ProTrek: Navigating the Protein Universe through Tri-Modal Contrastive Learning

Jin Su[1], Xibin Zhou[1], Xuting Zhang[1], Fajie Yuan[1]

[1]Westlake University.

Correspondence to: yuanfajie@westlake.edu.cn;

Abstract

ProTrek redefines protein exploration by seamlessly fusing sequence, structure, and natural language function (SSF) into an advanced tri-modal language model. Through contrastive learning, ProTrek bridges the gap between protein data and human understanding, enabling lightning-fast searches across nine SSF pairwise modality combinations. Trained on vastly larger datasets, ProTrek demonstrates quantum leaps in performance: (1) Elevating protein sequence-function interconversion by 30-60 fold; (2) Surpassing current alignment tools (i.e., Foldseek and MMseqs2) in both speed (100-fold acceleration) and accuracy, identifying functionally similar proteins with diverse structures; and (3) Outperforming ESM-2 in 9 of 11 downstream prediction tasks, setting new benchmarks in protein intelligence. These results suggest that ProTrek will become a core tool for protein searching, understanding, and analysis.

Proteins, the fundamental molecular machines of life, orchestrate a vast array of biological processes. Deciphering their intricate structures, diverse functions, and complex interactions lies at the heart of modern biochemistry, molecular biology, and pharmacological innovation. Biologists are driven to unravel the tripartite complexity of a protein's sequence, structure, and function (SSF)—a molecular enigma that holds the key to unlocking the fundamental principles of life.

To address this challenge, the protein analysis landscape has evolved significantly, employing various computational approaches. Alignment-based tools like BLAST [3], MUSCLE [12] MMseqs2 [26], TM-align [37] and Foldseek [32] have made substantial contributions. However, these tools often focus on local alignment for efficiency, potentially overlooking global insights, and typically operate within a single modality—either sequence or structure—limiting their capacity to align multiple modalities concurrently. This limitation is further exacerbated by the fact that approximately 30% of proteins in UniProt [28] remain unannotated, primarily due to their sequences being phylogenetically distant from known functional homologs [4, 14]. The advent of neural network-based functional annotation tools [4, 38, 25] has ushered in a new era of protein characterization, enabling the identification of corresponding annotations for given proteins. Nevertheless, these annotation tools, primarily reliant on predefined labels, lack the sophistication to comprehend natural language, thereby impeding their ability to generate nuanced textual descriptions of protein functions or identify proteins from extensive textual narratives.

Recent breakthroughs in large language models, exemplified by ChatGPT [1] and LlaMA [30, 31], have demonstrated unprecedented capabilities across various natural language processing tasks. Concurrently, the field of computational biology has witnessed the emergence of protein language models (PLM) as a burgeoning area of exploration [19, 13, 2, 21, 36, 22]. Building upon these advancements, we posit the feasibility of establishing a foundational protein model capable of comprehensively representing protein SSF through sophisticated language modeling techniques.

Here, we introduce ProTrek, an innovative tri-modal PLM that jointly models protein SSF modalities. ProTrek leverages contrastive learning [24] through three core alignment training strategies (Fig. 1a): (1) bidirectional supervision between protein structure and sequence, (2) reciprocal supervision between protein functions and structures, and (3) mutual supervision between protein functions and sequences. This tri-modal alignment training paradigm enables ProTrek to forge tight associations among SSF by converging genuine sample pairs (sequence-structure, structure-function, and sequence-function) while simultaneously diverging negative samples within the latent space.

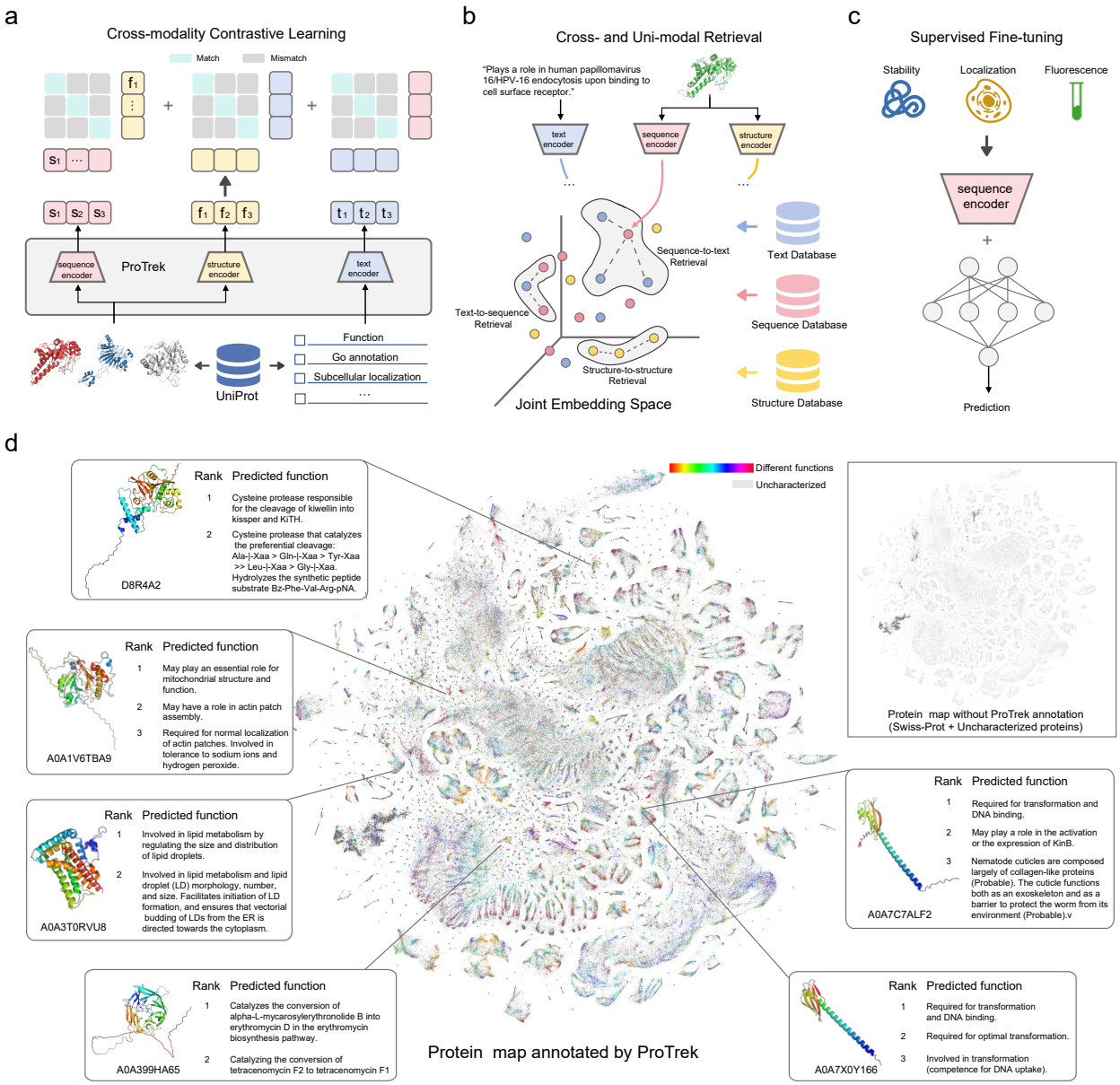


Fig. 1: **Illustration of ProTrek. a,** ProTrek architecture and tri-modal contrast learning. **b,** Cross-modal and uni-modal retrieval. ProTrek supports nine searching tasks. **c,** After the tri-modal contrast learning, the protein sequence encoder encodes almost universal representation of proteins, which can be fine-tuned to predict diverse downstream tasks, such as protein fitness and stability prediction. **d,** Using ProTrek's natural language capabilities to decode the protein universe. Each cluster represents proteins with close embedding distances. Over 99% of protein entries in UniProt remain unreviewed, as shown in the top right.

