z., Ding, X., Xu, Y., and Li, Y. Unidl4biopep: a universal deep learning architecture for binary classification

in peptide bioactivity. *Briefings in Bioinformatics*, 24(3): bbad135, 2023.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Fan, X., Pan, H., Tian, A., Chung, W. K., and Shen, Y. Shine: protein language model-based pathogenicity prediction for short inframe insertion and deletion variants. *Briefings in Bioinformatics*, 24(1):bbac584, 2023.

Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Flamholz, Z. N., Biller, S. J., and Kelly, L. Large language models improve annotation of prokaryotic viral proteins. *Nature Microbiology*, 9(2):537–549, 2024.

Hasani, M., Trost, C. N., Timmerman, N., and Jin, L. Acrtransact: Pre-trained protein transformer models for the detection of type i anti-crispr activities. In *Proceedings of The 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, pp. 6, Houston, TX, USA, 2023. ACM.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735. IEEE Computer Society, 2020.

Hermosilla, P. and Ropinski, T. Contrastive representation learning for 3d protein structures. *arXiv preprint arXiv:2205.15675*, 2022.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456. PMLR, 2015.

Jiang, F., Li, M., Dong, J., Yu, Y., Sun, X., Wu, B., Huang, J., Kang, L., Pei, Y., Zhang, L., et al. A general temperature-guided language model to design proteins of enhanced stability and activity. *Science Advances*, 10(48):eadr2641, 2024.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980, 2014.

Kulandaisamy, A., Sakthivel, R., and Gromiha, M. M. Mptherm: database for membrane protein thermodynamics for understanding folding and stability. *Briefings in Bioinformatics*, 22(2):2119–2125, 2021.

Ledesma-Fernandez, A., Velasco-Lozano, S., Santiago-Arcos, J., López-Gallego, F., and Cortajarena, A. L. Engineered repeat proteins as scaffolds to assemble multi-enzyme systems for efficient cell-free biosynthesis. *Nature Communications*, 14(1):2587, 2023.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

Lima, B., Ricci, M., Garro, A., Juhász, T., Szigyártó, I. C., Papp, Z. I., Feresin, G., de la Torre, J. G., Cascales, J. L., Fülöp, L., et al. New short cationic antibacterial peptides. synthesis, biological activity and mechanism of action. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1863(10):183665, 2021.

Lin, W., Wells, J., Wang, Z., Orengo, C., and Martin, A. C. Enhancing missense variant pathogenicity prediction with protein language models using varipred. *Scientific Reports*, 14(1):8136, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Liu, B., Liu, X., Jin, X., Stone, P., and Liu, Q. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

Liu, S., Chen, T., Zhang, Z., Chen, X., Huang, T., Jaiswal, A. K., and Wang, Z. Sparsity may cry: Let us fail (current) sparse neural networks together! In *International Conference on Learning Representations*, 2023.

Livesey, B. J. and Marsh, J. A. Interpreting protein variant effects with computational predictors and deep mutational scanning. *Disease Models & Mechanisms*, 15(6): dmm049510, 2022.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2022.

Loshchilov, I., Hutter, F., et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, pp. 3–8, 2013.

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., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Matthews, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2): 442–451, 1975.

McBride, J. M., Polev, K., Abdirasulov, A., Reinharz, V., Grzybowski, B. A., and Tlusty, T. Alphafold2 can predict single-mutation effects. *Physical Review Letters*, 131(21): 218401, 2023.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Modarres, H. P., Mofrad, M., and Sanati-Nezhad, A. Protein thermostability engineering. *RSC advances*, 6(116): 115252–115270, 2016.

Mosca, R., Céol, A., and Aloy, P. Interactome3d: adding structural details to protein networks. *Nature methods*, 10(1):47–53, 2013.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, pp. 807–814, 2010.

Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D., and Gromiha, M. M. Prothermdb: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic acids research*, 49(D1):D420–D424, 2021.

Nobeli, I., Favia, A. D., and Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nature biotechnology*, 27(2):157–167, 2009.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023a.

Notin, P., Weitzman, R., Marks, D., and Gal, Y. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023b.

Olsen, T. H., Yesiltas, B., Marin, F. I., Pertseva, M., García-Moreno, P. J., Gregersen, S., Overgaard, M. T., Jacobsen, C., Lund, O., Hansen, E. B., et al. Anoxpepred: using deep learning for the prediction of antioxidative properties of peptides. *Scientific reports*, 10(1):21471, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint*, arXiv:1807.03748, 2018.

Pandi, A., Adam, D., Zare, A., Trinh, V. T., Schaefer, S. L., Burt, M., Klabunde, B., Bobkova, E., Kushwaha, M., Foroughijabbari, Y., et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nature Communications*, 14(1): 7197, 2023.

Petsko, G. A. and Ringe, D. *Protein structure and function*. New Science Press, 2004.

Pinacho-Castellanos, S. A., García-Jacas, C. R., Gilson, M. K., and Brizuela, C. A. Alignment-free antimicrobial peptide predictors: improving performance by a thorough analysis of the largest available data set. *Journal of Chemical Information and Modeling*, 61(6):3141–3157, 2021.

Ramakrishnan, G., Baakman, C., Heijl, S., Vroling, B., van Horck, R., Hiraki, J., Xue, L. C., and Huynen, M. A.

Understanding structure-guided variant effect predictions using 3d convolutional neural networks. *Frontiers in molecular biosciences*, 10:1204157, 2023.

Rasmussen, H. B., Branner, S., Wiberg, F. C., and Wagtmann, N. Crystal structure of human dipeptidyl peptidase iv/cd26 in complex with a substrate analog. *Nature structural biology*, 10(1):19–25, 2003.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Shanker, V. R., Bruun, T. U., Hie, B. L., and Kim, P. S. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science*, 385 (6704):46–53, 2024.

Smith, C. A. and Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of molecular biology*, 380(4):742–756, 2008.

Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 2017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., Mazurenko, S., and Bednar, D. Fireprotdb: database of manually curated protein stability data. *Nucleic acids research*, 49(D1):D319–D324, 2021.

Su, J., Zhou, X., Zhang, X., and Yuan, F. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, pp. 2024–05, 2024.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology*, 40(7):1023–1025, 2022.

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., and Doğan, T. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4 (3):227–245, 2022.

Van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vilas Boas, L. C. P., Campos, M. L., Berlanda, R. L. A., de Carvalho Neves, N., and Franco, O. L. Antiviral peptides as promising therapeutic drugs. *Cellular and Molecular Life Sciences*, 76:3525–3542, 2019.

Wells, J. A. and McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, 450(7172):1001–1009, 2007.

Wu, K. E., Chang, H., and Zou, J. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, 2024–05, 2024.

Xu, J., Sun, X., Zhang, Z., Zhao, G., and Lin, J. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

Yan, J., Zhang, B., Zhou, M., Kwok, H. F., and Siu, S. W. Multi-branch-cnn: Classification of ion channel interacting peptides using multi-branch convolutional neural network. *Computers in Biology and Medicine*, 147:105717, 2022.

Yan, J., Zhang, B., Zhou, M., Campbell-Valois, F.-X., and Siu, S. W. A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against escherichia coli using multi-branch-cnn and attention. *Msystems*, 8(4):e00345–23, 2023.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning.

Advances in neural information processing systems, 33: 5824–5836, 2020.

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., and Zhao, H. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *International Conference on Learning Representations*, 2022.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, 490(7421): 556–560, 2012.

Zhang, Y., Venkitasamy, C., Pan, Z., Liu, W., and Zhao, L. Novel umami ingredients: Umami peptides and their taste. *Journal of food science*, 82(1):16–23, 2017.

Zhang, Z., Lu, J., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Structure-informed protein language model. *arXiv preprint arXiv:2402.05856*, 2024.

Zhao, J., Liu, H., Kang, L., Gao, W., Lu, Q., Rao, Y., and Yue, Z. deep-amppred: A deep learning method for identifying antimicrobial peptides and their functional activities. *Journal of Chemical Information and Modeling*, 2025.

Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.-C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Zhou, H., Yin, M., Wu, W., Li, M., Fu, K., Chen, J., Wu, J., and Wang, Z. Protclip: Function-informed protein multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22937–22945, 2025.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Zou, T.-B., He, T.-P., Li, H.-B., Tang, H.-W., and Xia, E.-Q. The structure-activity relationship of the antioxidant peptides from natural proteins. *Molecules*, 21(1):72, 2016.

A. Related works

Several prior studies (Xu et al., 2023; Zhang et al., 2024; Hayes et al., 2025; Wu et al., 2024; Su et al., 2024) have explored pretraining protein foundation models using pairs of modalities — for example, combining protein sequences with literature texts (Xu et al., 2023), or protein sequences with structural information (Zhang et al., 2024). ProteinCLIP (Wu et al., 2024) adopts a dual-encoder architecture, aligning protein sequences with functional annotations via contrastive learning in a shared latent space, using a pretrained sequence encoder and a transformer-based text encoder. ProtCLIP (Zhou et al., 2025) builds upon this by introducing three types of contrastive objectives: (i) sequence-text alignment, (ii) intra-sequence consistency across overlapping fragments, and (iii) prototype-based alignment for known functional regions. ProtST (Xu et al., 2023) also uses contrastive learning between sequences and texts, incorporating curriculum-based negative sampling and a margin-based loss. ESM-S (Zhang et al., 2024) extends ESM-2 (Lin et al., 2023) by injecting structural information through fine-tuning on a remote homology detection task. It predicts fold classes directly from sequence embeddings, thereby enriching the model with implicit structural knowledge. However, pretraining on all three modalities remains largely underexplored.

Recently, two works — ESM-3 (Hayes et al., 2025) and ProTrek (Su et al., 2024) — conducted independently and concurrently with ours, leverage all three modalities for pretraining. ESM-3 (Hayes et al., 2025) integrates amino acid sequences, 3D structural coordinates, and textual annotations by converting each modality into a unified token track processed jointly by a transformer backbone. During pretraining, it employs a masked language modeling objective in which tokens across all modalities are randomly masked, and the model is trained to predict the missing elements, thereby learning cross-modal representations. ProTrek (Su et al., 2024) encodes sequences, structures, and functional text using three separate encoders whose outputs are projected into a shared latent space. Its training strategy combines bidirectional InfoNCE (Oord et al., 2018) losses with masked language modeling objectives applied separately to the sequence and structure modalities.

ESM-3 has two major limitations. First, its pretraining relies exclusively on masked token prediction and does not incorporate contrastive learning (He et al., 2020; Hermosilla & Ropinski, 2022; Oord et al., 2018), which limits its ability to align biologically equivalent inputs such as protein sequences, structures, and functional annotations. This absence of contrastive supervision can lead to fragmented and less transferable representations that perform poorly on tasks requiring integrated biological reasoning. Second, ESM-3 compresses structural information into discrete tokens using a VQ-VAE (Razavi et al., 2019) encoder, introducing quantization noise that discards fine-grained biophysical details such as side-chain orientations, solvent accessibility, and subtle backbone perturbations. Similarly, ProTrek also exhibits key limitations. Although it combines masked token prediction and contrastive learning during pretraining, this multitask setup can create optimization conflicts due to competing training signals (Kendall et al., 2018; Liu et al., 2021; Sener & Koltun, 2018; Yu et al., 2020), leading to unstable convergence and suboptimal alignment across modalities. In addition, ProTrek discretizes protein structures into tokens using Foldseek (Van Kempen et al., 2024), resulting in the loss of detailed geometric information critical for capturing functionally relevant structural features.

