

# How to formulate computational biology problem to an NLP problem

Sheng Wang  
University of Washington

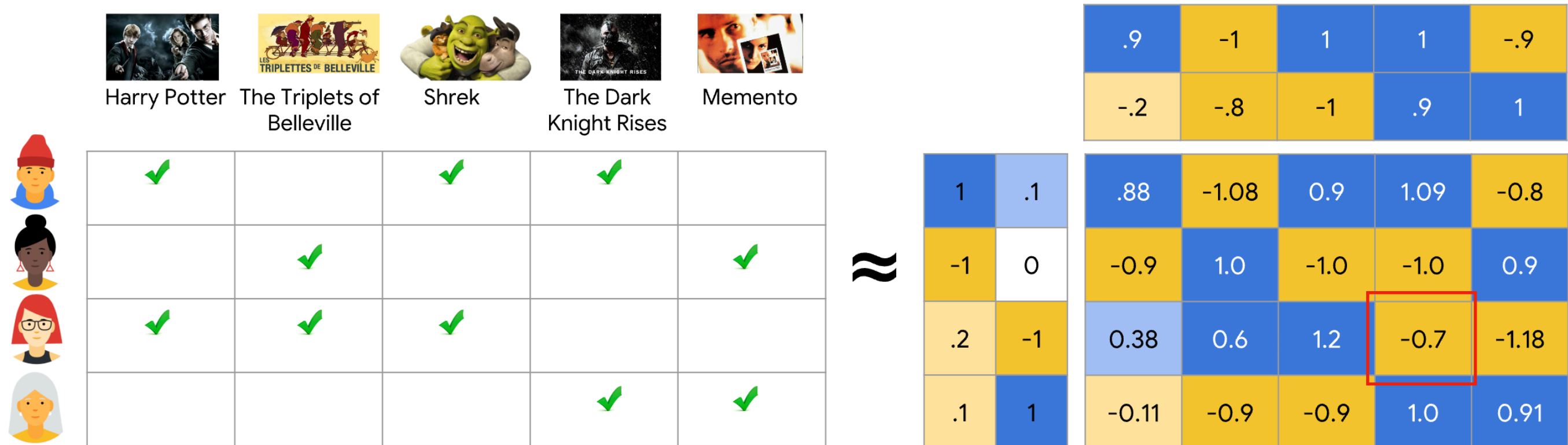
# Why do we want to formulate comp bio tasks to NLP problems

The diagram shows the equation for Bayes' theorem:  $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$ . Above the equation, three labels are positioned: 'Posterior' above  $P(A|B)$ , 'Likelihood' above  $P(B|A)$ , and 'Prior' above  $P(A)$ . Downward-pointing arrows connect each label to its corresponding term in the equation.

$$\begin{array}{ccc} \text{Posterior} & \text{Likelihood} & \text{Prior} \\ \downarrow & \downarrow & \downarrow \\ P(A|B) & = & \frac{P(B|A) * P(A)}{P(B)} \end{array}$$

- Prior:
  - NLP tasks can be solved by powerful language models (chatGPT, GPT4)
  - These models provide prior knowledge obtained from millions of scientific articles
- Likelihood:
  - Bio data is the likelihood. They often have small size.

# An example: movie recommendation system



Observed Only MF

1		1	1	
	1			1
1	1	1		
			1	1

Weighted MF

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

SVD

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

# More accurate recommendation if ratings are provided

Watch history

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

Rating

5	0	1	4	0
0	2	0	0	2
1	2	5	0	0
0	0	0	3	5

- Decompose rating matrix can more accurately recommend movies because ratings are more informative than watch history.





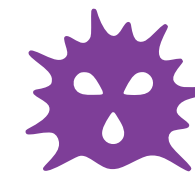
Cold



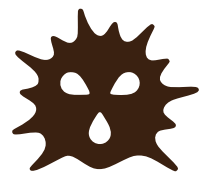
Flu



SARS



MERS



COVID-19

Less  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	N	Y	Y	Y	Y
Difficulty breathing?	N	N	Y	maybe	maybe
Runny nose?	Y	Y	maybe	N	maybe

More  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	5%	90%	95%	90%	95%
Difficulty breathing?	3%	10%	95%	50%	45%
Runny nose?	90%	90%	50%	2%	55%

- More fine-grained, informative data leads to better downstream performance



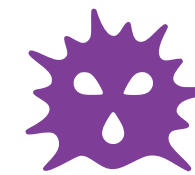
Cold



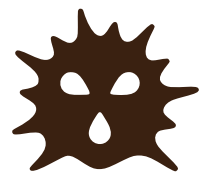
Flu



SARS



MERS



COVID-19

Less  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	N	Y	Y	Y	Y
Difficulty breathing?	N	N	Y	maybe	maybe
Runny nose?	Y	Y	maybe	N	maybe

More  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	5%	90%	95%	90%	95%
Difficulty breathing?	3%	10%	95%	50%	45%
Runny nose?	90%	90%	50%	2%	55%

Can we get more fine-grained information?

# NLP (language models) for providing fine-grained information



Is fever a symptom of SARS



Yes, fever is a common symptom of SARS (Severe Acute Respiratory Syndrome). SARS is a viral respiratory illness that emerged in 2002 and spread to many countries, causing a global outbreak. According to the World Health Organization (WHO), fever is one of the main symptoms of SARS, along with cough, difficulty breathing, and other flu-like symptoms. If you are experiencing these symptoms or are concerned about SARS, it is important to seek medical attention promptly.