The architecture of ProTrek incorporates a pre-trained ESM encoder [19] for amino acid (AA) sequence modeling and a pre-trained BERT encoder [15] for nuanced natural language function representation. Protein structures are also transformed into discrete-token sequences utilizing Foldseek [32], facilitating their encoding via a BERT-style sequential network. That is, ProTrek ensures that each modality of SSF is modeled by a dedicated language model, creating a harmonious tripartite representation.

ProTrek was trained on a dataset comprising 40 million examples (see Supplementary Table 1 and 2), the largest collection of protein-text pairs at preprint release, including 14 million precise protein-text pairs curated from Swiss-Prot [5] database and 25 million relatively noisy pairs meticulously filtered from 300 million UniRef50 [7] samples (also see Supplementary Fig. 1). The ProTrek model's training regimen incorporates 8 loss functions—6 dedicated to inter-modal alignment and 2 masked language modeling losses to preserve recognition fidelity at both the amino acid and structural (a.k.a. 3Di) token levels (see Methods).

Through an exquisite fusion of protein SSF modalities into a unified framework, ProTrek offers three pivotal capabilities that redefine our approach to decoding the protein universe: (1) As a zero-shot retrieval model, ProTrek enables precise exploration of intricate SSF interrelationships through *all* nine distinct

search tasks (Fig. 1b), namely, sequence-to-structure, sequence-to-function, sequence-to-sequence, structure-to-structure, structure-to-sequence, structure-to-function, function-to-function, function-to-sequence, and function-to-structure retrieval. Notably, it bridges the chasm between protein data and human comprehension, translating complex molecular landscapes into an intuitive, language-driven experience (Fig. 1d). This paradigm shift empowers researchers with unprecedented access to nuanced natural language descriptions of protein properties and functions. (2) Harnessing the power of 'global' representation learning, ProTrek effectively overcomes the 'local' constraints inherent in current sequence comparison tools. This enables the identification of proteins with convergent functions despite divergent structures and sequences (Fig. 2b)—a phenomenon potentially more ubiquitous in nature [18, 29]. (3) Through SSF cross-modal contrastive learning, ProTrek injects structural and functional information into amino acid sequences, thereby catalyzing effective transfer learning and enabling fine-tuning across a myriad of downstream tasks, rivaling and complementing leading protein language models(Fig. 1c,2e). This versatility extends its applicability across diverse domains of protein science.

ProTrek is available in two variants: a large-scale version featuring a 650M AA sequence encoder, 150M structure encoder, and 130M function encoder; and a more compact version with a 35M sequence encoder, 35M structure encoder, and 130M function encoder. These configurations are designed to accommodate varying computational resources and research objectives. During inference, ProTrek employs a max-inner product search (MIPS) algorithm [11] for efficient retrieval. This enables ProTrek to complete searches within seconds, even when querying databases containing billions of entries. Such rapid search capability significantly enhances the efficiency of protein research, providing a powerful tool for exploring the protein universe.

Here, we firstly evaluated ProTrek's performance on four emerging retrieval tasks, designed to bridge protein sequences/structures with their natural language functional descriptions. These tasks comprise bidirectional pairwise searches: sequence-to-text, structure-to-text, text-to-sequence, and text-to-structure (Fig. 2a). ProTrek's performance was benchmarked against two state-of-the-art methods, ProteinDT (also known as ProteinCLAP)[20] and ProtST [35], using the human-reviewed Swiss-Prot [5] dataset. This evaluation dataset comprises 4,000 proteins and their associated functional descriptions. For a fair evaluation, proteins sampled in the test set have less than 50% sequence identity with those in the training set. Additionally, we incorporated 100,000 randomly sampled proteins from UniProt as unknown negative examples to further assess ProTrek's generalization ability (see Methods).

Notably, ProTrek demonstrates superior performance across most functional categories, regardless of whether the AA sequence encoder or 3Di structure encoder is employed. It greatly outperforms both ProteinDT and ProtST, achieving remarkable improvements of over 30-fold and 60-fold in global retrieval tasks. Specifically, ProTrek attains scores of 0.233 in sequence-to-text and 0.190 in text-to-sequence search tasks, compared to ProtST's 0.007 and 0.001, and ProteinDT's 0.004 and 0.003, respectively (Fig. 2a). This exceptional performance can be attributed to ProTrek's extensive training dataset, which surpasses those of ProteinDT and ProtST by two orders of magnitude. Beyond basic search capabilities, this substantial advantage positions ProTrek as a highly promising biological tool with potential applications in text-based protein design and protein question-answering systems [9, 20, 34, 33]. Within ProTrek's architecture, the AA sequence encoder generally exhibits superior performance compared to the structure encoder, a discrepancy potentially linked to differences in encoder model size and pre-training status. Fig. 1d illustrates an example of how ProTrek embeddings and its natural language understanding capabilities can be utilized to decipher the vast protein universe.

Secondly, we assessed ProTrek's capability in searching functionally similar proteins, benchmarking it against several established alignment-based tools, including MMseqs2 [26], DIAMOND [6], BLASTP [3], and the cutting-edge structure aligner, Foldseek [32] (Fig. 2b,c). While these classical methods rely on local alignment through sequence or structure similarity, ProTrek employs a distinctive 'global' alignment approach via cross-modal contrastive learning. To illustrate the difference, we evaluated the functions of retrieved proteins across four search tasks: two cross-modal searches bridging sequences and structures (sequence-to-structure and structure-to-sequence), and two uni-modal searches within sequences (sequence-to-sequence) and within structures (structure-to-structure). We conducted an exhaustive all-versus-all search on the Swiss-Prot test dataset, comprising 4,000 proteins, and compared the performance of each method in identifying proteins sharing the same GO annotation [23] up to the 5th FP (false-positive) (see Methods).

ProTrek outperforms all these classical sequence alignment tools in terms of the total/average number of correct hits (Fig. 2c). While all methods can readily identify functionally similar proteins when they share higher TM-scores with query proteins, ProTrek's advantage becomes more pronounced when identifying functionally similar proteins with lower TM-scores. This is because proteins with similar functions can display distinct sequences or structures (Fig. 2b), as also exemplified by numerous isozymes [17] and functionally convergent proteins found in nature. Regarding ProTrek's four search modalities, sequence-to-sequence search exhibits the highest efficacy, followed by structure-to-structure and sequence-to-structure searches. Each search task, however, offers distinct advantages. For example, while the structure-to-sequence search shows slightly lower effectiveness, it has potential applications in filtering or ranking protein sequences

