nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-025-02836-0

A trimodal protein language model enables advanced protein searches

Received: 19 February 2025

Accepted: 29 August 2025

Published online: 02 October 2025

Check for updates

Jin Su[1,4], Yan He[1,4], Shiyang You[2,4], Shiyu Jiang[1], Xibin Zhou[1], Xuting Zhang[1], Yuxuan Wang[1], Xining Su[2], Igor Tolstoy[3], Xing Chang[1] ✉, Hongyuan Lu[2] ✉ & Fajie Yuan[1] ✉

ProTrek unifies protein sequence, structure and natural language function in a trimodal language model through contrastive learning, enabling comprehensive searches between any two modalities, including within modality. ProTrek surpasses current alignment tools (for example, Foldseek and MMseqs2) in speed and accuracy for identifying functionally related proteins. Computational and wet-lab experimental validations show that the ProTrek server (www.search-protrek.com), with precomputed embeddings for over 5 billion proteins, efficiently processes and analyzes large-scale protein repositories.

Proteins, the cell's principal molecular machines, drive a vast spectrum of biological processes. Deciphering how sequence dictates structure and function—the sequence–structure–function (SSF) relationship—is a central goal for the molecular sciences and pharmacology, yet remains a formidable challenge because of vast structural diversity, context-dependent activities and intricate molecular interactions.

To address this challenge, the protein analysis landscape has evolved substantially. Foundational alignment-based tools—spanning sequence (for example, BLAST[1] and MMseqs2 (ref. 2)) and structure (for example, TM-align[3] and Foldseek[4])—have made substantial contributions. However, these powerful tools are fundamentally constrained to pairwise comparisons within a single modality, either sequence or structure. This siloed approach inherently limits the discovery of cross-modal relationships. Furthermore, for the sake of computational efficiency, many search-oriented tools within this paradigm prioritize local similarities, which can overlook crucial global context. This methodological gap is particularly acute for the roughly 30% of proteins in UniProt[5] that remain unannotated, often because of phylogenetic distance from known homologs[6,7]. The advent of neural network-based tools[6,8–12] promised a new era, predicting functional labels from predefined vocabularies. Nevertheless, being reliant on these fixed labels, such tools cannot comprehend natural language, which prevents them from generating nuanced functional descriptions or identifying proteins from textual queries.

Recent advances in large language models, exemplified by ChatGPT[13], LlaMA[14,15] and DeepSeek[16], have demonstrated unprecedented capabilities across diverse natural language processing tasks. In parallel, computational biology has witnessed the rapid emergence of protein language models (pLMs) as a transformative research frontier[17–23]. Building upon these advancements, we propose developing a foundational pLM that can comprehensively represent all SSF modalities of proteins.

Here, we present ProTrek, a trimodal language model designed to jointly model protein sequence, structure and function modalities. ProTrek uses contrastive learning[24] through a bidirectional alignment strategy (Fig. 1a), which operates across three dimensions: (1) between protein structure and sequence; (2) between protein function and structure; and (3) between protein function and sequence (Methods). This trimodal alignment enables ProTrek to establish strong associations among SSF by bringing genuine sample pairs (sequence–structure, structure–function and sequence–function) closer together while pushing apart negative samples within the latent space.

The architecture of ProTrek incorporates a pretrained evolutionary-scale modeling (ESM) encoder[18] for amino acid sequence modeling and a pretrained BERT (bidirectional encoder representations from transformers)[25] for nuanced natural language function representation. Protein structures are transformed into discrete-token sequences using Foldseek[4], facilitating their encoding through 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 (Methods). ProTrek was trained on a large dataset of nearly 40 million protein–text pairs (Methods and Supplementary Tables 1 and 2). This dataset comprises two key components: 14 million high-precision

[1]Westlake University, Hangzhou, China. [2]The Hong Kong University of Science and Technology, Guangzhou, China. [3]Independent Scientist, Germantown, MD, USA. [4]These authors contributed equally: Jin Su, Yan He, Shiyang You. ✉e-mail: changxing@westlake.edu.cn; hongylu@hkust-gz.edu.cn; yuanfajie@westlake.edu.cn

https://doi.org/10.1038/s41587-025-02836-0

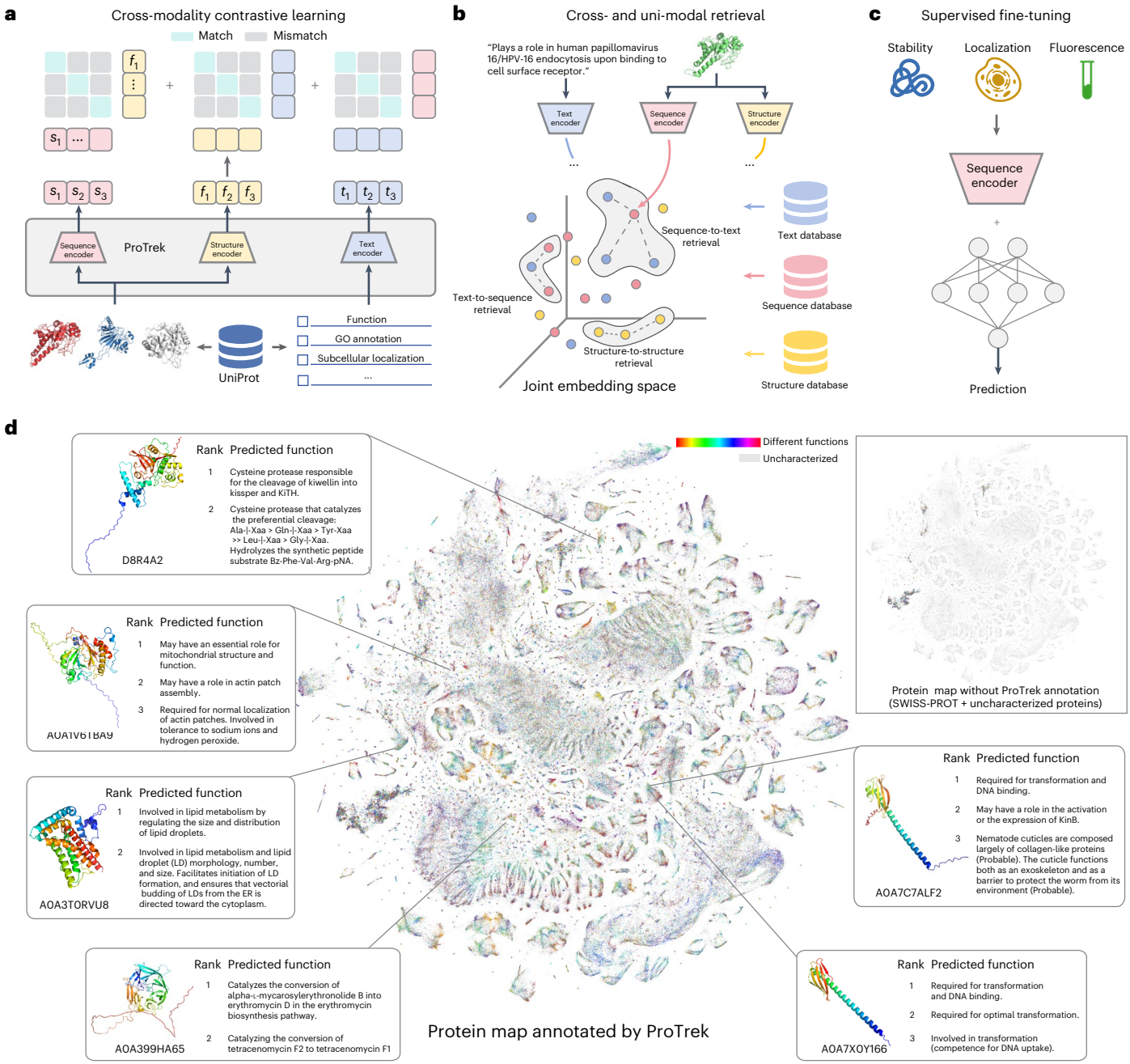


Fig. 1 | Illustration of ProTrek. a, ProTrek architecture and trimodal contrast learning. **b**, Cross-modal and unimodal retrieval. ProTrek supports nine searching tasks. **c**, After the trimodal contrast learning, the protein sequence encoder encodes general-purpose representation of proteins, which 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 (colored by protein function). Each cluster represents proteins with close sequence embedding distances. The plot illustrates the complex relationship between sequence and function as modeled by ProTrek. Top right, note over 99% of UniProt protein entries remain unreviewed.

pairs curated from the SWISS-PROT database[26] and 25 million noisy pairs. The latter were obtained by scoring and filtering a pool of 300 million pairs from TrEMBL50, a database of representative proteins derived by clustering TrEMBL[27] at 50% sequence identity. The ProTrek model's training regimen incorporates eight loss functions—six dedicated to intermodal alignment and two masked language modeling losses to preserve recognition fidelity at both the amino acid and the structural (3Di) token levels (Methods and Supplementary Fig. 1).

Through a fusion of protein SSF modalities into a unified framework, ProTrek offers three capabilities that redefine our approach to decoding the protein universe. Firstly, as a zero-shot retrieval model, ProTrek enables exploration of 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. It bridges the gap between protein data and human comprehension, translating complex molecular landscapes into an intuitive, language-driven experience (Fig. 1d and Supplementary Fig. 2). Secondly, 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 functional convergent evolution despite divergent structures and sequences (Fig. 2b,c)—a phenomenon potentially more ubiquitous in nature[28,29].

Thirdly, 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 downstream tasks, rivaling and complementing cutting-edge pLMs (Figs. 1c and 2e). This versatility extends its applicability across diverse domains of protein science.

ProTrek is available in two variants (Supplementary Fig. 3): a large-scale version featuring an amino acid sequence encoder with 650 million parameters, a structure encoder with 150 million parameters and a function encoder with 130 million parameters or a more compact version featuring corresponding encoders with 35 million, 35 million and 130 million parameters, respectively. These configurations are designed to accommodate varying computational resources and research objectives. During inference, ProTrek uses a maximum inner-product search (MIPS) algorithm[12,30] for efficient retrieval. This enables ProTrek to complete searches within seconds, even when querying databases containing hundreds of millions of entries. Such rapid search capability enhances the efficiency of protein research, providing a powerful tool for exploring the protein universe.

Here, we first evaluated ProTrek's performance on four standard retrieval tasks, designed to bridge protein sequences and structures with their natural language functional descriptions. These tasks comprise bidirectional sequence-to-text, structure-to-text, text-to-sequence and text-to-structure pairwise searches (Fig. 2a). ProTrek's performance was benchmarked against two state-of-the-art methods, ProteinDT (ProteinCLAP)[31] and ProtST[32], using the human-reviewed SWISS-PROT[26] dataset. This evaluation dataset comprises 4,000 proteins and their associated functional descriptions. For a fair evaluation, proteins sampled in the SWISS-PROT 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 (Methods 1.6, Supplementary Table 3).