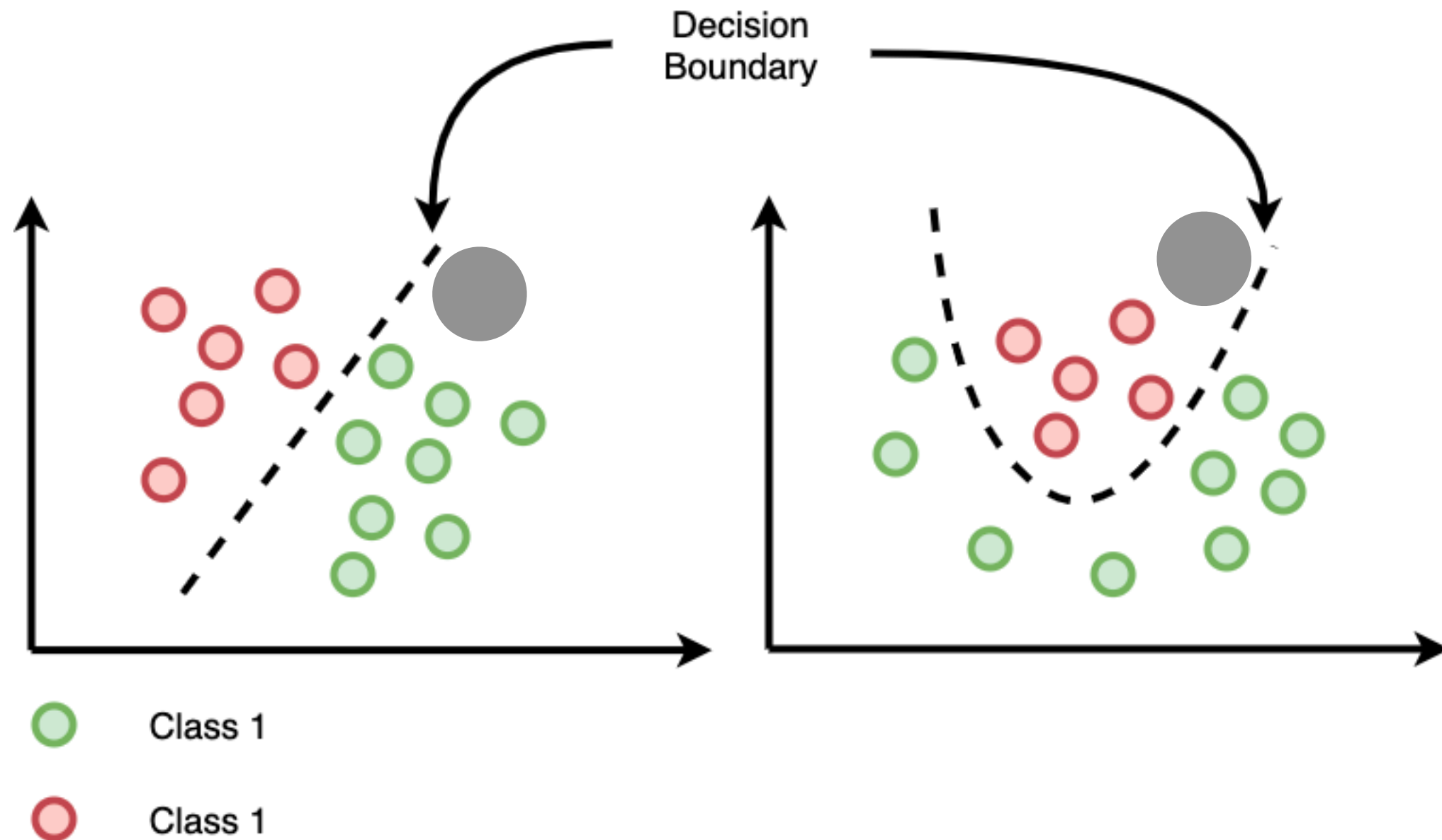
Is fever a symptom of COVID



Yes, fever is a common symptom of COVID-19. COVID-19 is a viral respiratory illness caused by the SARS-CoV-2 virus, and fever is one of the most frequently reported symptoms. According to the Centers for Disease Control and Prevention (CDC), a fever is defined as a temperature of 100.4°F (38°C) or higher. However, it's important to note that not everyone with COVID-19 will have a fever, and some people may have a fever that comes and goes. Other common symptoms of COVID-19 include cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. If you are experiencing any of these symptoms or are concerned about COVID-19, it is important to seek medical attention promptly.