B. ProteinAligner overview

ProteinAligner is a multimodal foundation model for protein representation learning, integrating three distinct modalities: amino acid (AA) sequences, 3D structures, and textual descriptions of proteins. ProteinAligner contains three encoders - a protein sequence encoder, a protein structure encoder, and a text encoder - each dedicated to learning representations for its corresponding modality (Fig. 1a). The protein sequence encoder is a protein language model that uses the transformer architecture (Vaswani et al., 2017) to extract a representation for the input AA sequence. It represents each AA as a token and employs self-attention (Vaswani et al., 2017) to capture long-range dependencies among AAs. The protein structure encoder utilizes Geometric Vector Perceptron Graph Neural Network (GVP-GNN) (Jing et al., 2020) layers for geometric representation learning of the input protein structure, followed by transformer layers that capture long-range interactions between atomic coordinates. The text encoder employs a transformer architecture, utilizing self-attention to capture long-range dependencies between language tokens. Specifically, we employed ESM-2 (650M) (Lin et al., 2023), a leading protein language model, as the protein sequence encoder, and ESM-IF1 (Hsu et al., 2022) as the protein structure encoder. ESM-2 (650M) consists of 33 transformer layers and 650 million parameters, pretrained on 65 million protein sequences. ESM-IF1 features 20 layers and 124 million parameters, pretrained on 12 million computed protein structures and 16,000 experimentally verified structures. The text encoder includes eight transformer layers with a total of 78 million parameters. ProteinAligner uses modality-specific linear projection modules to map the representations extracted by different encoders into a shared latent space with matching dimensions, ensuring that representations from different modalities are

directly comparable. ProteinAligner consists of 867 million model parameters in total.

ProteinAligner performs joint pretraining of the three encoders by leveraging a modality alignment strategy, using protein sequences as the anchor modality to align the other two modalities (Fig. 1a). Specifically, given a protein sequence and a protein structure, if they correspond to the same protein, ProteinAligner encourages their representations to be similar, and dissimilar otherwise. The same principle applies for protein text and protein sequences, with representations aligned if they refer to the same protein and separated if not. This alignment is accomplished by minimizing contrastive losses (He et al., 2020; Oord et al., 2018) defined on the representations of sequence-structure pairs and sequence-text pairs. ProteinAligner does not require all three modalities to be present simultaneously for each protein in the pretraining data. The alignment can be performed as long as the protein sequence and at least one additional modality - either structure or text - are available. We chose protein sequences as the anchor for alignment because they are the most prevalent data modality in protein databases; nearly every protein has an associated amino acid sequence, whereas information on structures or textual descriptions is often incomplete. By using sequences as the anchor, we can maximize data utilization, ensuring the inclusion of as many proteins as possible in the alignment process. With pretrained encoders in place, they can be fine-tuned on task-specific data to handle a variety of downstream tasks. During this process, the encoders are integrated with task-specific modules, creating models that are customized for specific prediction tasks.

We curated a large-scale pretraining dataset for ProteinAligner by integrating data from the UniProtKB/Swiss-Prot (Consortium, 2019) and RCSB PDB (Burley et al., 2023) databases. The dataset consists of 290,480 proteins, each with an amino acid sequence and a corresponding textual description. 133,726 of them are also associated with protein structures. In total, the dataset contains 133,726 sequence-structure pairs and 290,480 sequence-text pairs. The textual descriptions provide information about the proteins' functions. Both the structures and the functional descriptions were experimentally validated and reviewed by domain experts. Fig. 1b shows the distribution of protein taxonomy, functions, and types in the dataset.

C. More tasks

C.1. Zero-shot prediction of pathogenic missense variants and thermostability

To assess the generalization capability of ProteinAligner without task-specific fine-tuning, we conducted zero-shot evaluations on two representative tasks: pathogenic missense variant prediction and protein thermostability prediction. These tasks evaluate the model's ability to reason about the effects of single amino acid substitutions in a biologically meaningful manner.

For zero-shot pathogenic missense variant prediction, we followed the protocol of Meier et al. (Meier et al., 2021). Given a wild-type and a mutant sequence differing by a single amino acid substitution at a specific site, we used each protein foundation model to encode both sequences and extract the representation vectors at the mutation site. These vectors were passed through the pretrained ESM-2 prediction head, which outputs a probability distribution over the amino acid vocabulary. We then computed the log-probabilities of the wild-type and mutant residues at the mutation site, and used their difference (log-prob(mutant) – log-prob(wild-type)) as a pathogenicity score. A mutation was classified as pathogenic if this score exceeded a fixed threshold.

As shown in Fig. 9a, ProteinAligner achieved the highest area under the ROC curve (AUC = 0.240), outperforming ESM-2 650M (0.175), ProtST (0.215), ESM-S (0.212), and ProTrek (0.100). Notably, in the low false-positive rate regime (FPR $< 5\%$), which is especially relevant for clinical screening applications, ProteinAligner attained a substantially higher true positive rate than all baselines.

For zero-shot thermostability prediction, we adopted the evaluation strategy of Jiang et al. (Jiang et al., 2024), using mutation data from MPTherm (Kulandaisamy et al., 2021), FireProtDB (Stourac et al., 2021), and ProThermDB (Nikam et al., 2021), totaling 66 single-site mutation assays. As with the pathogenicity task, we calculated the log-probability difference between mutant and wild-type amino acids at the mutation site using the ESM-2 prediction head. This score was used as a proxy for the mutation's impact on melting temperature. The predicted scores were ranked and compared to experimentally measured rankings using Spearman's rank correlation. Fig. 9b shows that ProteinAligner achieved a Spearman correlation of 0.418, comparable to ESM-2 (0.433), ESM-S (0.433), ProtST (0.429), and ProTrek (0.408).

Together, these results demonstrate that ProteinAligner generalizes well to both functional and biophysical mutation effect prediction tasks in a zero-shot setting, highlighting the benefits of its contrastive tri-modal pretraining on sequence, structure, and function text.

C.2. Effect of prediction head on downstream performance

To assess the influence of prediction heads on downstream task performance, we conducted additional experiments on two representative tasks: pathogenic missense variant prediction and minimum inhibitory concentration (MIC) regression for antimicrobial peptides. These tasks were selected to reflect both classification and regression settings. For the pathogenicity prediction task, we compared two prediction head configurations: (1) a multi-layer perceptron (MLP) and (2) a linear Elastic Net (Zou & Hastie, 2005) model applied to protein embeddings after dimensionality reduction using principal component analysis (PCA), following the methodology in (Fan et al., 2023). PCA was employed to project the high-dimensional residue-level embeddings produced by protein language models into a lower-dimensional space, which was then used as input to the Elastic Net classifier. For the MIC regression task, we evaluated: (1) an MLP head, and (2) a multi-branch convolutional neural network (CNN) architecture proposed by (Yan et al., 2023; 2022), which has demonstrated strong performance in peptide-related predictions.

Results of these comparisons are shown in Fig. 10. Across all prediction head configurations, ProteinAligner consistently outperformed baseline models, indicating that its performance gains are robust to the choice of prediction head. While the architecture of the prediction head can affect the absolute level of performance, the improvements provided by ProteinAligner persisted regardless of the prediction head used.

C.3. ProteinAligner predicts protein fitness.

Protein fitness prediction aims to assign quantitative scores to single-mutation variants of a reference protein, reflecting changes in biochemical activity, stability, or binding. Such measurements are typically obtained through large-scale deep mutational scanning (DMS) assays, which combine systematic mutagenesis with high-throughput functional screening to map the sequence–function landscape (Chen et al., 2023a; Livesey & Marsh, 2022).

To predict fitness, we encoded each mutant sequence using ProteinAligner's sequence encoder and passed the resulting embeddings to a convolutional neural network (CNN) head (Notin et al., 2023b), followed by an independent multilayer perceptron (MLP) regressor for each mutation site (Fig. 8a). Experiments were conducted on a subset of the ProteinGym benchmark (Notin et al., 2023a), consisting of 15 randomly selected single-substitution DMS assays from a total of 217. The selected assays are listed in Table 2, with an average of 3,402 fitness measurements per assay. Each assay was evaluated using Spearman's rank correlation coefficient (Spearman), Pearson correlation coefficient (Pearson), and coefficient of determination (R^2), with higher values indicating better performance. All experiments were repeated five times to account for variability. ProteinAligner outperformed all baselines across all metrics, except for a slight decrease compared to ESM-3 in R^2 (Fig. 8b). These results indicate that ProteinAligner generalizes well to protein fitness prediction tasks.

D. Materials and methods

D.1. Dataset preprocessing

The pretraining data for ProteinAligner was sourced from the UniProtKB/Swiss-Prot (Consortium, 2019) and RCSB PDB (Burley et al., 2023) databases. UniProtKB/Swiss-Prot is a well-curated repository containing high-quality protein sequences across a wide variety of organisms, along with detailed annotations on protein functions and properties. We utilized version UniProt 2023_02, which was released on May 2, 2023. The RCSB PDB database offers a comprehensive collection of experimentally determined 3D protein structures, derived from methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). It includes protein structures from a wide range of proteins, such as enzymes, receptors, and antibodies, originating from diverse organisms.

From these databases, we collected sequence-text pairs and sequence-structure pairs. The sequence-text pairs were sourced from UniProtKB/Swiss-Prot. We first obtained a collection of protein entries from Swiss-Prot that included textual descriptions of their functions, by filtering for entries where the `commentType` field was set to 'Function'. We then retrieved the corresponding sequence for each protein in this collection. Specifically, we accessed the UniProt ID from the `primaryAccession` field and used it to retrieve the corresponding protein FASTA file from the UniProt website, which contains the protein's sequence. We downloaded all available PDB files from the May 2, 2023 dataset release (Burley et al., 2023), comprising 200,734 experimentally determined protein structures. We then employed the UniProt ID mapping tool[1] to link the structures in the PDB files to their corresponding amino acid sequences in the FASTA files. To address memory

[1]https://www.uniprot.org/id-mapping

Table 1. Data split for bioactive peptide identification

Bioactive peptides	Train (Pos./Neg.)	Test (Pos./Neg.)
DPP IV inhibitory peptides	532 / 532	133 / 133
Neuropeptide	1,940 / 1,940	485 / 485
Antiviral peptides	2,321 / 2,321	623 / 623
Antioxidant peptides	582 / 541	146 / 135
Umami peptides	112 / 241	28 / 61
Blood–brain barrier peptides	100 / 100	19 / 19
TTCA peptides	470 / 318	122 / 75

constraints during pretraining, we excluded protein sequences longer than 300 residues, yielding 133,726 sequence-structure pairs and 290,480 sequence-text pairs.