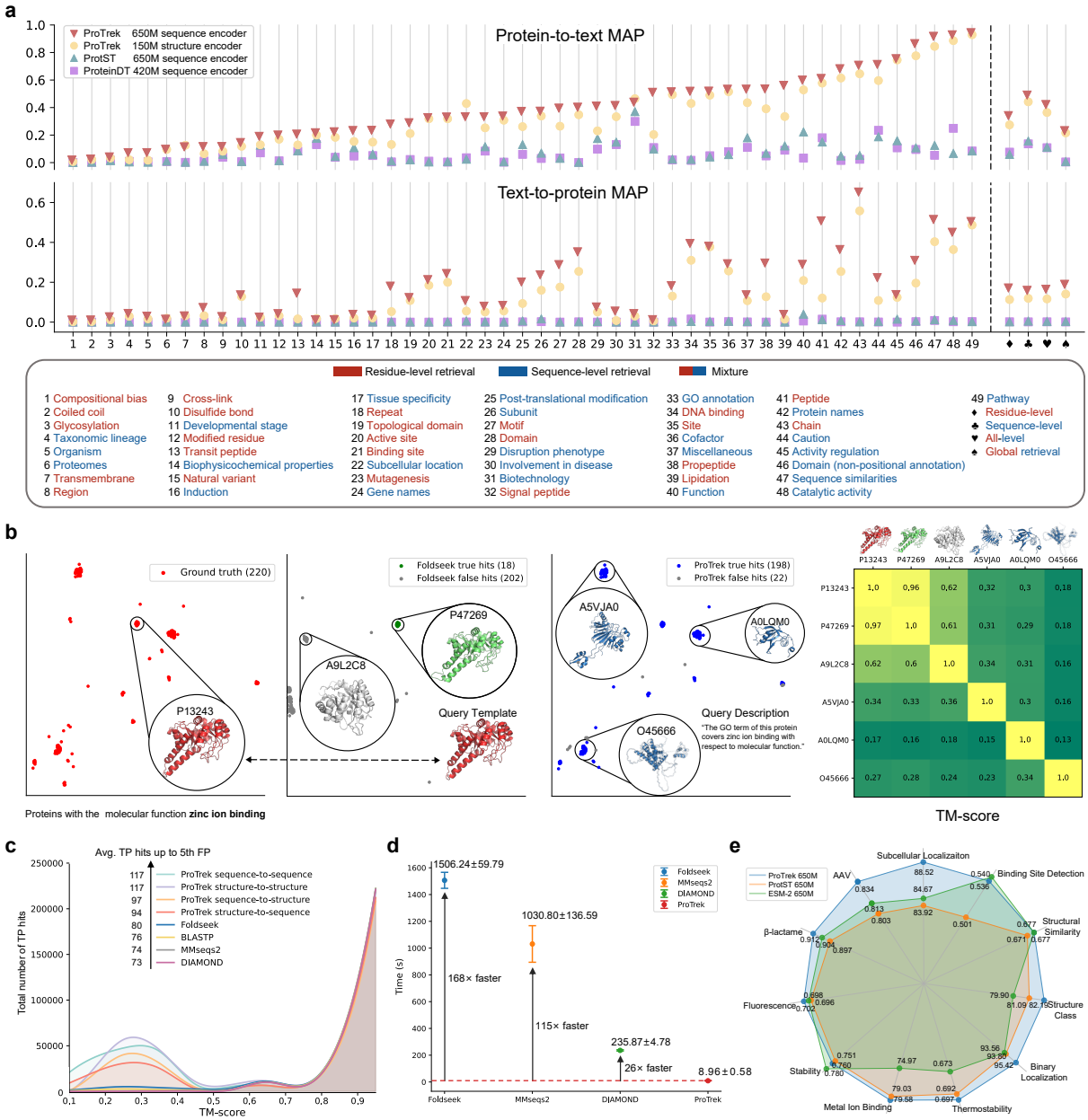


Fig. 2: **ProTrek performance on protein search and representation tasks. a**, Top chart: Search protein functional descriptions using sequences/structures. Bottom chart: Search protein sequences/structures using textual descriptions. x-axis: Specific protein function categories left of the dashed line; aggregated categories (residue-, protein-, all-level) right of it. "Global retrieval" indicates a search across the entire database, not within individual categories. The y-axis is MAP (mean average precision), a commonly used ranking metric for searching tasks. **b**, ProTrek employs "zinc ion binding" as the query term, while Foldseek utilizes P13243 as a query template, which is the protein with the most hits. In the testing set, 220 proteins share similar functional annotations with P13243. Foldseek identified 18 true hits, whereas ProTrek discovered 198 true hits. The TM-score results in the right subfigure reveal that proteins with similar functions can exhibit diverse structures. Conversely, proteins with similar structures (e.g., A9L2CB) may encode different functions. **c**, Searching proteins with similar functions using protein sequence/structure as input. TP (true-positive): Matches sharing ≥1 GO term. FP: Matches sharing no GO terms. **d**, Comparing alignment speed (CPU time) for 100 query proteins on UniRef50 with 50 million candidate proteins, utilizing 24 CPU cores. **e**, Evaluating the protein representation ability of the ProTrek AA sequence encoder.

generated by AI-based protein design models (such as ProteinMPNN [10]). This versatility underscores ProTrek's comprehensive approach to protein analysis and its potential to enhance various aspects of protein research and design.

ProTrek's inference speed, bolstered by the max-inner product search (MIPS) algorithm, represents another significant advantage. This algorithmic prowess enables ProTrek to execute lightning-fast searches and sorting operations on billion-scale databases within seconds, outpacing Foldseek and MMseqs2 by over two orders of magnitude (Fig. 2d). Correspondingly, it demonstrates a speed enhancement of approximately 400,000-fold compared to TM-align [37] and Dali [16] (as reported in [32]). This exceptional search velocity

empowers researchers to swiftly navigate and analyze the vast protein universe with unprecedented efficiency. Furthermore, ProTrek's unique cross-modal search capabilities set it apart from conventional alignment tools, underscoring its pivotal role as a transformative addition to the bioinformatics toolkit.

Beyond its exceptional search capabilities, ProTrek's amino acid and 3Di sequence encoders serve as universal representation models, owing to two unsupervised loss functions: contrastive learning and masked language modeling. We rigorously evaluated these models across a diverse spectrum of over 10 downstream protein tasks, encompassing both protein-level and residue-level analyses, as well as regression and classification challenges. Through downstream supervised fine-tuning, ProTrek demonstrated superior performance, surpassing the established ESM-2 model and the state-of-the-art ProtST model in 9 out of 11 tasks (Fig. 2e). Notably, in several tasks, ProTrek exhibited a remarkable performance boost, exceeding ESM-2 or ProtST by relative margins of 5% to 7%. These results underscore ProTrek's robust transfer learning capabilities and its potential as a powerful representation model for a wide array of protein-related applications.

We have made the ProTrek model weights publicly available (https://huggingface.co/westlake-repl/ProTrek_650M_UniRef50) and developed a web server[1] (https://huggingface.co/spaces/westlake-repl/Demo_ProTrek_650M_UniRef50) for multi-database searches across Swiss-Prot, AlphaFoldDB, UniRef50, and PDB. We've also released protein sequence embeddings for all proteins within these databases. To enhance accessibility, we've created a Google Colab implementation of ProTrek (https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/ColabProTrek.ipynb) for biologist-friendly fine-tuning on new data, inspired by the OPMC concept in [27].

ProTrek's high accuracy and efficiency in cross-modal and uni-modal searches make it a valuable tool for protein database analysis, empowering biologists to navigate the vast protein universe more effectively. Its sophisticated natural language understanding capabilities transcend traditional keyword-matching limitations, enabling nuanced and context-aware searches (see Supplementary Fig.4). As a robust foundational model, ProTrek facilitates complex protein-text interactions, paving the way for emerging applications such as text-guided protein design [9, 20] and advanced protein ChatGPT systems [33]. These developments have the potential to significantly advance protein research and engineering.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[3] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[4] Maxwell L Bileschi, David Belanger, Drew H Bryant, Theo Sanderson, Brandon Carter, D Sculley, Alex Bateman, Mark A DePristo, and Lucy J Colwell. Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6):932–937, 2022.