ProTrek demonstrates strong performance across most functional categories, regardless of whether the amino acid sequence encoder or 3Di structure encoder is used (Fig. 2a). It outperforms both ProteinDT and ProtST, achieving improvements of over 30-fold and 60-fold in global retrieval tasks. Specifically, ProTrek achieves mean average precision (MAP; Methods) scores of 0.233 for sequence-to-text and 0.190 for text-to-sequence search tasks, outperforming ProtST (0.007 and 0.001) and ProteinDT (0.004 and 0.003) in the respective tasks (Fig. 2a, Supplementary Figs. 4 and 5 and Supplementary Table 4). This performance can be largely attributed to ProTrek's extensive training dataset (Supplementary Fig. 6), which surpasses those of ProteinDT and ProtST by two orders of magnitude. In terms of ProTrek's own architecture, the amino acid sequence encoder generally exhibits higher performance than the structure encoder, a discrepancy potentially linked to differences in encoder model size and pretraining status. Figure 1d illustrates an example of how ProTrek embeddings and its natural language understanding capabilities can be used.

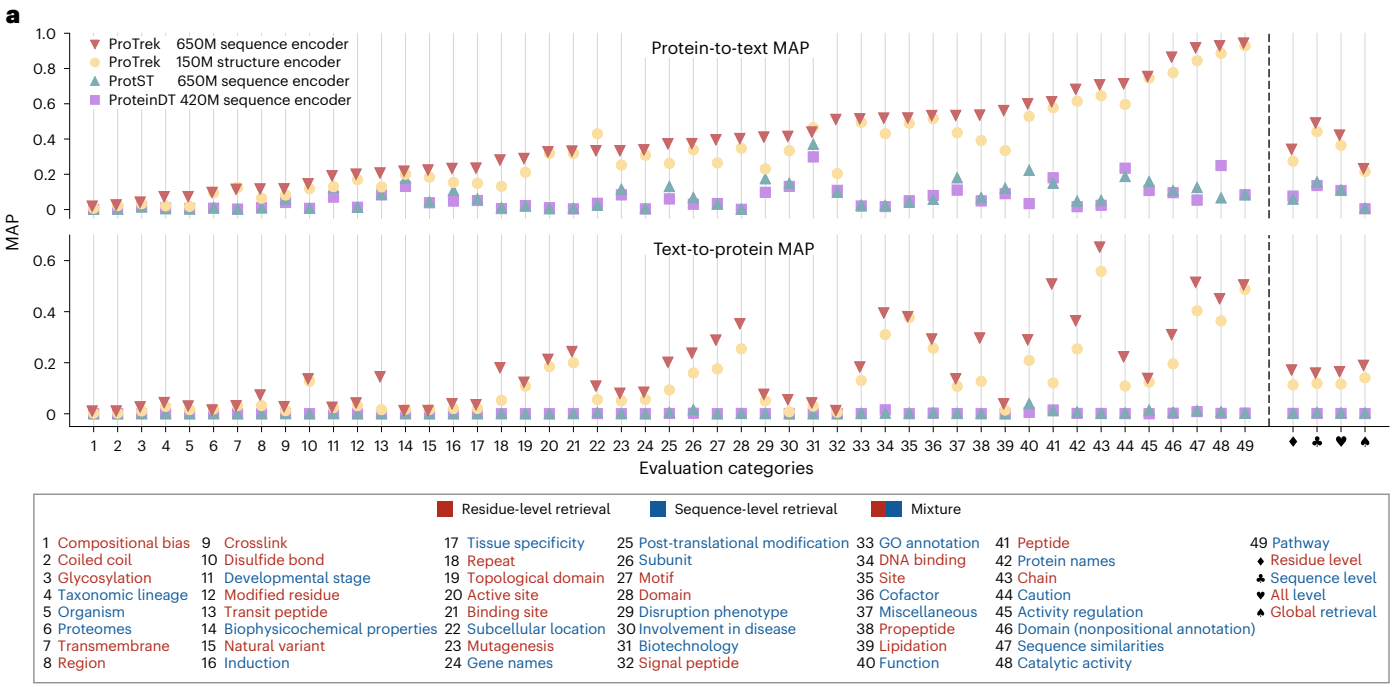
We further evaluated ProTrek's text–protein translation capabilities using recent literature cases (Methods), encompassing newly characterized enzymes such as Om1Cas9 (ref. 33), SlugCas9 (ref. 34), dsPETase[33], TTHA0338 (ref. 35), SsFlA[35] and $SsdA_{tox}$[36]. These proteins were not included in ProTrek's training data. Moreover, the research papers describing Om1Cas9, dsPETase and $SsdA_{tox}$ were published after ProTrek's model release, ensuring a more realistic test of the model's generalization capabilities. For protein-to-text retrieval, ProTrek's top-ranked protein descriptions consistently align with their corresponding experimental or functional details (Supplementary Tables 5 and 6). In text-to-protein retrieval, ProTrek accurately identifies target proteins from 45 million candidate proteins, assigning them high rankings solely on the basis of their natural language descriptions as summarized from the literature (Supplementary Table 7). To further assess ProTrek's proficiency in processing protein descriptions, we designed experiments showcasing its fine-grained text-to-sequence retrieval capacity (Methods). As shown in Supplementary Table 8, for single-category retrieval, ProTrek achieved a hit rate of 95.33%, demonstrating precise matching to these annotated descriptions. In more complex combinatorial queries, ProTrek maintained robust performance, attaining a hit rate of 65.83% for proteins characterized by dual functions or properties. This demonstrates ProTrek's ability to handle complex queries and effectively capture multifaceted protein characteristics. ProTrek also exhibits strong performance in task-specific applications. When benchmarked against CLEAN[35] using enzyme annotation datasets (released after the ProTrek training data collection date), ProTrek achieved comparable results (Fig. 2d) by retrieving top-ranked Enzyme Commission (EC) function descriptions from a curated textual EC database using the maximum separation method[35] (Methods and Supplementary Table 9). Additional comparative results demonstrating ProTrek's sequence-to-text capabilities relative to nearest neighbor approaches based on ESM-2 (ref. 18) are presented in Supplementary Fig. 7. These findings underscore ProTrek's versatility as both a specialized and general-purpose annotation tool.

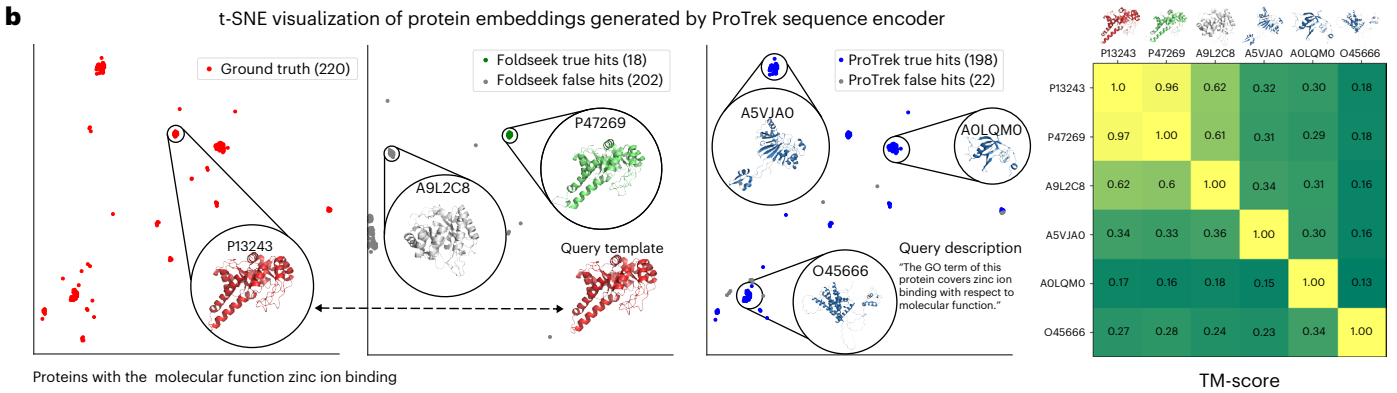
Existing alignment tools such as Foldseek[4] and BLASTp[1] excel at detecting proteins with notable structural or sequence similarities, which typically suggests shared ancestry through homologous evolution. ProTrek, however, transcends this conventional homology-based approach through its global alignment approach driven by cross-modal contrastive learning. This enables ProTrek to identify proteins with shared functional characteristics despite lacking a common evolutionary origin—a hallmark of convergent evolution[37,38]. To demonstrate this capability, we curated comprehensive text descriptions, covering enzyme catalytic activities and CRISPR-associated protein functions, to retrieve top-ranked proteins from the UniProt50 database (excluding those present in the training set; Supplementary Tables 11 and 12) and investigated convergent evolution through pairwise template modeling score (TM-score) matrix calculations (Methods). Our analysis revealed that, for enzyme-related proteins, ProTrek achieved a high hit rate of 89% and effectively captured both structurally similar and distinct proteins sharing similar enzyme functions (Extended Data Figs. 1a and 2). For CRISPR-associated proteins, ProTrek exhibited robust recognition across diverse types of CRISPR–Cas systems (average hit rate 67%) while maintaining the ability to retrieve convergent functional analogs (Extended Data Figs. 1b and 3). These findings highlight ProTrek's distinctive capacity to reveal convergent evolution by aligning functionally analogous proteins based on detailed textual cues (case study in Fig. 2b).

Beyond text–protein retrieval, ProTrek enables intramodal and intermodal sequence–structure searches. We benchmarked this against established alignment-based tools (Fig. 2c), including MMseqs2 (ref. 2), DIAMOND[39], BLASTp[1] and Foldseek[4] for structure-to-structure search. We evaluated the sequence-to-structure, structure-to-sequence, sequence-to-sequence and structure-to-structure search modalities for ProTrek. Using the 4,000 SWISS-PROT testing proteins, we performed exhaustive all-versus-all searches and compared retrieval effectiveness for proteins sharing the same Gene Ontology (GO) annotations[40] up to the fifth false positive (Methods).