For datasets used in downstream tasks, we followed the train–test split strategies specified in the original papers from which the datasets were obtained whenever applicable. In cases where no predefined split was available, we adopted a standard 80:20 random split. Below, we provide a detailed description of the data splitting procedures for each dataset. For type I anti-CRISPR activity detection, we used a dataset (Hasani et al., 2023) comprising 227 Acr–CRISPR-Cas system pairs, including 132 positive (inhibitory) and 95 negative (non-inhibitory) examples. Following the setup in (Hasani et al., 2023), we applied a random 80:20 split. For peptide bioactivity prediction, we used the train–test splits provided by (Du et al., 2023), who curated and partitioned the datasets for each specific bioactivity task. These datasets include a range of short peptides annotated with diverse functional labels. Table 1 summarizes the distribution of examples across training and test sets for each task. For minimum inhibitory concentration (MIC) prediction of antimicrobial peptides, the original dataset (Pandi et al., 2023) for Gram-negative AMPs included only a training set. We therefore performed an 80:20 random split, yielding 3,695 peptides for training and 924 for testing. For pathogenic missense variant prediction, we followed the predefined balanced 50:50 split from the original source (Lin et al., 2024), consisting of 100 mutations in the training set and 100 in the test set. Finally, for thermostability prediction, we used the same train–test split as in (Liu et al., 2023), which contains 936 proteins for training and 104 for testing.

D.2. Encoders in ProteinAligner

ProteinAligner utilizes ESM-2 (650M) (Lin et al., 2023) to learn representations for protein sequences. ESM-2 (650M), a protein language model, was pretrained on 65 million protein sequences from UniRef50 (Suzek et al., 2015) by predicting masked amino acids. The model features 33 transformer layers and an embedding dimension of 1280, allowing it to effectively capture the complexities inherent in protein sequences. To encode protein structures, ProteinAligner employs ESM-IF1 (Hsu et al., 2022), a model trained to address the inverse folding problem - predicting the amino acid sequence from a protein's backbone atom coordinates. ESM-IF1 comprises an encoder and a decoder, where the encoder extracts a representation vector from the input structure, which is then fed into the decoder to generate the corresponding sequence. ProteinAligner utilizes only the encoder from ESM-IF1, omitting the decoder component. The encoder is composed of four Geometric Vector Perceptron Graph Neural Network (GVP-GNN) (Jing et al., 2020) layers for geometric feature extraction, followed by eight transformer encoder layers to capture long-range interactions between these features. ESM-IF1 was trained on 12 million AlphaFold2 (Jumper et al., 2021) computed protein structures and 16,000 experimentally verified structures, along with their associated sequences from the UniRef50 dataset (Suzek et al., 2015). The text encoder is a transformer model comprising eight layers and a total of 78 million parameters.

To aggregate per-position embeddings into a single representation for the entire protein sequence, we use the output embedding of the beginning-of-sequence (BOS) token from the ESM-2 encoder. The BOS token is prepended to each input sequence and attends to all positions during encoding, thereby serving as a global summary that captures information across the full sequence. For structure-based representations extracted by the ESM-IF1 encoder, we compute the mean of the per-residue embeddings across all positions, providing an overall summary of the 3D structural context. For functional text, following (Cherti et al., 2023), we use the embedding of the end-of-sequence (EOS) token, which is appended to the end of the tokenized input, to summarize the textual information. Each modality-specific summary is then passed through a lightweight projection head consisting of a LayerNorm (Ba et al., 2016) and a single-layer multi-layer perceptron, yielding a fixed-size representation vector. These vectors are used for contrastive learning across modalities.

D.3. ProteinAligner pretraining

Given a protein sequence S and a protein structure R, we employ the sequence encoder $E_s(\cdot)$ and structure encoder $E_r(\cdot)$ to extract representation vectors $\mathbf{s} = E_s(S)$ and $\mathbf{r} = E_r(R)$ for S and R, respectively. To ensure that the representations are similar when S and R belong to the same protein, and dissimilar when they do not, we minimize the InfoNCE (Oord et al., 2018) contrastive learning loss:

$$\mathcal{L}_{s,r} = -\log \frac{\exp\left(\mathbf{s}_i^\top \mathbf{r}_i/\tau\right)}{\exp\left(\mathbf{s}_i^\top \mathbf{r}_i/\tau\right) + \sum_{j\neq i} \exp\left(\mathbf{s}_i^\top \mathbf{r}_j/\tau\right)}, \tag{1}$$

where $\mathbf{s}_i$ and $\mathbf{r}_i$ represent the sequence and structure representations of the same protein i, while $\mathbf{s}_i$ and $\mathbf{r}_j$ represent the representations of different proteins i and j. This loss function encourages the alignment of $\mathbf{s}_i$ and $\mathbf{r}_i$ and discourages the similarity between $\mathbf{s}_i$ and $\mathbf{r}_j$. The temperature parameter τ controls the sharpness of the softmax distribution.

Similarly, given a protein sequence S and a protein text description T, we use the sequence encoder $E_s(\cdot)$ and the text encoder $E_t(\cdot)$ to obtain representation vectors $\mathbf{s} = E_s(S)$ and $\mathbf{t} = E_t(T)$ for S and T, respectively. To ensure that the representations are similar when S and T correspond to the same protein, and dissimilar when they do not, we minimize the following InfoNCE contrastive learning loss:

$$\mathcal{L}_{s,t} = -\log \frac{\exp\left(\mathbf{s}_i^\top \mathbf{t}_i/\tau\right)}{\exp\left(\mathbf{s}_i^\top \mathbf{t}_i/\tau\right) + \sum_{j\neq i} \exp\left(\mathbf{s}_i^\top \mathbf{t}_j/\tau\right)}. \tag{2}$$

For specific experimental settings, we minimized the sum of the two loss functions with equal weights. The temperature parameter τ was configured to 0.07. Pretraining was carried out over 20 epochs with a total batch size of 80 using 40 A100 GPUs. For distributed training, we employed Distributed Data Parallel (DDP) (Li et al., 2020) rather than Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023), due to the increased inter-device communication overhead associated with FSDP. In our experiments, DDP offered a favorable trade-off between performance and scalability. On the other hand, our framework is fully compatible with FSDP and can be adapted to it if larger-scale protein encoders are used. During pretraining, each batch contains both sequence–structure and sequence–text pairs. For example, when the batch size is set to 80, we sample 80 sequence–structure pairs and 80 sequence–text pairs per iteration. The InfoNCE loss is then computed separately for each modality pair and averaged. The sampling of pairs is performed uniformly at random from the available paired data. We optimized the model weights using the AdamW optimizer (Loshchilov et al., 2017), with an initial learning rate of 5×10^{-6}, weight decay of 1×10^{-4}, and betas of $(0.9, 0.95)$. The learning rate was dynamically adjusted throughout pretraining via the CosineAnnealingLR scheduler (Loshchilov & Hutter, 2022). Figure 12 illustrates the training dynamics of the contrastive learning losses for sequence–structure pairs, sequence–text pairs, and their combined loss. The combined loss decreases smoothly and stabilizes around 0.1, indicating stable convergence. The sequence–structure loss shows a rapid and consistent decline, reaching approximately 0.05 within the first 20,000 training steps. The sequence–text loss exhibits moderate fluctuations in the early stages, which gradually diminish and stabilize near 0.15.

D.4. Pathogenic missense variants prediction

The overall model architecture is depicted in Fig. 11a. For this task, the sequence encoder pretrained by ProteinAligner was fine-tuned. The classification module was based on a multi-layer perceptron, which comprises a fully connected layer with a hidden state size of 1280, a dropout layer with a probability of 0.5, a leaky ReLU activation (Maas et al., 2013), and a second fully connected layer with a softmax activation function. We employed the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3}, training the model for a maximum of 200 epochs with a batch size of 32. To mitigate overfitting, we applied an early stopping strategy: if validation performance did not improve over 10 consecutive epochs, training was halted, and the model checkpoint with the best validation performance was retained as the final model. The metrics used to evaluate model performance in this task include precision, recall rate, and F1 score.

D.5. Thermostability prediction

The overall model architecture is depicted in Fig. 11b. The structure encoder, pretrained using ProteinAligner, was fine-tuned for this task. The classification module, implemented as a multi-layer perceptron (MLP), includes a fully connected layer with a hidden dimension of 128, followed by layer normalization (Xu et al., 2019), a ReLU activation, and a second fully connected layer to produce the classification logits.

D.6. Type I anti-CRISPR activity detection

The overall model architecture for this task is illustrated in Fig. 11c. The sequence encoder pretrained by ProteinAligner was fine-tuned using data specific to this task. The classification module was based on a CNN, which was composed of two 1D convolutional layers, each with a stride of 1 and a kernel size of 7. The first convolutional layer takes an input of 1280 channels and outputs 4 channels. The second convolutional layer maintains the same input and output dimensions. Batch normalization (Ioffe & Szegedy, 2015) and ReLU activation (Nair & Hinton, 2010) are applied after each convolutional layer. Following the convolutional layers, two fully connected layers, each with a hidden size of 4, are employed for final classification.

During training, we optimized model weights using the Adam (Kingma, 2014) optimizer with an initial learning rate of 3×10^{-3}, over a maximum of 250 epochs with a batch size of 32. To prevent overfitting, we employed early stopping when the decrease in training loss fell below 0.005 and applied weight decay, starting at 0.01 and gradually reducing to 0.001. We also performed a hyperparameter sweep on the dropout rate (Srivastava et al., 2014), exploring values between 0.3 and 0.5. Additionally, we implemented a learning rate reduction strategy, decreasing the rate by a factor of 0.9 if validation performance did not improve for 10 consecutive epochs. The model's performance was evaluated using accuracy and F1 score.

D.7. Identification of potent bioactive peptides

The overall model architecture is illustrated in Fig. 11d. The sequence encoder, pretrained by ProteinAligner, was fine-tuned for each of the eight tasks. The classification module employs a convolutional neural network (CNN) with six layers, structured as follows: a 1D convolutional layer (kernel size = 3, stride = 1, padding = 2), followed by BatchNorm and ReLU activation; a max pooling layer (kernel size = 2, padding = 1) and a dropout layer with a probability of 0.15; another 1D convolutional layer (kernel size = 3, stride = 1, padding = 2), followed by BatchNorm and ReLU activation; a max pooling layer (kernel size = 2, padding = 1) and a dropout layer with a probability of 0.15; a fully connected layer with a hidden state size of 64, followed by ReLU activation and a dropout layer (probability = 0.15); and finally, a fully connected binary classification layer with sigmoid activation.