[5] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[6] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368, 2021.

[7] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

[8] Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, pages 2024–08, 2024.

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

[1] Initial startup may take a few minutes due to sleep mode. The ProTrek server will support larger databases like BFD (https://bfd.mmseqs.com/), MGnify (https://www.ebi.ac.uk/metagenomics), and OMG [8], containing billions of protein sequences.

[10] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[11] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[12] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[13] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, 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.

[14] A Gane, ML Bileschi, D Dohan, E Speretta, A Héliou, L Meng-Papaxanthos, H Zellner, E Brevdo, A Parikh, MJ Martin, et al. Protnlm: model-based natural language protein annotation. *Preprint at https://storage. googleapis. com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft. pdf*, 2022.

[15] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[16] Liisa Holm. Using dali for protein structure comparison. *Structural Bioinformatics: Methods and Protocols*, pages 29–42, 2020.

[17] Ro L Hunter and CL Markert. Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science*, 125(3261):1294–1295, 1957.

[18] Julia Koehler Leman, Pawel Szczerbiak, P Douglas Renfrew, Vladimir Gligorijevic, Daniel Berenberg, Tommi Vatanen, Bryn C Taylor, Chris Chandler, Stefan Janssen, Andras Pataki, et al. Sequence-structure-function relationships in the microbial protein universe. *Nature communications*, 14(1):2351, 2023.

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

[20] Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023.

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

[22] Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks. *bioRxiv*, pages 2024–02, 2024.

[23] Mikhail Pomaznoy, Brendan Ha, and Bjoern Peters. Gonet: a tool for interactive gene ontology analysis. *BMC bioinformatics*, 19:1–8, 2018.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[26] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[27] Jin Su, Zhikai Li, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, Dacheng Ma, Sergey Ovchinnikov, Fajie Yuan, et al. Saprothub: Making protein modeling accessible to all biologists. *bioRxiv*, 2024.

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

[29] Annabelle E Todd, Christine A Orengo, and Janet M Thornton. Evolution of protein function, from a structural perspective. *Current opinion in chemical biology*, 3(5):548–556, 1999.

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

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

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

[34] Chentong Wang, Yannan Qu, Zhangzhi Peng, Yukai Wang, Hongli Zhu, Dachuan Chen, and Longxing Cao. Proteus: exploring protein structure generation for enhanced designability and efficiency. *bioRxiv*, pages 2024–02, 2024.

[35] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pages 38749–38767. PMLR, 2023.

[36] Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

[37] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

[38] Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20:1–23, 2019.

Methods

Protein-function pair construction

For a given protein, we created protein-function pairs using the descriptive information from almost all subsections in the UniProt database [7]. We categorized the subsection information of a protein into two types: sequence-level and residue-level (see Supplementary Table 2). Sequence-level information consists of one or multiple sentences that describe the overall characteristics of the protein, such as its function and catalytic activity. Residue-level information contains phrases describing specific information about certain residues in the protein sequence, which cannot be used directly, such as binding sites and active sites. For residue-level subsections, we employ GPT-4 [39] to generate multiple specific template sentences, thus organizing the information into a coherent sentence. For sequence-level subsections, we also utilize GPT-4 to paraphrase the descriptions and generate multiple alternative sentences to enhance the model's robustness to textual input. In the end, each subsection in a given protein results in a text sentence, and these sentences are paired with the protein to form protein-function pairs, which are used for training or evaluation purposes.

Pre-training dataset construction

We first performed a 50% sequence similarity clustering on the human-reviewed Swiss-Prot database [5]. We designated 1000 clusters for validation and another 1000 clusters for testing, using the remaining data as the training set. For each protein, we constructed the protein-function pairs as described above, resulting in a final training set of 14 million protein-function pairs. This high-quality dataset was used to train an initial version of ProTrek. The initial ProTrek model, a 35M version, was pre-trained on 12 NVIDIA 80G A100 GPUs over 100K steps.

Next, we used the initial ProTrek model to score and filter 300 million protein-function pairs from UniRef50 [7]. We retained all protein-function pairs with model scores higher than the average score on Swiss-Prot, resulting in 25 million pairs. These 25 million pairs, combined with the original 14 million pairs, form the final pre-training dataset.

Pre-training loss function

We adopted InfoNCE loss [43] for protein sequence–structure-function contrastive learning. The InfoNCE loss can be detailed as:

$$\mathcal{L}_{InfoNCE} = -log \frac{exp(f(z_i, z_j)/\tau)}{\sum_{k=1}^{N} exp(f(z_i, z_k)/\tau)} \tag{1}$$

Here, z_i and z_j represent the embeddings of any two modalities. $f(z_i, z_j)$ denotes a similarity score between the embeddings. N is the total number of pairs in a batch and τ is a learnable temperature parameter. We calculated the InfoNCE loss for the sequence-structure, sequence-function and structure-function pairs from two directions, resulting in 6 contrastive learning functions.

We additionally added 2 Masked Language Modeling (MLM) [40] loss functions to the ProTrek sequence and structure encoder to maintain model's recognition at the amino acid and 3Di token levels. The training objective is to predict masked tokens by capturing dependencies between masked positions and surrounding context, with the loss function formally described as:

$$\mathcal{L}_{MLM} = \sum_{i \in T} -logP(s_i | S_{\setminus T}) \tag{2}$$

where T is a set of token positions to be masked and $S_{\setminus T}$ represents a protein/structural sequence with tokens at specific positions masked. We formed the final loss function by averaging 2 MLM and 6 InfoNCE loss functions.

Pre-training setting

We initialized the ProTrek sequence encoder from the ESM-2 650M [19] and the text encoder from the PubMedBERT [15], and we randomly initialized the ProTrek structure encoder as there was no available pre-trained model. For the optimization, we used the DeepSpeed strategy [44] and employed the AdamW optimizer [42], setting $\beta_1 = 0.9, \beta_2 = 0.98$ and we utilized L_2 weight decay of 0.01. We gradually increased the learning rate from 0 to 4e-4 over the first 2000 steps decreased it to 4e-5 using the cosine annealing schedule [41]. The overall training phase lasted approximately 100K steps trained on 20 NVIDIA 80G A100 GPUs. We truncated protein and structural sequences to a maximum of 512 tokens, and truncated the test descriptions to a maximum of 100 tokens. Our total batch size consisted of 1280 protein sequence-structure-function pairs. Additionally, we employed mixed precision training to train ProTrek.

Protein-function retrieval benchmark

We utilized 4,000 proteins from the Swiss-Prot test set to construct a benchmark for the protein-function task. To evaluate ProTrek's generalization performance with proteins in other databases, we included 100,000 randomly sampled proteins from UniProt as negative samples in the test set. The textual descriptions of all these proteins were added to construct the text collection. For the protein-to-text search task, the model needed to retrieve the functional descriptions that mostly match the proteins in the Swiss-Prot test set. Conversely, for the text-to-protein task, the model had to find Swiss-Prot proteins that match the query function description (only human-reviewed descriptions were used as query texts) from a set of nearly 104,000 negative sample proteins. We selected the top 33 function subsections based on the number of protein-text pairs for evaluation. This approach aims to mitigate the bias impact of subsections with fewer pairs.