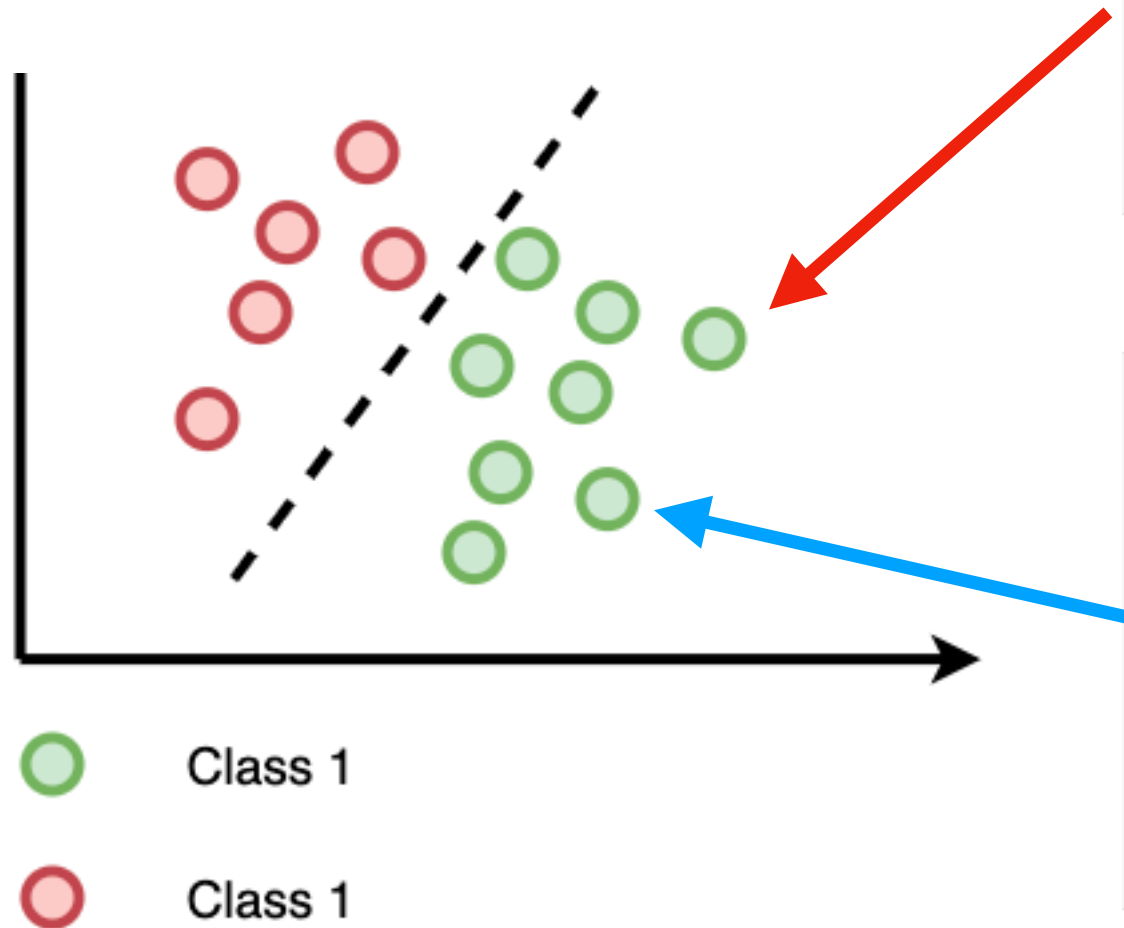


# Supervised classification



- Features: determine where the data point locates
- Class label: determine the “color” of the data points
- Goal: find a decision boundary using features + class label

# Training of a text classifier



Is fever a symptom of SARS



Yes, fever is a common symptom of SARS (Severe Acute Respiratory Syndrome). SARS is a viral respiratory illness that emerged in 2002 and spread to many countries, causing a global outbreak. According to the World Health Organization (WHO), fever is one of the main symptoms of SARS, along with cough, difficulty breathing, and other flu-like symptoms. If you are experiencing these symptoms or are concerned about SARS, it is important to seek medical attention promptly.



Is fever a symptom of COVID



Yes, fever is a common symptom of COVID-19. COVID-19 is a viral respiratory illness caused by the SARS-CoV-2 virus, and fever is one of the most frequently reported symptoms. According to the Centers for Disease Control and Prevention (CDC), a fever is defined as a temperature of 100.4°F (38°C) or higher. However, it's important to note that not everyone with COVID-19 will have a fever, and some people may have a fever that comes and goes. Other common symptoms of COVID-19 include cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. If you are experiencing any of these symptoms or are concerned about COVID-19, it is important to seek medical attention promptly.



- Every text paragraph (a pair of disease and symptom) is a data point



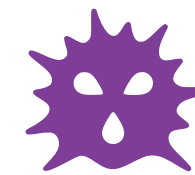
Cold



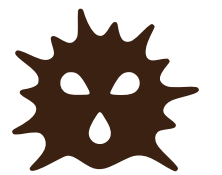
Flu



SARS



MERS



COVID-19

Less  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	N	Y	Y	Y	Y
Difficulty breathing?	N	N	Y	maybe	maybe
Runny nose?	Y	Y	maybe	N	maybe