ProTrek outperforms all these sequence and structure alignment tools in terms of the total or average number of correct hits (Fig. 2c, Supplementary Fig. 8 and Supplementary Table 13). While all methods can effectively identify proteins with analogous functions when they share high TM-scores with query proteins, ProTrek's advantage becomes particularly evident when detecting functional similarities in proteins with lower TM-scores, especially in twilight zones. This finding aligns with the diverse TM-scores observed in the protein structures retrieved from the above text-to-sequence searches (Fig. 2b). Among ProTrek's four search modalities, the sequence-to-sequence search shows the highest efficacy, followed by structure-to-structure and sequence-to-structure searches. Each modality serves distinct purposes; for instance, the structure-to-sequence search, despite

https://doi.org/10.1038/s41587-025-02836-0





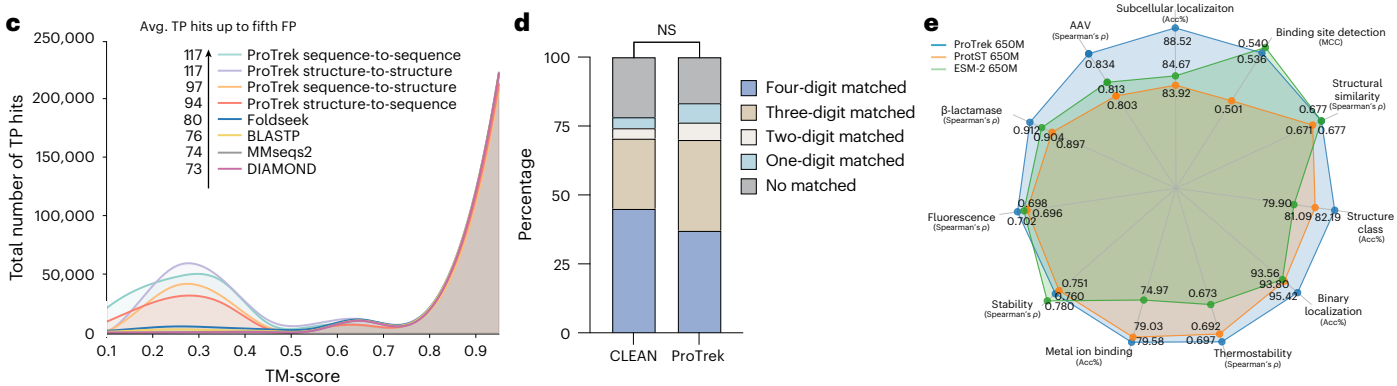


Fig. 2 | ProTrek performance on protein search and representation tasks. a, Top: search protein functional descriptions using sequences and structures. Bottom: search protein sequences and structures using textual descriptions. The *x* axis denotes specific protein function categories left of the dashed line and aggregated categories (residue, protein and all levels) to the right of it. 'Global retrieval' indicates a search across the entire database, not within individual categories. The *y* axis denotes the MAP. ProtST is the abbreviation of ProtST-ESM-1b. M, million parameters. **b**, Case study: ProTrek uses 'zinc ion binding' as the query term, while Foldseek uses P13243 as a query template, which is the protein with the most hits. In the testing set, 220 proteins share similar functional annotations to 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 (for example, A9L2CB) may encode different functions. t-SNE, *t*-distributed stochastic neighbor embedding. **c**, Searching proteins with similar functions using protein sequence or structure as input. TP (true positive), matches sharing ≥1 GO term; FP (false positive), matches sharing no GO terms. **d**, Comparison of CLEAN and ProTrek in EC number annotation. The performance difference is not statistically significant (NS), as assessed by the Wilcoxon test. **e**, Evaluating the protein representation ability of the ProTrek amino acid sequence encoder (more results in Supplementary Table 10). Acc, accuracy; MCC, Matthews correlation coefficient.

https://doi.org/10.1038/s41587-025-02836-0

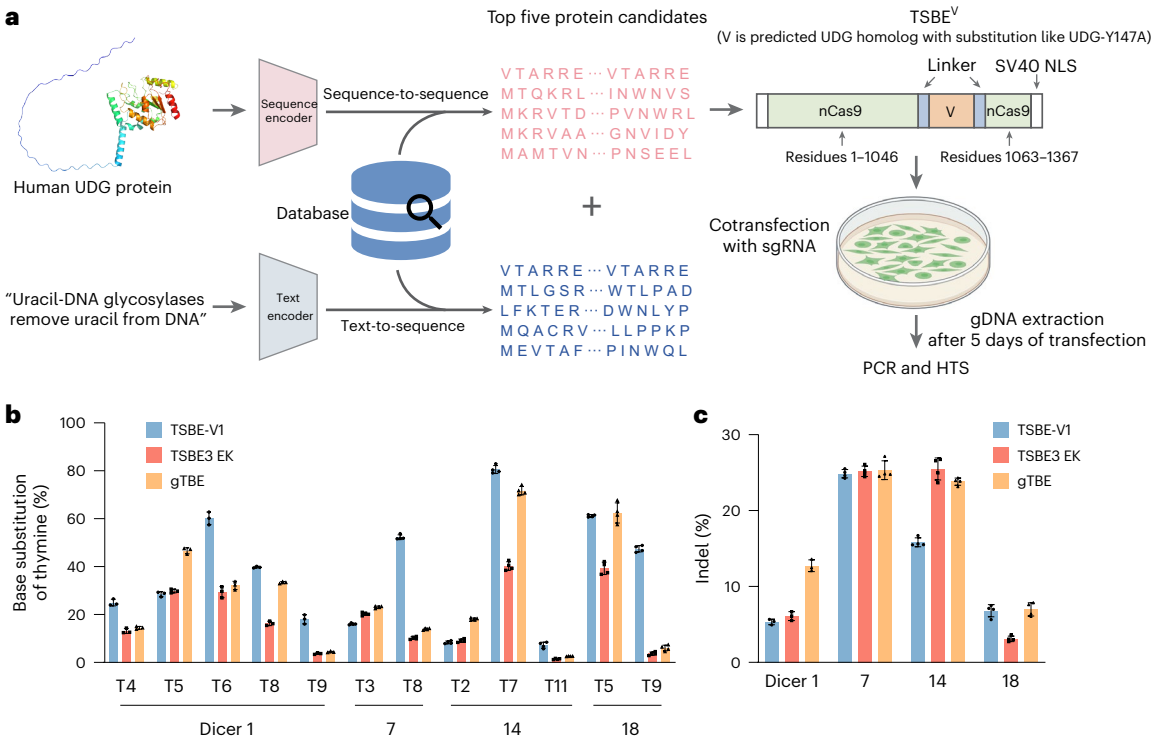


Fig. 3 | ProTrek-enabled identification of proteins functionally analogous to UDG. a, Schematic representation of the experimental validation workflow. Variant V, a predicted UDG homolog identified among the top five hits from both sequence-to-sequence and text-to-sequence searches conducted by ProTrek, carries a substitution analogous to UDG-Y147A. This variant was inserted into the central region of spCas9n to generate fusion proteins, which were cotransfected with sgRNA into HeLa cells. gDNA was extracted 5 days after transfection and nucleotide substitutions (mutant frequency > 0.1%) were analyzed by HTS. **b**, HeLa cells were transfected with TSBE-V1, TSBE3 EK or gTBE, along with the specified sgRNAs (Dicer 1 and sites 7, 14 and 18). Variant V1 represents the top-ranking protein identified by both sequence-to-sequence and text-to-sequence searches. Nucleotide substitutions of thymine in target sites were quantified by HTS; the positions of mutated thymines are numbered from the 5′ end of the protospacer. Data represent the mean ± s.d. of $n = 3$ independent biological replicates at site Dicer 1 and $n = 4$ independent biological replicates at sites 7, 14 and 18. **c**, Indel frequencies generated by TSBE-V1 were assessed using HTS and compared to those generated by TSBE3 EK and gTBE (where lower values indicate improved performance). Data represent the mean ± s.d. of $n = 3$ independent biological replicates at Dicer 1 and $n = 4$ independent biological replicates at sites 7, 14 and 18.

relatively lower overall effectiveness, shows promise in evaluating protein sequences generated by artificial intelligence (AI) design models such as ProteinMPNN[41] (Methods and Supplementary Fig. 9).

To experimentally validate ProTrek, we sought to identify proteins sharing functional characteristics with uracil DNA glycosylase (UDG). Recently, a novel class of base editors, including TSBE3 (ref. 42), gTBE[43], and DAF-TBE[44], based on the human UDG mutant has been developed. However, because of the limited activity of human UDG mutant, the editing efficiency of these thymine-targeting base editors was relatively low. With the goal of finding novel proteins that function similarly or outperform human UDG mutant, we aimed to identify proteins functionally analogous to UDG through ProTrek. Using the OMG_prot50 (ref. 45) database with 200 million proteins, we performed parallel searches with sequence-to-sequence and text-to-sequence modes, using the human UDG sequence and the functional description 'UDGs remove uracil from DNA' as respective queries (Methods and Supplementary Table 14). Although the returned candidates exhibited low overall sequence identity, they shared conserved enzymatically active regions (Extended Data Fig. 4a). On the basis of the established principle that the Y147A substitution in human UDG alters its substrate recognition from uracil to thymine[46], we introduced analogous substitutions into the active sites of all identified candidate proteins to modify their substrate specificity. Through this approach, we engineered novel proteins capable of recognizing and cleaving normal thymine bases in DNA. After codon optimization, we fused these mutant proteins with spCas9n to create fusion proteins (Methods and Supplementary Table 15), which were subsequently cotransfected with single guide RNA (sgRNA) (primers in Supplementary Table 16) into HeLa cells for thymine-editing assays (Methods and Fig. 3a). All identified proteins exhibited varying levels of efficiency in editing thymine after fusion with spCas9n, whereas the negative controls demonstrated no detectable thymine-editing activity (Methods and Extended Data Fig. 4b). Notably, the V1 protein, which ranked first in both sequence-to-sequence and text-to-sequence methods, displayed higher editing efficiency at multiple sites compared to existing thymine base editors such as TSBE3 EK and gTBE (Fig. 3b). Furthermore, V1 exhibited substantially lower indel frequencies at Dicer 1 and site 14 compared to TSBE3 EK and gTBE (Fig. 3c).

ProTrek's inference speed, powered by the MIPS algorithm, represents another advantage. This algorithmic design enables ProTrek to process and search billion-scale databases with rapid response times, achieving over 100-fold speed improvement compared to Foldseek and MMseqs2 (Methods and Extended Data Fig. 5).

While ProTrek's training offers broad coverage, some protein families may be underrepresented, limiting fine-grained prediction. In such cases, its sequence encoder can be fine-tuned on task-specific data. We evaluated the fine-tuning ability of ProTrek on 11 downstream tasks, including both protein-level and residue-level analyses (Methods). Overall, ProTrek's performance surpassed prior methods ESM-2 and ProtST and was comparable to the concurrent work ESM-3 (ref. 47) (Fig. 2e and Supplementary Table 10). To boost usability, we developed ColabProTrek, an interactive Google Colab notebook that lets researchers fine-tune ProTrek on custom datasets without machine learning or programming expertise (see Code Availability).

