NL2Prot: Protein Database Retrieval with Natural Language Queries

Qianyu Zheng, Wentao Yue

Georgia Institute of Technology

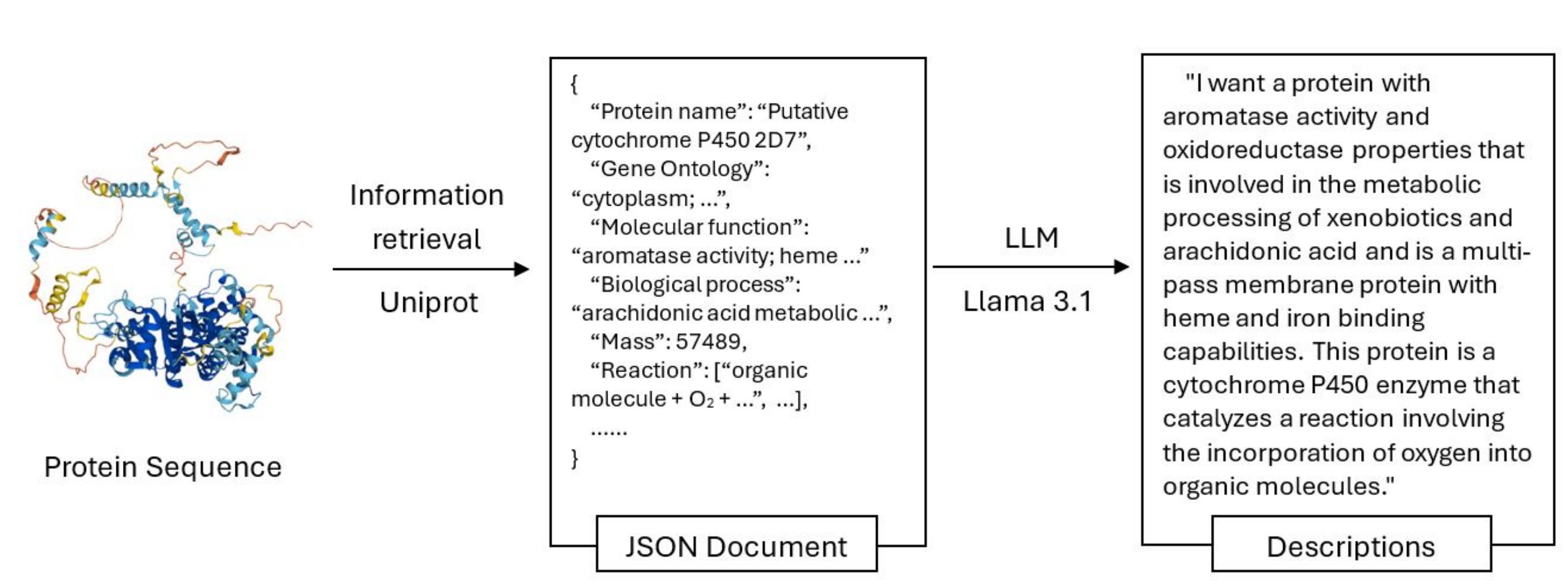
Introduction

Protein databases like UniProt[1], PDB[2] and GenBank[3] are fundamental repositories for biological research. They drive advances in biological discoveries by storing accurate information about protein sequences. However, accessing these resources often requires specialized knowledge of both biological terminology and query languages.

Our project aims to democratize protein data access by developing a versatile interface that accepts both everyday language and technical descriptions. This dual-functionality interface will streamline research workflows for scientists and make protein information more accessible to curious individuals, from students to healthcare consumers.

