

Language-guided biology

James Zou

4/16/2025





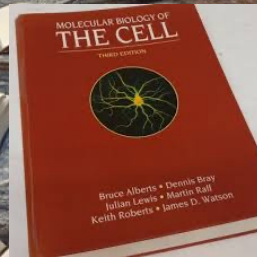
ChatGPT



PubMed®

bioRxiv

arXiv



LLMs are trained over tremendous corpus of knowledge.

How many cell types are in the human brain? Be concise



The human brain contains approximately 100 different cell types, including various types of neurons, glial cells (such as astrocytes, oligodendrocytes, and microglia), and other specialized cells.



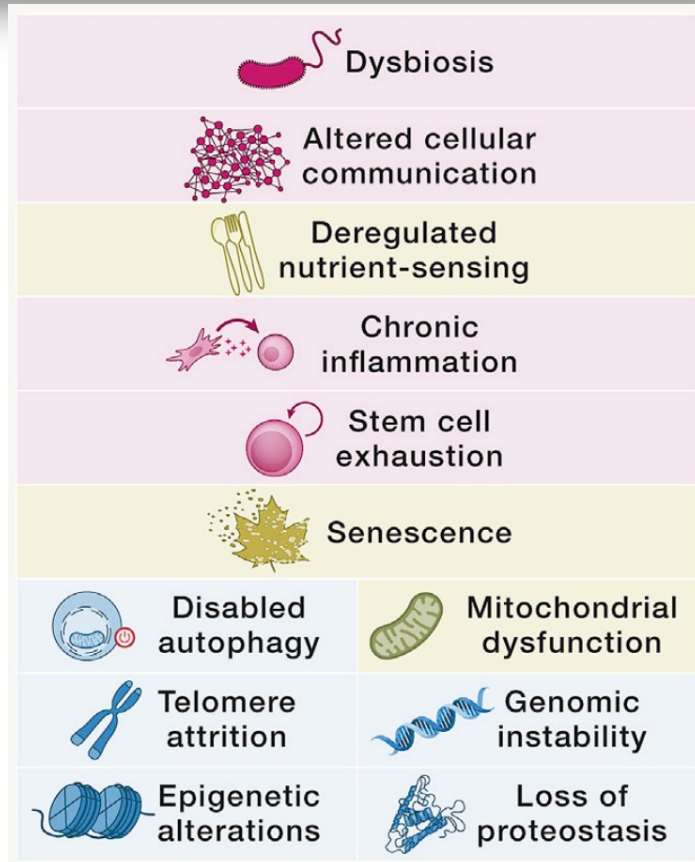
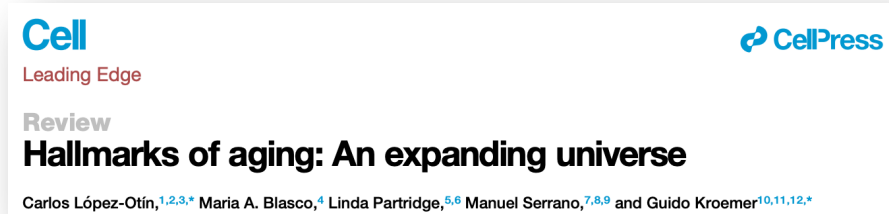
Which one is the most common?



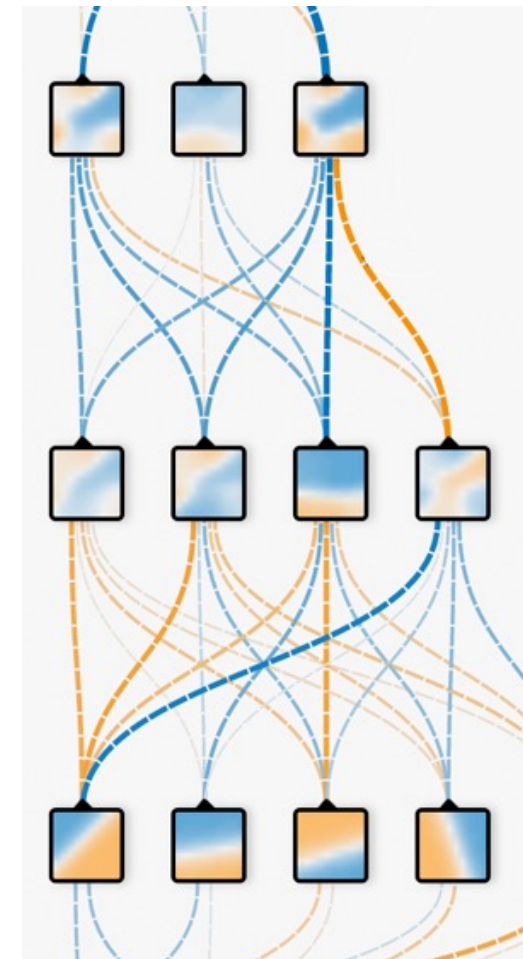
The most common cell type in the human brain is the **astrocyte**, a type of glial cell. Astrocytes outnumber neurons and play crucial roles in supporting neurons, maintaining the blood-brain barrier, and regulating the brain's environment.