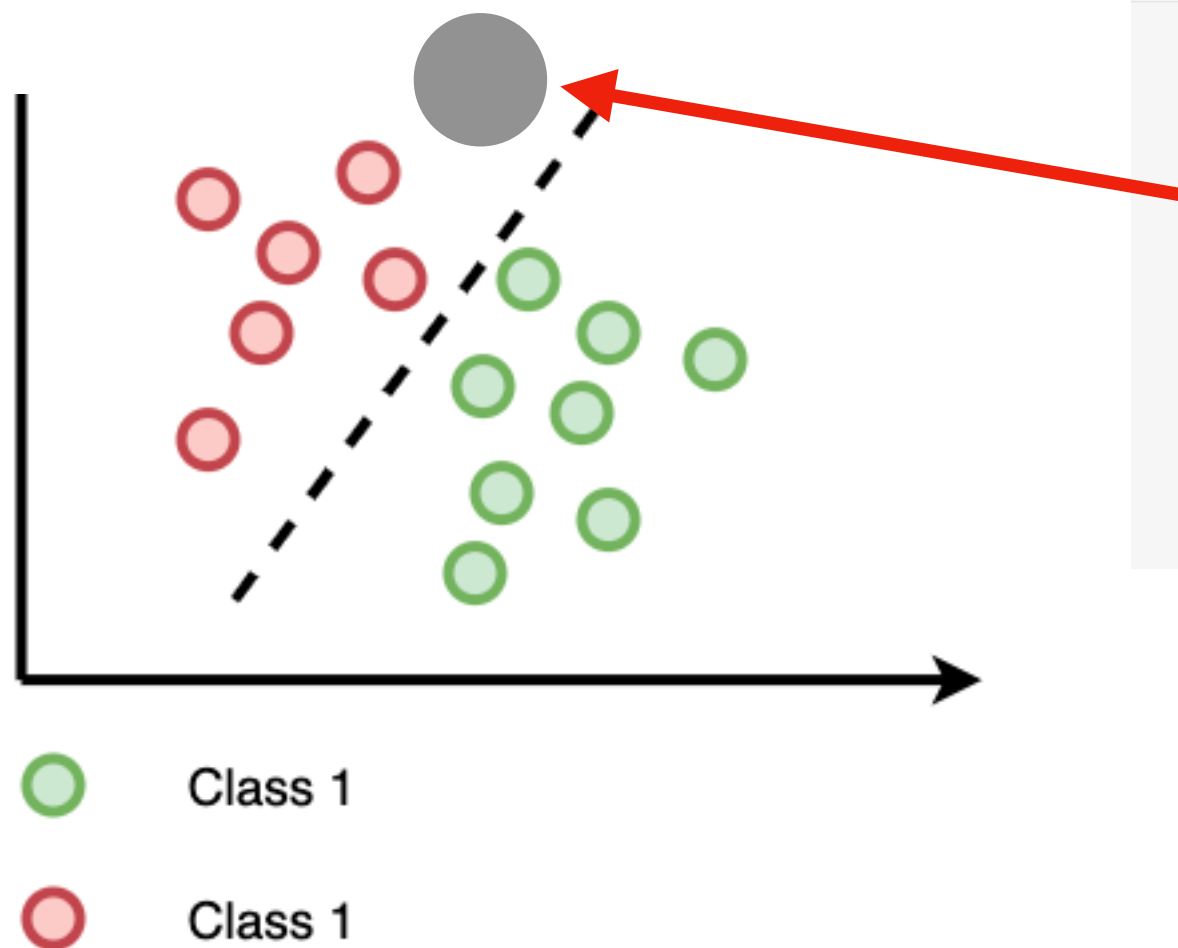
More  
informative

	Cold	Flu	SARS	MERS	COVID-19
Fever?	5%	90%	95%	90%	95%
Difficulty breathing?	3%	10%	95%	50%	45%
Runny nose?	90%	90%	50%	2%	55%

Can we get more fine-grained information?

# Predict if running nose is a symptom of SARS?

- Step 1: obtain the feature by asking a question to chatGPT
- Step 2: use this feature to obtain the label based on the decision boundary



If running nose a symptom of SARS



Yes, a runny nose can be a symptom of SARS (Severe Acute Respiratory Syndrome), although it is not one of the main symptoms. According to the World Health Organization (WHO), the primary symptoms of SARS are fever, cough, and difficulty breathing. However, some people with SARS may also experience other symptoms such as headache, muscle aches, and a runny nose.

It's worth noting that a runny nose is a very common symptom of many other respiratory infections, including the common cold and the flu, and is not a specific indicator of SARS. If you are experiencing symptoms of a respiratory infection, it's important to seek medical attention to determine the cause and appropriate treatment.

# Recap and discussion

- Original problem: we observe the associations between some symptoms and some diseases. We want to identify the missing ones.
- NLP formulation: use chatGPT to obtain a text sentence between every pair of symptom and disease and then classify sentence.
- Why it works? chatGPT might provide more informative features from millions of scientific papers. (This is a hypothesis)



# What other comp bio problems can be formulated to NLP tasks?

- If we have feature names, we can use chatGPT to append the feature
- If we have class names, we can use chatGPT to convert class labels to class feature vectors.
- If we have both, we can use chatGPT to obtain more labelled data.

# What other comp bio problems can be formulated to NLP tasks?

	TP53	KMT2A	EGFR	MAPK1	ERBB2
Breast cancer	1	0	0	1	1
Lung cancer	0	0	1	0	0
Breast cancer	1	1	0	0	0
?	1	1	1	0	0



what is the role of TP53 in cancer



TP53 (also known as p53) is a tumor suppressor gene that plays a critical role in preventing the development and progression of cancer. When functioning properly, TP53 helps to regulate cell division and promote the death of cells that are damaged or abnormal, including cells that have the potential to become cancerous.



Mutations in the TP53 gene can disrupt its normal function and lead to the development of cancer. In fact, TP53 mutations are among the most common genetic alterations found in human cancers. When TP53 is mutated, it can no longer properly regulate cell division or promote cell death, allowing abnormal cells to continue to grow and divide uncontrollably. This can lead to the formation of tumors and the spread of cancer to other parts of the body.

Given its critical role in preventing cancer, TP53 is a major focus of cancer research. Researchers are working to better understand the mechanisms underlying TP53 function, as well as to develop new therapies that can target TP53 mutations and restore its normal function in cancer cells.

# Find problems with the same problem setting



Harry Potter



The Triplets of Belleville



Shrek



The Dark Knight Rises



Memento

✓		✓	✓	
	✓			✓
✓	✓	✓		
			✓	✓

≈

1	.1
-1	0
.2	-1
.1	1

.9	-1	1	1	-.9
-.2	-.8	-1	.9	1

.88	-1.08	0.9	1.09	-0.8
-0.9	1.0	-1.0	-1.0	0.9
0.38	0.6	1.2	-0.7	-1.18
-0.11	-0.9	-0.9	1.0	0.91

## THE BAT AND THE CAT... FALL FLAT!

### The Dark Knight Rises

Rating: PG-13

Length: 2 hr. 45 min.

Genre: Drama, Action & Adventure.

Directed By: Christopher Nolan.

Written By: David S. Goyer, Christopher Nolan, Jonathan Nolan, Bob Kane.

Starring: Christian Bale, Tom Hardy, Anne Hathaway, Michael Caine, Gary Oldman, Morgan Freeman, Joseph Gordon-Levitt.

Four years after *The Dark Knight* knocked our socks off with the late Heath Ledger's dark portrayal of the Joker, Batman (Bale) is back on our screens. Complete with old favourites such as Alfred (Caine) and Commissioner Gordon (Oldman), *The Dark Knight* puts them alongside newbie's to the franchise like Hathaway as Selina/Catwoman and Gordon-Levitt as Blake. While Ledger's scene-stealing performance is no longer here, *The Dark Knight Rises* offers the audience an equally as disturbed character in the form of Bane (Hardy), a masked bad guy set on destroying Gotham City, eight years after Batman brought down the Joker.



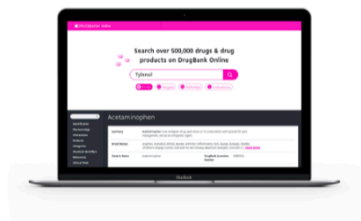
Content-based  
recommendation

# What other comp bio problems can be formulated to NLP tasks?

- If we have feature names, we can use chatGPT to append the feature
- If we have class names, we can use chatGPT to convert class labels to class feature vectors.
- If we have both, we can use chatGPT to obtain more labelled data.

