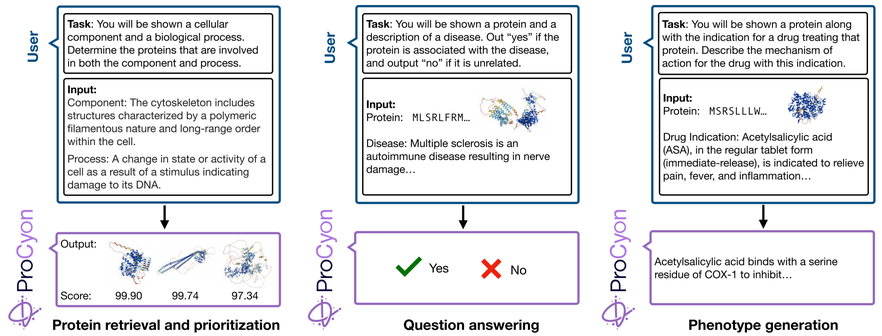
ProCyon : A multimodal foundation model for protein phenotypes

Owen Queen[1,*], Yepeng Huang[1,*], Robert Calef[1,2,*], Valentina Giunchiglia[1,3,4], Tianlong Chen[1,2], George Dasoulas[1], LeAnn Tai[2], Yasha Ektefaie[1], Ayush Noori[1], Joseph Brown[5], Tom Cobley[2,6], Karin Hrovatin[7,8], Tom Hartvigsen[9], Fabian J. Theis[7,10], Bradley L. Pentelute[5,14], Vikram Khurana[11,12,14], Manolis Kellis[2,14], Marinka Zitnik[1,13,14,15,‡]

[1]Department of Biomedical Informatics, Harvard Medical School; [2]Computer Science and Artificial Intelligence Laboratory, MIT; [3]Department of Brain Sciences, Imperial College London; [4]Centre for Neuroimaging Sciences, King's College London; [5]Department of Chemistry, MIT; [6]Department of Computing, Imperial College London; [7]Institute of Computational Biology, Computational Health Center, Helmholtz Munich; [8]TUM School of Life Sciences Weihenstephan, Technical University of Munich; [9]School of Data Science, University of Virginia; [10]School of Computation, Information and Technology, Technical University of Munich; [11]Department of Neurology, Brigham and Women's Hospital; [12]Harvard Stem Cell Institute; [13]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University; [14]Broad Institute of MIT and Harvard; [15]Harvard Data Science Initiative, *Co-first authors; ‡Corresponding author: marinka@hms.harvard.edu

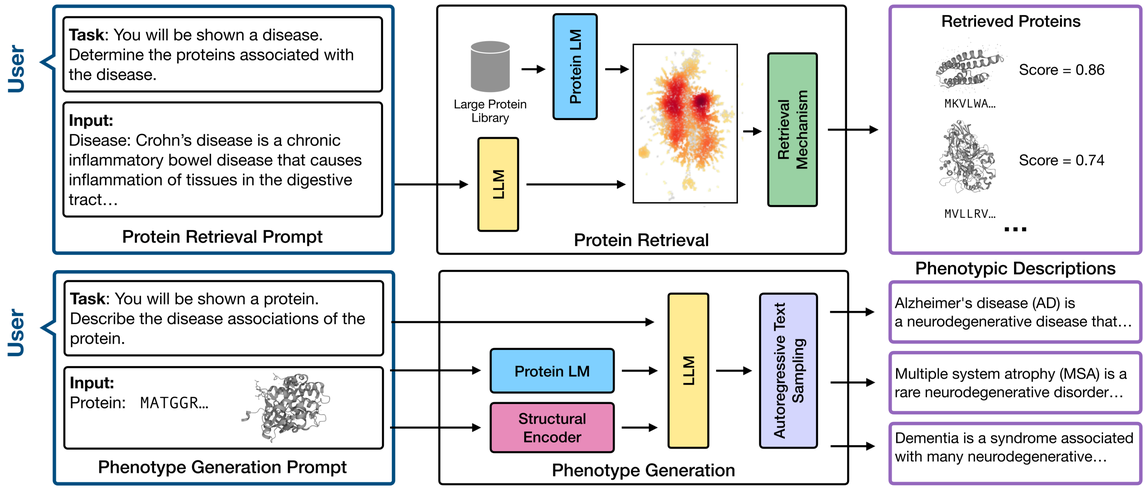
Motivation and Model Overview

- Characterization of protein sequence and structure has rapidly outpaced functional characterization
- 20% of human protein-coding genes lack known functions
- 95% of life science publications focus on ~5000 well-characterized proteins



