Decoding the Molecular Language of Proteins with Evolla

Xibin Zhou, Chenchen Han, Yingqi Zhang, Huan Du, Jiayuan Tian, Yang Liu, Hongyuan Lu, Fajie Yuan

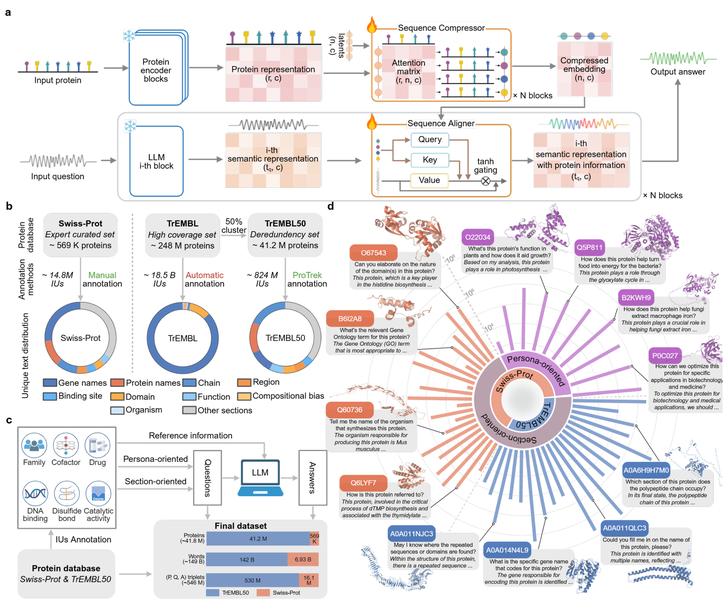
INTRODUCTION

Proteins, nature's intricate molecular machines, are the products of billions of years of evolution and play fundamental roles in sustaining life. Yet, deciphering their molecular language—understanding how protein sequences and structures encode biological functions—remains a cornerstone challenge in modern biology. While traditional homology-based tools fail for divergent proteins and classifiers offer narrow, label-centric answers, a generative understanding of functional mechanisms is missing.

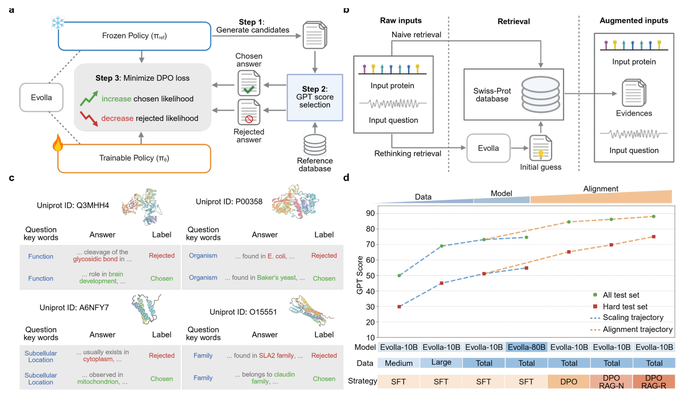
Here, we introduce Evolla, an **80-billion-parameter** generative model designed to decode protein function through natural language dialogue. Evolla integrates information from protein sequences, structures, and user queries to generate precise, mechanistically-detailed insights. A key innovation lies in its training on an unprecedented AI-generated dataset: **546 million protein question-answer pairs (150 billion tokens)**, enabling it to capture the immense complexity of the protein universe. Post-pretraining, Evolla integrates Direct Preference Optimization (DPO) and Retrieval-Augmented Generation (RAG) to align the model for functional veracity and factual accuracy.