Detecting proteins with similar functions

For ProTrek and all baseline methods, we conducted an all-versus-all search using the same Swiss-Prot test dataset consisting of 4,000 proteins and compared their performance for finding proteins of the same GO annotation. for a given query protein, G_q represents the set of GO annotations associated with that protein. Similarly, for each hit retrieved from the database, G_h represents the set of GO annotations assigned to that hit. We defined a hit as correct if there was at least one common GO annotation between G_q and G_h, denoted as $G_q \cap G_h \neq \emptyset$. To generate the y-axis values in Fig. 2c, we counted the number of correct hits across all query proteins in the test set and summed them up.

Mean average precision

Mean Average Precision (MAP), a commonly used metric for information retrieval tasks, is the Average Precisions (APs) for each query. For a given query, AP is calculated by the following formula:

$$AP = \frac{1}{R} \sum_{i=1}^{N} P(i) r(i) \tag{3}$$

where R is the number of relevant results and N denotes the total number of results. For the ith result, $P(i)$ is the precision at the i-th position in the ranked list, and $r(i)$ indicates whether the result is relevant to the query (1 if relevant and 0 if not relevant). The MAP is calculated by averaging the APs from all queries:

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q) \tag{4}$$

Deep learning baseline models

For protein-function searches, we compared ProTrek to ProtST-ESM-1b from ProtST [35] and ProtBERT_BFD-512-1e-5-1e-1-text-512-1e-5-1e-1-InfoNCE-0.1-batch-9-gpu-8-epoch-5 from ProteinDT [20]. For the downstream protein tasks, we used the ESM-2 650M version [19] and ProtST-ESM-1b as baselines.

Downstream task fine-tuning

We strictly followed SaprotHub's fine-tuning settings [27], including dataset pre-processing, splitting, fine-tuning hyper-parameters, etc. We fine-tuned all model parameters until convergence and selected the best checkpoints based on their performance on the validation set.

Tool execution commands

We ran each alignment-based tool according to its standard usage example:

Foldseek

We used Foldseek with the **ef4e960ab84fc502665eb7b84573dfff9c2aa89d** version. The command line was executed with default parameters: **foldseek easy-search pdb_dir targetDB aln.m8 tmpFolder**

MMseqs2

We used MMseqs2 with the **edb8223d1ea07385ffe63d4f103af0eb12b2058e** version. The command line was executed with default parameters: **mmseqs easy-search seqs.fasta targetDB alnRes.m8 tmp**

BLASTP

We used BLASTP with the **Protein-Protein BLAST 2.15.0+** version. We ran BLASTP from the command line: **blastp -query seqs.fasta -db db -outfmt 6 -out blastp_result**

DIAMOND

We used DIAMOND with the **v2.1.9.163** version. We ran DIAMOND in very-sensitive mode following the standard case: **diamond blastp -q seqs.fasta -d db -o result.tsv –very-sensitive -k 0**

Data availability

The pre-trained ProTrek weights can be downloaded from https://huggingface.co/westlake-repl/ProTrek_650M_UniRef50. All pre-computed protein sequence, structure and text embeddings for Swiss-Prot database are available at https://huggingface.co/datasets/westlake-repl/faiss_index_ProTrek_650M_UniRef50 (will release protein embeddings for other database soon). The structural 3Di sequences are available at https://github.com/steineggerlab/foldseek. The text descriptions are available at https://www.uniprot.org/

Code availability

ProTrek is open-sourced under the MIT license. The code repository is available at https://github.com/westlake-repl/ProTrek. The ProTrek webserver is located at https://huggingface.co/spaces/westlake-repl/Demo_ProTrek_650M_UniRef50. The ColabProTrek is available at https://colab.research.google.com/drive/1On2xQU0d7351bIBgZpz2T0VUp2gZium0?usp=sharing.

References

[39] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[42] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[44] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

Acknowledgements

We thank Sergey Ovchinnikov and Chentong Wang for their valuable suggestions to improve this paper.

Supplementary information

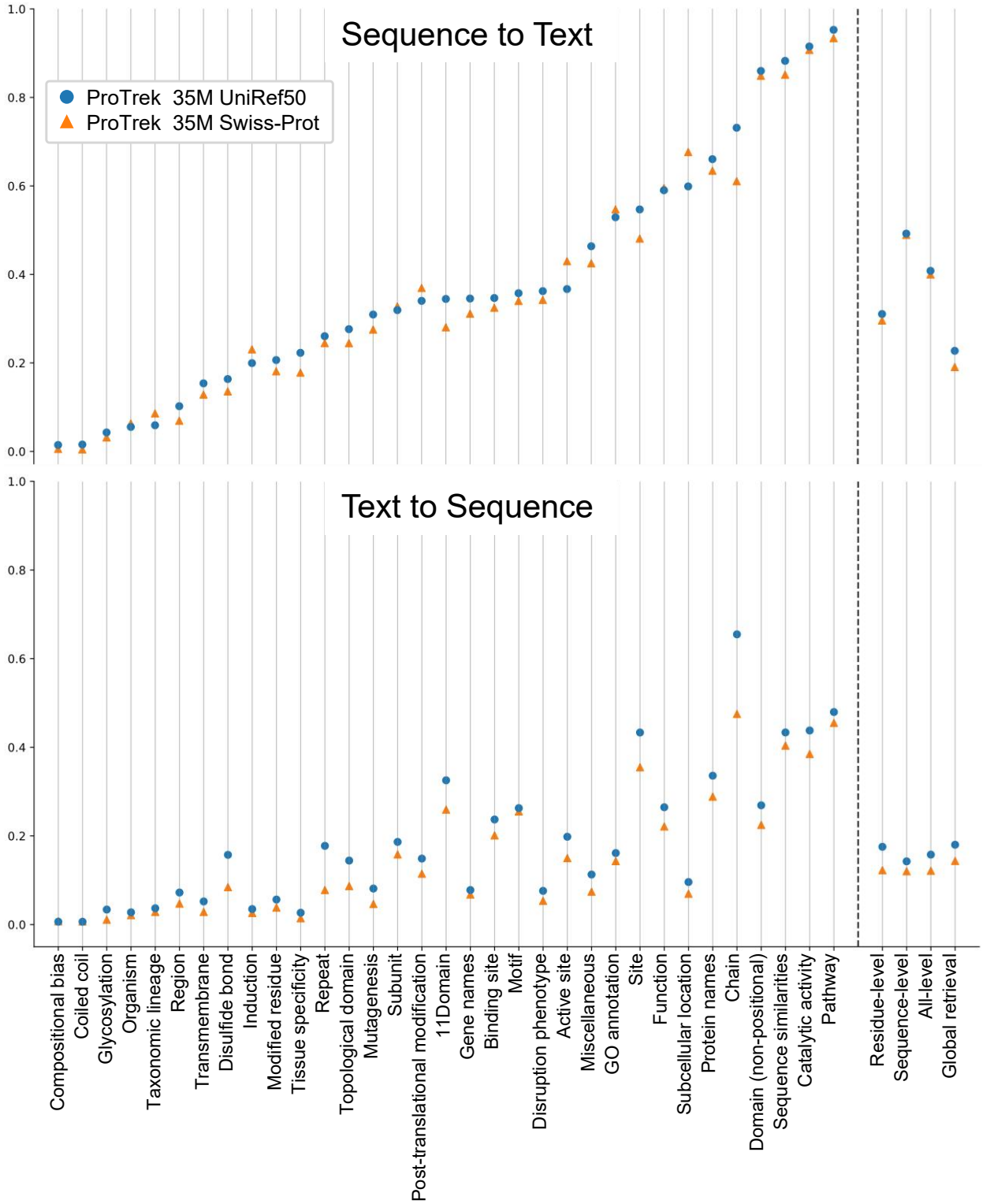
UniRef50							
Sequence-level				Residue-level			
Subsection	**Num(F)**	**Num(A)**	**Num(F) / Num(A) (%)**	**Subsection**	**Num(F)**	**Num(A)**	**Num(F) / Num(A) (%)**
Gene names	403362	930060	43.37	Domain	2929458	5885898	49.77
Protein names	57883	3979530	1.45	Chain	585920	1769394	33.11
Organism	20076	372085	5.40	Compositional bias	143116	801190	17.86
GO annotation	11371	3621371	0.31	Region	102923	1512529	6.80
Sequence similarities	6861	2145953	0.32	Repeat	75579	700456	10.79
Function	4405	509670	0.86	Binding site	61385	584807	10.50
Catalytic activity	2672	617558	0.43	Coiled coil	41573	154788	26.86
Proteomes	1673	12260	13.65	DNA binding	19364	242336	7.99
Subunit	1297	124955	1.04	Transmembrane	11971	840833	1.42
Pathway	726	161197	0.45	Active site	9607	124509	7.72
Taxonomic lineage	652	39964	1.63	Disulfide bond	8434	33640	25.07
Subcellular location	468	92646	0.51	Site	7064	39743	17.77
Cofactor	296	99807	0.30	Motif	5309	35961	14.76
Domain (non-positional annotation)	248	38189	0.65	Modified residue	4257	125752	3.39
Post-translational modification	119	9353	1.27	Topological domain	882	1508	58.49
Miscellaneous	61	8654	0.70	Cross-link	238	290	82.07
Activity regulation	45	4462	1.01	Lipidation	130	409	31.78
Caution	14	19553	0.07	Signal peptide	36	58736	0.06
Induction	9	129	6.98	Propeptide	33	73	45.21
				Intramembrane	8	10	80.00
				Peptide	7	19	36.84
				Glycosylation	4	8	50.00
Total	512,238	12,787,396		Total	4,007,298	12,912,889	