# Annotation serves as basis for biological systems analysis

## Chemical Compound



### DrugBank Online at your fingertips

Search our knowledge base for drug interactions, pharmacology, chemical structures, targets, metabolism, & more.

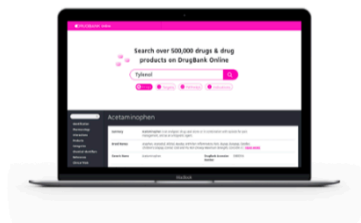
Download limited datasets, free for academic & non-commercial researchers.

[Search now >](#)

DrugBank

# Annotation serves as basis for biological systems analysis

## Chemical Compound



### DrugBank Online at your fingertips

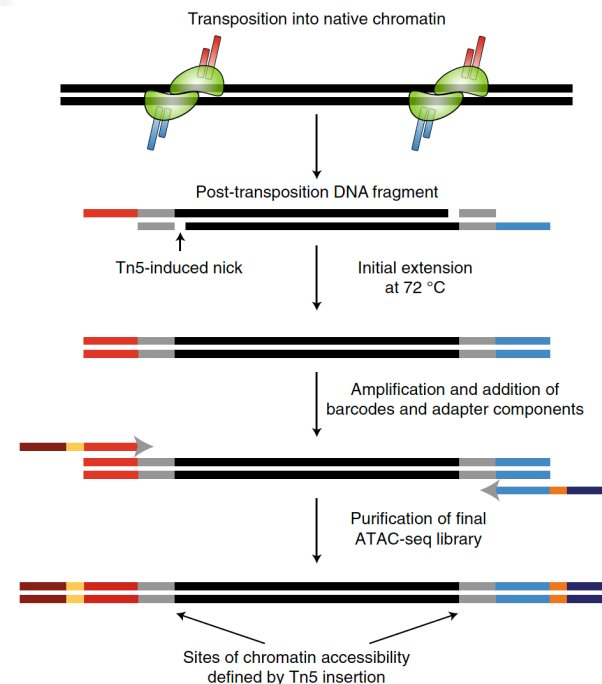
Search our knowledge base for drug interactions, pharmacology, chemical structures, targets, metabolism, & more.

Download limited datasets, free for academic & non-commercial researchers.

[Search now >](#)

DrugBank

## Transcription



Chromatin accessibility annotation  
(Nature Protocols 2022)

# Annotation serves as basis for biological systems analysis

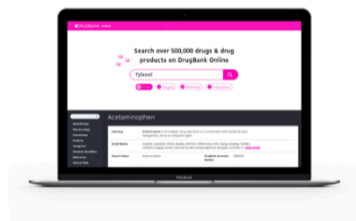
## Chemical Compound



## Transcription



## Histology



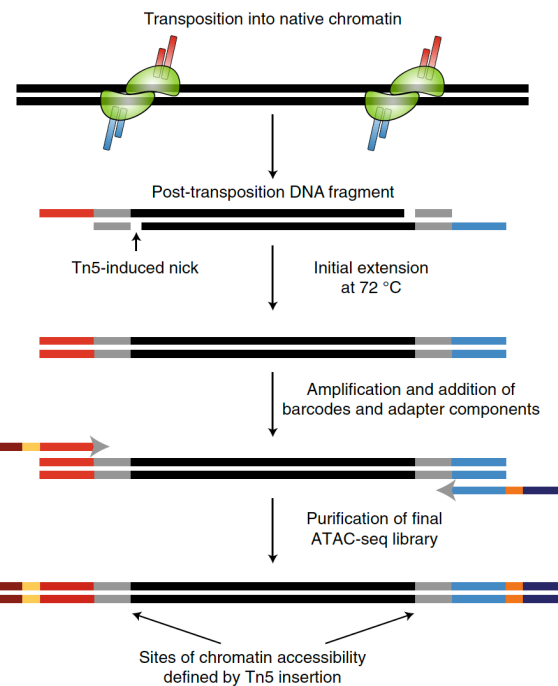
### DrugBank Online at your fingertips

Search our knowledge base for drug interactions, pharmacology, chemical structures, targets, metabolism, & more.

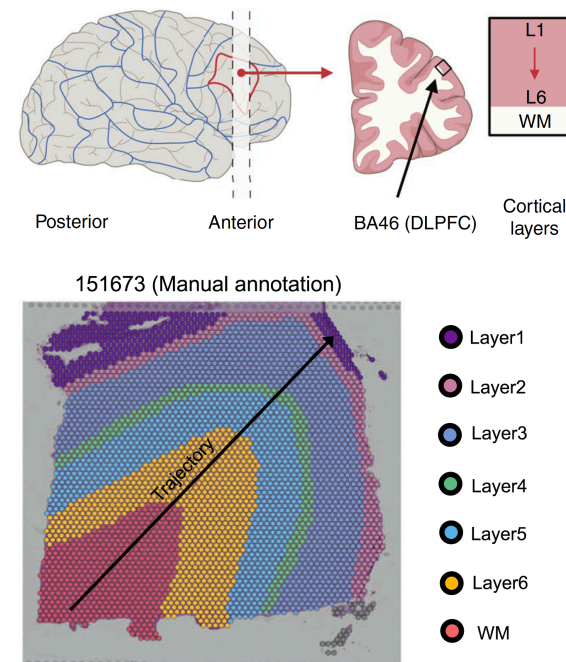
Download limited datasets, free for academic & non-commercial researchers.

