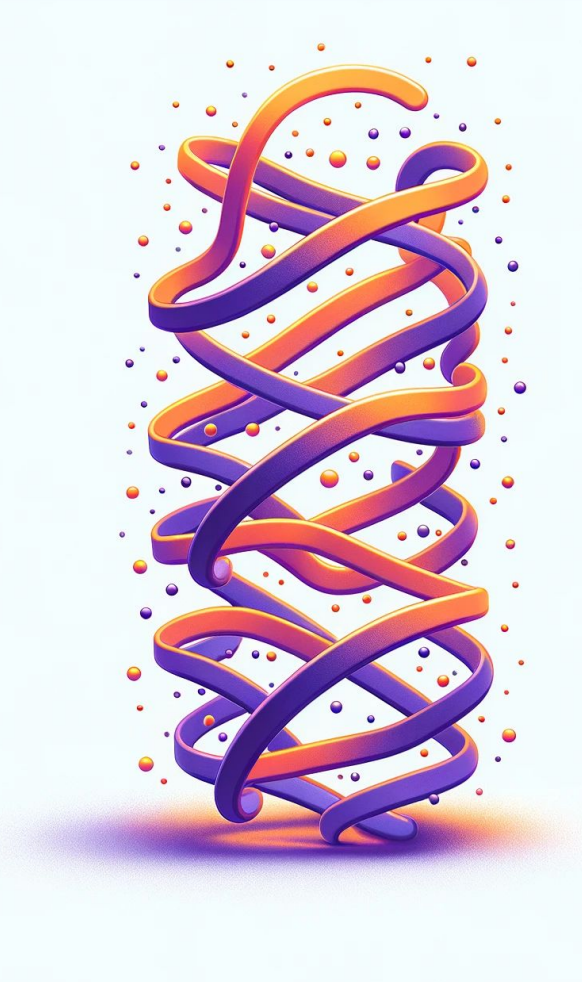


PESS:

Protein Embedding & Semantic Search

Ishaan Javali

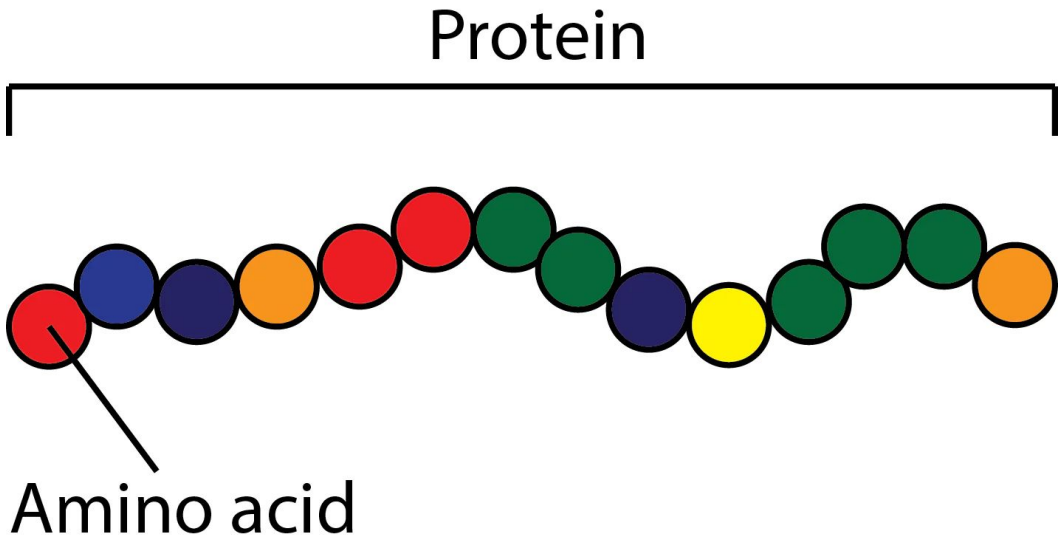
COS 597A

Background

Proteins are molecules that are critical for all processes of life

- Critical for immune response, transport and storage, muscle contraction…
- Relevant for disease prevention and drug development

Amino acid sequence determines protein structure and thus, function



Data Collection

InterPro

Classification of protein families

Database of 200M+ proteins.

Example sequence:

MRLLALSGLLCMLLLCFCIFSSEGRRHPAKSLKL…

(FASTA Format)

Mus musculus

17k proteins

Dataset:

- Accession (ID)
- Name
- Protein Sequence
- Sequence Length
- Description (if exists)

Data Processing

2 vector search indices (with cosine similarity metric):

Functional Descriptions
~15k proteins

↓

OpenAI's Ada Model
(350M parameters)

↓

1536-dimensional vector

Protein Sequences
~17k proteins

↓

Meta's ESM Model
(650M parameters)

↓

1280-dimensional vector

Examples

textSearch("Cytokine involved with immune system that targets blood cells")

1 Granulocyte-macrophage colony-stimulating factor (0.871313214)
Cytokine that stimulates the growth and differentiation of hematopoietic precursor cells from various lineages, including granulocytes, macrophages, eosinophils and erythrocytes...

2 Interleukin-22b (0.867893875)
Cytokine that contributes to the inflammatory response in vivo

3 Interleukin-7 (0.866275311
Hematopoietic cytokine that plays an essential role in the development, expansion, and survival of naive and memory T-cells and B-cells

esmSearch(the first 85/380 amino acids in *human* Cytochrome B protein)

1 Cytochrome b (0.980121553)
... the b-c1 complex mediates electron transfer from ubiquinol to cytochrome c...

esmSearch(all 380 amino acids in *human* Cytochrome B)

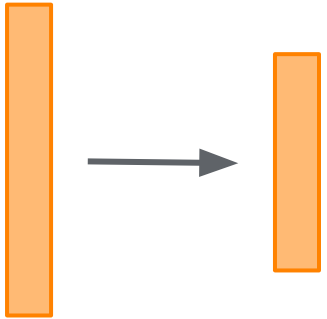
2 Protein transport protein Sec61 subunit alpha isoform 1 (0.959146738)
...complex that mediates transport of signal peptide-containing precursor polypeptides across the endoplasmic...

3 Protein SERAC1 (0.979960322)
unrelated description

4 Phospholipid-transporting ATPase ABCA1 (0.978927851)
Catalyzes the translocation of specific phospholipids from the cytoplasmic to the extracellular/lumenal leaflet of membrane coupled to the hydrolysis of ATP. Thereby, participates in phospholipid transfer to apolipoproteins to form nascent high density lipoproteins/HDLs. Transports...

Mapping from Text → ESM Embedding Space

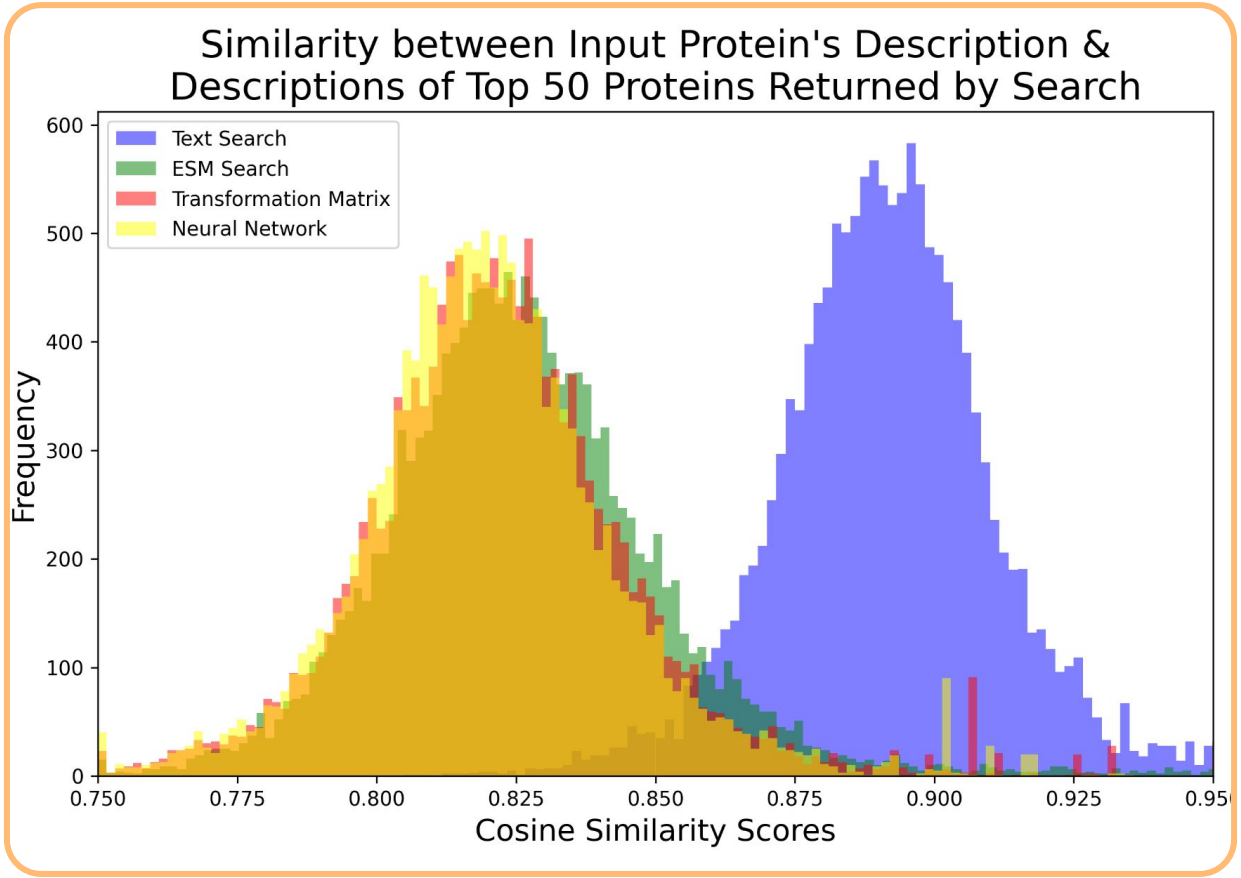
1536-dimensional vector → 1280-dimensional vector



Purpose: Not all proteins in InterPro have descriptions

Extract structural meaning from text queries

	Avg Euclidean Distance	Avg Cosine Similarity	Avg MAE of Dot Products	Avg MAE per vector element
Transformation Matrix	2.0793	0.9554	6.0050	0.0379
NN w/ MSE Loss	2.1792	0.9560	6.5735	0.0404
NN w/ Cos Similarity Loss	73.9164	0.9594	535.2528	0.6336



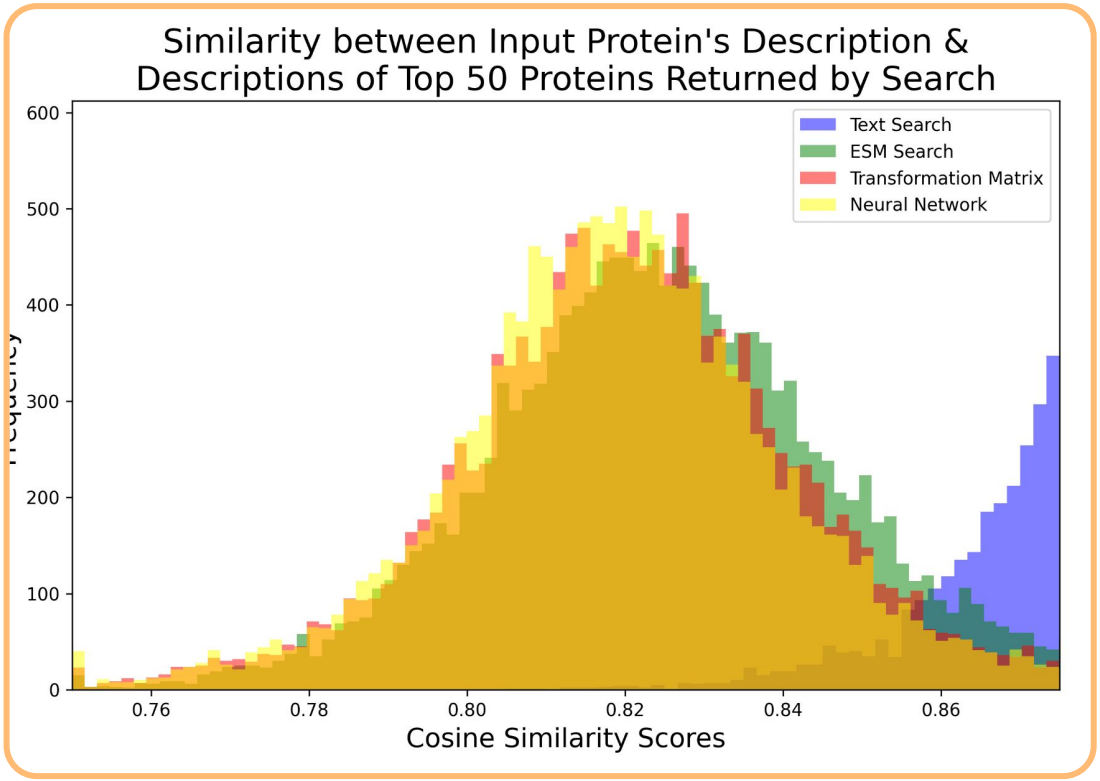
***Purpose**: Evaluate how well the search results' descriptions relate to original protein's description*

Procedure

For each protein in dataset:

1. Do search (each of the 4 methods)
2. Get top 50 proteins' IDs (in metadata)
3. Using IDs, retrieve text embedding for each of the 50 result proteins (precomputed)
4. Calculate avg cosine similarity between the query protein's description & top 50 proteins' descriptions

Text → ESM Transformation Evaluation



Zoomed in view

Examples

```
textToESMSearch("Involved with transport to cell membrane")
```

The text → ESM vector transformation retains structural & functional info from the original text input

1 Glycerophosphodiester phosphodiesterase domain-containing protein (0.982093632)

Glycerophosphodiester phosphodiesterase that promotes neurite formation … removes the GPI-anchor of RECK, leading to release RECK from the plasma membrane (By similarity). May contribute to the osmotic regulation of cellular glycerophosphocholine…

2 Protein SERAC1 (0.979960322)

unrelated description

3 Phospholipid transporting ATPase ABCA1 (0.978927851)

Catalyzes the translocation of specific phospholipids from the cytoplasmic to the extracellular/lumenal leaflet of membrane coupled to the hydrolysis of ATP. Thereby, participates in phospholipid transfer to apolipoproteins to form nascent high density lipoproteins/HDLs. Transports…

Takeaways

Semantic search tool: Researchers / Doctors can search for proteins via desired, natural-language descriptions

- **Studying proteins with *known* functions, *unknown* sequences**
  (via text embedding search)

- **Studying proteins with *unknown* functions, *known* sequences**
  (via ESM search)

To be continued…

- **Extracting latent, structural meaning from text queries**
  (via fine-tuning or better text → ESM embedding transformation)