Swiss-Prot							
Sequence-level				Residue-level			
Subsection	**Num(F)**	**Num(A)**	**Num(F) / Num(A) (%)**	**Subsection**	**Num(F)**	**Num(A)**	**Num(F) / Num(A) (%)**
Gene names	666,632	1,121,127	59.46	Chain	252,675	575,915	43.87
Protein names	301,073	1,476,325	20.39	Region	119,507	319,532	37.40
Function	98,147	487,765	20.12	Binding site	108,215	1,186,207	9.12
Subunit	44,603	295,608	15.09	Domain	107,195	214,158	50.05
Tissue specificity	31,165	50,620	61.57	Mutagenesis	89,827	90,626	99.12
GO annotation	29,283	3,165,677	0.93	Compositional bias	75,066	174,232	43.08
Organism	26,684	881,200	3.03	Repeat	68,638	108,887	63.04
Subcellular location	19,972	509,901	3.92	Natural variant	64,392	64,820	99.34
Disruption phenotype	17,298	20,051	86.27	Topological domain	62,960	149,000	42.26
Post-translational modification	17,291	63,799	27.10	Disulfide bond	41,989	134,091	31.31
Miscellaneous	14,926	45,515	32.79	Modified residue	32,358	259,509	12.47
Catalytic activity	14,509	335,116	4.33	Transmembrane	24,967	380,879	6.56
Induction	14,210	25,198	56.39	Motif	17,405	47,202	36.87
Sequence similarities	14,161	518,327	2.73	Site	16,906	64,740	26.11
Developmental stage	11,576	14,119	81.99	Coiled coil	15,720	22,423	70.11
Domain (non-positional annotation)	10,159	57,737	17.60	Active site	12,870	176,303	7.30
Activity regulation	7,498	18,196	41.21	Glycosylation	9,309	123,475	7.54
Involvement in disease	6,359	6,845	92.90	Peptide	9,183	12,462	73.69
Taxonomic lineage	5,714	569,793	1.00	Propeptide	5,991	14,843	40.36
Caution	5,163	14,333	36.02	Cross-link	5,538	24,680	22.44
Biophysicochemical properties	3,398	7,835	43.37	DNA binding	4,465	12,152	36.74
Cofactor	2,017	230,702	0.87	Lipidation	2,172	13,723	15.83
Pathway	1,454	143,674	1.01	Intramembrane	1,254	2,969	42.24
Polymorphism	1,246	1,481	84.13	Transit peptide	383	8,162	4.69
RNA Editing	882	3,254	27.11	Signal peptide	185	43,855	0.42
Biotechnology	875	1,823	48.00				
Toxic dose	712	851	83.67				
Proteomes	600	490,045	0.12				
Allergenic properties	284	947	29.99				
Pharmaceutical use	154	165	93.33				
Total	1,368,045	10,558,029		Total	1,149,170	4,224,845	

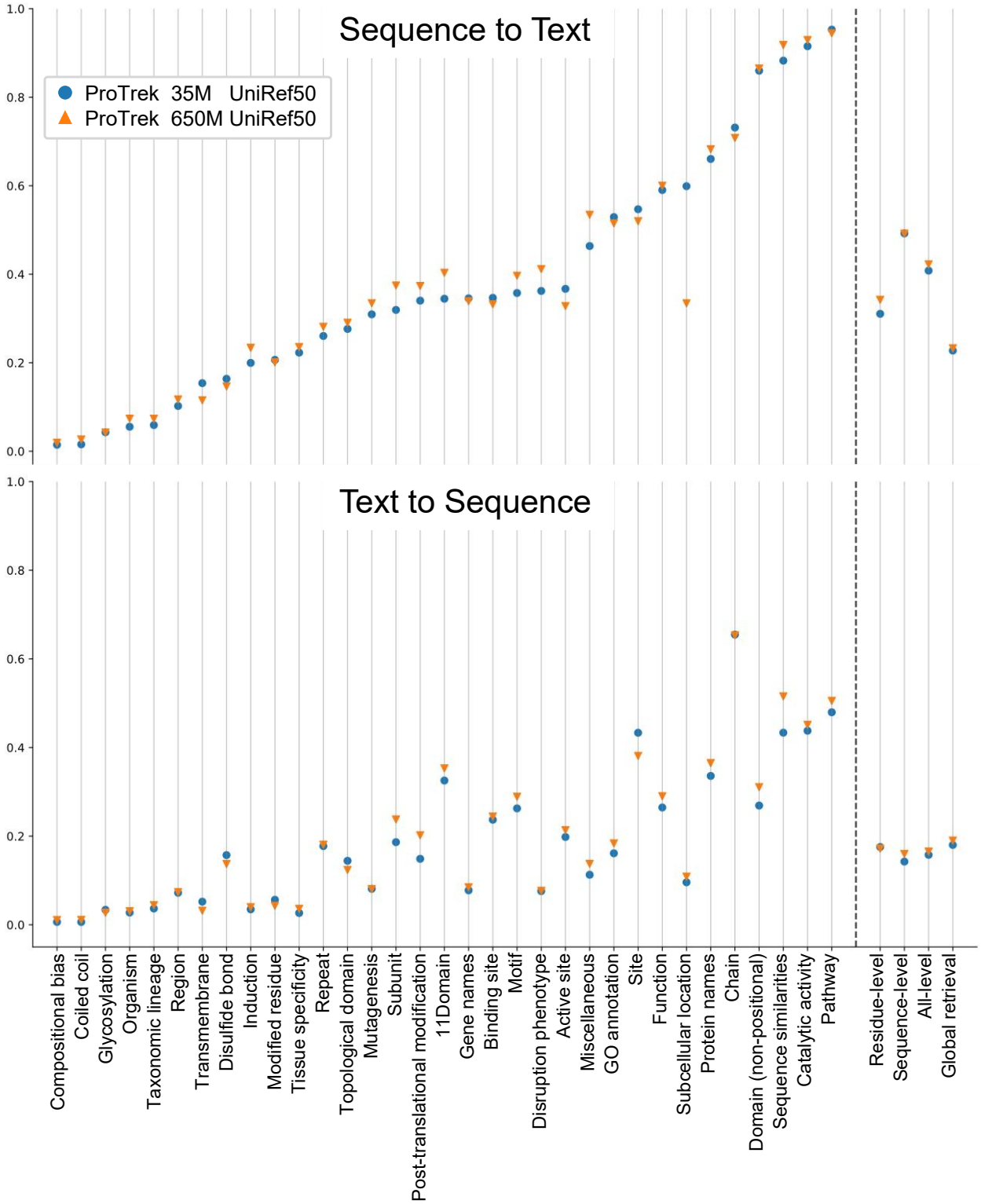
Supplementary Table 1: Statistics of the pre-training dataset. **Num(F)** represents the number of unique descriptions, while **Num(A)** indicates the total number of descriptions, including replicates.