For the dipeptidyl peptidase IV (DPP-IV) inhibitory peptide prediction task, the goal is to identify peptides that inhibit DPP-IV activity (Charoenkwan et al., 2020a). We used the iDPPIV-SCM dataset (Charoenkwan et al., 2020a), containing 532 inhibitory peptides and 532 non-inhibitory peptides for training, and 133 inhibitory peptides and 133 non-inhibitory peptides for testing. In the neuropeptide (NP) prediction task, the aim is to classify peptides as neuropeptides or non-neuropeptides (Bin et al., 2020). We used the PredNeuroP dataset (Bin et al., 2020), containing 1940 neuropeptides and 1940 non-neuropeptides for training, and 485 neuropeptides and 485 non-neuropeptides for testing. For the antiviral peptide prediction task, the objective is to predict whether a peptide has antiviral activity (preventative and therapeutic against viral infections) (Vilas Boas et al., 2019). We utilized the ABPDiscover dataset (Pinacho-Castellanos et al., 2021), which includes 2321 antiviral peptides and 2321 non-antiviral peptides for training, and 623 antiviral peptides and 623 non-antiviral peptides for testing. In the antioxidant peptide prediction task, the aim is to classify peptides based on their antioxidant properties (Zou et al., 2016). We used the AnOxPePred dataset (Olsen et al., 2020), containing 582 antioxidative peptides and 241 non-antioxidative peptides for training, with a test set comprising 28 antioxidative peptides and 61 non-antioxidative peptides. In the umami peptide prediction task, the goal is to determine whether a peptide elicits an umami taste (Zhang et al., 2017). For this task, we used the iUmami-SCM dataset (Charoenkwan et al., 2020c), with a training set of 112 umami peptides and 241 non-umami peptides, and a test set of 28 umami peptides and 61 non-umami peptides. In the blood–brain barrier peptide (BBP) prediction task, the objective is to classify whether a peptide can penetrate the blood–brain barrier (i.e., BBP) (Dai et al., 2021). We employed the BBPpred dataset (Dai et al., 2021), consisting of 100 BBPs and 100 non-BBPs for training, and 19 BBPs and 19 non-BBPs for testing. The tumor T cell antigen prediction task aims to classify peptides capable of inducing a T-cell immune response (Charoenkwan et al., 2020b). We used the iTTCA-Hybrid dataset (Charoenkwan et al., 2020b), including 470 antigenic peptides and 318 non-antigenic peptides for training, with 122 antigenic peptides and 75 non-antigenic peptides for testing.

During training, we optimized the model using stochastic gradient descent (SGD) with a learning rate of 1×10^{-2}, momentum of 0.5, and no weight decay, over 200 epochs. Additionally, we applied step decay for learning rate adjustment and utilized early stopping based on validation accuracy, halting training if no improvement was observed for 40 consecutive epochs. Model performance was assessed using several metrics, including accuracy (ACC), balanced accuracy (BACC) (He & Garcia, 2009), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC) (Matthews, 1975), and area

Table 2. List of assays randomly selected from the ProteinGym benchmark (15 out of 217 total assays).

Data source	Num of observed fitness
A0A140D2T1_ZIKV_Sourisseau_2019	9576
A0A247D711_LISMN_Stadelmann_2021	1653
A4D664_9INFA_Soh_2019	14421
CAS9_STRP1_Spencer_2017_positive	8117
D7PM05_CLYGR_Somermeyer_2022	1169
ENV_HV1B9_DuenasDecamp_2016	375
GLPA_HUMAN_Elazar_2016	245
KCNE1_HUMAN_Muhammad_2023_function	2315
ODP2_GEOSE_Tsuboyama_2023_1W4G	669
PKN1_HUMAN_Tsuboyama_2023_1URF	1301
Q59976_STRSQ_Romero_2015	2999
REV_HV1H2_Fernandes_2016	2147
SBI_STAAM_Tsuboyama_2023_2JVG	1025
TNKS2_HUMAN_Tsuboyama_2023_5JRT	1118
TPK1_HUMAN_Weile_2017	3181
YNZC_BACSU_Tsuboyama_2023_2JVD	714

under the ROC curve (AUC).

D.8. Minimum inhibitory concentration (MIC) value prediction

The model architecture is illustrated in Fig. 11e. The sequence encoder, pretrained with ProteinAligner, was fine-tuned to address this task. The classification module is a multi-layer perceptron (MLP) consisting of two fully connected layers, with a hidden size of 256 and a ReLU activation function. We employed the Adam optimizer with an initial learning rate of 1×10^{-4}, training the model for 200 epochs. Throughout the training process, the learning rate was dynamically adjusted at each epoch using the LambdaLR (Smith, 2017) scheduler. To assess the model's performance, we used mean squared error (MSE) as the evaluation metric.

D.9. Supervised protein fitness prediction

We evaluated ProteinAligner and baseline models on the ProteinGym Deep Mutational Scanning (DMS) benchmark, which comprises 217 assays with quantitative fitness measurements (Notin et al., 2023a). In the supervised setting, each model was trained on a subset of variants from each assay and evaluated on held-out single-mutation variants from the same assay. Predictive performance on unseen variants was assessed using Spearman's rank correlation coefficient(Spearman), Pearson correlation coefficient (Pearson), and coefficient of determination (R^2). To reduce computational cost, we randomly selected 15 single-substitution assays (listed in Table 2). The model architecture used for this task is shown in Fig. 11f: the sequence embedding produced by ProteinAligner is passed through a one-dimensional convolutional layer (kernel size 7, stride 1, same padding), followed by dropout (rate 0.1), a ReLU activation, and a single-layer multilayer perceptron. Fine-tuning was performed using the AdamW optimizer (learning rate 3×10^{-4}, weight decay 5×10^{-2}, batch size 64) with a cosine annealing schedule and 100 warm-up steps, for a total of 10,000 training steps.

E. Discussions

ProteinAligner introduces a comprehensive approach to protein representation learning by integrating amino acid sequences, continuous 3D structures, and literature texts into a unified framework. This multimodal design allows the model to capture complementary information from each modality, providing a richer and more holistic understanding of proteins. By employing contrastive alignment, ProteinAligner learns representations that incorporate structural and functional attributes alongside contextual knowledge from experimental literature, enabling superior performance across a range of challenging protein-related tasks (Figs. 2, 3, 4, 5, 6, 7, 8, 9, and 10). Its ability to bridge critical gaps in existing models demonstrates its potential for advancing research in protein biology, drug development, and biotechnology, highlighting the importance of multimodal frameworks in addressing complex biological challenges.

The superior performance of ProteinAligner over ESM-2 (Figs. 2, 4, 5, 6, 7, 8, 9, and 10) can be primarily attributed to the differences in their pretraining strategies. ESM-2 is pretrained exclusively on large-scale protein sequences using a masked language modeling objective, which allows it to capture local and global sequence patterns but lacks exposure to structural

or functional context. In contrast, ProteinAligner adopts a multi-modal pretraining framework that jointly leverages protein sequences, protein 3D structures, and textual descriptions of protein function. This integration of multiple biologically relevant modalities enables ProteinAligner to learn richer and more biologically grounded representations. Structural data encode critical spatial and physicochemical properties, such as residue-residue proximity, binding pocket geometry, and overall protein folding, which are not directly inferable from sequence alone. Additionally, functional text provides high-level semantic information about biological roles, molecular mechanisms, and cellular processes, complementing the syntactic patterns learned from sequence and the geometric insights from structure. By aligning representations across sequence, structure, and function during pretraining, ProteinAligner is able to internalize correspondences between syntax, shape, and semantics in protein biology. This cross-modal alignment enhances its ability to generalize across a broad range of downstream prediction tasks, including those requiring inference of higher-order biological properties. As a result, ProteinAligner consistently outperforms ESM-2, especially in tasks where structural context or functional semantics are essential for accurate prediction. This highlights the advantage of incorporating multimodal biological knowledge during pretraining and underscores the importance of moving beyond sequence-only approaches when modeling protein function and behavior.

ProteinAligner outperforms ProtST (Figs. 2, 4, 5, 6, 7, 8, 9, and 10) largely due to the incorporation of structural information during pretraining. While both models utilize protein sequences and functional text, ProtST does not leverage 3D structural data, limiting its ability to capture the spatial and physicochemical context critical to understanding protein function. ProteinAligner's multi-modal framework integrates sequence, structure, and function, enabling it to align these three complementary modalities and learn more holistic protein representations. Structural information provides key insights into residue interactions, conformational flexibility, and binding site geometry — factors that often underpin functional behavior but are not readily apparent from sequence or text alone. By incorporating this additional modality, ProteinAligner can better generalize to tasks that require nuanced understanding of protein conformation or interactions. Moreover, the joint modeling of structure with sequence and text enables the model to associate semantic functional descriptors with both linear and spatial features of proteins, which enhances interpretability and predictive power. This comprehensive representation explains ProteinAligner's consistent advantage over ProtST across a range of downstream tasks.

The performance advantage of ProteinAligner over ESM-IF1 and ESM-S (Figs. 2, 4, 5, 6, 7, 8, 9, and 10) can be attributed to its incorporation of functional text during pretraining, in addition to protein sequences and structures. While ESM-IF1 and ESM-S capture rich structural and sequential patterns, they are not exposed to explicit functional semantics, which limits their ability to align molecular features with biological meaning. In contrast, ProteinAligner leverages textual descriptions of protein function as an additional modality, enabling it to associate structural and sequence features with high-level functional attributes described in natural language. This grounding in functional text enhances the model's ability to recognize biologically relevant patterns that may not be evident from sequence or structure alone. For instance, two proteins with divergent sequences or conformations may share a similar function — a relationship that functional text helps to bridge. By jointly pretraining on sequence, structure, and function, ProteinAligner develops semantically informed representations that improve generalization across diverse downstream tasks, particularly those involving function prediction or annotation transfer. This integrated approach gives ProteinAligner an edge over structure-sequence-only models like ESM-IF1 and ESM-S, which lack direct exposure to the linguistic and conceptual framing of protein function.

ProteinAligner outperforms ESM-3 in most downstream tasks and across most evaluation metrics (Figs. 2, 5, 6, and 7), which can be primarily attributed to two key factors. First, ESM-3's pretraining strategy relies exclusively on masked token prediction and does not incorporate a contrastive learning objective. Without contrastive supervision, the model is not explicitly encouraged to bring representations of biologically equivalent inputs — such as a protein's sequence and its corresponding structure or functional annotation — closer together in the embedding space. As a result, ESM-3 may fail to establish coherent cross-modal alignments, leading to fragmented and less transferable representations that underperform on tasks requiring integrated biological reasoning, such as pathogenic variant classification (Fig. 2) and minimum inhibitory concentration (MIC) prediction (Fig. 7). In contrast, our method employs contrastive learning to align modalities explicitly, encouraging the model to capture shared semantics and fine-grained relations between different modalities, which significantly enhances generalization to downstream tasks. Second, the representational fidelity of ESM-3 is further limited by quantization noise introduced by its compression of each residue's continuous 3D neighborhood into a single discrete 'structure token' using a VQ-VAE (Razavi et al., 2019) encoder. This discretization process discards critical biophysical information — such as side-chain orientations (Lima et al., 2021), solvent accessibility (Ramakrishnan et al., 2023), and sub-Ångström backbone perturbations (McBride et al., 2023; Smith & Kortemme, 2008) — that are essential for capturing fine structural determinants of protein behavior. In contrast, our method's structure encoder, based

on the inverse-folding model ESM-IF1 (Hsu et al., 2022), operates directly on full backbone atomic coordinates without quantization. By leveraging rotation- and translation-equivariant geometric vector perceptron (GVP) graph layers (Jing et al., 2020), ProteinAligner inherits a strong geometric inductive bias that faithfully captures the spatial symmetries of protein folds (Jing et al., 2020). Aligning these geometry-aware structural embeddings with protein sequence embeddings propagates physically meaningful constraints that ESM-3 is unable to exploit, as its quantized structure consists of discretized tokens lacking built-in equivariance.