Can we use this as a general prior for biology?

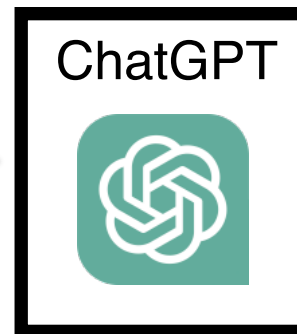
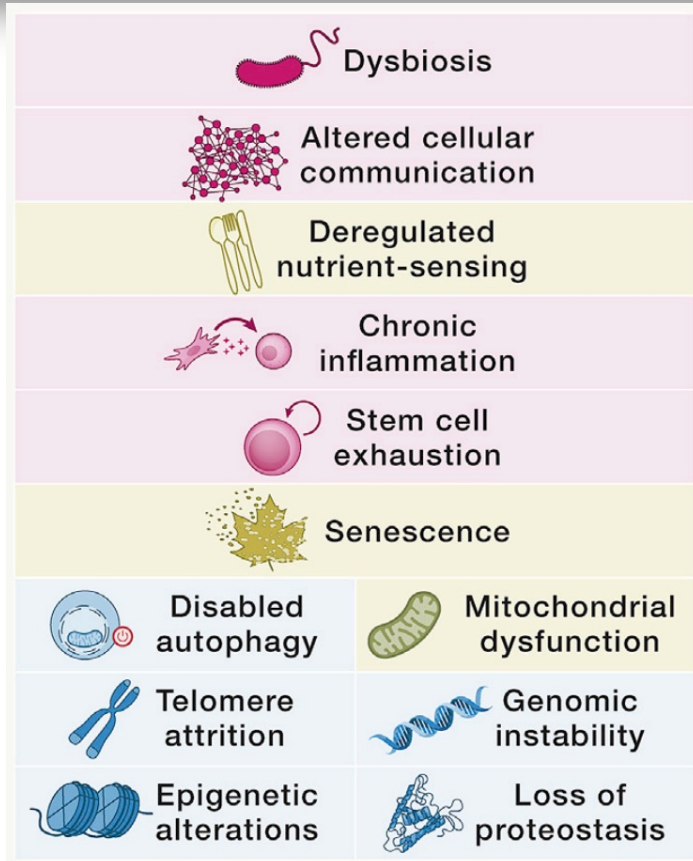
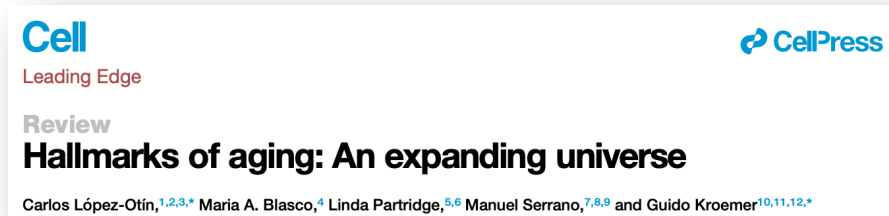
Currently, difficult to integrate unstructured knowledge into models



Deep learning model



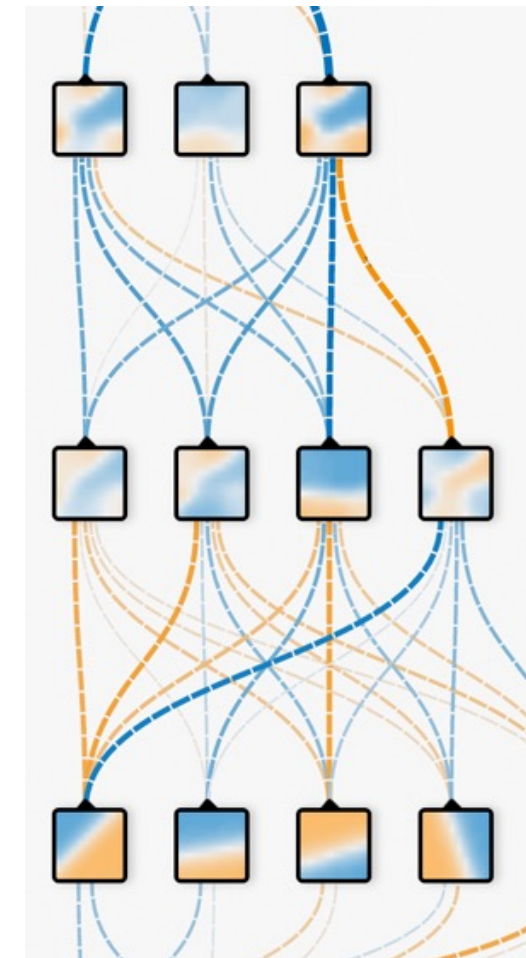
Language-guided biology: use LLM to generate biological priors for models



0.2
0.7
1.2
3.5
0.1
0.2
0.6



Deep learning model



What are possible representations for a gene?