Sequence-level		Residue-level	
Subsection	**Example**	**Subsection**	**Example**
Gene names	The nucleotide sequence rpsI corresponds to the synthesis of this protein.	Chain	The protein's polypeptide chain at position 1 to 294 is recognized as Diaminopimelate epimerase.
Protein names	This protein is identified as Small ribosomal subunit protein uS9.	Region	This protein is labeled with Interaction with BTRC when in the region at position 1 to 151.
Function	Regulates a signal transduction pathway linking plasma membrane receptors to the assembly of focal adhesions and actin stress fibers.	Binding site	The residue at position 460 forms a bond with Mg(2+), indicating a strong binding affinity.
Subunit	Heterooligomer composed of large and small subunits.	Domain	At position 214 to 375, the domain is characterized as TrmE-type G.
Tissue specificity	Mostly expressed in stems and, to a lower extent, in roots and leaves.	Mutagenesis	The effect of altering the amino acid in this protein from I to A at position 582 is decreases kinase activity.
GO annotation	The GO term for this protein involving biological process incorporates leaf senescence.	Compositional bias	At position 299 to 314, the predominating amino acid type is Polar residues.
Organism	The organism's name Finless sole is linked to the protein sequence.	Repeat	The repeated sequence motif or domain at position 29 to 127 falls into the classification Solcar.
Subcellular location	The protein can be spotted at Cellular thylakoid lumen.	Natural variant	The protein's natural variant, known as in RNA edited version; does not alter GABA receptor activity;
Disruption phenotype	Reduced mating efficiency. Defective sperm. Abnormal distribution of pkd-2.	Topological domain	The protein's non-membrane segment is located at Cytoplasmic when considering its position 482 to 505.
Post-translational modification	Can be phosphorylated to form H3S10ph, H3T11ph and H3S28ph.	Disulfide bond	Cysteine residues are crucial in the formation of disulfide bonds at positions 113 and 341.
Miscellaneous	Present with 799 molecules/cell in log phase SD medium.	Modified residue	Phosphoserine designates the occurrence of a modified residue type at position 657 in this protein.
Catalytic activity	The reaction: ATP + H(+) + nicotinate beta-D-ribonucleotide = deamido-NAD(+) + diphosphate undergoes catalysis due to the presence of this protein.	Transmembrane	The protein's membrane-spanning segment at position 291 to 311 is identified as Helical.
Induction	Up-regulated by karrikins and strigolactone treatments.	Motif	The sequence motif 'HIGH' region occurs at the specific position 7 to 17 in this protein.
Sequence similarities	Belongs to the eukaryotic GSH synthase family.	Site	The residue found at 44 possesses the attribute of interacts with H-NS.
Developmental stage	Expressed during the parasite blood stage, in schizonts (at protein level).	Coiled coil	Position 115 to 135 in this protein is characterized by a coiled coil.
Domain (non-positional annotation)	The first C2 domain mediates Ca(2+)-dependent phospholipid binding.	Active site	Catalysis involves the residue at position 102 playing the role of Phosphoserine intermediate.
Activity regulation	Inhibited by C/VIF1 and C/VIF2.	Glycosylation	Position 230 hosts a glycan group that is covalently connected, and its type is N-linked (GlcNAc...) asparagine.
Involvement in disease	Isovaleric acidemia (IVA) is the disease tied to genetic variations in this protein.	Peptide	The active peptide at position 1 to 34 in this protein goes by the name Subtilosin-A.
Taxonomic lineage	In the taxonomic hierarchy, the source organism of this protein is classified within Bacteria.	Propeptide	The protein has been described to include a propeptide at position 181 to 203, which can be identified as Removed in mature form.
Caution	In strain K12 / CR63, finO is interrupted by an IS3 element.	Cross-link	The residues at positions 240 and 244, specifically H and Y, respectively, take part in an intrachain covalent linkage referred to as 1'-histidyl-3'-tyrosine (His-Tyr).
Biophysicochemical properties	Optimum pH is 8.0. The pH has no influence on the tropine to pseudotropine product ratio.	DNA binding	The DNA-binding domain occurring at position 265 to 340 is nuclear receptor.
Cofactor	Without pyruvate acting as a cofactor, this protein's function is compromised.	Lipidation	S-palmitoyl cysteine is the specific lipid group that forms a covalent attachment at position 19.
Pathway	The metabolic pathway that this protein is associated with consists of Carbohydrate biosynthesis; Calvin cycle.	Intramembrane	At position 121 to 127 in Pore-forming, one can observe the presence of the region embedded within a membrane.
Polymorphism	There are several alleles. The sequence shown is that of IMGT allele TRAV41*01.	Transit peptide	Through the transit peptide at 1 to 38, the protein is guided to Mitochondrion.
RNA Editing	This protein experiences RNA editing specifically at 494.	Signal peptide	The position 1 to 26 in this protein contains an N-terminal signal peptide.
Biotechnology	Could be a valuable biocatalyst used in the industrial field for an efficient biotransformation of mevastatin.		
Toxic dose	LD(50) is 0.06 ug/kg by intracerebroventricular injection into mice		
Proteomes	A proteome, containing Partially assembled WGS sequence, is composed of this protein.		
Allergenic properties	Causes an allergic reaction in human. Binds to IgE.		
Pharmaceutical use	Available under the name BeneFix (Baxter and American Home Products). Used to treat hemophilia B.		