Evolla's performance surpasses traditional methods, particularly in their blind spots. We demonstrate it can **reconstruct entire functional systems** via semantic search; for instance, it identified **all 36 partners of an archaeal ribosome from a single protein query where homology tools fail**. Furthermore, we used Evolla to drive novel, experimentally-validated discoveries: identifying **four eukaryotic-like VPS4 proteins in Asgard archaea** and **discovering a novel PETase from a deep-sea bacterium**. These results prove its ability to generate testable, real-world hypotheses, shifting the paradigm from simple annotation to generative discovery. The online demo is available at **http://www.chat-protein.com/.**

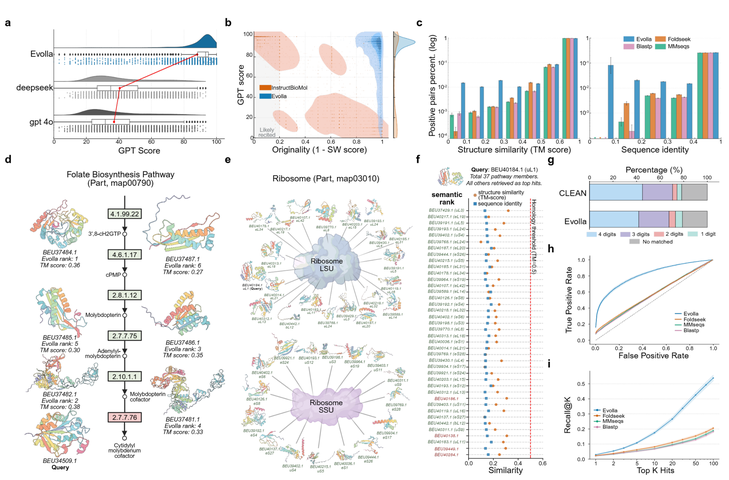
Methods & Dry-lab Experiments



Evolla's architecture and the training data. (a) Evolla architecture. A frozen protein encoder generates a representation that is distilled into a fixed-length protein functional codebook by the Sequence Compressor. This codebook is then injected into a frozen Large Language Model (LLM) via trainable, interleaved Sequence Aligner blocks to generate a textual answer. **(b)** Sources and curation of protein Information Units (IUs). Swiss-Prot IUs (569K proteins, 14.8M IUs) are expert-reviewed and directly used. TrEMBL sequences were clustered at 50\% identity to generate TrEMBL50 representatives (41.2M proteins), which were re-annotated using ProTrek (824M IUs). The final dataset combines Swiss-Prot and ProTrek-annotated TrEMBL50. Donut charts show unique IU text distributions for top categories. **(c)** Dataset construction pipeline. IUs from the protein database (Swiss-Prot & TrEMBL50) are used to generate questions via persona-oriented and section-oriented strategies. Questions and their corresponding IUs serve as reference information for LLM-based answer generation. The final dataset contains 546M protein-question-answer triplets, covering 41.8M proteins and 150B word tokens. **(d)** Data distribution of top 15 sections for both Swiss-Prot and TrEMBL50 dataset and 10 persona for Swiss-Prot dataset.