Nucleotide sequences

```
TGCTCTTGTT GCCCAGGCTG
GAATGCAATG GCCTGATCTC
TGCTCACTGC AACCTCCACC
TCTCAGGTTC AAGCTATTCT
CCTGCCTCAG CCTCCCAAGT
AGCTGGGATT ACAGGTATGA
GCAACCATGC ATGGCTAATG
TTTGGATTTT TAGTAGAGAC
TAAATTTTAG TAGAGACTAGA
TTTTAGGTTT CACCATGTTG
GTCAGGTTGG TCTGGAACCC
TGACCTCAGG TGATCAGCCC
```

Ji et al. (2021); Poli et
al. (2023); ...

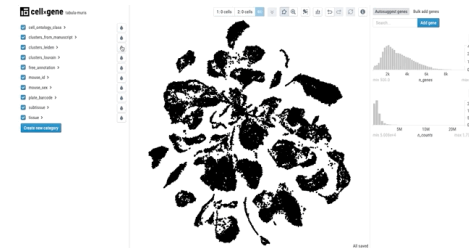
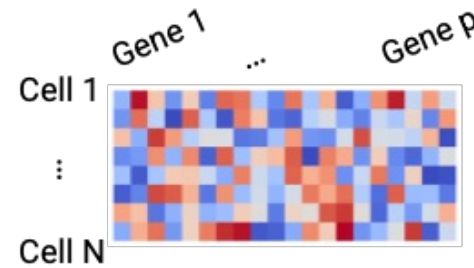
What are possible representations for a gene?

Nucleotide sequences

```
TGCTCTTGTT GCCCAGGCTG
GAATGCAATG GCCTGATCTC
TGCTCACTGC AACCTCCACC
TCTCAGGTTC AAGCTATTCT
CCTGCCTCAG CCTCCCAAGT
AGCTGGGATT ACAGGTATGA
GCAACCATGC ATGGCTAATG
TTTGGATTTT TAGTAGAGAC
TAAATTTT TAGAGACTAGA
TTTTAGGTTT CACCATGTTG
GTCAGGTTGG TCTGGAACCC
TGACCTCAGG TGATCAGCCC
```

Ji et al. (2021); Poli et al. (2023); ...

Gene expression data



Theodoris et al. (2023); Cui et al. (2023); ...

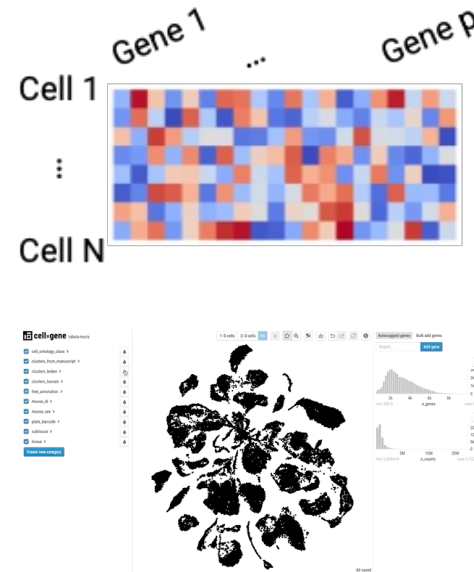
What are possible representations for a gene?

Nucleotide sequences

```
TGCTCTTGTT GCCCAGGCTG
GAATGCAATG GCCTGATCTC
TGCTCACTGC AACCTCCACC
TCTCAGGTTC AAGCTATTCT
CCTGCCTCAG CCTCCCAAGT
AGCTGGGATT ACAGGTATGA
GCAACCATGC ATGGCTAATG
TTTGGATTTT TAGTAGAGAC
TAAATTTTAG TAGAGACTAGA
TTTTAGGTTT CACCATGTTG
GTCAGGTTGG TCTGGAACCC
TGACCTCAGG TGATCAGCCC
```

Ji et al. (2021); Poli et al. (2023); ...

Gene expression data



Theodoris et al. (2023); Cui et al. (2023); ...

Text descriptions for a gene 📖

CD24 (CD24 Molecule) is a **Protein Coding gene**. Diseases associated with CD24 include Erythroleukemia and Multiple Sclerosis. Among its related pathways are **Nervous system development and L1CAM interactions**. Gene Ontology (GO) annotations related to this gene include protein kinase binding and carbohydrate binding.

Our proposal: leverage LLMs directly

GenePT: natural language embeddings of genes and cells

Text descriptions for a gene 📖

CD24 (CD24 Molecule) is a **Protein Coding gene**. Diseases associated with CD24 include Erythroleukemia and Multiple Sclerosis. Among its related pathways are **Nervous system development and L1CAM interactions**. Gene Ontology (GO) annotations related to this gene include protein kinase binding and carbohydrate binding.

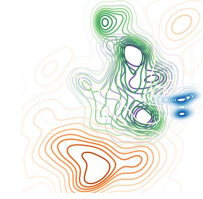
ChatGPT



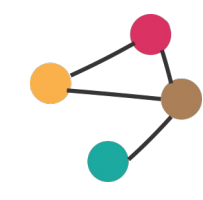
0.5
-0.8
0.4
0.7
0.1
-0.2
0.3

GenePT

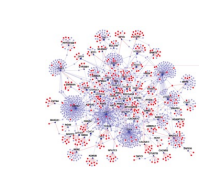
LLM embeds each gene's text description into a vector.