[Search now >](#)

DrugBank



Chromatin accessibility annotation  
(Nature Protocols 2022)



Human cortex annotation  
(Nature Neuroscience 2021)



# Annotation serves as basis for biological systems analysis

Chemical Compound



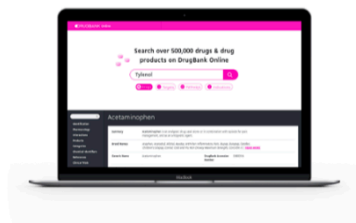
Transcription



Histology



Humans

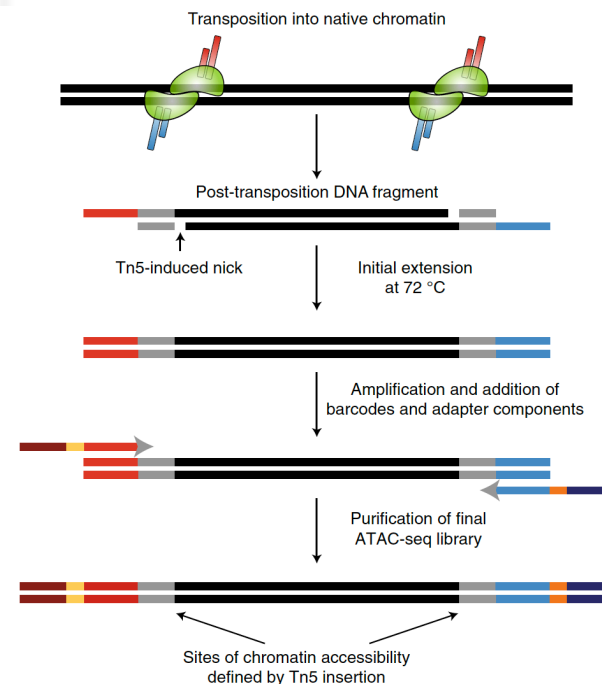


## DrugBank Online at your fingertips

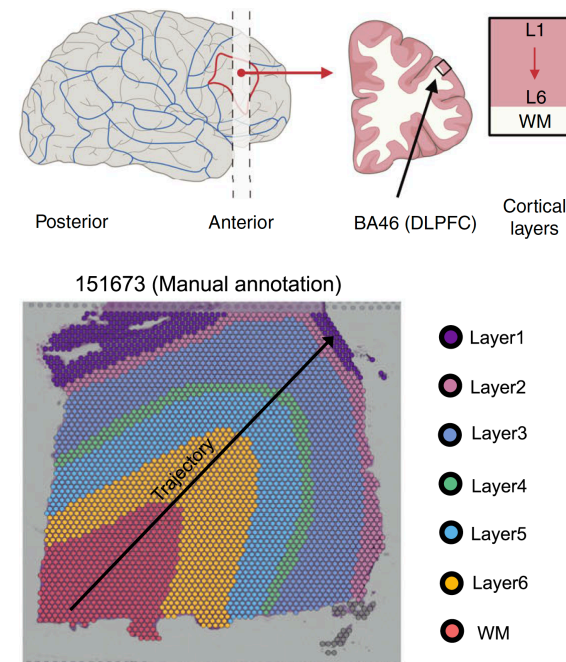
Search our knowledge base for drug interactions, pharmacology, chemical structures, targets, metabolism, & more.  
Download limited datasets, free for academic & non-commercial researchers.

[Search now >](#)

DrugBank



Chromatin accessibility annotation  
(Nature Protocols 2022)



Human cortex annotation  
(Nature Neuroscience 2021)



Electronic Health Records  
(Nature Outlook 2019)

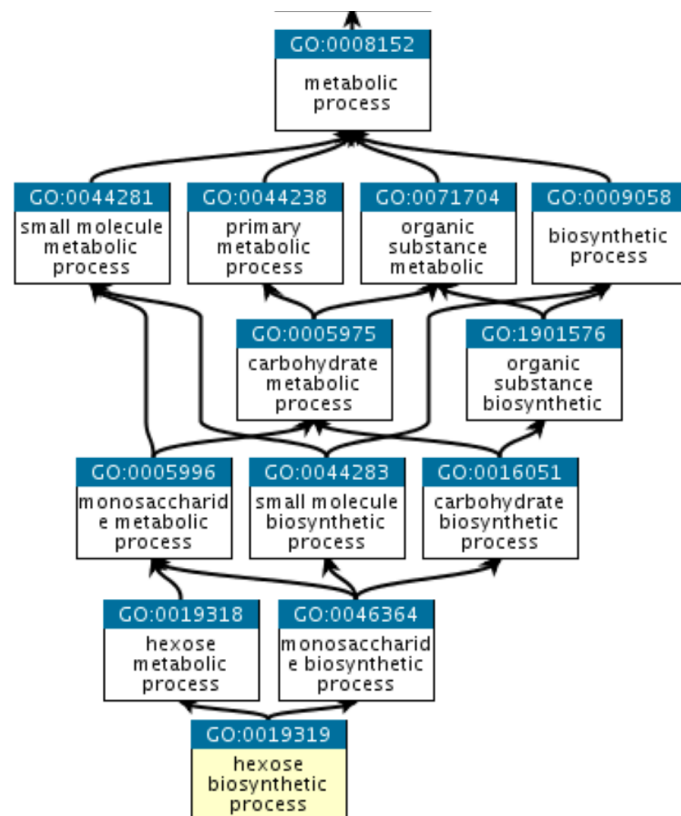
- Annotation is vital for biological systems analysis across multiple levels