The ProTrek web platform (www.search-protrek.com) enables comprehensive protein searches across an unprecedented collection of over 5 billion proteins—ten times larger than UniProt—integrating seven major databases: SWISS-PROT, UniRef50, Protein Data Bank[48], Open MetaGenomic (OMG)[45], MGnify[49], global ocean microbiome protein catalog[33] and National Center for Biotechnology Information[50]. This represents a major computational undertaking, with protein embedding generation requiring over 5 years of NVIDIA A100 GPU computing time. In the near future, we will expand ProTrek to encompass the entire global protein data, targeting a search capacity of 10 billion proteins. In addition, for users requiring private database integration, we provide command-line tools to build embeddings for local deployment (see Code Availability).

ProTrek enables advanced protein searches with semantic understanding beyond simple keyword matching (Supplementary Fig. 2), demonstrating a strong sensitivity to functionally relevant keywords over template-based text structures (Extended Data Fig. 6). Its robust protein–text comprehension serves two additional crucial functions: First, it supports emerging research by generating extensive, high-quality synthetic protein–text pairs for training large-scale foundation models, such as Pinal[51] with 16 billion parameters for natural language-guided protein design and Evola[20] with 80 billion parameters for protein question answering. Second, it provides essential reference metrics for evaluating protein–text alignment, offering valuable insights for cutting-edge research[51–57].

As an AI model, ProTrek has its limitations. Having been trained exclusively on natural proteins, it may not perform well for certain de novo designed proteins and is less sensitive to subtle sequence variations—akin to AlphaFold2—which means that it cannot precisely predict specific values such as fluorescence wavelengths or protein stability. Achieving this level of accuracy usually requires training on specialized mutation datasets (such as through ColabProTrek). Nevertheless, ProTrek's accessibility, search efficiency and broad predictive power make it a valuable tool for generating biological hypotheses, from discovering novel proteins and characterizing previously unknown ones to elucidating patterns of convergent evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-025-02836-0.

References

1. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
2. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
3. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
4. Van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
5. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
6. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
7. Gane, A. 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).
8. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
9. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 1–23 (2019).
10. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
11. Liu, W. et al. PLMSearch: protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **15**, 2775 (2024).
12. Hong, L. et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nat. Biotechnol.* **43**, 983–995 (2025).
13. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
14. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
15. Touvron, H. et al. LLaMA 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
16. Guo, D. et al. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning. Preprint at https://arxiv.org/abs/2501.12948 (2025).
17. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
18. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
19. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2021).
20. Zhou, X. et al. Decoding the molecular language of proteins with Evolla. Preprint at *bioRxiv* https://doi.org/10.1101/2025.01.05.630192 (2025).
21. Peng, F. Z. et al. PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks. *Nat. Methods* **22**, 945–949 (2025).
22. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. In *Proc. 12th International Conference on Learning Representations* (ICLR, 2024); https://openreview.net/forum?id=6MRm3G4NiU
23. Su, J. et al. SaprotHub: making protein modeling accessible to all biologists. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.24.595648 (2024).
24. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
25. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
26. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
27. UniProt Consortium UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
28. Koehler Leman, J. et al. Sequence–structure–function relationships in the microbial protein universe. *Nat. Commun.* **14**, 2351 (2023).
29. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* **3**, 548–556 (1999).
30. Douze, M. et al. The Faiss library. Preprint at https://arxiv.org/abs/2401.08281 (2024).

31. Liu, S. et al. A text-guided protein design framework. *Nat. Mach. Intell.* **7**, 580–591 (2025).
32. Xu, M., Yuan, X., Miret, S. & Tang, J. ProtST: multi-modality learning of protein sequences and biomedical texts. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 38749–38767 (PMLR, 2023).
33. Chen, J. et al. Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).
34. Hu, Z. et al. Discovery and engineering of small SlugCas9 with broad targeting range and high specificity and activity. *Nucleic Acids Res.* **49**, 4008–4019 (2021).
35. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
36. Kweon, J. et al. Efficient DNA base editing via an optimized DYW-like deaminase. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.15.594452 (2024).
37. Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. & Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **372**, 817–845 (2007).
38. Doolittle, R. F. Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15–18 (1994).
39. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
40. Pomaznoy, M., Ha, B. & Peters, B. GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics* **19**, 470 (2018).
41. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
42. He, Y. et al. Protein language models-assisted optimization of a uracil-*N*-glycosylase variant enables programmable T-to-G and T-to-C base editing. *Mol. Cell* **84**, 1257–1270 (2024).
43. Tong, H. et al. Development of deaminase-free T-to-S base editor and C-to-G base editor by engineered human uracil DNA glycosylase. *Nat. Commun.* **15**, 4897 (2024).
44. Ye, L. et al. Glycosylase-based base editors for efficient T-to-G and C-to-G editing in mammalian cells. *Nat. Biotechnol.* **42**, 1538–1547 (2024).
45. Cornman, A. et al. The OMG dataset: an Open MetaGenomic corpus for mixed-modality genomic language modeling. In *Proc. 13th International Conference on Learning Representations* (ICLR, 2025); https://openreview.net/forum?id=jlzNb1iWs3
46. Kavli, B. et al. Excision of cytosine and thymine from DNA by mutants of human uracil-DNA glycosylase. *EMBO J.* **15**, 3442–3447 (1996).
47. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
48. Burley, S. K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).
49. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
50. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
51. Dai, F. et al. Toward de novo protein design from natural language. Preprint at *bioRxiv* https://doi.org/10.1101/2024.08.01.606258 (2024).
52. Liu, N. et al. Protein design with dynamic protein vocabulary. Preprint at https://arxiv.org/abs/2505.18966 (2025).
53. Kuang, J., Liu, N., Sun, C., Ji, T. & Wu, Y. PDFBench: a benchmark for de novo protein design from function. Preprint at https://arxiv.org/abs/2505.20346 (2025).
54. Ko, Young Su. Using ProTrek for protein binder design. *Twitter* https://x.com/youngsuko9/status/1865845977673834595 (2024).
55. Gitter, A. Using ProTrek to retrieve proteins with desired function. *Twitter* https://x.com/anthonygitter/status/1827760237194920435 (2024).
56. Gitter, A. Using ProTrek to retrieve proteins with desired function. *Twitter* https://x.com/anthonygitter/status/1813427191000035330 (2024).
57. Gitter, A. Using ProTrek to retrieve proteins with desired function. *Twitter* https://x.com/anthonygitter/status/1882642214624678193 (2025).

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[27]. We categorized the subsection information of a protein into two types: sequence level and residue level (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 use GPT-4 (ref. 13) to generate multiple specific template sentences, thus organizing the information into a coherent sentence. For sequence-level subsections, we also use 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. The prompt for GPT-4 to generate templates and paraphrase sentences is shown below.

You are an expert biologist who is good at paraphrasing sentences from biological area while keeping their semantic meaning the same. Now given the sentence quoted by [], you have to try your best to create as much as possible variants of the sentence. Remember that these sentences should have the same meaning. Here's the original sentence: [sentence]. Please give me as many as 10 variants of the sentence with the same meaning. Note that you should keep variants as diverse as possible as you can. Here's the format:

1. *Paraphrased sentence 1*
2. *Paraphrased sentence 2*

...

Our initial attempts to use GPT-4 for generating paraphrases revealed that the model occasionally introduced hallucinations when processing longer texts. Specifically, it tended to partially paraphrase certain sentences while omitting others. To mitigate this issue, we divided the text descriptions into shorter sequences on the basis of sentence boundaries. For each sequence, we individually prompted GPT-4 to generate diverse paraphrases. The resulting paraphrases were then randomly selected and concatenated in their original order to reconstruct a complete text description. Upon thorough examination, we observed that, when provided with shorter text sequences, GPT-4 was able to produce high-quality paraphrases that preserved the original semantics while exhibiting greater diversity.

Pretraining dataset construction

We collected protein sequences and function descriptions from the UniProt database[27] and downloaded corresponding protein structures from AlphaFoldDB[58]. Proteins without predicted structures were removed. We first performed a 50% sequence similarity clustering on the human-reviewed SWISS-PROT database[26]. We designated 1,000 clusters for validation and another 1,000 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 version with 35 million parameters, was pretrained on 12 NVIDIA 80-GB A100 GPUs over 100,000 steps.

Next, we used the initial ProTrek model to score and filter a pool of 300 million noisy protein–function pairs from TrEMBL50. TrEMBL50 was generated by clustering the TrEMBL database at 50% sequence identity and retaining the representative proteins along with their functional descriptions. 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, formed the final pretraining dataset.

Pretraining loss function

We adopted InfoNCE loss[59] for protein SSF contrastive learning. The InfoNCE loss can be described as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(f(x_i, y_i)/\tau)}{\sum_{k=1}^{N} \exp(f(x_i, y_k)/\tau)} \quad (1)$$

Here, x_i and y_i represent the ith embeddings of any two modalities in a batch. $f(x_i, y_i)$ represents a similarity score between the embeddings, which, in this study, is defined as the cosine similarity between x_i and y_i. N is the total number of pairs in a batch and τ is a learnable temperature parameter used to control the probability distribution of the logits predicted by the model (τ is set to 0.07 as an initial value[24]). During pretraining, samples within each batch were randomly selected. In each batch, the positive sample corresponds to the ground-truth pair (for example, (x_i, y_i)), while the remaining unknown pairs are treated as negative samples (that is, (x_i, y_k), where $k \neq i$). This constitutes a standard contrastive learning framework. We calculated the InfoNCE loss for the sequence–structure, sequence–function and structure–function pairs from two directions, resulting in six contrastive learning functions.

We additionally added two masked language modeling (MLM)[60] 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 follows:

$$\mathcal{L}_{\text{MLM}} = \sum_{i \in T} -\log P(s_i | S_{\setminus T}) \quad (2)$$

where T is a set of token positions to be masked and $S_{\setminus T}$ represents a protein or structural sequence with tokens at specific positions masked. The final loss was constructed by averaging two MLM losses and six InfoNCE losses. To examine whether the MLM and InfoNCE losses are of comparable magnitudes, we plotted their respective loss trajectories during pretraining of the ProTrek model with 35 million parameters (Supplementary Fig. 1). As shown in the figure, the sum of InfoNCE losses and the sum of MLM losses quickly converge to the same order of magnitude after about 2,500 steps. This observation suggests that averaging the MLM and InfoNCE losses provides a simple yet effective approach to construct the final loss.

Selection of ProTrek encoders

For the protein sequence encoder, we selected ESM-2 (650 million parameters)[18] as the initial model weights. This choice was based on the established reputation of the ESM-2 series, which have consistently demonstrated superior performance in representing protein sequences across diverse applications.