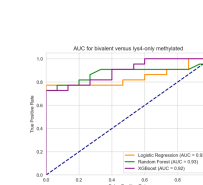
Gene clustering



Gene-gene interaction

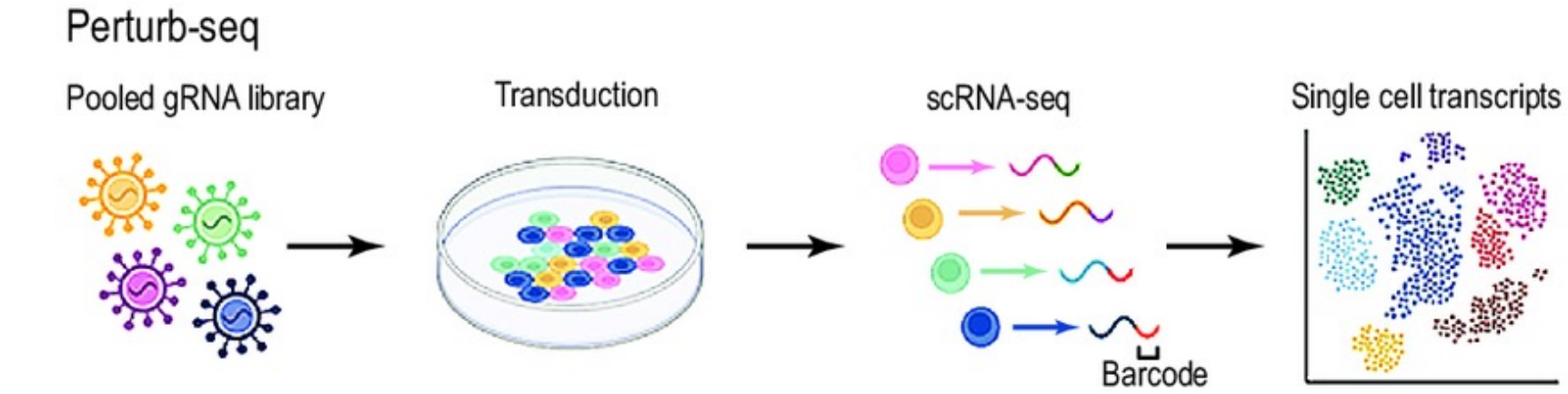


Protein network

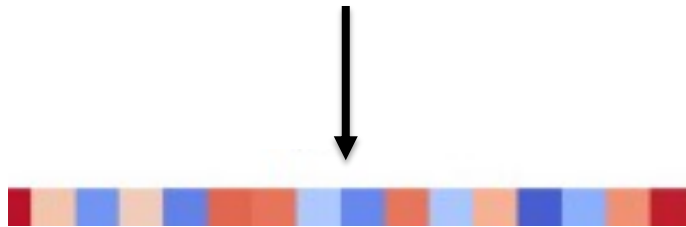


Function prediction

GenePT embeddings predict gene perturbations



$f(\text{embed}(\text{perturbed gene}))$



Predict expression of other genes

f = linear regression or MLP

Train on subset of perturbed genes
and test on held-out genes

GenePT embeddings predict gene perturbations

$f(\text{embed}(\text{perturbed gene}))$



Predict expression of other genes

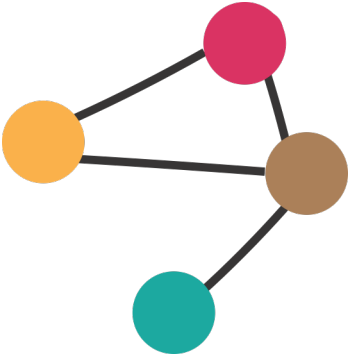
f = linear regression or MLP

Train on subset of perturbed genes
and test on held-out genes

	Norman CRISPRi	Replogle K562	Replogle RPE1	Adamson CRISPRi	Dixit CRISPRi	Wessels Cas13d
GenePT	0.85	0.51	0.65	0.79	0.79	0.76
scGPT	0.79	0.44	0.63	0.76	0.29	0.75
Geneformer	0.71	0.42	0.63	0.76	0.29	0.73
GEARS	0.63	0.38	0.52	0.76	--	--