ProteinAligner also outperforms ProTrek in most tasks (Figs. 2, 4, 5, 6, 7, 8, 9, and 10), likely due to two main factors. First, while ProTrek employs a combination of masked token prediction and contrastive learning during pretraining, this multitask setup introduces inherent optimization conflicts. Masked language modeling (MLM) encourages the model to focus on local contextual reconstruction, optimizing for token-level accuracy. In contrast, contrastive learning promotes global alignment between modalities by pulling together semantically related representations and pushing apart unrelated ones. These two objectives often operate at different granularities and may impose competing gradient signals during training, which can lead to unstable convergence, diminished alignment quality, and suboptimal representation learning. Empirical studies have shown that when multitask objectives are not carefully balanced, they can interfere with each other, reducing the effectiveness of both (Yu et al., 2020; Liu et al., 2021; Kendall et al., 2018; Sener & Koltun, 2018). ProteinAligner avoids this issue by adopting a streamlined pretraining objective based solely on contrastive learning. This choice allows the model to concentrate on learning globally consistent, modality-aligned representations without the interference of reconstruction-based losses, resulting in more coherent and transferable embeddings for downstream tasks. Second, ProteinAligner's structural encoder preserves continuous 3D geometry using a geometric vector perceptron (GVP) (Jing et al., 2020) based architecture, whereas ProTrek represents protein structures by discretizing them into tokens using Foldseek (Van Kempen et al., 2024). This discretization introduces a loss of fine-grained spatial information — such as torsion angles, side-chain orientation, and atomic-level packing — that are crucial for modeling functionally relevant features.

ProteinAligner's ability to perform a wide range of prediction tasks presents promising applications across biology, drug discovery, and medicine. In drug discovery, ProteinAligner's accurate identification of bioactive peptides, such as DPP-IV inhibitors (Fig. 5a), is particularly relevant for developing treatments for metabolic disorders like diabetes. Its capability to predict antimicrobial peptide properties, such as minimum inhibitory concentration (MIC) (Fig. 7b), is critical for advancing new antimicrobial therapies, especially in addressing the challenge of antibiotic-resistant pathogens. This has important implications for the global fight against antimicrobial resistance. The model's ability to detect type I anti-CRISPR activities supports the design of more efficient and precise CRISPR-based tools for both research and therapeutic applications (Fig. 4b). Anti-CRISPR systems could be used to enhance the safety of gene editing by mitigating off-target effects or enabling reversible gene modifications. In precision medicine, the prediction of pathogenic missense variants aids in the early detection and diagnosis of genetic disorders. By identifying harmful mutations that may lead to diseases, ProteinAligner can contribute to personalized treatment strategies, improving patient outcomes (Fig. 2b). Protein fitness prediction accelerates enzyme engineering and therapeutic protein design by pinpointing mutations that enhance catalytic efficiency, stability, or specificity (Fig. 8b). Additionally, ProteinAligner's accurate prediction of protein thermostability is vital for protein engineering, biopharmaceutical development, and industrial biotechnology, where stable proteins are necessary for drug formulations and biocatalysts (Fig. 3b). Overall, ProteinAligner's diverse prediction capabilities position it as a valuable tool that can accelerate innovation in multiple fields, enabling faster therapeutic discoveries, more precise gene-editing tools, and advancements in personalized medicine and protein engineering.

Despite the advantages of ProteinAligner, the model has several limitations. One of the key challenges is the dependency on high-quality structural and textual data, which is not always available for all proteins. While ProteinAligner can perform pretraining even when only sequences and one additional modality (either structure or text) are present, the absence of full multimodal data for many proteins can limit the model's ability to learn comprehensive representations. Additionally, ProteinAligner's reliance on contrastive loss for modality alignment may not fully capture subtle biological nuances in cases where sequence, structure, and text data are not perfectly aligned. Another limitation is the computational cost associated with training multimodal models, especially when dealing with large-scale protein datasets that involve high-dimensional structural information and large text corpora. Finally, while ProteinAligner improves upon previous models by integrating structure, sequence, and text, it still does not account for other potentially informative modalities, such as protein-protein interactions or functional annotations from various databases, which could further enhance its predictive capabilities.

Future work on ProteinAligner could focus on several key directions to further enhance its performance and applicability. One promising area is the incorporation of additional modalities, such as protein-protein interaction networks and post-translational modifications. These additional data sources could provide deeper insights into protein behavior and interactions,

leading to even more robust and comprehensive protein representations. Another direction for future work is to improve the model's ability to handle incomplete or noisy data by developing more sophisticated alignment strategies that better tolerate inconsistencies between modalities. Enhancing the interpretability of ProteinAligner's predictions is also a critical area for future research, which could involve incorporating explainability techniques to make the model's decision-making process more transparent, particularly in cases where sequence, structure, and text data converge. Lastly, expanding ProteinAligner's applications beyond protein function and property prediction - such as protein design and structure prediction - could broaden its impact across a wide range of biological and biomedical challenges.

F. Figures

Due to the space limitation, all the figures are presented in the Appendix.

G. Code availability

The source code for ProteinAligner pretraining, along with the pretrained checkpoints, is available at https://github.com/Alexiland/ProteinAligner. Additionally, links to the code for downstream tasks can be found in the README file.

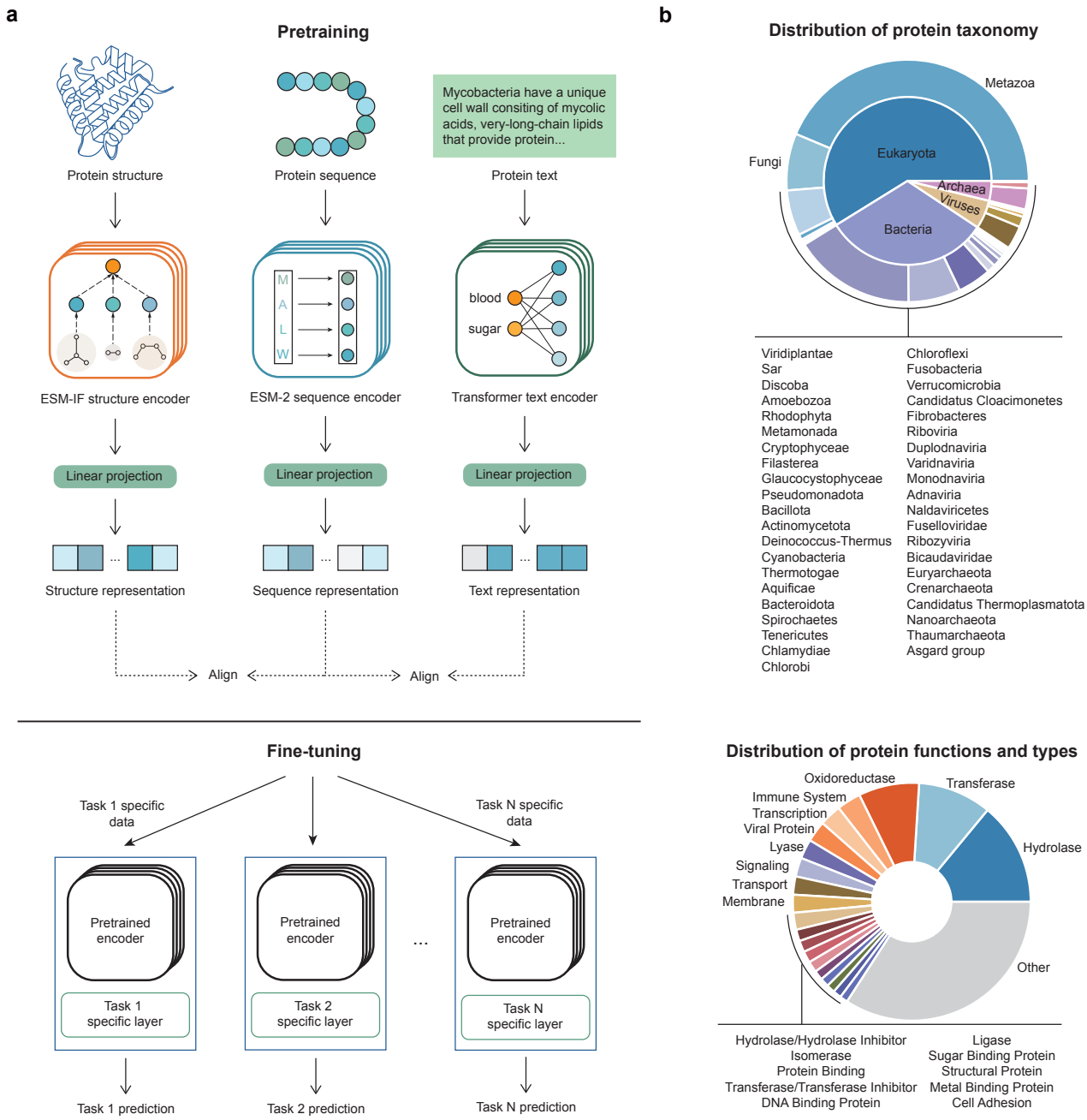

Figure 1. **ProteinAligner facilitates multimodal pretraining of protein foundation models by integrating diverse modalities including amino acid sequences, 3D structures, and textual data. a**, ProteinAligner consists of three encoders: a protein sequence encoder based on ESM-2 (650M), a protein structure encoder based on ESM-IF1, and a transformer-based protein text encoder. These encoders learn representations for protein sequences, structures, and text, respectively. Modality-specific projection modules then transform these representations into a shared latent space, enabling direct comparison across modalities. Using protein sequences as the anchor, ProteinAligner aligns the other two modalities by minimizing contrastive losses. After pretraining, the encoders can be fine-tuned with task-specific data for various downstream applications. **b**, Our curated pretraining data for ProteinAligner spans a diverse range of proteins from various taxonomic groups, functions, and types. The upper chart displays the distribution of protein taxonomy, with the inner ring representing superkingdoms and the outer ring representing kingdoms. The lower chart illustrates the distribution of protein functions and types.