For the protein structure encoder, we used Foldseek[4] to transform protein structures into 3Di sequences, which were subsequently modeled using the BERT architecture[60]. The decision to use Foldseek was based on the discovery from SaProt[22] that directly encoding the three-dimensional coordinates of protein structures generated by AlphaFold2 (ref. 61) might lead to a data leakage problem. In contrast, encoding protein structures with Foldseek could effectively alleviate this issue. Given the absence of a pretrained model for the structure encoder, we initiated pretraining of the structure encoder from scratch. Our adoption of the BERT architecture was driven by its proven effectiveness in handling large-scale sequential data.

For the text encoder, we used PubMedBERT[25], a well-known model specifically pretrained on a large-scale biomedical text corpus, as the initial weights. We posited that, as opposed to general language

models pretrained on text data spanning diverse domains, this specialized model would achieve a more profound understanding of protein-related descriptions, thereby enhancing both training efficiency and model performance.

Setting for pretraining

We used the DeepSpeed strategy[62] and AdamW optimizer[63], setting $\beta_1 = 0.9, \beta_2 = 0.98$ and an L_2 weight decay of 0.01. We gradually increased the learning rate from 0 to 4×10^{-4} over the first 2,000 steps decreased it to 4×10^{-5} using the cosine annealing schedule[64]. The overall training phase lasted approximately 100,000 steps trained on 20 NVIDIA 80G A100 GPUs, taking around 2 weeks to complete. We truncated proteins and structural sequences to a maximum of 512 tokens. When the length of a protein exceeded 512 aa, we randomly selected a starting position and extracted a consecutive segment of 512 aa as input for the model. Likewise, we also truncated the text descriptions to a maximum of 100 tokens using the same strategy. Our total batch size consisted of 1,280 protein SSF pairs. Additionally, we used mixed precision training to train ProTrek.

Protein–function mutual retrieval benchmark

We used 4,000 proteins from the SWISS-PROT test set to construct a benchmark for the protein–function mutual retrieval task. To evaluate ProTrek's generalization performance with proteins in other databases, we included 100,000 randomly sampled proteins from UniProt and TrEMBL as 'negative' samples in the test set. The textual descriptions of all these proteins were added to construct the text collection. Notably, a single protein could be associated with multiple textual descriptions, each representing different functional aspect of the protein. Likewise, in the test set, each textual description may correspond to multiple proteins that share similar functions.

The evaluation criteria for retrieval tasks were defined as follows: for text-to-sequence or text-to-structure retrieval tasks, a retrieved protein was considered correct if it appeared in the ground-truth set of proteins associated with the query text. For sequence-to-text or structure-to-text retrieval tasks, a retrieved textual description was considered correct if it belonged to the ground-truth pool of functional descriptions linked to the query protein. Here, each text description is assigned an identifier and the ground-truth pool is the collection of these text identifiers.

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 $G_q \cap G_h \neq \varnothing$. To generate the y axis in Fig. 2c, we counted the number of correct hits across all query proteins in the test set and summed them.

MAP

MAP, a commonly used metric for the information retrieval task, comprises the AP for each query. The MAP not only evaluates whether there are correct hits in the retrieval results but also takes into account the ranks of these correct hits. For a given query, the AP is calculated by the following formula:

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

Here, 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 ith 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:

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

Deep learning baseline models

For protein–function searches, we compared ProTrek to ProtST-ESM-1b from ProtST[32] 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[31]. ProtST offers multiple versions of fine-tuned models. We adopted ProtST-ESM-1b for protein–text retrieval, as it was the model used by the authors in their analysis. Furthermore, we conducted a comparative evaluation of ProtST-ESM-1b and ProtST-ESM-2 on the protein–text retrieval task (see Supplementary Fig. 5). The results demonstrate that ProtST-ESM-1b greatly outperformed ProtST-ESM-2 in protein–text retrieval tasks.

For the downstream fine-tuning tasks, we used the ESM-2 650M version[18] and still used ProtST-ESM-1b as the baseline in the main text for consistency. Furthermore, ProtST-ESM-2 and ESM-3 were also included as baselines for comparison in Supplementary Information. The detailed experimental results are presented in Supplementary Table 10. As indicated in the table, ProtST-ESM-2 demonstrated higher performance than ProtST-ESM-1b in 8 of 11 tasks. However, the improvements are generally modest across most tasks, with the exception of obvious gains observed in the binding site detection and AAV tasks. Furthermore, ESM-3 and ProTrek showed the best overall performance, exhibiting roughly comparable performance across these downstream tasks, although ESM-3 contained twice the number of parameters as ProTrek.

ProTrek structure-to-sequence scores are a useful reference for the inverse fold task

We randomly sampled pairs of two proteins from the test set. For each pair, the structure embedding was computed from one protein, while the sequence embedding was derived from the other. Subsequently, we calculated their structure-to-sequence ProTrek score and obtained the TM-score between the AlphaFold2-predicted structures of the two proteins, thereby forming a (TM-score, ProTrek score) pair. Following this procedure, we randomly sampled 4,000 such pairs and generated the scatter plot (Supplementary Fig. 9).

As illustrated in the figure, the ProTrek structure-to-sequence ProTrek score demonstrates a notable correlation with the TM-score. When the structure-to-sequence ProTrek score is below 20, the TM-score between two proteins is correspondingly very low, indicating their lack of structure similarity.

For protein reverse-folding tasks, generative AI models such as ProteinMPNN or other models are expected to generate protein sequences that fold into the desired structure, that is, high TM-scores with the designed structures. Assuming that the expected TM-score (y axis) is greater than 70, 80 or 90, as discussed in a previous study[65], to meet this requirement, the ProTrek structure-to-sequence score (x axis) should be roughly at least greater than 25. One can also set the threshold to 30 or 35 for a higher standard. Compared to these accurate structure prediction tools such as AlphaFold, ProTrek is much faster and can perform large-scale screening when the researcher designs thousands of candidates, effectively narrowing down the range of potential candidates. Subsequently, other advanced tools such as AlphaFold can be used for further selection and analysis. Here, we preliminarily demonstrate that the structure-to-sequence score of ProTrek may serve as a fast screening method for protein design (more specifically, protein inverse folding) models. However, more rigorous theoretical analysis and validation are still necessary, which is beyond the scope of this paper.

Foldseek
We used Foldseek with the ef4e960ab84fc502665eb7b84573dfff-9c2aa89d 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 version 2.1.9.163. 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.

Literature validation
We validated ProTrek's text-to-protein translation capabilities using literature-based evidence. To simulate data mining on a novel database, we incorporated the studied protein sequences into the UniProt50 database and ranked them together. By assessing the ranking performance using the textual descriptions of the proteins derived from literature, we demonstrated ProTrek's proficiency in text-to-protein retrieval, as indicated by its high ranking accuracy. For protein-to-text retrieval, we inputted the protein sequence to perform functional annotation by retrieving relevant text snippets. By identifying the highest-ranked text, we demonstrated ProTrek's effectiveness in this task. As the retrieved texts in ProTrek are constrained to existing database entries, novel proteins from literature often exhibit multifaceted characteristics that may not correspond to complete verbatim descriptions in the existing database. Consequently, we focused on manually evaluating the functional relevance of retrieved descriptions rather than pursuing exact textual matches. The protein sequences used for this evaluation are listed in Supplementary Table 6.

Fine-grained retrieval
We identified key protein categories, such as CRISPR-associated proteins, deaminases engineered for base editing, various protein families and specific organisms to demonstrate ProTrek's capability for fine-grained text-to-sequence retrieval. To extend our analysis, we defined composite categories such as 'CRISPR + organism' and 'deaminase + organism', facilitating combinatorial text-to-sequence retrieval experiments aimed at simulating the investigation of heterologous protein expression. Each major category was further subdivided to test ProTrek's ability for precise retrieval based on detailed textual descriptions (Supplementary Table 8).

We compiled brief introductions for each item, encompassing diverse aspects of relevant information, and used them as queries to search the UniProt50 database. By analyzing the top 20 retrieved proteins excluded from the training dataset, we assessed whether their descriptions aligned with the input text. The criteria for determining matches for each category are outlined below.

- CRISPR: Large CRISPR-associated proteins included those related to Cas12 and Cas9. Small CRISPR-associated proteins encompassed Cas4, Cas6, Cas14, Cse3, etc. and related proteins.
- Deaminase: Matches were identified on the basis of keywords related to adenosine deaminase and cytosine deaminase.
- Family: Matches were determined by the presence of the family name.
- Organism: Matches were assessed on the basis of the protein's taxonomic lineage.

The annotations for unreviewed proteins in this experiment are generated by the UniProt's automatic annotation pipeline (https://www.uniprot.org/help/automatic_annotation). This pipeline integrates a suite of various predictive models, including sequence analysis methods (SAM), rule-based methods (UniRule and ARBA), feature extraction tools (InterPro) and machine learning models (ProtNLM), to infer functional domains and annotations. By leveraging multiple predictive toolkits, UniProt's pipeline facilitates cross-verification of annotation results, supporting statistical robustness and serving as a reliable preliminary reference. Although these methods provide scalable and statistically reliable annotations, they may yield less accurate descriptions compared to reviewed entries, which are manually annotated based on experimental results, computed features and scientific conclusions (https://www.uniprot.org/help/manual_curation).

Therefore, on the basis of this experiment alone, we cannot definitively claim that ProTrek achieves high accuracy in predicting actual protein properties. What we can conclude is that ProTrek accurately predicts UniProt annotations, potentially with even higher accuracy than UniProt itself, although this cannot be statistically verified. Despite this limitation, we argue that such predictions remain valuable for analyzing the billions of uncharacterized proteins in metagenomic databases, given that UniProt entries represent extensive research efforts and generally demonstrate reliable statistical accuracy. Given ProTrek's demonstrated ability to efficiently process and annotate proteins at scale while maintaining comparable accuracy to UniProt, we can potentially provide statistically reliable descriptions for tens of billions of uncharacterized proteins worldwide and easily establish a UniProt-like database for billion-scale metagenomic data.

Protein-to-EC retrieval evaluation
We curated and standardized an EC textual description database, enabling precise mapping of proteins to enzymatic functions. To construct this database, we obtained the most up-to-date EC number list from the International Union of Biochemistry and Molecular Biology (IUBMB) enzyme database[66] (version 2024_03) and the corresponding descriptive categories for each EC number from the Kyoto Encyclopedia of Genes and Genomes (KEGG) enzyme database[67] (version 2024_09). Using selected categories from KEGG, including EC number, name, class, systematic name, reaction (IUBMB), substrate, product and comment, we generated standardized textual EC descriptions using general large language model (LLM) gpt-4o-2024-11-20. Representative examples and LLM prompts are provided in Supplementary Table 9.