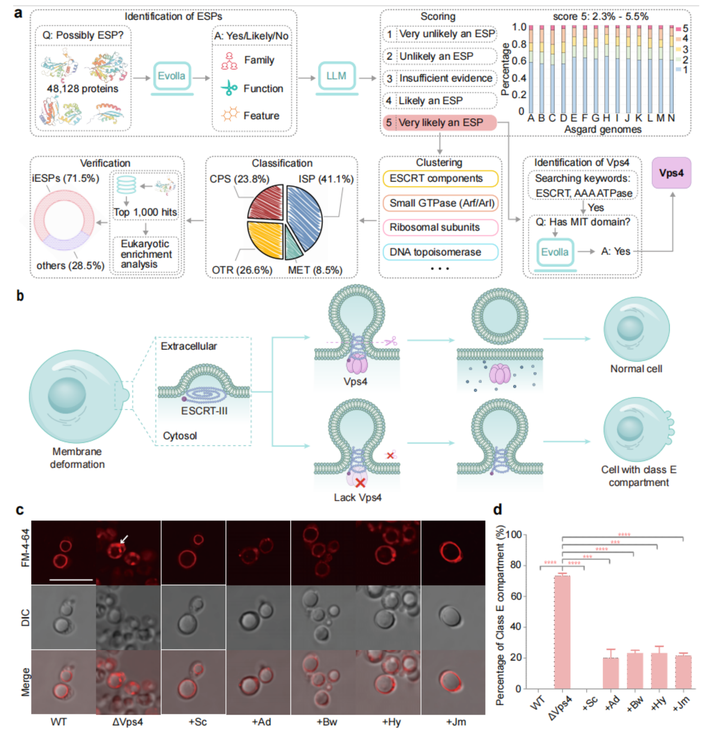


Enhancing Evolla with advanced strategies. (a) Direct preference optimization (DPO) workflow. Evolla is duplicated into a frozen policy (π_{ref}) and a trainable policy (π_θ). The frozen policy generates candidate answers (Step 1), which are scored by GPT against a reference database (Step 2), yielding chosen and rejected answers. These preference pairs are used to minimize the DPO loss (Step 3), increasing the likelihood of chosen answers while decreasing that of rejected ones.**(b)** Retrieval augmented generation (RAG) workflow. Two retrieval strategies are employed: naive retrieval directly queries the Swiss-Prot database using the input protein and question, while rethinking retrieval first generates an initial guess via Evolla, then retrieves relevant information. Retrieved evidences are combined with the input protein and question as augmented inputs for Evolla to generate the final answer. **(c)** Example of a factual hallucination. **(d)** Models are trained at different parameter scale, data scale and strategies.