Pearson corr w/ Perturb-seq

GenePT-derived features encode gene-gene interactions

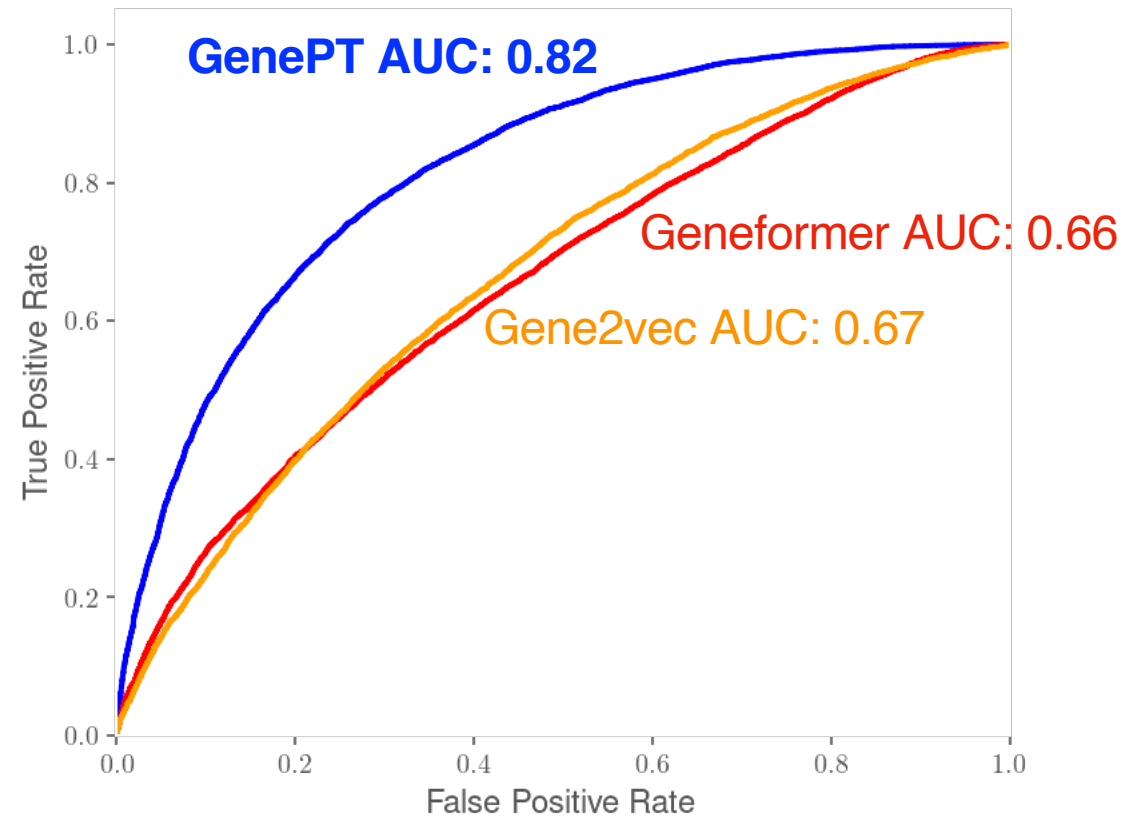


Gene-gene interactions from Du et al. (2019)

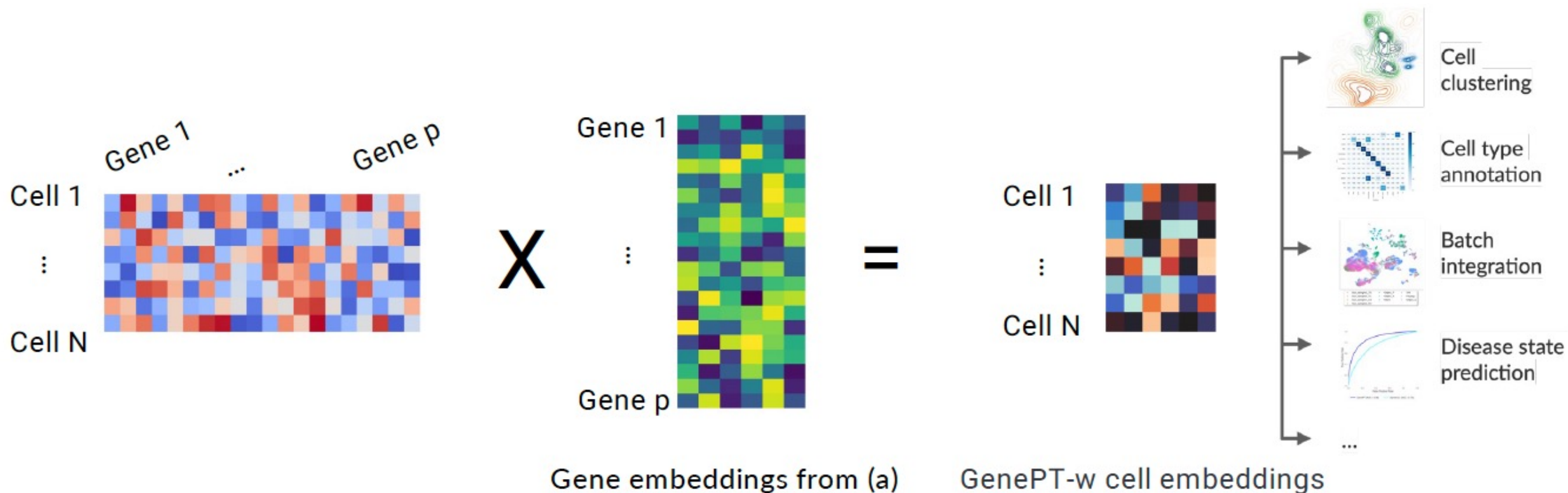
- predict pairs of genes known to share functions
- state-of-the-art AUC: 0.77