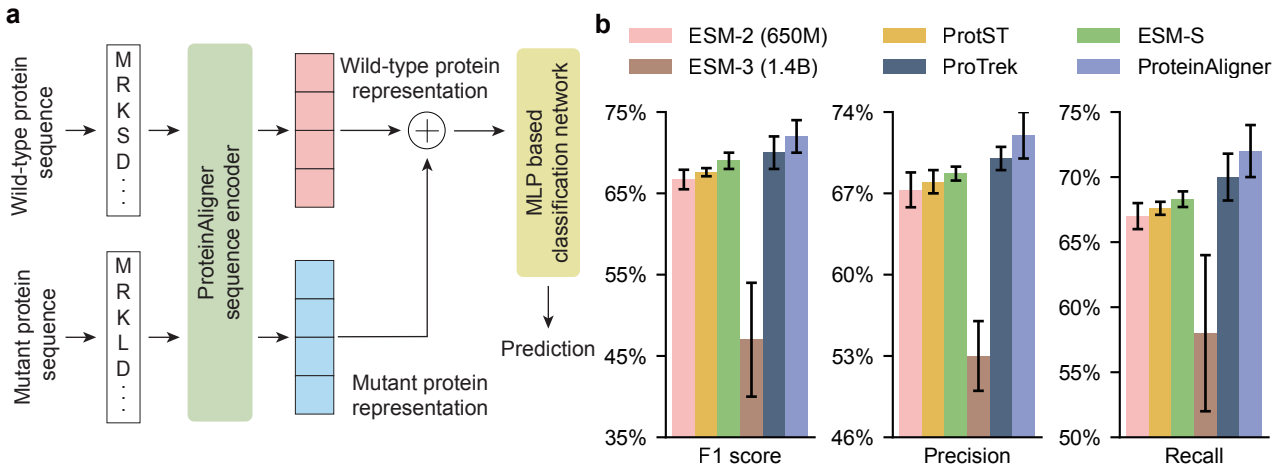

Figure 2. **ProteinAligner excels in predicting pathogenic missense variants compared to existing protein foundation models. a**, Model architecture used to fine-tune the pretrained ProteinAligner sequence encoder for this task. **b**, ProteinAligner achieves higher performance than ESM-2 (650M), ProtST, ESM-S, ESM-3 (1.4B), and ProTrek

, as measured by F1 score, precision, and recall. Error bars in all result figures indicate the standard deviation over five independent runs.

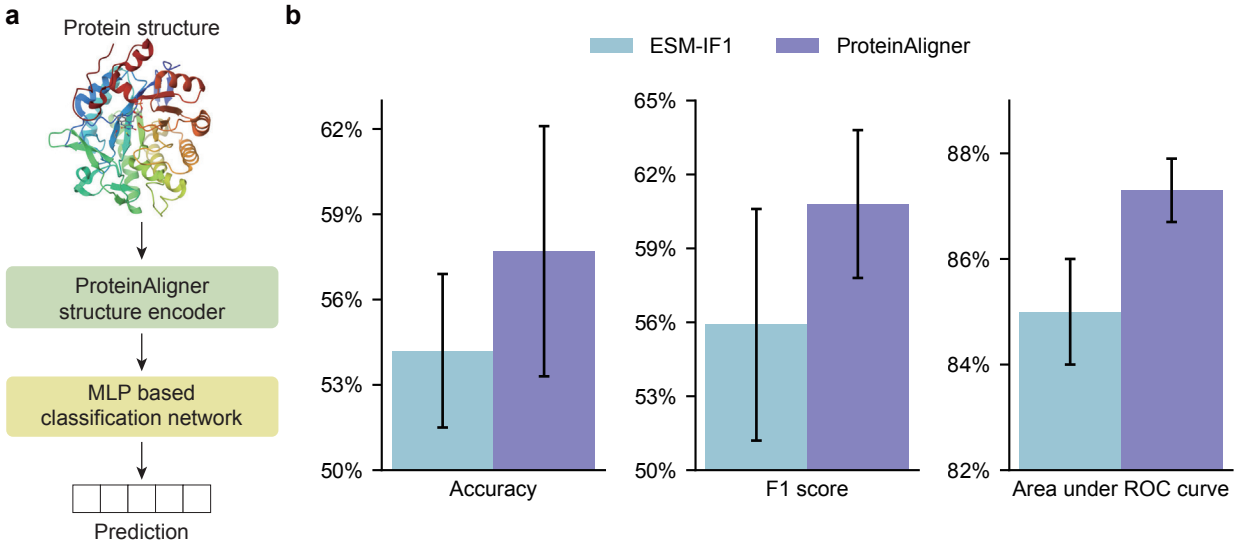

Figure 3. **ProteinAligner demonstrates superior performance in predicting protein thermostability compared to existing protein foundation models. a**, Model architecture used for fine-tuning the pretrained ProteinAligner structure encoder for this task. **b**, ProteinAligner outperforms ESM-IF1, achieving higher accuracy, F1 score, and area under ROC curve.

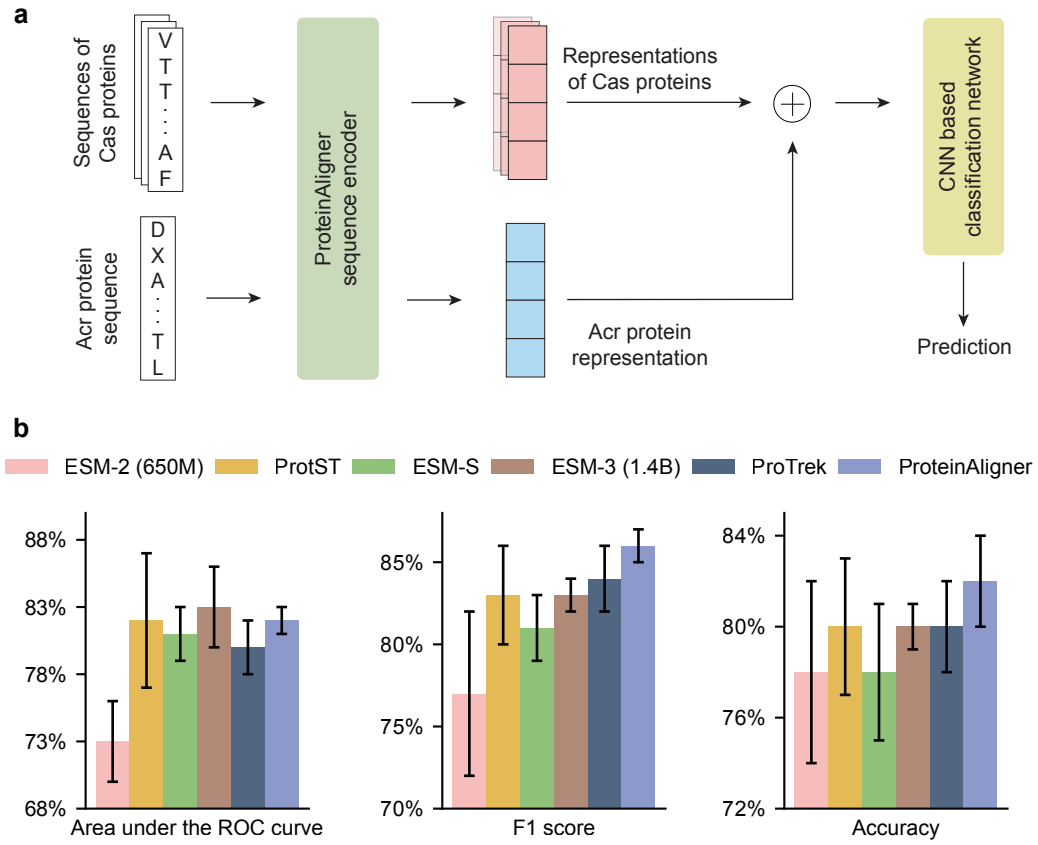

Figure 4. **ProteinAligner demonstrates strong performance in detecting type I anti-CRISPR activities. a**, Model architecture for fine-tuning the pretrained ProteinAligner sequence encoder for predicting type I anti-CRISPR activities. **b**, ProteinAligner outperformed all baselines in terms of F1 score and accuracy. For area under the ROC curve (AUC), it achieved the second-highest performance, slightly trailing ESM-3 (1.4B).

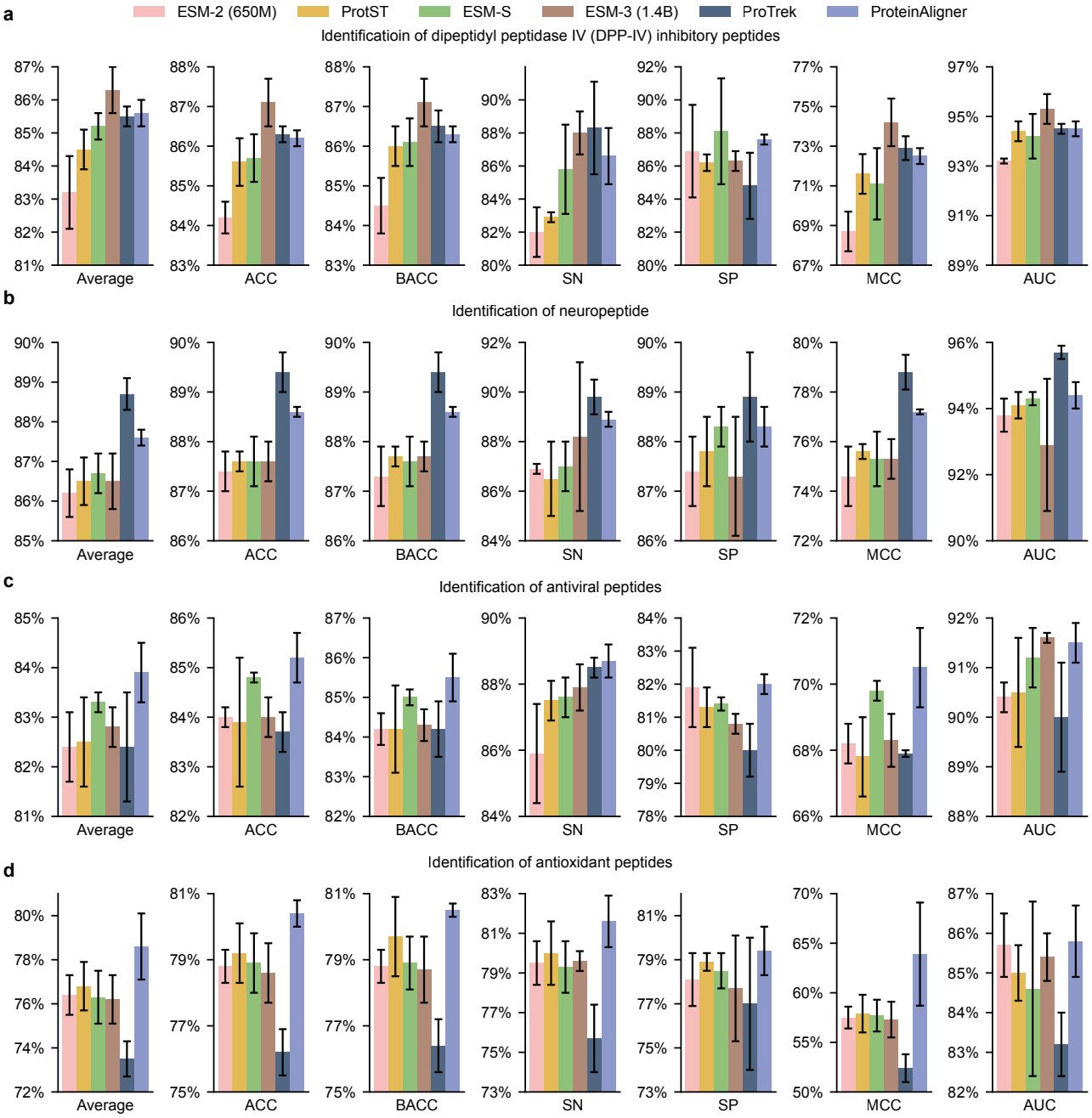

Figure 5. **ProteinAligner demonstrates superior performance in identifying potent bioactive peptides.** Across four tasks — predicting inhibition of dipeptidyl peptidase IV (DPP-IV) (**a**), modulation of brain activity (**b**), antiviral properties (**c**), and antioxidant activity (**d**) — ProteinAligner outperformed ESM-2 (650M), ProtST, and ESM-S in all tasks, and surpassed ESM-3 (1.4B) and ProTrek in three out of four tasks. Model performance was evaluated using accuracy (ACC), balanced accuracy (BACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), and the area under the ROC curve (AUC). "Average" refers to the mean value across these six metrics.