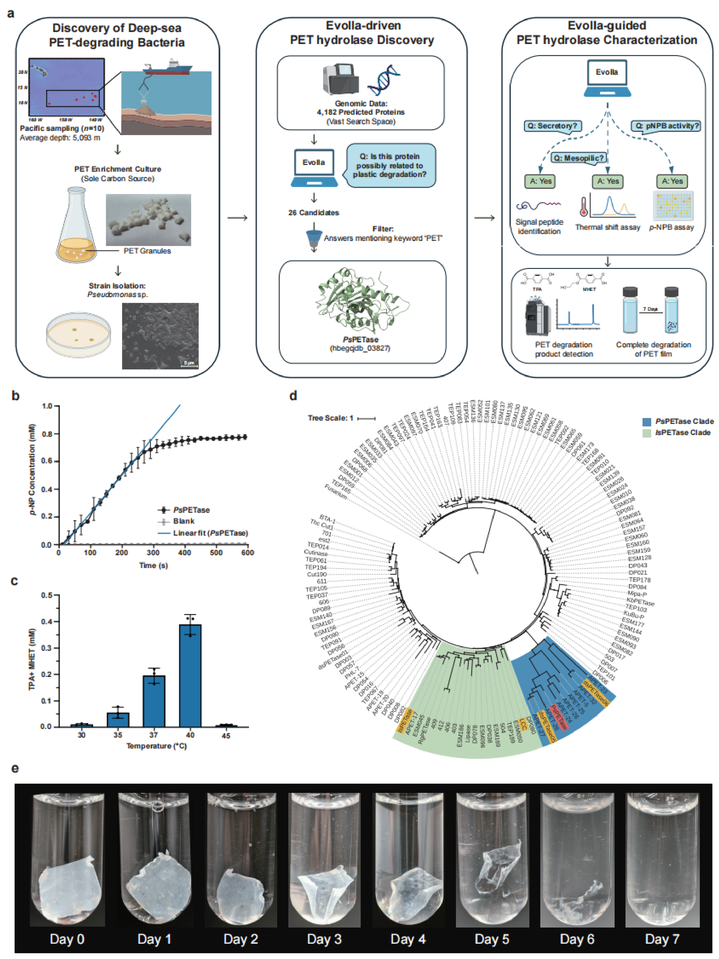


Benchmark. (a) Performance against general LLMs. Raincloud plot of factual accuracy (GPT Score) for generated protein descriptions, showing the superior performance of the domain-specialized Evolla over general LLMs. **(b)** Performance against InstructBioMol. 2D density plot of factual accuracy versus generative originality, showing Evolla resolves the performance trade-off that confines InstructBioMol to either accurate recitation or inaccurate novelty. **(c)** Performance on Gene Ontology annotation. Retrieval of functionally related pairs (by shared GO terms) binned by structural (left) and sequence (right) similarity. Evolla maintains high performance in the low-similarity regime, a known blind spot for homology-based tools. **(d)** Reconstructing a metabolic pathway. A search using a single enzyme from the Folate Biosynthesis pathway semantically retrieves its direct metabolic partners, despite their low structural similarity. **(e, f)** Assembling a molecular machine. A search for the archaeal ribosome complex (e) using protein uL1 as a query. The semantic rank plot (f) of Evolla's top 40 hits shows near-perfect retrieval, identifying all 36 other ribosomal proteins (green text) with only four non-ribosomal hits (red text). The correctly retrieved proteins exhibit low sequence (orange dots) and structural (blue dots) similarity to the query. **(g)** Zero-shot EC number prediction. On a temporal hold-out set of newly characterized enzymes, Evolla's zero-shot performance in predicting full four-digit EC numbers is comparable to the state-of-the-art specialist model CLEAN (41.85\% vs 44.98\%; ns, not significant). **(h, i)** Performance on pathway annotation for retrieving KEGG pathway co-members in a novel archaeon proteome. (h) ROC curves comparing Evolla, Foldseek, MMseqs and Blastp. (i) Recall@K curves up to top 100 hits comparing Evolla, Foldseek, MMseqs and Blastp.

Wet-lab Experiments



Global identification of Asgard archaeal eukaryotic signature proteins (ESPs) and functional characterization of Vps4. (a) Computational pipeline for ESP identification. Evolla screened 48,128 protein structures across 14 Asgard genomes, providing classification decisions based on protein family, function, and features. Responses were scored (1–5) by an LLM; high-confidence hits (score 5; 2.3–5.5% per genome) were functionally clustered into four categories (ISP, CPS, MET, OTR) and structurally verified via Foldseek. **(b)** Schematic of Vps4 function. Vps4 drives ATP-dependent ESCRT-III disassembly and membrane scission; its loss leads to class E compartment formation. **(c)** Vacuolar morphology by FM 4-64 staining. The Δ*vps4* strain exhibited class E compartments (white arrow), fully rescued by *S. cerevisiae* Vps4 (+Sc) and partially rescued by Asgard homologs (+Ad, +Bw, +Hy, +Jm). Scale bar, 10 μm. **(d)** Quantification of class E compartments. Asgard Vps4 expression reduced compartment accumulation from ~70% (Δ*vps4*) to ~20%. Mean ± SD, n = 3 biological replicates. ****P < 0.0001, unpaired two-tailed t-test.



Discovery and Evolla-guided functional mining of the deep-sea PET hydrolases, *Ps*PETase. (a) Schematic of the Evolla-guided discovery workflow. A *Pseudomonas sp.* isolate was obtained from deep-sea sediments (avg. depth 5,093 m) via PET-enrichment culture. Evolla screened 4,182 predicted proteins through natural language interrogation, identifying *Ps*PETase as a top candidate, and further guided experimental validation design. **(b)** Esterase activity of *Ps*PETase measured by *p*-NPB hydrolysis at 37 °C (pH 8.0), showing a linear initial velocity (v_0) of 0.164 mM min^{-1}. Data: mean ± s.d., n = 3. **(c)** Temperature-dependent solvent-cast PET (scPET) films depolymerization profile. Released monomers (TPA + MHET) were quantified by HPLC after 24 h. Optimal activity was observed at 40 °C, consistent with Evolla's mesophilic prediction. Mean ± s.d., n = 3. **(d)** Maximum-likelihood phylogeny of PsPETase among 137 validated PET hydrolases. PsPETase defines a distinct marine clade (blue), evolutionarily divergent from terrestrial archetypes such as IsPETase (green). **(e)** Time-course degradation of scPET films by PsPETase (1,000 nM, buffer/enzyme refreshed daily), achieving complete degradation within 7 days.