Gene-gene interactions not contained in text used to generate GenePT embeddings.

$$f_{Random\ Forest}(\Phi_{GenePT}(gene\ 1), \Phi_{GenePT}(gene\ 2))$$



GenePT for cell embeddings



GenePT better at annotating cells than directly asking ChatGPT

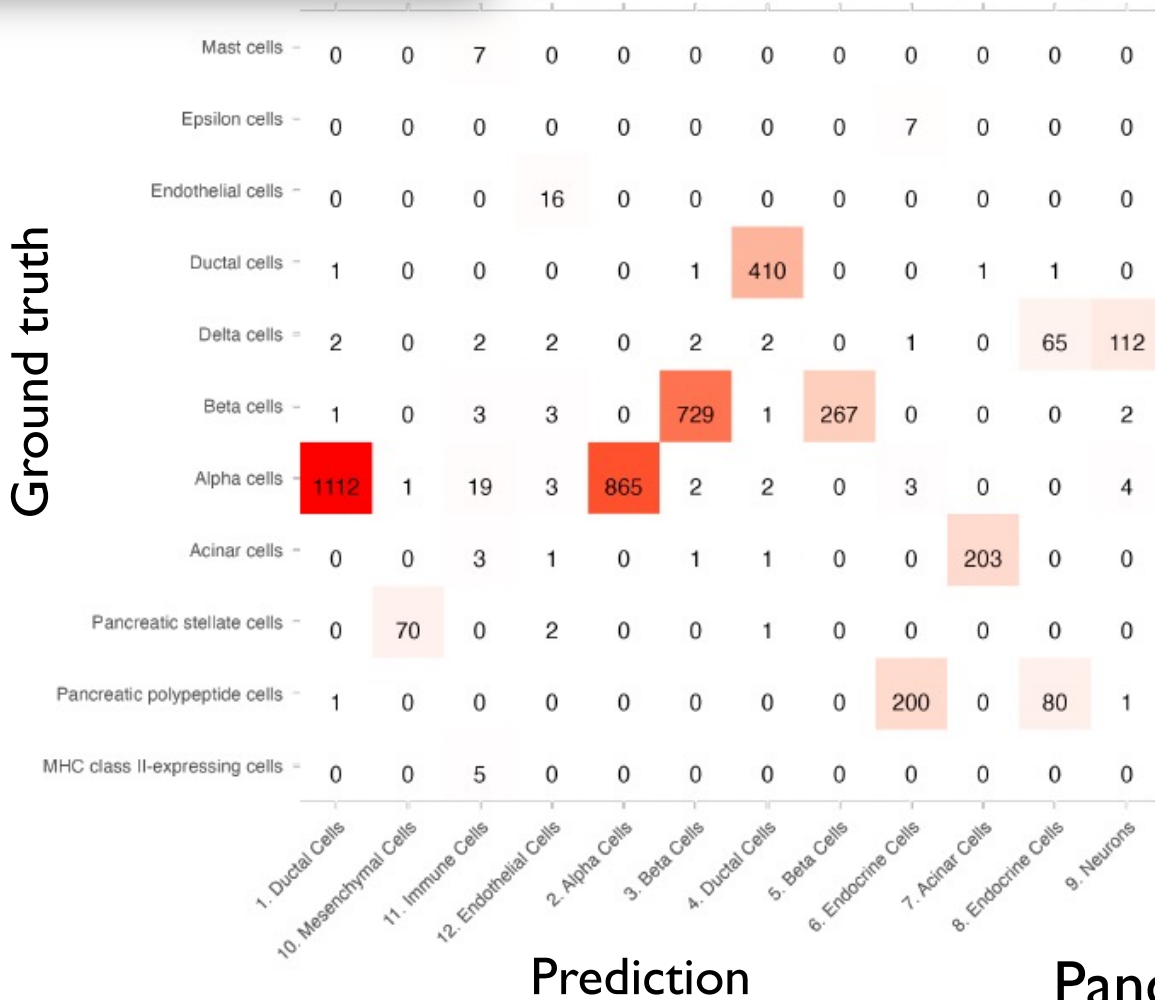
nature methods

Brief Communication

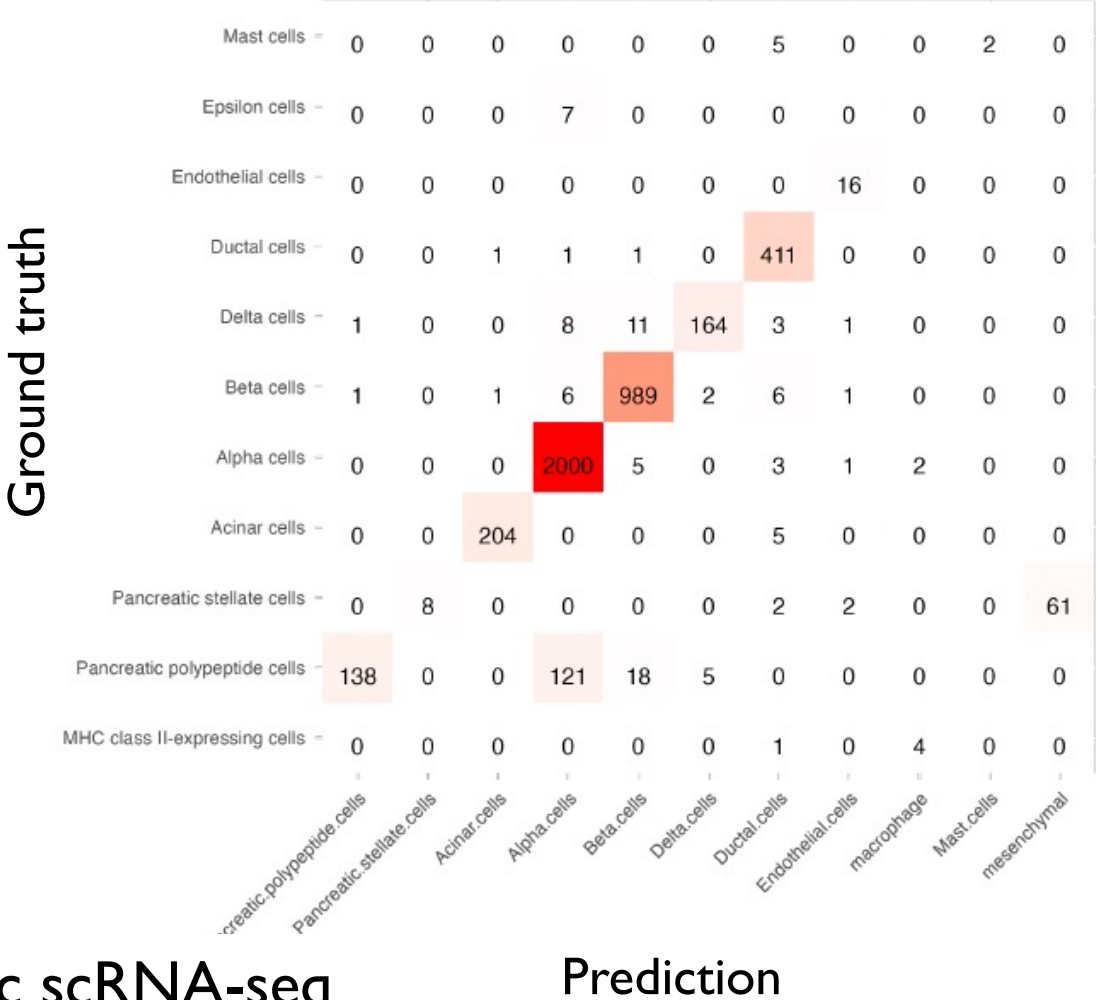
Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis

<https://doi.org/10.1038/s41592-024-02235-4>

Direct query GPT-4



GenePT



Pancreatic scRNA-seq

GenePT better at annotating cells than directly asking ChatGPT

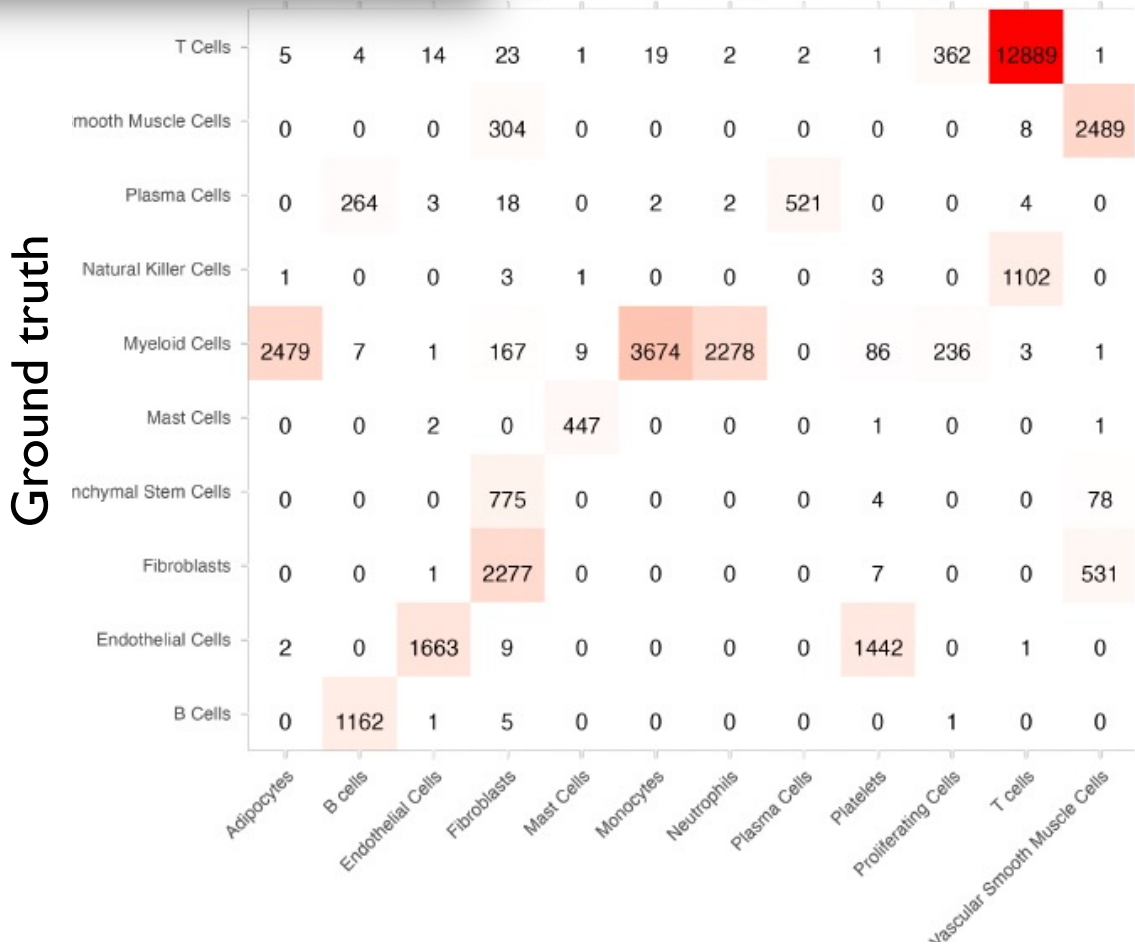
nature methods

Brief Communication

Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis

<https://doi.org/10.1038/s41592-024-02235-4>

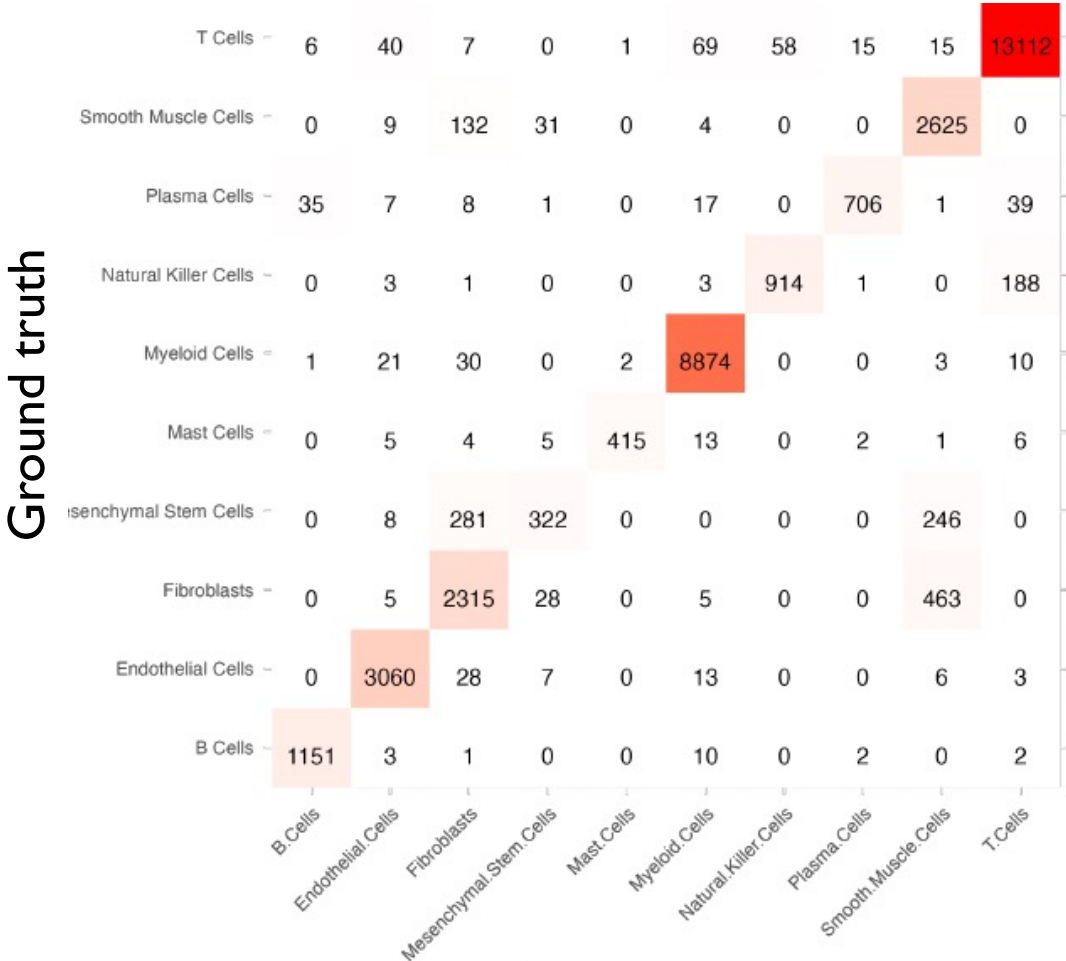
Direct query GPT-4



Prediction

Artery scRNA-seq

GenePT



Prediction

GenePT cell embeddings capture disease cells