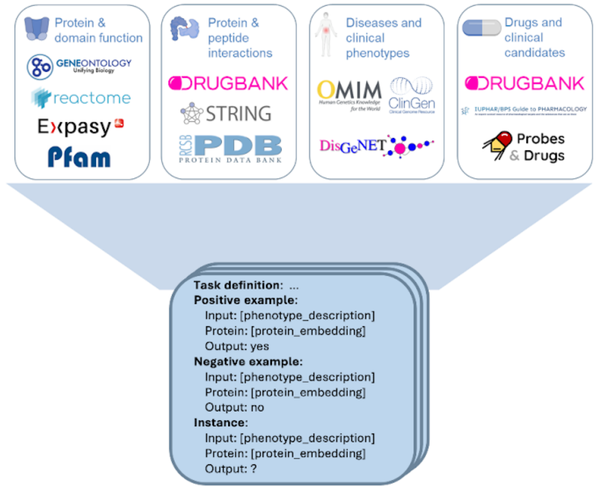
- **Phenotype:** Observable characteristic of protein function, e.g. enzymatic activity, pathway involvement, disease association
- Comprehensive phenotype annotation requires:
  - Integration of knowledge across disparate domains
  - Generalizing beyond pre-defined categories (GO)
- Natural language as a unifying modality to solve these problems

- Solution: Co-train large language model and protein representation models
- Result: Model that operates over **interleaved natural language and protein inputs**, enabling **universal annotation of proteins**

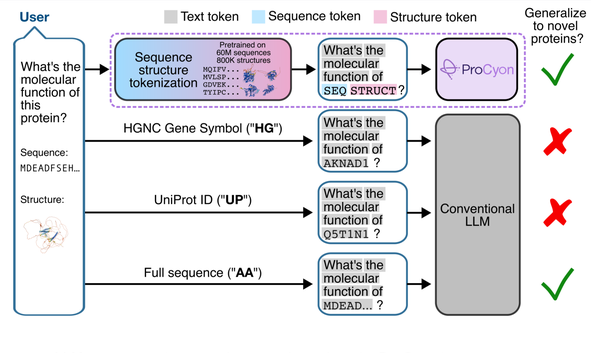


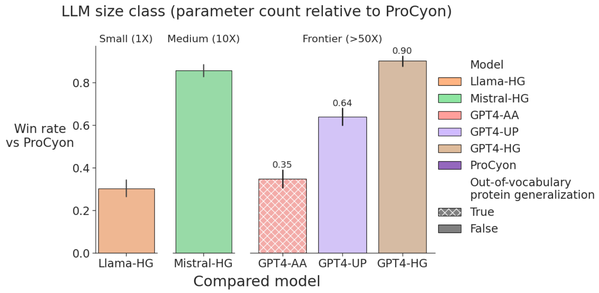
ProCyon-Instruct Dataset

- Source protein phenotypes from broad range of sources
- 18k human proteins + 48k phenotypes →1.85M protein-phenotype pairs → LLM rephrasing → **33M samples**
- Protein and phenotype are transformed into prompts for **instruction tuning** of the model

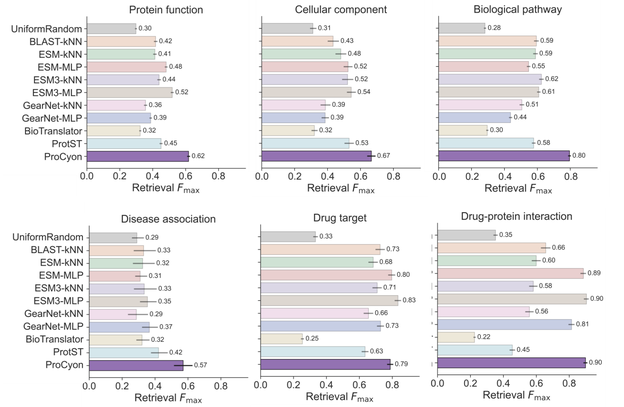


Comparison to Conventional LLMs

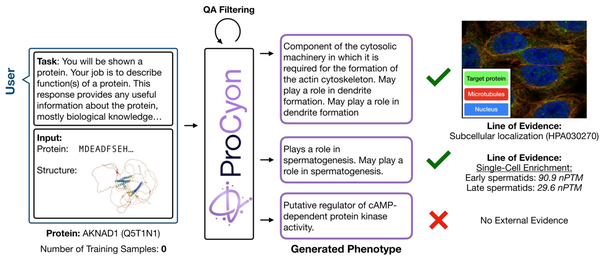




Benchmarking



Poorly Characterized Protein Annotation



ProCyon generated phenotypes for the protein AKNAD1. Supporting external evidence found for 2 out of 3 predicted phenotypes.