To benchmark this task, we followed the EC number selection strategy introduced in the CLEAN literature[35], which proposed a maximum separation method for confident functional annotation. This greedy algorithm identifies EC numbers whose distances to the protein query are maximally separated from background noise. It assumes a background noise level γ, where incorrect EC numbers cluster around γ within a small deviation ϵ, whereas correct EC numbers are distinctly separated by at least δ ($\epsilon \ll \delta$). The optimal EC description set EC_i is selected by fulfilling the following conditions:

$$|s_{i'} - \gamma| \leq \epsilon, \quad |s_i - \gamma| \geq \delta, \quad |\mathrm{EC}_i| \leq \frac{n}{2},$$

where s represents the Euclidean distance between the protein query and EC descriptions and n is set to 10. The approach maximizes the distinction between correct and incorrect EC descriptions, ensuring reliable functional annotation.

For protein-to-EC retrieval, ProTrek retrieves top EC number with its textual descriptions from the EC textual description database given an query protein. The function can be found in ProTrek online sever by selecting 'EC number' under the 'subsection of text' in the output settings.

ProTrek for protein convergent evolution

We evaluated the ProTrek text-to-sequence search capacity in the presence or absence of clear homology. We focused on two extensively studied categories: enzymes and CRISPR-associated proteins. For our analysis, we selected one representative EC number from each of the six EC classes, alongside six key types of CRISPR-associated proteins. Supplementary Tables 11 and 12 detail the functional descriptions compiled for these selected categories, which were subsequently used as query inputs for the UniProt50 database, selecting the top 100 proteins that were excluded from the training dataset. To assess the retrieval outcomes, we adopted a coarse-grained criterion to determine whether the retrieved proteins matched the input descriptions.

For CRISPR-associated proteins, matches were identified by the presence of CRISPR–Cas-related keywords within the UniProt entries, such as Cas3, Cas9, Cas10, Cas12, Cas13 and Csm/Cmr. For the evaluation of enzyme retrievals, we assessed matches on the basis of hierarchical EC number alignment, recognizing matches at various digit levels.

Because of the large quantity of retrieved proteins, manual review for precise protein identification is infeasible. Instead, we adopted the keyword-matching evaluation approach, whereby proteins were identified on the basis of the presence of specific terms in their associated UniProt text descriptions. For example, a protein was classified as Cas3 if its UniProt annotation included the term 'Cas3'. However, similar to these fine-grained retrieval tasks, the majority of these proteins remain unreviewed. Consequently, the results should be regarded as a preliminary indication rather than definitive validation.

Supervised fine-tuning tasks

We performed supervised fine-tuning on 11 widely adopted downstream tasks and datasets, as detailed previously[23]. For datasets with structural information (subcellular localization, binding site detection, structural similarity, structure class, binary localization, thermostability and metal ion binding), we used the suggested 70% local distance difference test similarity threshold (refer to Proteinshake[68]) for dataset partitioning. For datasets without protein structures (adeno-associated virus (AAV), β-lactamase, fluorescence and stability), we retained the original splits provided in the official literature, as these datasets consist of only mutational variants of one single protein. We used the same experimental settings as described previously[23] (for example, train, validation and test set division, training hyperparameters and evaluation metrics).

UDG mining through ProTrek

ProTrek searches were performed from the OMG_prot50 database in both sequence-to-sequence and text-to-sequence modes. For the sequence-to-sequence search, the human UDG sequence (Supplementary Table 14) was used as the input query. For the text-to-sequence search, the functional description 'UDGs remove uracil from DNA' was the query. The search results were ranked on the basis of the ProTrek matching score and the top k results, where $k = 5$, were returned for analysis. The outcomes of the sequence-to-sequence search were designated as S-V1 through S-V5, representing the top five hits. Similarly, the top five hits from the text-to-sequence search were labeled as T-V1 through T-V5.

Plasmid construction

The base editors used in this study were cloned into a pCMV plasmid containing blasticidin resistance, while sgRNAs were cloned into a pSuper-sgRNA plasmid equipped with puromycin resistance. Protein sequences (S-V1 to S-V5 and T-V1 to T-V5; eGFP as negative control), searched using ProTrek, were codon-optimized and synthesized by Qingke. These sequences served as templates for subsequent site-directed mutagenesis using overlap PCR to introduce point mutations corresponding to UNG2-Y147A. The mutated sequences were then fused to the SpCas9 protein as specified. All primers and plasmid protein sequences used in this study are provided in Supplementary Tables 15 and 16.

Human cell culture and cell transfection

HeLa cells were cultured in DMEM supplemented with 10% FBS (Cellmax). Cells were seeded into 48-well plates (Corning) at a density of 2×10^5 cells per well in 500 µl of complete growth medium. Then, 16 h after seeding, cells were transfected at 70–80% confluency using 2.5 µl of polyethylenimine (linear, molecular weight: 40,000; Yeasen), along with 700 ng of base editor plasmid and 350 ng of sgRNA plasmid. Following transfection, puromycin (2 µg ml^{-1}) and blasticidin (20 µg ml^{-1}) were added to the culture medium for resistance selection. Cells were harvested 5 days after transfection and lysed using genomic DNA (gDNA) lysis buffer (10 mM Tris-HCl pH 8.0, 0.05% SDS and 25 µg ml^{-1} proteinase K (New England BioLabs)) at 37 °C for 1 h, followed by enzyme inactivation at 85 °C for 30 min.

High-throughput DNA sequencing and data analysis

Genomic target sites were amplified from gDNA and analyzed by high-throughput sequencing (HTS) as described previously. Briefly, a primary PCR was performed to amplify the target genomic regions using site-specific primers with universal bridging sequences (5′-GGAGTGAGTACGGTGTGC-3′ and 5′-GAGTTGGATGCTGGATGG-3′) appended to the 5′ ends. The primary amplification was carried out in a 25-µl reaction volume containing 50 ng of gDNA, 0.4 µM locus-specific forward and reverse primers and 12.5 µl of Hieff Canace Plus PCR master mix (Yeasen). The resulting PCR products were subjected to a second round of amplification using primers containing unique barcode sequences. Subsequently, barcoded PCR products were pooled and sequenced on an Illumina HiSeq platform.

Amplicon sequences were aligned to the reference sequence using CRISPResso2. A 10-bp window centered around the middle of the 20-bp protospacer sequence was used to quantify nucleotide modifications, with all other parameters set to default. For base substitution analysis, a 40-bp quantification window was also used, with the 20-bp protospacer centered in the middle of the window. Base substitution frequencies were calculated as the number of base substitution reads divided by the total number of reads, based on the output file 'Quantification_window_nucleotide_percentage_table.txt.' Additionally, the indel frequency for each target site was determined using the output file 'CRISPResso_quantification_of_editing_frequency.txt' and calculated as the number of reads containing indels divided by the total number of reads.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The pretrained model weights (650 million and 35 million parameters) of ProTrek can be obtained online (https://huggingface.co/westlake-repl/ProTrek_650M and https://huggingface.co/westlake-repl/ProTrek_35M, respectively). All precomputed protein sequence, structure and text embeddings for the SWISS-PROT database are available online (https://huggingface.co/datasets/westlake-repl/faiss_index). Larger protein databases (TB-scale embeddings) are accessible through ProTrek's web server (http://search-protrek.com) and official GitHub (https://github.com/westlake-repl/ProTrek), which contains billions of entries. The structural 3Di sequences are available from GitHub (https://github.com/steineggerlab/foldseek). The text descriptions are available from UniProt (https://www.uniprot.org).

Code availability

ProTrek is open-sourced under the MIT license. The code repository is available from GitHub (https://github.com/westlake-repl/ProTrek).

The ProTrek web server can be accessed online (http://search-protrek.com). ColabProTrek v1 and v2 are available online (https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/ColabProTrek.ipynb and https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/ColabSeprot.ipynb?hl=en, respectively). For users requiring private database integration, we provide command-line tools to build embeddings for local deployment through GitHub (https://github.com/westlake-repl/ProTrek#add-custom-database).

References

58. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
59. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2018).
60. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
61. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
62. Rasley, J., Rajbhandari, S., Ruwase, O. & He, Y. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Gupta, R. & Liu, Y.) 3505–3506 (Association for Computing Machinery, 2020).
63. Loshchilov, I. and Hutter, F. Fixing weight decay regularization in Adam. *OpenReview.net* https://openreview.net/forum?id=rk6qdGgCZ (2018).
64. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations* (ICLR, 2017); https://openreview.net/forum?id=Skq89Scxx
65. Xu, J. et al. Protein inverse folding from structure feedback. Preprint at https://arxiv.org/abs/2506.03028 (2025).
66. *Enzyme Nomenclature* (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, 2024); https://iubmb.qmul.ac.uk/enzyme/
67. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
68. Kucera, T., Oliver, C., Chen, D., and Borgwardt, K. ProteinShake: building datasets and benchmarks for deep learning on protein structures. In *Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) (NeurIPS, 2023).

Acknowledgements

We thank P. Beltrao, S. Ovchinnikov, J. Zeng, C. Wang and J. Huang for their valuable suggestions to improve this paper. We thank L. Hong and M. Li for providing the predownloaded OMG and MGnify databases. This work is supported by Zhejiang Province Leading Geese Plan (2025C01094), the National Natural Science Foundation of China (32471547, U21A20427, 82450102 and 32025016), the Ministry of Science and Technology of the People's Republic of China (2022YFA0807300), the National Key Research and Development Program of China (2022ZD0115100), the Zhejiang Key Laboratory of Low-Carbon Intelligent Synthetic Biology (2024ZY01025), the Guangzhou Science and Technology Program City–University Joint Funding Project (2023A03J0001), the Nansha Key Science and Technology Project (2023ZD015), the Guangdong Education Department (2023ZDZX2073), the Hong Kong University of Science and Technology 20 for 20 Cross-Campus Collaborative Research Scheme (G051) and the Westlake Center of Synthetic Biology and Integrated Bioengineering. We also thank N. Li and the Westlake HPC Center for computing resources and technical support.