Dataset & Materials

Dataset curation: UniProt + LLM



```
{
  “Protein name”: “Putative cytochrome P450 2D7”,
  “Gene Ontology”: “cytoplasm; …”,
  “Molecular function”: “aromatase activity; heme …”
  “Biological process”: “arachidonic acid metabolic …”,
  “Mass”: 57489,
  “Reaction”: [“organic molecule + O₂ + …”, …],
  ……
}
```

"I want a protein with aromatase activity and oxidoreductase properties that is involved in the metabolic processing of xenobiotics and arachidonic acid and is a multi-pass membrane protein with heme and iron binding capabilities. This protein is a cytochrome P450 enzyme that catalyzes a reaction involving the incorporation of oxygen into organic molecules."

Models: BERT & ESM

We use Small-BERT and several variants of ESM2 model as the two encoders for descriptions and sequences in this project.

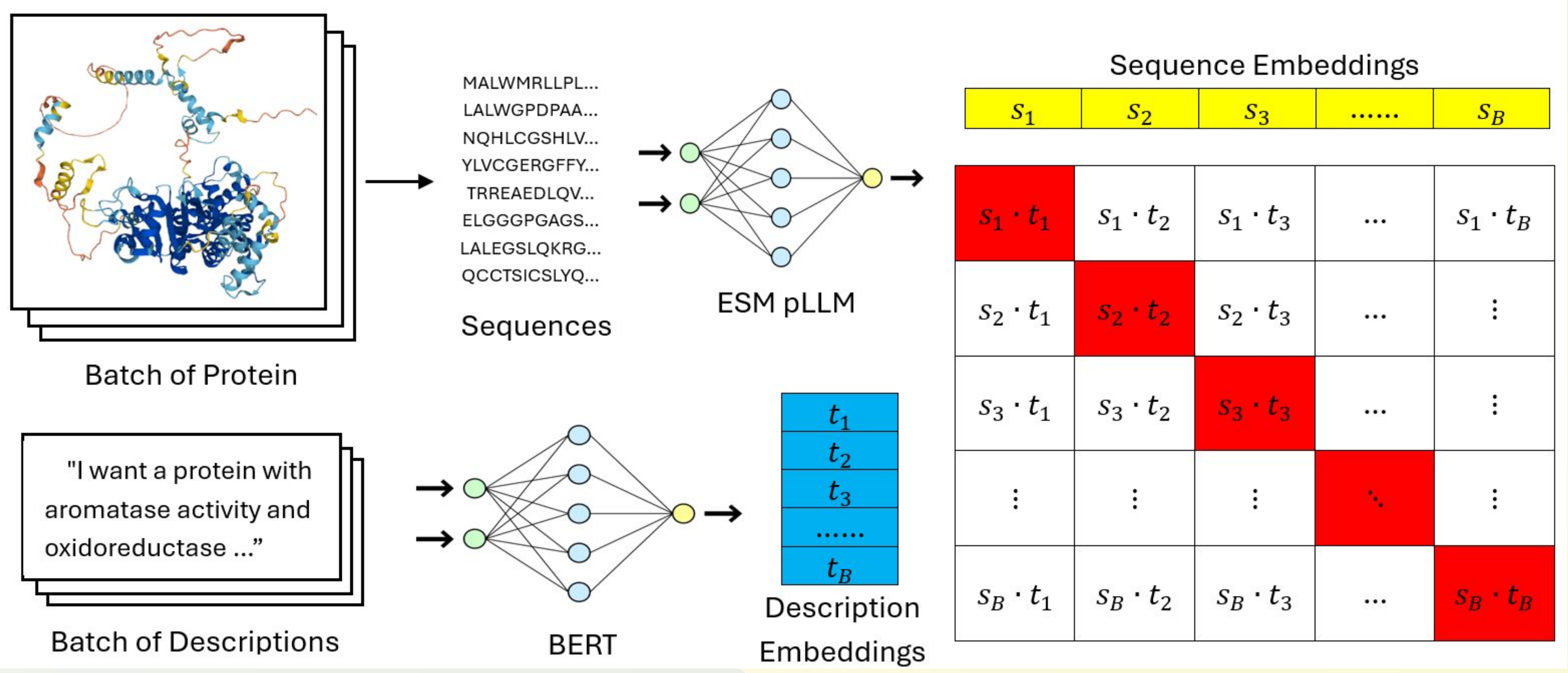
Small-BERT[4]: 28M, 6 layers compared to 109M, 12 layers of BERT.