# Annotation with pre-defined Controlled Vocabularies

## Gene Ontology

43,329 terms

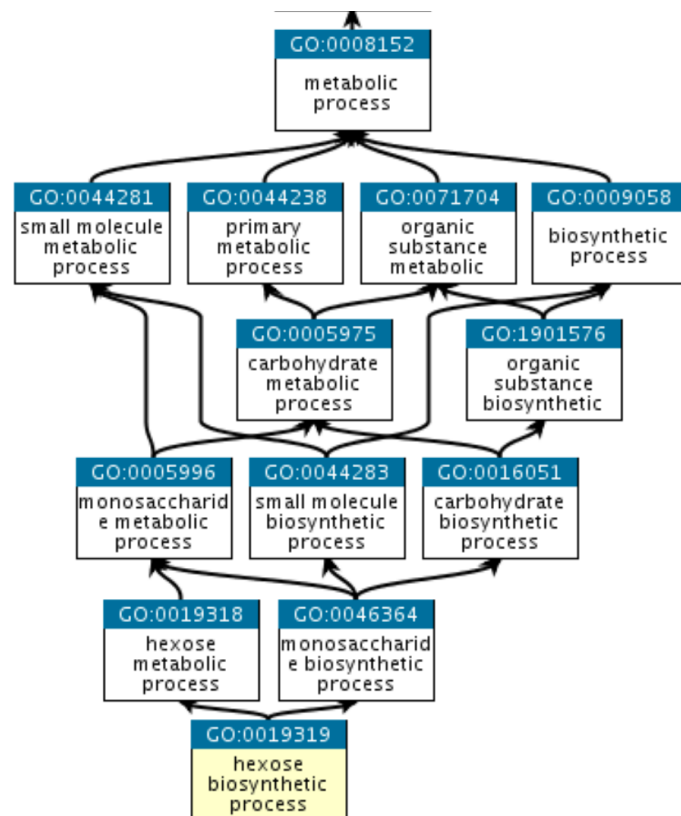


(Nature Genetics 2020)

# Annotation with pre-defined Controlled Vocabularies

## Gene Ontology

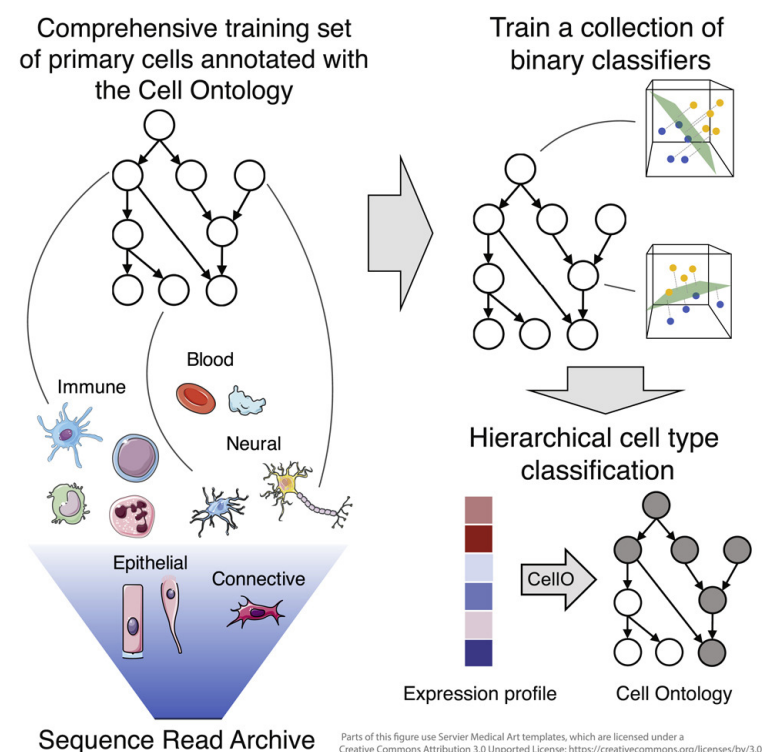
43,329 terms



(Nature Genetics 2020)

## Cell Ontology

2,311 terms

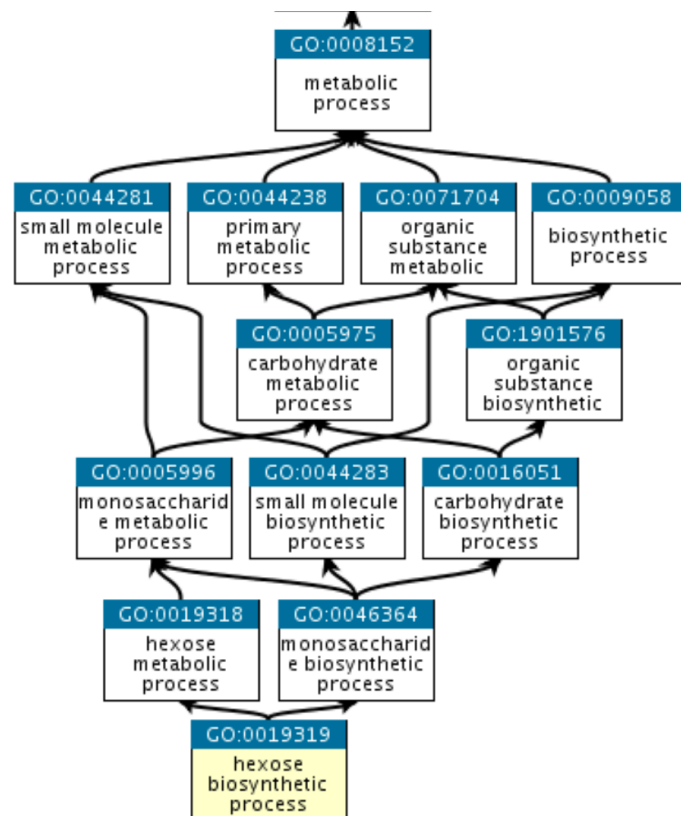


(iScience 2021,  
Genome Biology 2020)