Author contributions

F.Y. conceptualized and led this research. J.S., Y.H. and S.Y. performed the main research. H.L. and X.C. supervised and led the wet-lab experiments and analyses. J.S. and F.Y. designed the ProTrek network architecture. J.S. conducted the machine learning modeling and implementation. J.S. and X. Zhou collected and curated the training datasets. J.S., S.J. and I.T. performed the computational experimental analyses. I.T., H.L. and X.C. provided bioinformatics guidance. X. Zhang developed the ColabProTrek. Y.H., S.Y., X.S., H.L. and X.C. conducted the wet-lab experiments and evaluations. Y.W. explored an early wet-lab validation experiment. F.Y., J.S., Y.H., X.C. and H.L. wrote and revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

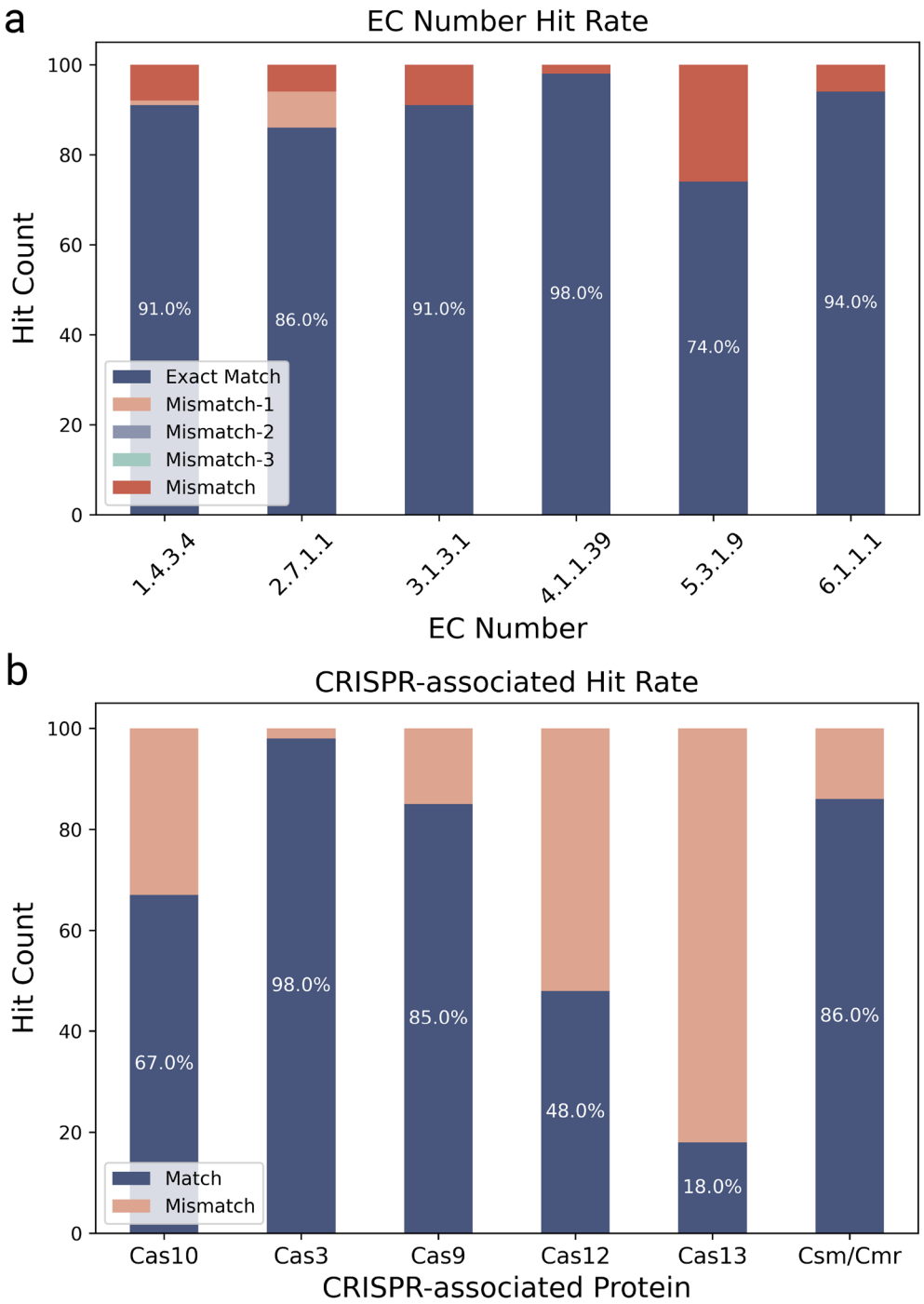
Extended data is available for this paper at https://doi.org/10.1038/s41587-025-02836-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02836-0.

Correspondence and requests for materials should be addressed to Xing Chang, Hongyuan Lu or Fajie Yuan.

Peer review information *Nature Biotechnology* thanks Wei Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-025-02836-0

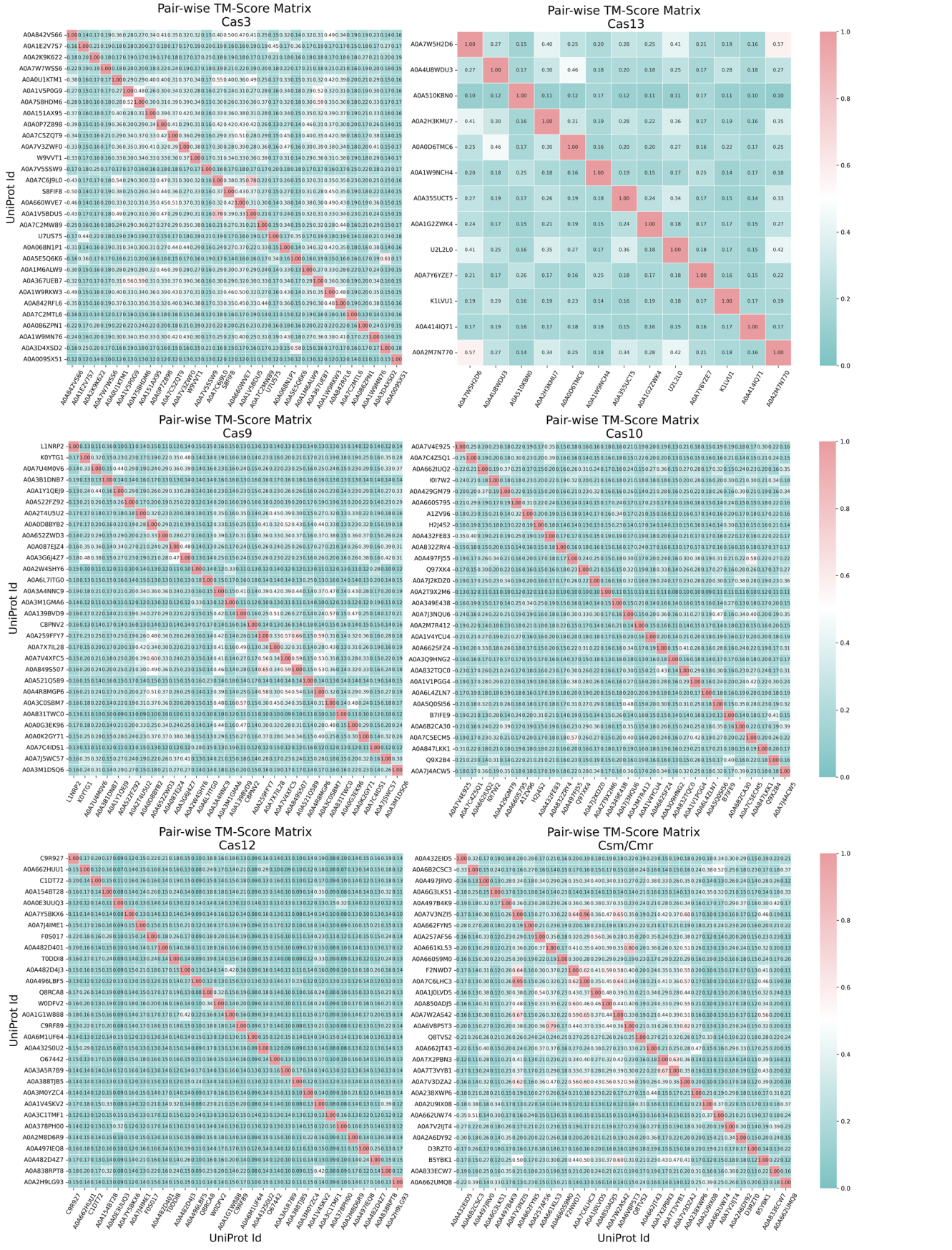


Extended Data Fig. 1 | Evaluation of Hit Rates for Enzyme Commission (EC) Numbers and CRISPR-associated Proteins Using ProTrek's Text-to-Sequence Retrieval Function. Textual descriptions were used to query the UniProt50 database and extract the top 100 proteins not included in the training data. **a**, Bar plot depicts hit rate for six EC numbers selected from six top-level EC classes, categorized into exact matches and varying levels of mismatches (Mismatch-1: mismatch in the last digit; Mismatch-2: mismatch in the last two digits; Mismatch-3: only match the first digit; Mismatch: mismatch all digits). Each bar shows the proportion of exact matches and progressively less accurate matches. The results demonstrate ProTrek's performance in identifying enzyme categories, with variations in match accuracy across different EC numbers. **b**, Bar plot illustrates hit rate for six extensively studied CRISPR-associated proteins (Cas10, Cas3, Cas9, Cas12, Cas13, and Csm/Cmr). Each bar is segmented into matches and mismatches, showcasing ProTrek's retrieval performance across these protein categories. If the corresponding keywords such as 'Cas10', 'Cas3', 'Cas9', 'Cas12', or 'EC:3.4.15.1' appear in the protein's ground truth, it is considered a true hit.

https://doi.org/10.1038/s41587-025-02836-0

Extended Data Fig. 2 | Structural similarity matrix of the top 30 exact matched enzymes. The matrix shows the pair-wise TM-scores for the top 30 exact matched enzymes that were shown in Extended Data Fig. 1a. The value of each cell in the matrix denotes the TM-score, indicating the degree of structural similarity for each protein pair. The color scale indicates the TM-score value, where warmer colors represent higher structural similarity and cooler colors represent lower similarity.