ESM2[5]: a family of large language model pre-trained on evolutionary-scale protein sequence data that learns protein patterns to predict structure and function based on raw sequences.

Methods

Proposed Method: CLIP-based Contrastive Learning

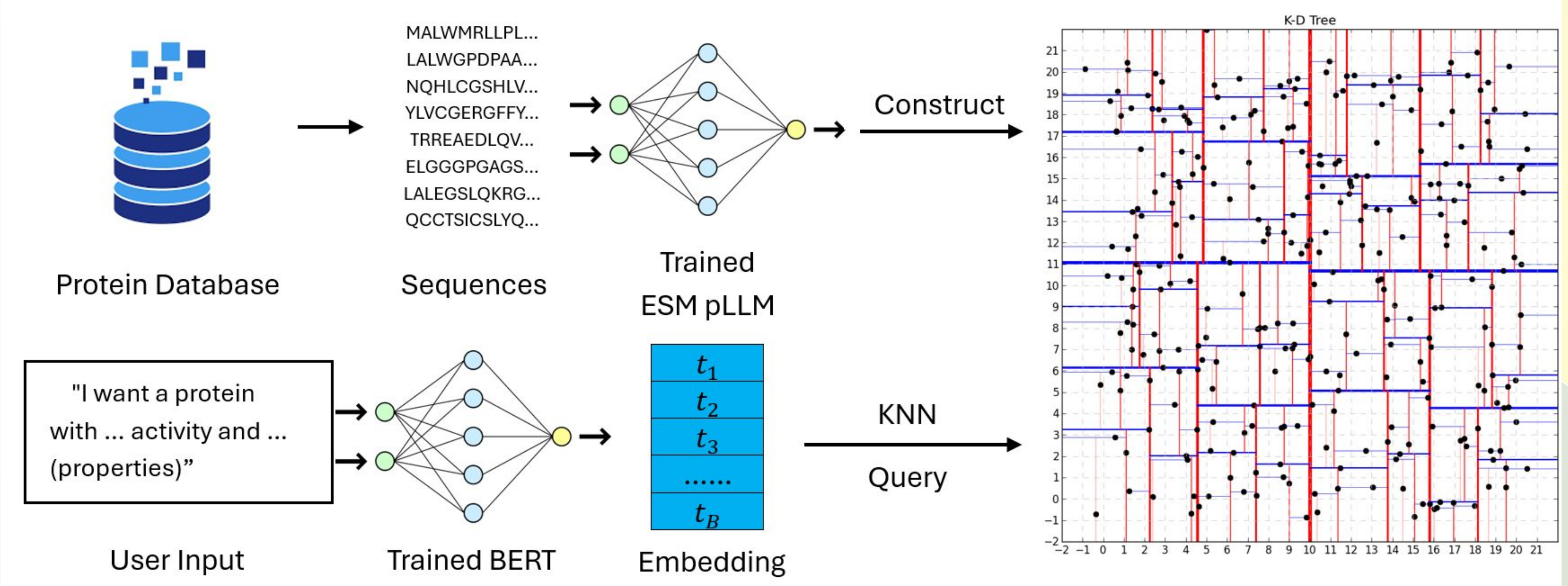
Training procedure:



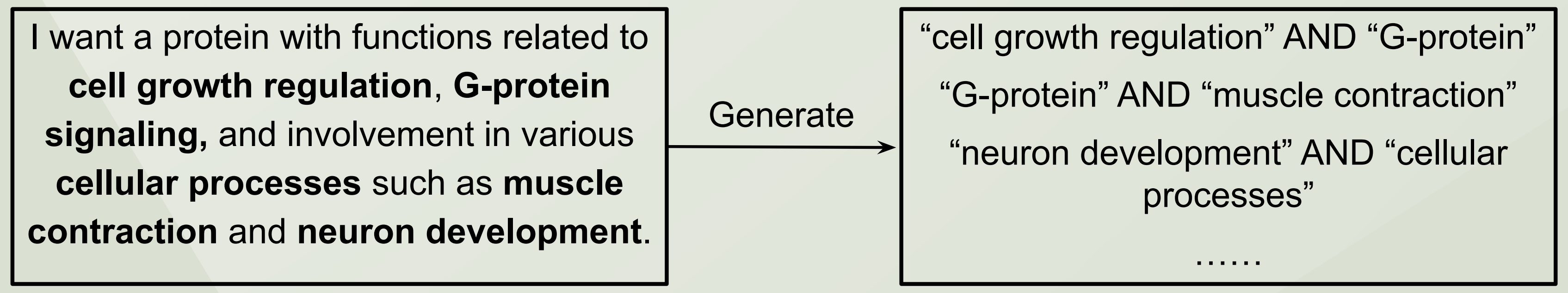
s_1	s_2	s_3	……	s_B
$s_1 \cdot t_1$	$s_1 \cdot t_2$	$s_1 \cdot t_3$	…	$s_1 \cdot t_B$
$s_2 \cdot t_1$	$s_2 \cdot t_2$	$s_2 \cdot t_3$	…	⋮
$s_3 \cdot t_1$	$s_3 \cdot t_2$	$s_3 \cdot t_3$	…	⋮
⋮	⋮	⋮	⋱	⋮
$s_B \cdot t_1$	$s_B \cdot t_2$	$s_B \cdot t_3$	…	$s_B \cdot t_B$

t_1, t_2, t_3, ……, t_B

Inference procedure:



Baseline Method: Keyword Extraction



I want a protein with functions related to **cell growth regulation**, **G-protein signaling**, and involvement in various **cellular processes** such as **muscle contraction** and **neuron development**.

“cell growth regulation” AND “G-protein”
“G-protein” AND “muscle contraction”
“neuron development” AND “cellular processes”
……

Extract keywords for generating UniProt queries:

- NER (SpaCy) + RE for keyword extraction.
- Take combinations of 2 - 3 keywords to construct query.
- UniProt API for querying.

Conclusions & Outcomes

Accuracy: for both layman & professional inputs

Professional inputs: I want a protein with a cytoplasmic localization, involved in the positive regulation of cellular response to growth factor stimulus and the regulation of signal transduction processes. This protein is a phosphoprotein with alternative splicing variants, playing a role in cellular signaling pathways.

Model	Top-10 Acc	Top-20 Acc	Top-50 Acc	Time*
150M-ESM2	78.25%	**87.10%**	**94.05%**	39.81 ± 0.11 s
35M-ESM2	**78.80%**	86.75%	93.90%	
8M-ESM2	75.65%	84.55%	92.90%	
Baseline	67.33%			> 24 h

Layman inputs: I want a protein with functions that help regulate the body's immune response to infections and potentially control cell death in response to certain signals.

Model	Top-50 Acc	Top-100 Acc	Time*
150M-ESM2	**75.10%**	**83.95%**	29.05 ± 0.15 s
35M-ESM2	74.80%	83.25%	
8M-ESM2	73.45%	82.70%	
Baseline	24.55%		~ 17 h

*Time: inference time for processing 2,000 description queries with the deployed configuration on an Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz with 12 cores. Experiments (besides baseline) are repeated five times, and the mean and std are reported above.

References

[1] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 11 2016.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) *Nucleic Acids Research* 28: 235-242 https://doi.org/10.1093/nar/28.1.235.

[3] Eric W Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, Ilene Karsch-Mizrachi, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D84–D86, https://doi.org/10.1093/nar/gkz956

[4] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962, 2019

[5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Acknowledgements

We would like to thank PACE for providing GPU resources, Professor Xu and TA Wei Zhou for their support in developing ideas and providing feedback during the project.