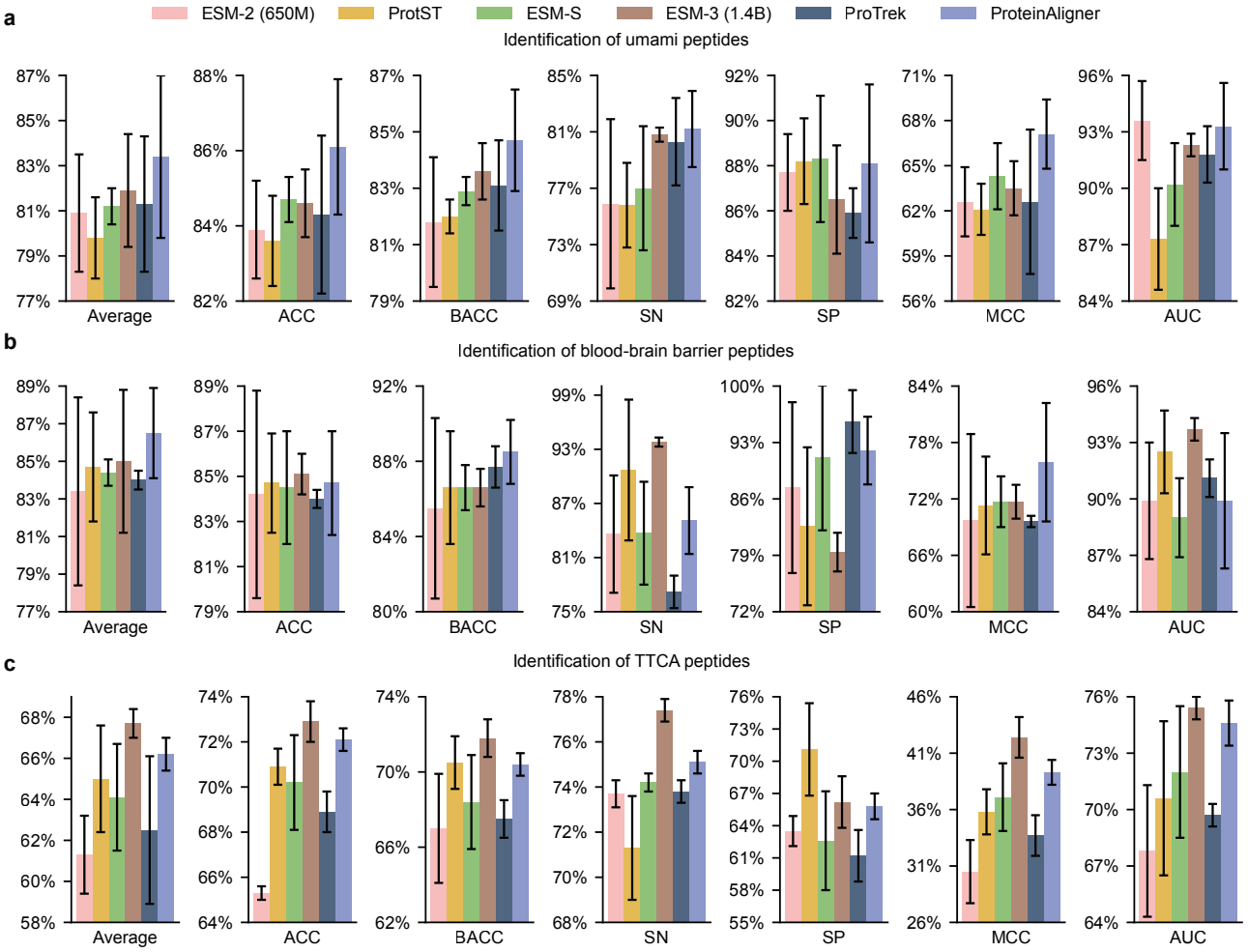

Figure 6. **ProteinAligner also demonstrates superior performance in three additional tasks related to bioactive peptide identification.** Specifically, in predicting umami taste induction (**a**), blood-brain barrier penetration (**b**), and T-cell immune response induction (**c**), ProteinAligner outperformed ESM-2 (650M), ProtST, ESM-S, and ProTrek across all tasks, and exceeded ESM-3 (1.4B) in two out of the three tasks. Model performance was assessed using accuracy (ACC), balanced accuracy (BACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), and the area under the ROC curve (AUC). "Average" refers to the mean value across these six metrics.

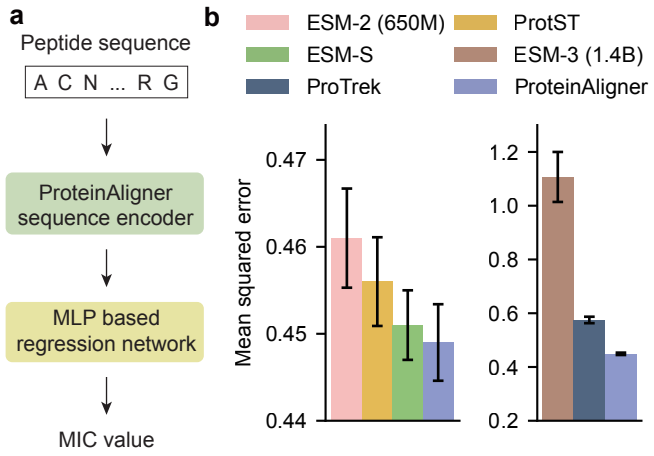

Figure 7. **ProteinAligner outperforms existing protein foundation models in predicting the minimum inhibitory concentration (MIC) of antimicrobial peptides. a**, Model architecture designed for this prediction task. **b**, ProteinAligner achieved a lower mean squared error than ESM-2 (650M), ProtST, ESM-S, ESM-3 (1.4B), and ProTrek.

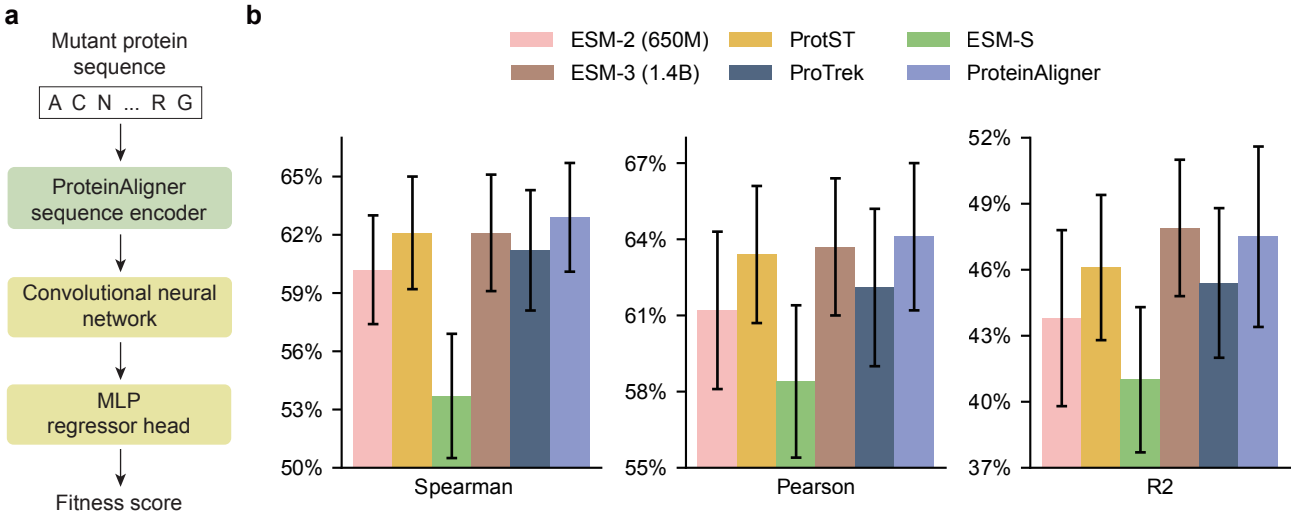

Figure 8. **ProteinAligner demonstrates competitive performance in protein fitness prediction. a**, Fine-tuning architecture for using the pretrained ProteinAligner sequence encoder to predict protein fitness. **b**, ProteinAligner outperformed all baselines across all evaluation metrics — including Spearman's rank correlation coefficient (Spearman), Pearson correlation coefficient (Pearson), and coefficient of determination (R^2) — with the exception of a slight deficit to ESM-3 in R^2.

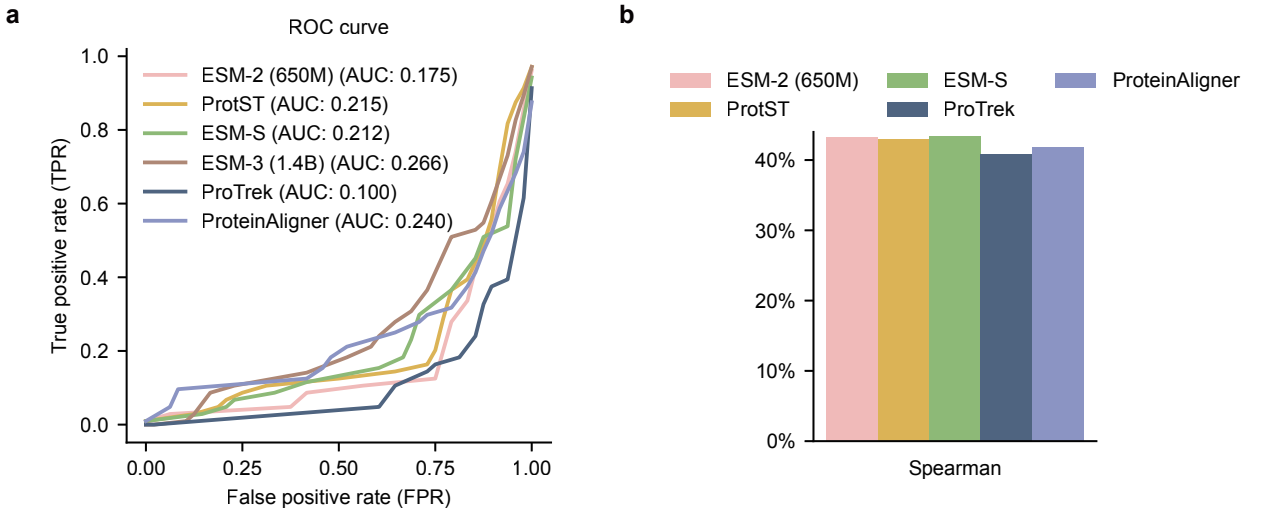

Figure 9. **ProteinAligner achieves competitive performance in zero-shot prediction settings. a**, ProteinAligner outperformed all baseline methods in zero-shot pathogenic missense variant prediction, as measured by area under the ROC curve (AUC). **b**, ProteinAligner achieved Spearman correlation scores comparable to those of baseline methods in zero-shot protein thermostability prediction.

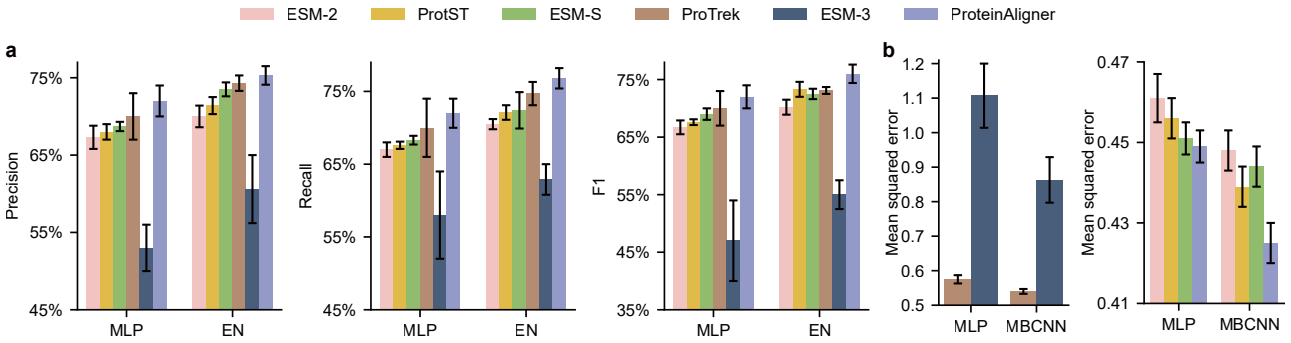

Figure 10. **ProteinAligner demonstrates robust performance across different prediction head architectures. a**, ProteinAligner achieved higher precision, recall, and F1 score than baseline models in pathogenic missense variant prediction when using either a multi-layer perceptron (MLP) or an Elastic Net (EN) classifier as the prediction head. **b**, ProteinAligner yielded lower prediction errors than baselines in minimum inhibitory concentration (MIC) regression for antimicrobial peptides when using either a multi-layer perceptron (MLP) or a multi-branch convolutional neural network (MBCNN) as the prediction head.

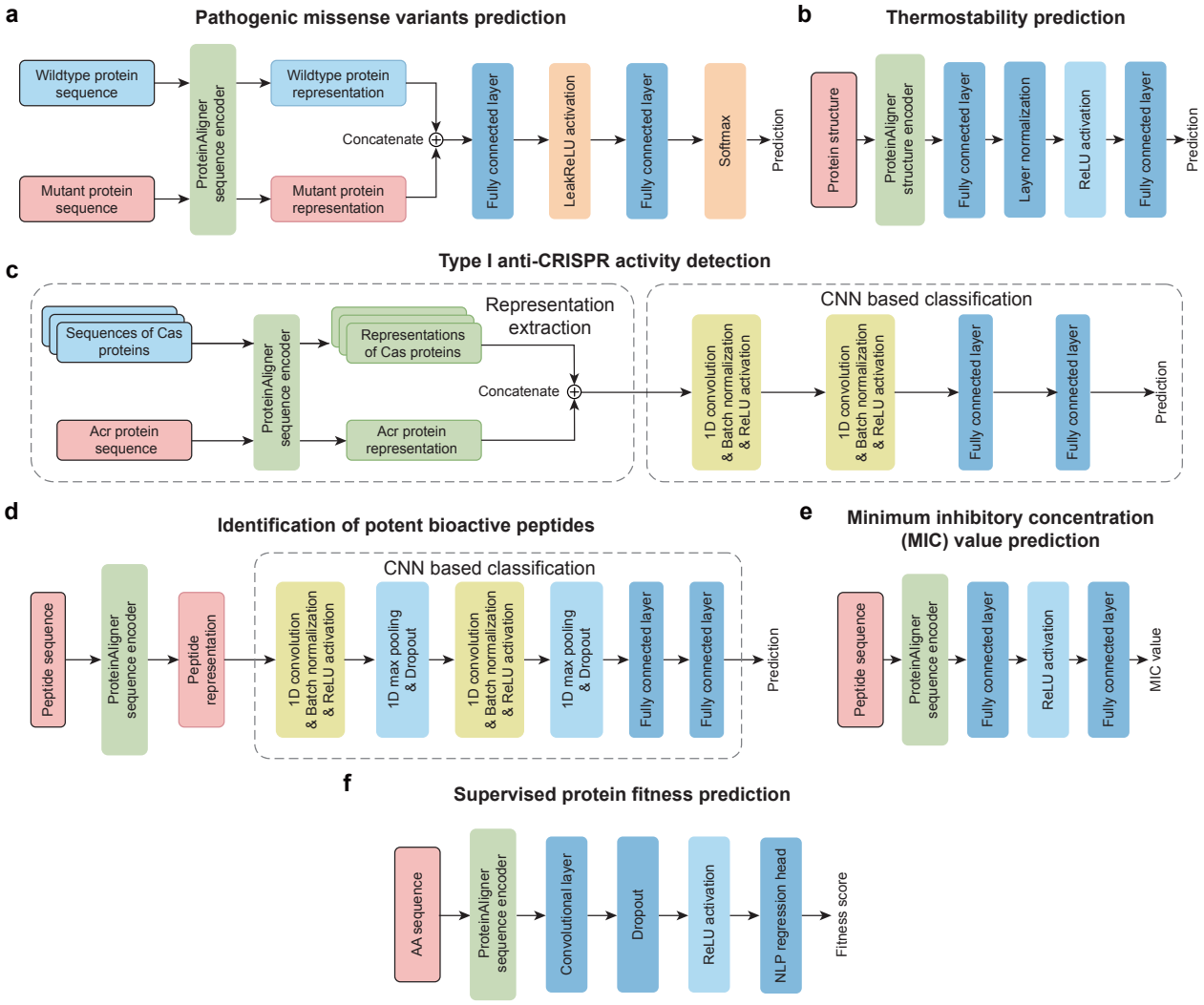

Figure 11. **Model architectures used in downstream tasks. a**, Model architecture used in pathogenic missense variants prediction. **b**, Model architecture used in thermostability prediction. **c**, Model architecture used in type I anti-CRISPR activity detection. **d**, Model architecture used for identifying potent bioactive peptides. **e**, Model architecture used for predicting minimum inhibitory concentration values. **f**, Model architecture used for supervised protein fitness prediction.

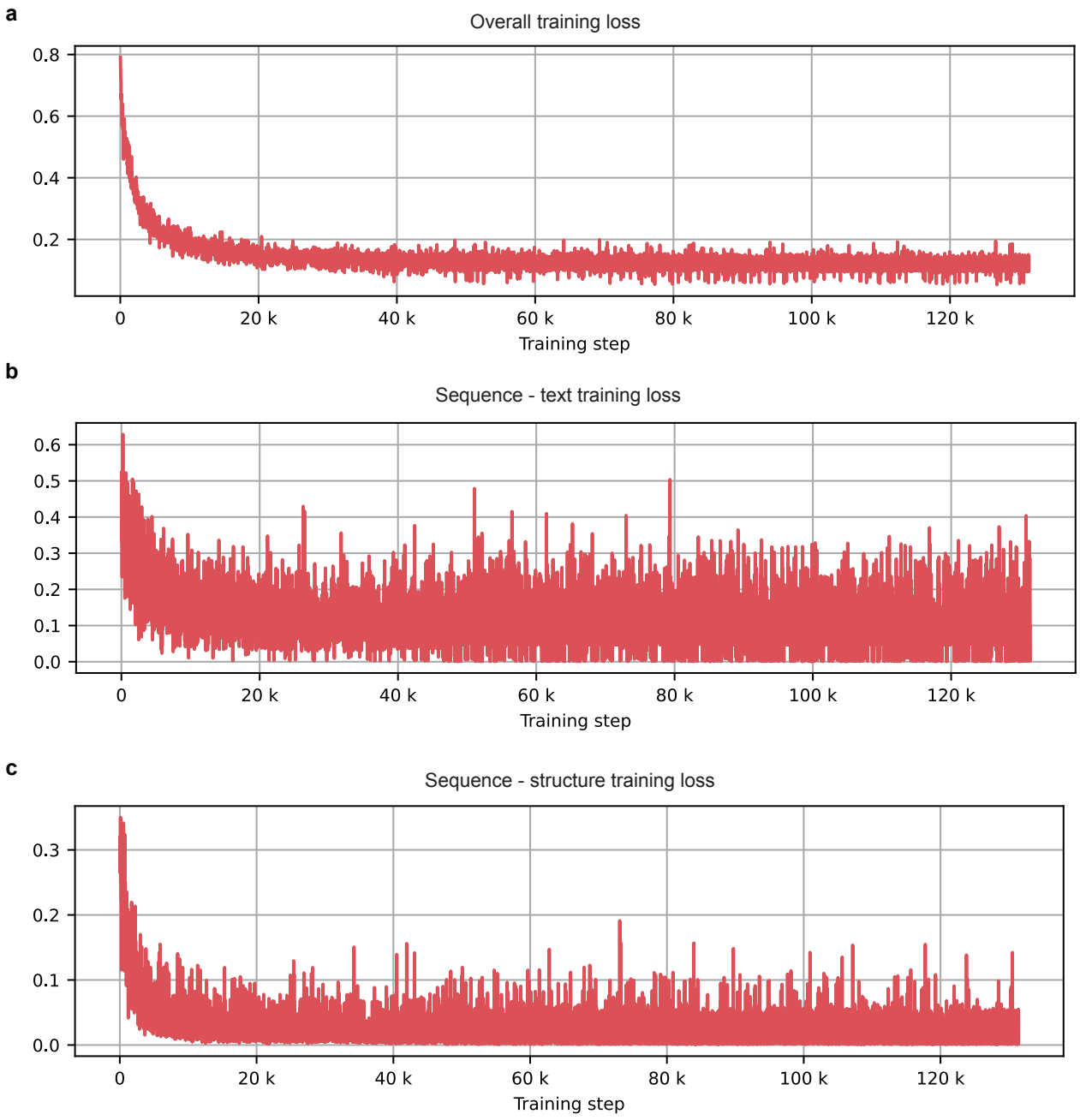

Figure 12. **Training dynamics of ProteinAligner. a**, Overall loss curve during pretraining. **b**, Sequence-text loss curve during pretraining. **c**, Sequence-structure loss curve during pretraining.