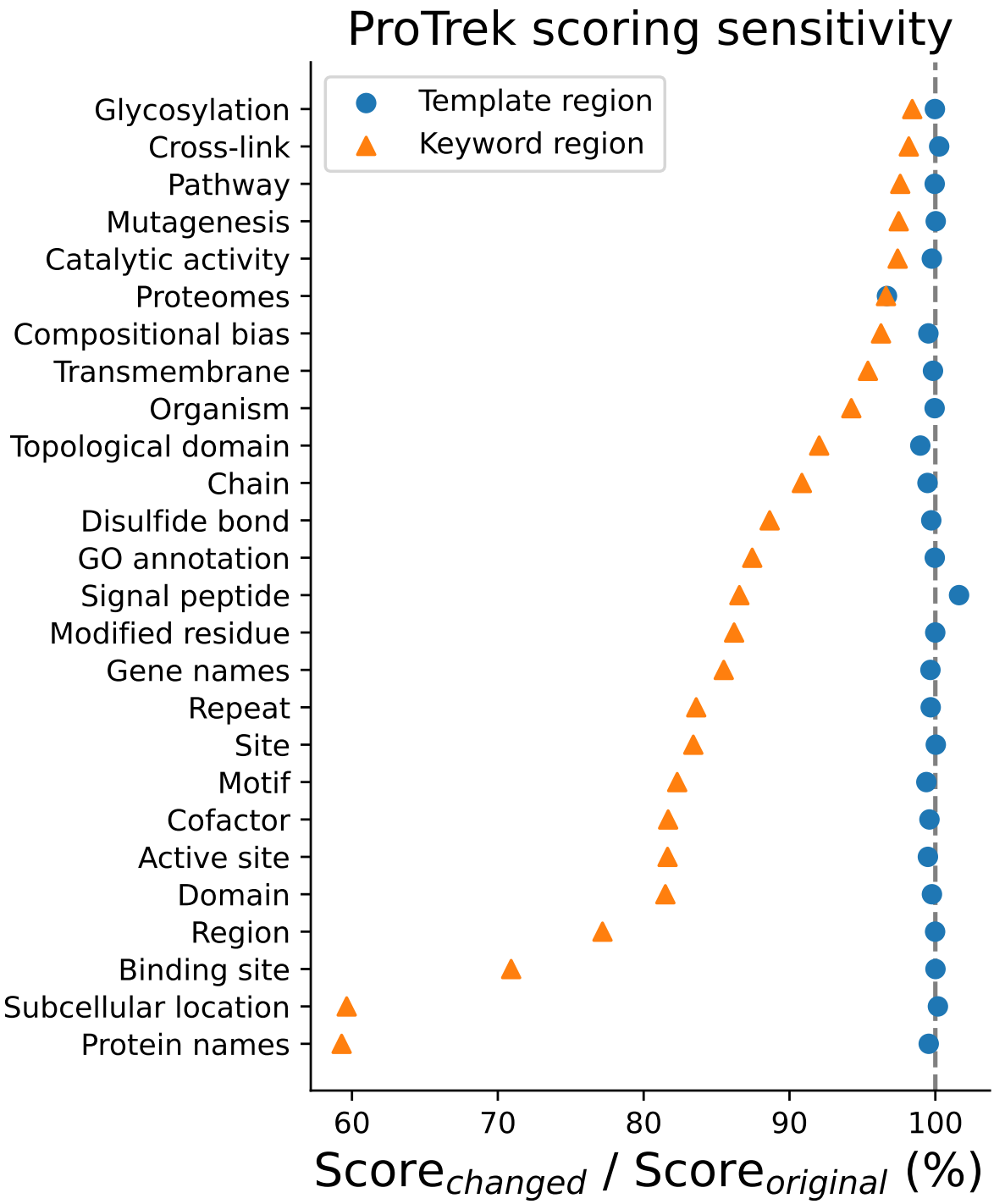
Supplementary Table 2: Examples of functional descriptions for each subsection. The examples in the residue-level subsections are created by filling in the keywords (e.g. position) in the templates.



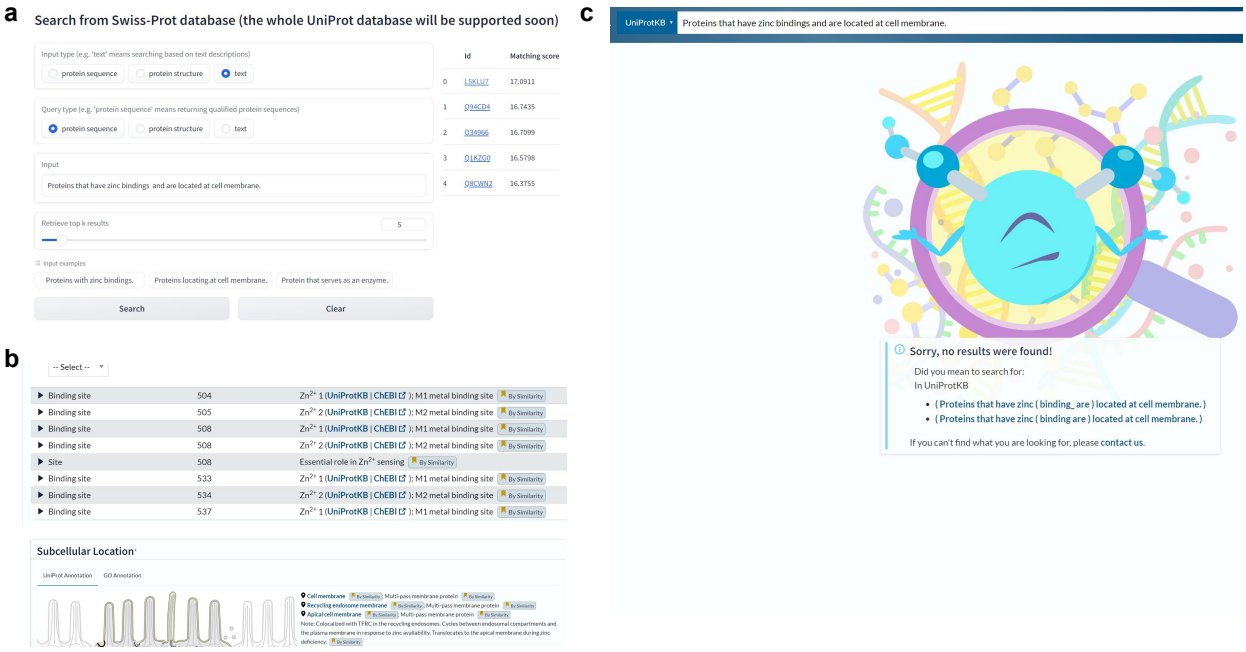
Supplementary Fig. 1: Comparison of protein-text retrieval between models trained on different datasets. "ProTrek 35M Swiss-Prot" was trained on Swiss-Prot data only and "ProTrek 35M UniRef50" was trained on a mixture of Swiss-Prot and UniRef50 data.



Supplementary Fig. 2: Performance comparison of the ProTrek 35M and ProTrek 650M.



Supplementary Fig. 3: Scoring sensitivity for keyword regions and template regions in residue-level functional descriptions. Initially, we computed the similarity score "Score$_{original}$" between a protein sequence and a textual description. Subsequently, we removed each word in the description and calculated a new similarity score with the protein. For the template region (highlighted with a blue circle), we averaged all the new scores whose deleted words were in the template region, denoted as "Score$_{changed}$". Similarly, the "Score$_{changed}$" of the keyword region (indicated by an orange triangle) was calculated by averaging all the new scores whose deleted words were in that region.



Supplementary Fig. 4: A case study using ProTrek to identify proteins with a long functional description. a, Using ProTrek demo to search proteins with the long functional description "Proteins that have zinc bindings and are located at cell membrane." **b,** "L5KLU7," a human-reviewed protein, is returned as the top-ranked result by ProTrek, matching the corresponding description exactly. **c,** Using the same description to search against the UniProt database. The UniProt database only supports keyword searches and did not return any results for the long natural sentence.