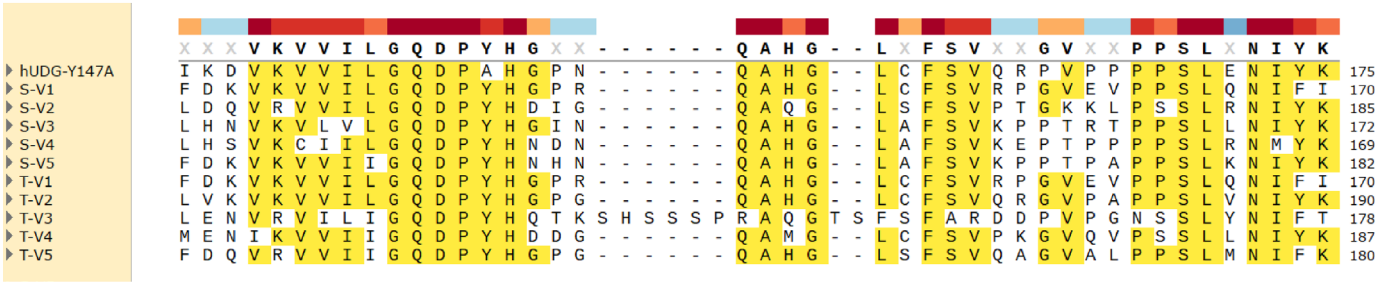
https://doi.org/10.1038/s41587-025-02836-0



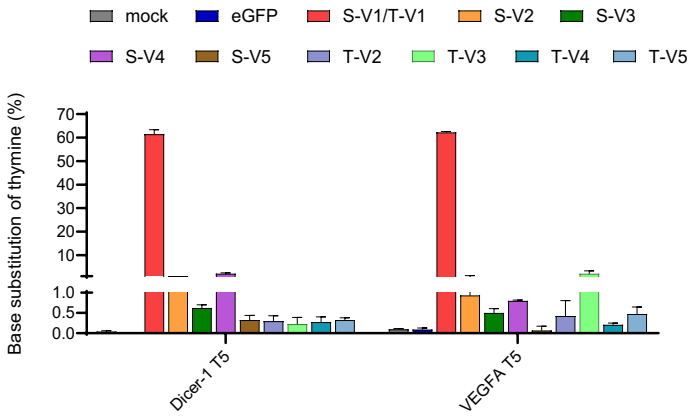
Extended Data Fig. 3 | Structural similarity matrix of the top 30 identified CRISPR-associated proteins. The matrix displays the pair-wise TM-scores for the top 30 matched CRISPR-associated proteins that were shown in Extended Data Fig. 1b. The value of each cell in the matrix denotes the TM-score, indicating the degree of structural similarity for each protein pair. The color scale indicates the TM-score value, where warmer colors represent higher structural similarity and cooler colors represent lower similarity.

https://doi.org/10.1038/s41587-025-02836-0

a

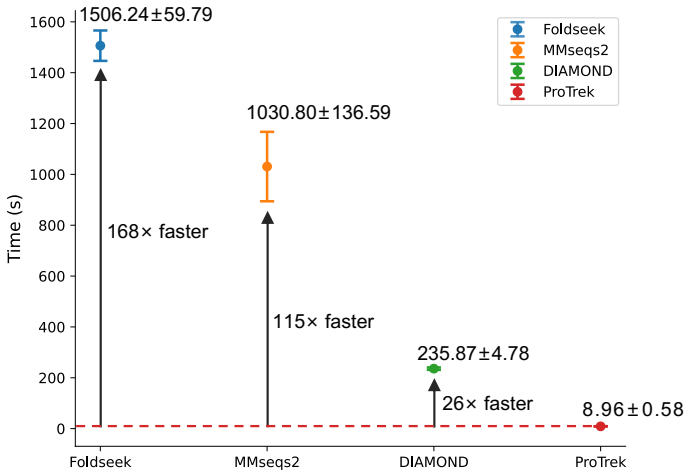


b



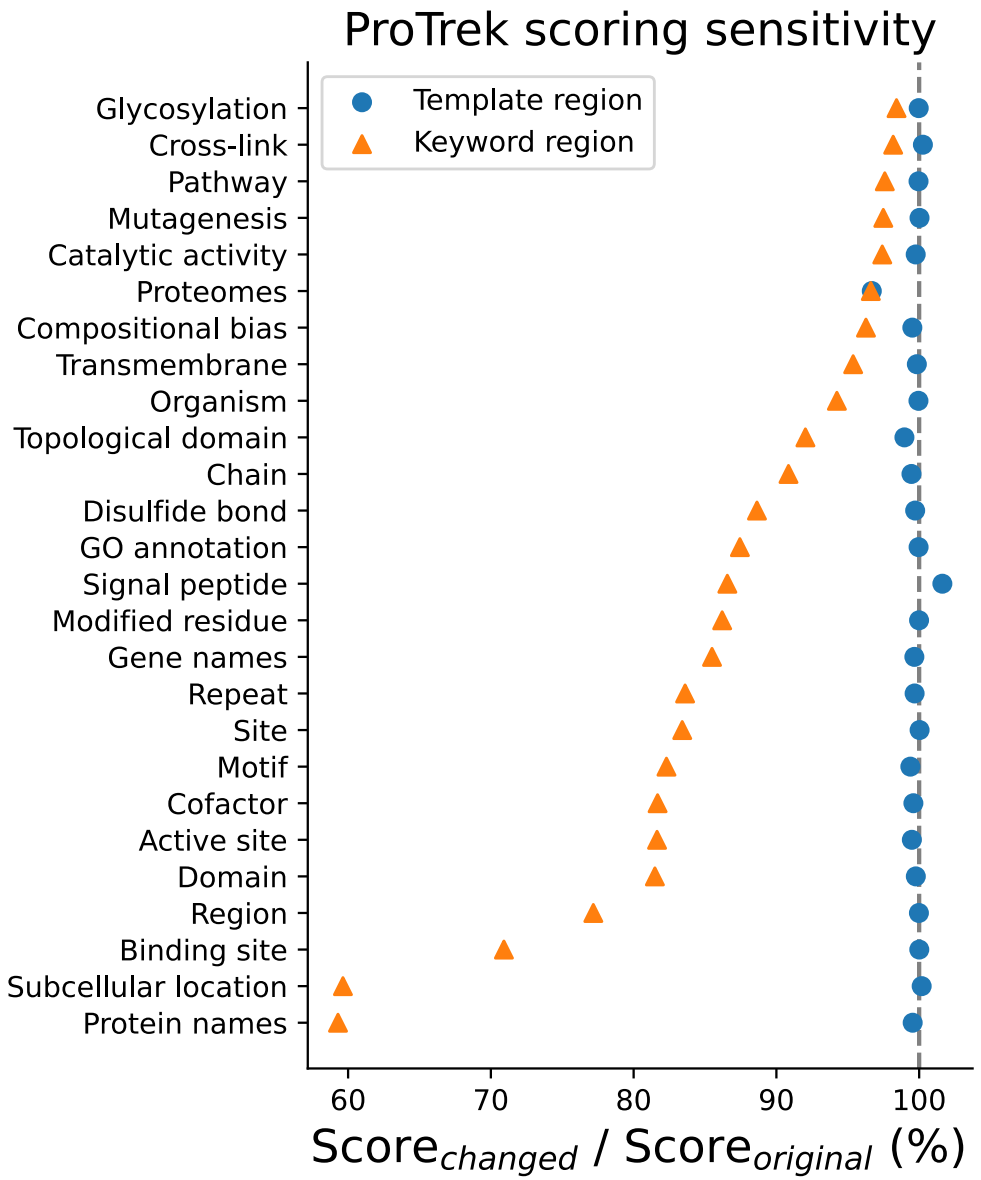
Extended Data Fig. 4 | Experimental validation of proteins identified through ProTrek searches. a, Sequence alignment of hUDG-Y147A with proteins identified via ProTrek searches. S-V1 to S-V5 represent the top five hits obtained from the sequence-to-sequence search, while T-V1 to T-V5 represent the top five hits obtained from the text-to-sequence search. Yellow-highlighted regions denote highly conserved sequences. Sequence alignment was performed using Clustal Omega. **b**, Proteins S-V1 to S-V5 and T-V1 to T-V5 were fused with Cas9n following the introduction of a mutation analogous to UDG-Y147A. Notably, S-V1 and T-V1 refer to the same protein. eGFP was used as a negative control, while the mock group represented cells without any treatment. HeLa cells were transfected with the base editor constructs alongside specific sgRNAs targeting Dicer-1 and VEGFA. Five days post-transfection, thymine nucleotide substitutions at the target sites were quantified using high-throughput sequencing (HTS), with the mutated nucleotide positions annotated relative to the 5′ end of the protospacer. Data are presented as mean ± s.d. from two independent experiments (n = 2).

https://doi.org/10.1038/s41587-025-02836-0



Extended Data Fig. 5 | Alignment speed comparison (CPU time) for 100 query proteins against the UniRef50 database, using 24 CPU cores. ProTrek demonstrates efficient database processing and searching capabilities with rapid response times, achieving over 100-fold speed improvement compared to Foldseek and MMseqs2.

https://doi.org/10.1038/s41587-025-02836-0



Extended Data Fig. 6 | ProTrek score sensitivity analysis for keyword and template regions in functional descriptions. We first computed the similarity score ('Score_original') between a protein sequence and its complete textual description. Next, we systematically removed each word from the description and recalculated the similarity score with the protein. For the template region (blue circle), we averaged all similarity scores obtained after deleting words within this region to derive 'Score_changed'. Similarly, for the keyword region (orange triangle), 'Score_changed' was calculated by averaging all similarity scores resulting from the removal of words in that specific region.

nature portfolio

Corresponding author(s): Fajie Yuan

Last updated by author(s): Aug 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All natural language descriptions and protein sequences are available at https://www.uniprot.org. All protein structures are available at https://alphafold.ebi.ac.uk. Protein structures were encoded using Foldseek with the version ef4e960ab84fc502665eb7b84573dfff9c2aa89d. The downstream task datasets were from the paper https://doi.org/10.1101/2024.05.24.595648.
Data analysis	All python-generated figures were made using matplotlib==3.8.3. TM-scores were calculated using TMalign with the version 20220412. We compared ProTrek to these baselines: Foldseek (version: ef4e960ab84fc502665eb7b84573dfff9c2aa89d), MMseqs2 (version: edb8223d1ea07385ffe63d4f103af0eb12b2058e), BLASTP (version: Protein-Protein BLAST 2.15.0+) and DIAMOND (version: v2.1.9.163).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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. Larger protein databases (around TB-scale embeddings) are accessible through ProTrek's web server http://search-protrek.com, which contains billions of entries, covering a wide range of proteins found in life on earth. 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 web server is located at http://search-protrek.com. The ColabProTrek v1 is available at https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/ColabProTrek.ipynb, and v2 is available at https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/ColabSeprot.ipynb?hl=en. For users requiring private database integration, we provide command-line tools to build embeddings for local deployment via https://github.com/westlake-repl/ProTrek#add-custom-database.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For downstream tasks, ProTrek was evaluated on standard benchmark datasets that are publicly available, so we do not consider the sample size in this case.
Data exclusions	No data was excluded.
Replication	All experiments were conducted with a fixed random seed to ensure the reproducibility.
Randomization	Not applicable.
Blinding	Not applicable.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing	
Data exclusions	
Non-participation	
Randomization	

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Validation

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines
(See ICLAC register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Dating methods

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	
Study protocol	
Data collection	
Outcomes	

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	
Novel plant genotypes	
Authentication	

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Files in database submission

Genome browser session
(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Acquisition

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

Correction

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis