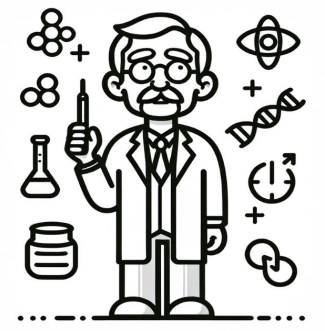
Biomedical generative pre-trained based transformer language model for age-related disease target discovery

By Aarohi Chopra

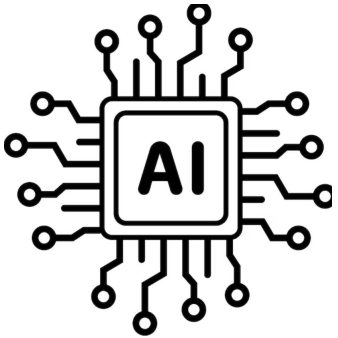
Background

Motivation

Why



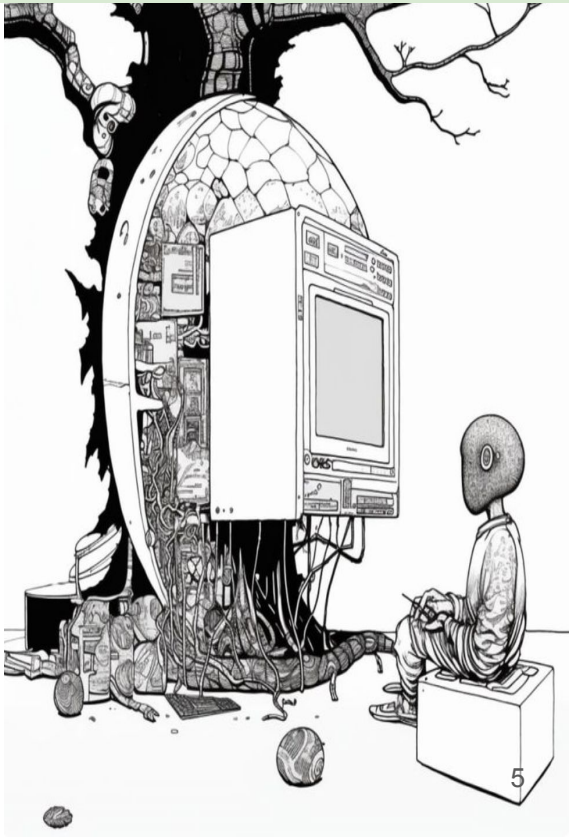
It’s complicated



Goal

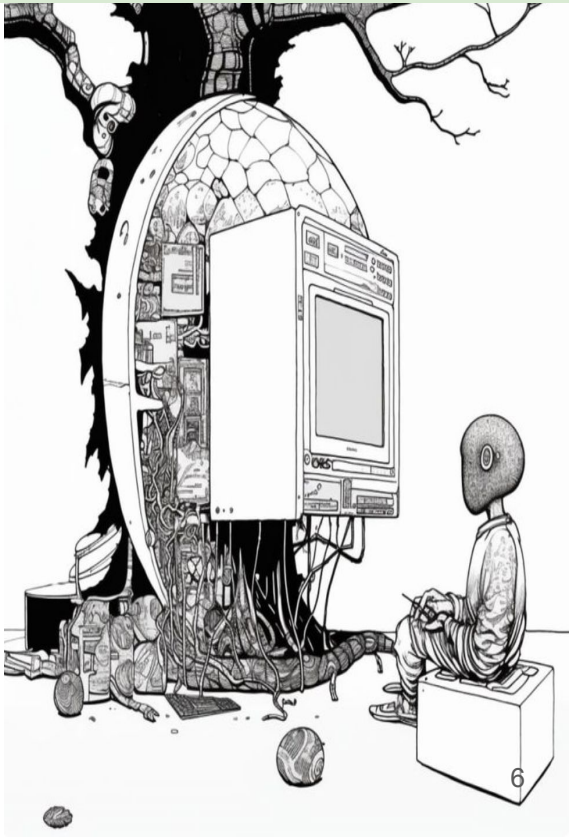
Large Language Models

1. LLMs?
2. Idea: Next word prediction
3. Medical Context: Disease to genes



Large Language Model Usage

1. Construct Prompt
2. Tokenize
3. Predict Gene



Data

Training Data:

- National Institutes of Health Grants: 9
- PubMed: 15 million

Reference Data

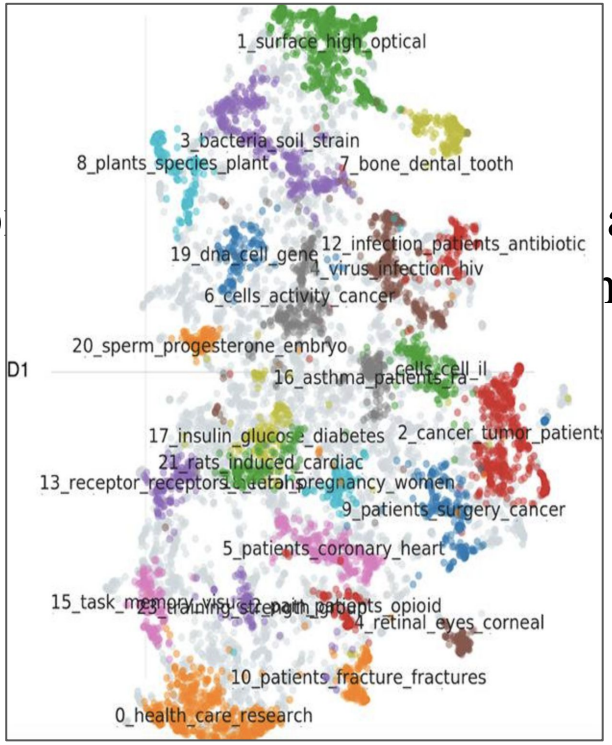
- HUGO Gene Nomenclature Committee
- GeneAge
- ClinicalTrials.gov
- DrugAge
- PubMedQA

	Number of parameters	Texts for training
BioGPT + Grants	349 million	15M PubMed abstracts + 900K biomedical grants full texts
BioGPT	349 million	15M PubMed abstracts
BioGPT Large PubMedQA	1.5 billion	15M PubMed abstracts + PubMedQA dataset
BioGPT Large	1.5 billion	15M PubMed abstracts

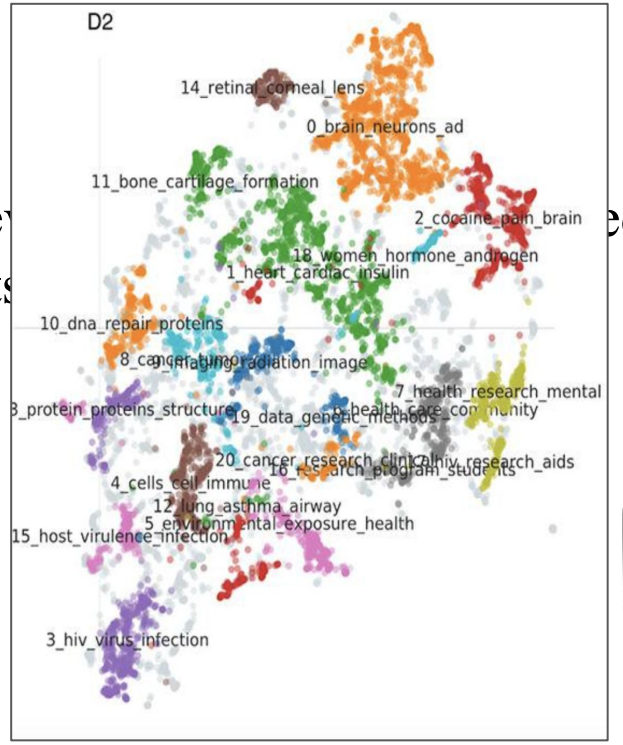
Exploratory Data Analysis

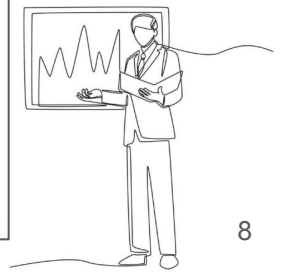
BERTop at prev ed and nt texts

Grants



PubMed





Methods

Embeddings

1. Sample: PubMed abstracts + grant texts

2. Processing: Lowercase + Removing stopwords(NLTK)

3. Embedding: “all-mpnet-base-v2” sentence-transformers model

Training

1. Base: BioGPT (347 million parameters)pre-trained on 15 million PubMed abstracts
2. Task: Maximize log-likelihood of a next token given the context
3. Customize: BioGPT-G additional: 900 thousand grant abstracts
4. Details:

   Training time: 40 hours
   Batch size: 16
   Gradient accumulation step: 64 per device on four A5000 GPUs
   Adam algorithm with 100 warm-up steps
   Learning rate: 5e-5

$$def\ 1: \ maximize \sum_{t=1}^{M} log_2 p(y_t \ | \ y_t),$$

Prompt Optimization

1. Prompt: “Human gene targeted by a drug for treating {DISEASE} is”
2. Task: Next word prediction
3. Construction parameters settings
4. Efficiency Estimation
5. Findings:
    - Larger prompt length BAD
    - Addition of Articles GOOD
6. Brute Result: Total $\text{Prob}_{\text{Target Gene}}$ = Multiplication of all tokens probabilities

Validation

1. Analyze the top K values

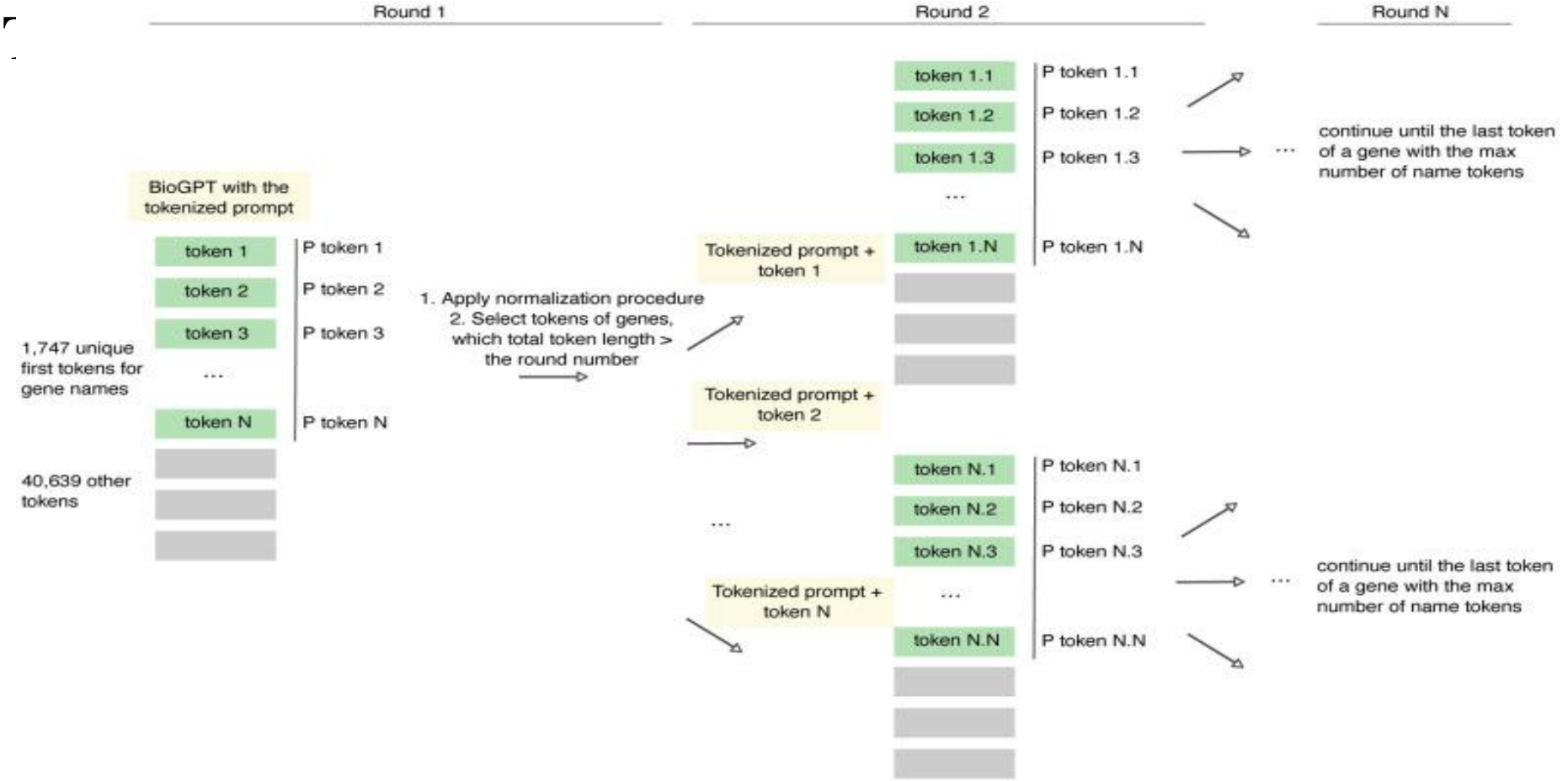
2. ELFC - Log Fold Change of Enrichments (occurrence of targets relative to their general frequency the dataset)

$$ELFC(score) = log2\left(\frac{targets_k * N}{k * targets_N}\right)$$

3. HGPV - Hypergeometric p-value (occur by chance or unusually high)

4. Higher valu

$$HGPV(score) = -log10\left(1 - hgcdf(targets_k, k, targets_N, N)\right),$$


Round 1
Round 2
Round N
BioGPT with the tokenized prompt
token 1
token 2
token 3
...
token N
P token 1
P token 2
P token 3
P token N
1,747 unique first tokens for gene names
40,639 other tokens
1. Apply normalization procedure
2. Select tokens of genes, which total token length > the round number
Tokenized prompt + token 1
Tokenized prompt + token 2
Tokenized prompt + token N
token 1.1
token 1.2
token 1.3
token 1.N
P token 1.1
P token 1.2
P token 1.3
P token 1.N
token N.1
token N.2
token N.3
token N.N
P token N.1
P token N.2
P token N.3
P token N.N
continue until the last token of a gene with the max number of name tokens
continue until the last token of a gene with the max number of name tokens

Processing Tokenization

1. Maximum token limit
2. Filtering longer genes
3. Prevent skew for abundant non-gene data
4. Use cases (based on token length)
    A. Gene length <= iterations(N)
    B. Gene length > iterations(N)
5. Varying gene length normalization (separate and combined)
6. Normalization of Final Probabilities with longer lengths

Results

1. Product/token length (Figure C 1.3 & 2.4)

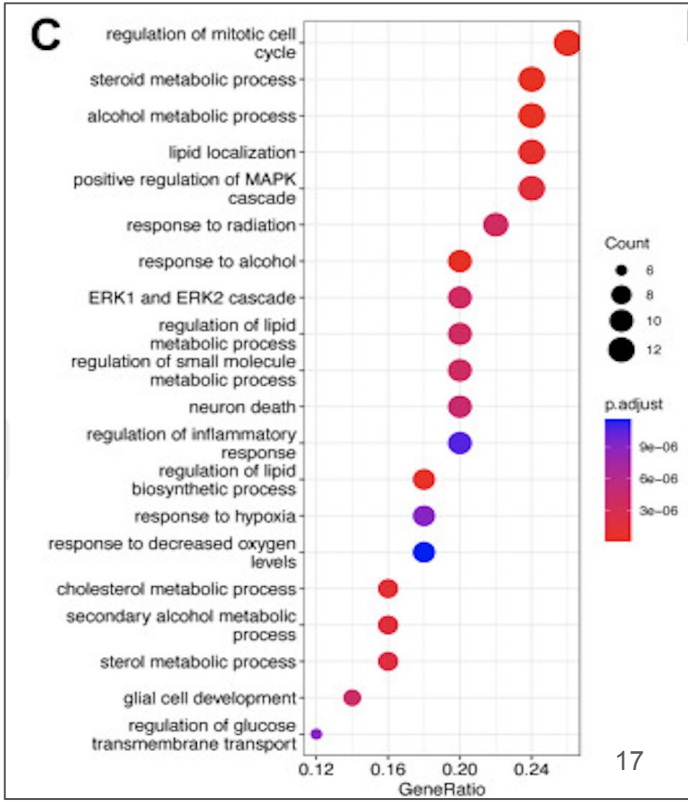
2. Normalizing individual tokens(each iteration) given $\sum_{\text{all tokens}} = 1$ (Figure C 3.1)

3. Parameter selection

Normalization version	HGPV	ELFC	AUROC
1. Probabilities normalization on each iteration			
Total sum	5.93	4.11	0.59
Separate sum	4.53	3.78	0.51
Not sum	1.39	1.73	0.63
2. Final probabilities normalization according to number of tokens in a gene name			
$\sqrt[\text{length}]{\text{Total sum}}$	0.78	0.61	0.51
Total sum / length	5.94	4.02	0.59
Separate sum / length	5.59	3.83	0.57
Not sum / length	1.71	1.97	0.63
3. Variation of the parameter for the final normalization			
Total sum / length*parameter (=4)	5.50	3.75	0.58
Total sum / length * parameter (=4)**-1	5.49	3.76	0.57
Total sum / length ** length	5.37	3.67	0.57
4. Apply cut-off for the max length of tokens in a gene name			
Total sum / token limit (=5)	5.15	3.56	0.56
Total sum / length + token limit (=5)	5.49	3.76	0.57

Target Discovery for Aging

1. Top 200 genes were selected
2. ∩ GenAge database ($p < 0.001$)
3. ∩ PubMed database ($p < 0.001$)
4. Gene Ontology (GO) enrichment analysis (FDR adjusted $p < 0.01$)

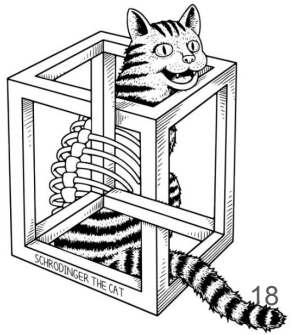


LLM Explainability

1. Task: Protein Embeddings and graphs

2. Hypothesis: Learn not only probabilities also internal associations of word similarities

3. Result: Protein [pubmed(~ aging) and BioGPT(aging)] ---- Protein [pubmed(aging) and BioGPT(aging)]



Protein Graphs:

1. "The human age-associated gene is the" gene name —> Tokenized$_{\text{mean output pooling}}$ —> PyTorch tensor$_{1024}$ —> Individual proteins

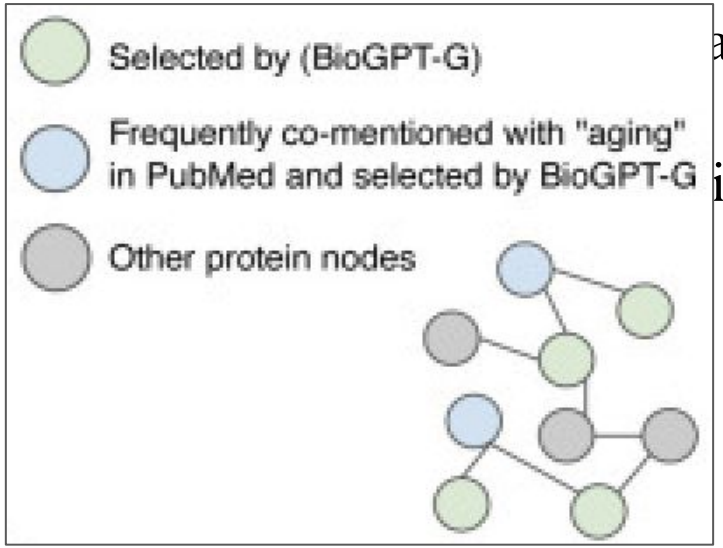
2. Source Nodes: Proteins in (PubMed abstract co-mentions) AND (BioGPT aging)

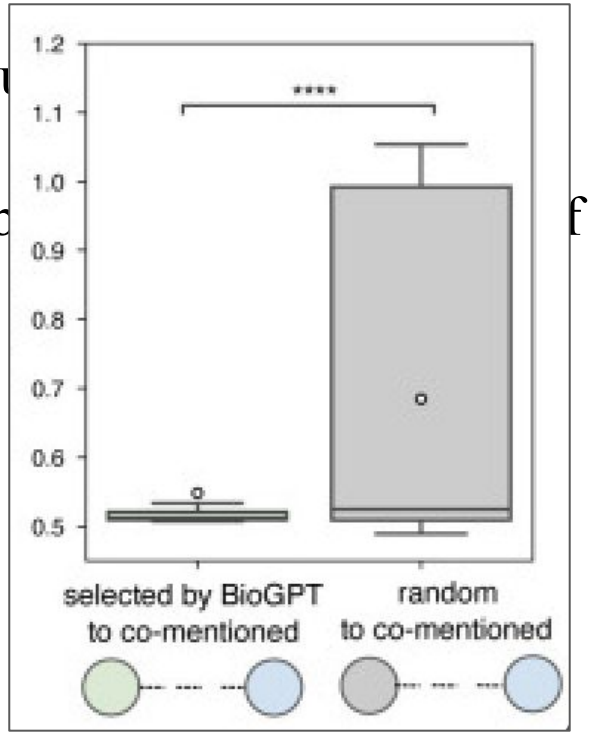
3. Target Nodes: Pro(BioGPT aging) & NOT(PubMed abstract co-mentions)

4. Random Nodes: Random proteins for main experiment and control values

Graphs Continued

1. as constru
2. ilarities b f of 0.507
3.





Dual-purpose Disease and Age-Related Targets

Targets	Protein family	Clinical trial status[1]	Known as age-related genes[2]	Potential dual-purpose candidates[3]
BRCA1	Acyltransferase	No	Yes	No
CCR5	GPCR	Yes	No	Yes
EGF	Growth factor	No	Yes	No
MIP	Generic protein	No	No	No
PTH	Generic protein	No	No	Yes
RET	Receptor kinase	Yes	Yes	Yes
SRC	Tyrosine kinase	Yes	Yes	Yes
TNF	Tumor necrosis factor	Yes	Yes	Yes
VHL	Ligase	No	Yes	No

Discussion

1. Novel: CCR5, [illegible], and PTH (Previous Table)
2. After filtering TNF, SRC and RET, and two novel genes, CCR5 and PTH
3. TNF -> Age associated inflammation
4. SRC -> Targeted by Dasatinib(Senolytic)
5. RET -> Higher levels causes thyroid cancer + Increases with age
6. CCR5 -> Neuroinflammation + Alzheimer's disease
7. PTH -> Osteoporosis + Frailty

References

1. https://pubmed.ncbi.nlm.nih.gov/30269508
2. https://pubmed.ncbi.nlm.nih.gov/36936271
3. https://pubmed.ncbi.nlm.nih.gov/35837482
4. https://pubmed.ncbi.nlm.nih.gov/32534441
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10564439/

Thank You :)