# Annotation with pre-defined Controlled Vocabularies

## Gene Ontology

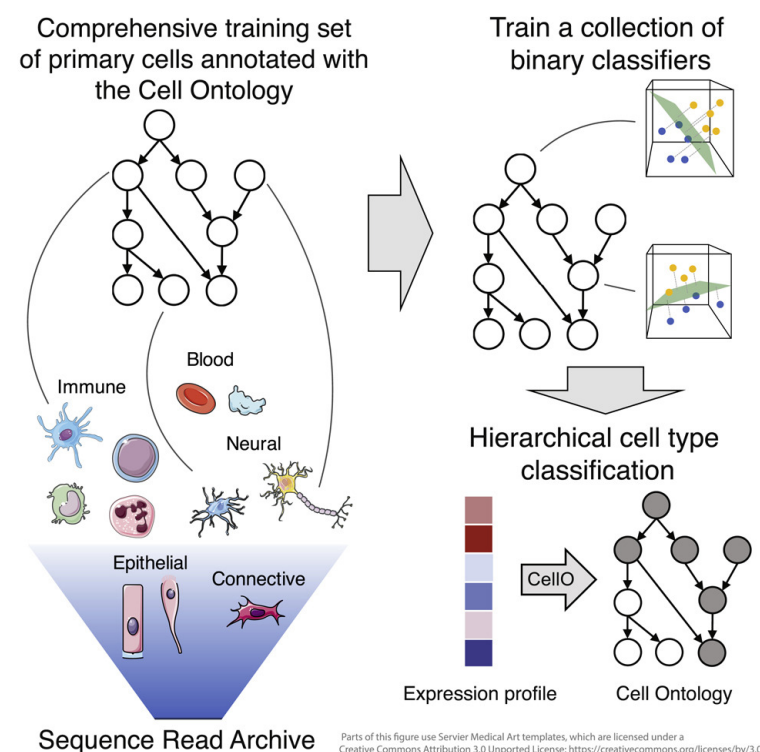
43,329 terms



(Nature Genetics 2020)

## Cell Ontology

2,311 terms



(iScience 2021,  
Genome Biology 2020)

## Enzyme Commission Ontology

8,067 terms

### EC numbers

Enzyme Commission

each enzyme has a 4-part numerical ID more than just a name, it gives you information about what the enzyme does & how

**EC 1.1.1.1**

**main class**

1. oxidoreductases
2. transferases
3. hydrolases
4. lyases
5. isomerases
6. ligases
7. translocases

**subclass**

often tells you often what type of compound or bond it acts on

**sub-subclass**

further classification

**serial number**

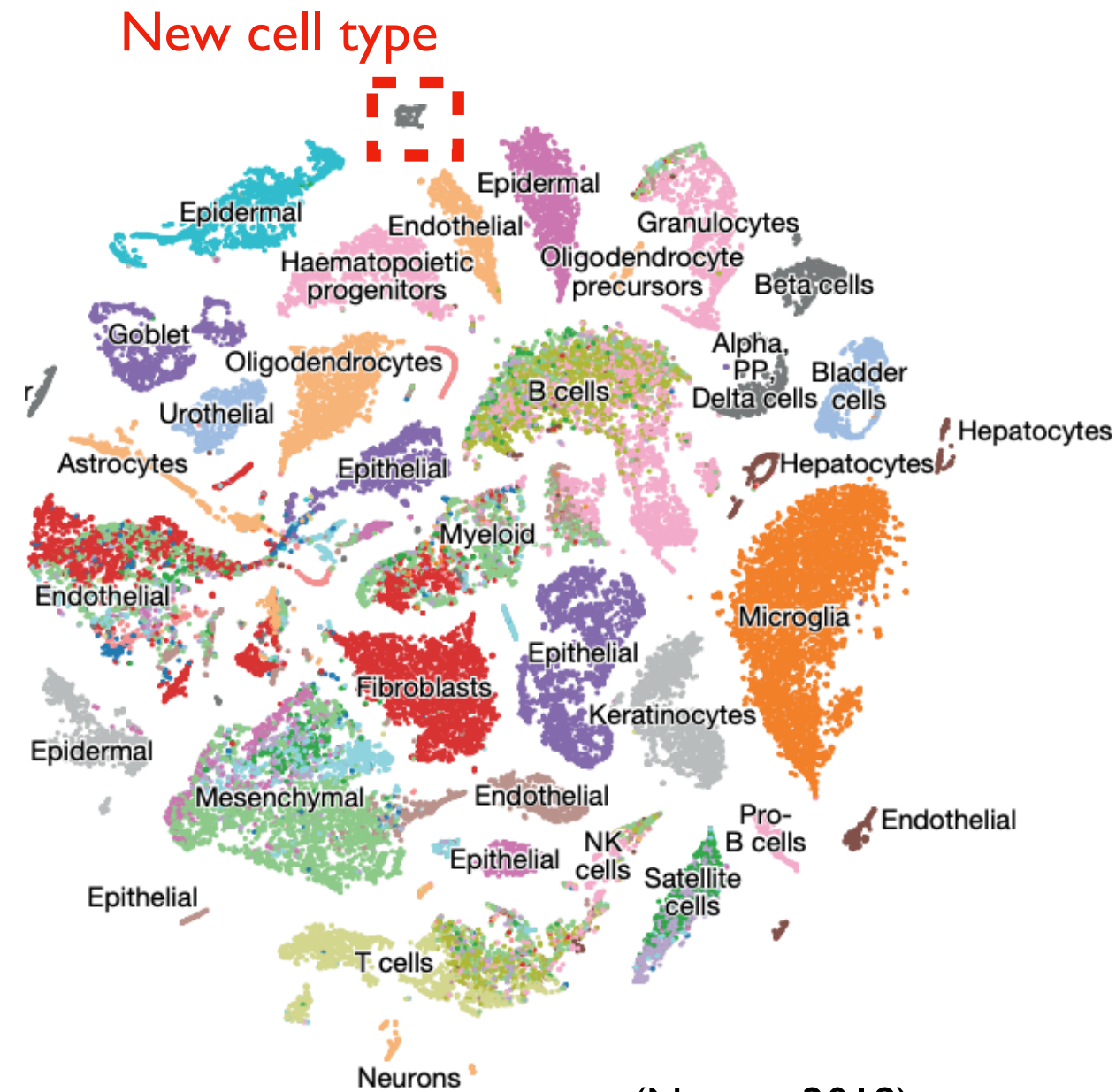
"just" makes it so each enzyme has its own unique EC number

(Genomics 2001,  
BMC Bioinformatics 2000)

- Current Annotation paradigm classifies instances into **Controlled Vocabularies**

# Limitations of using controlled vocabularies (CV)

- CVs cannot annotate new discoveries
  - New cell types are not in the cell type CVs

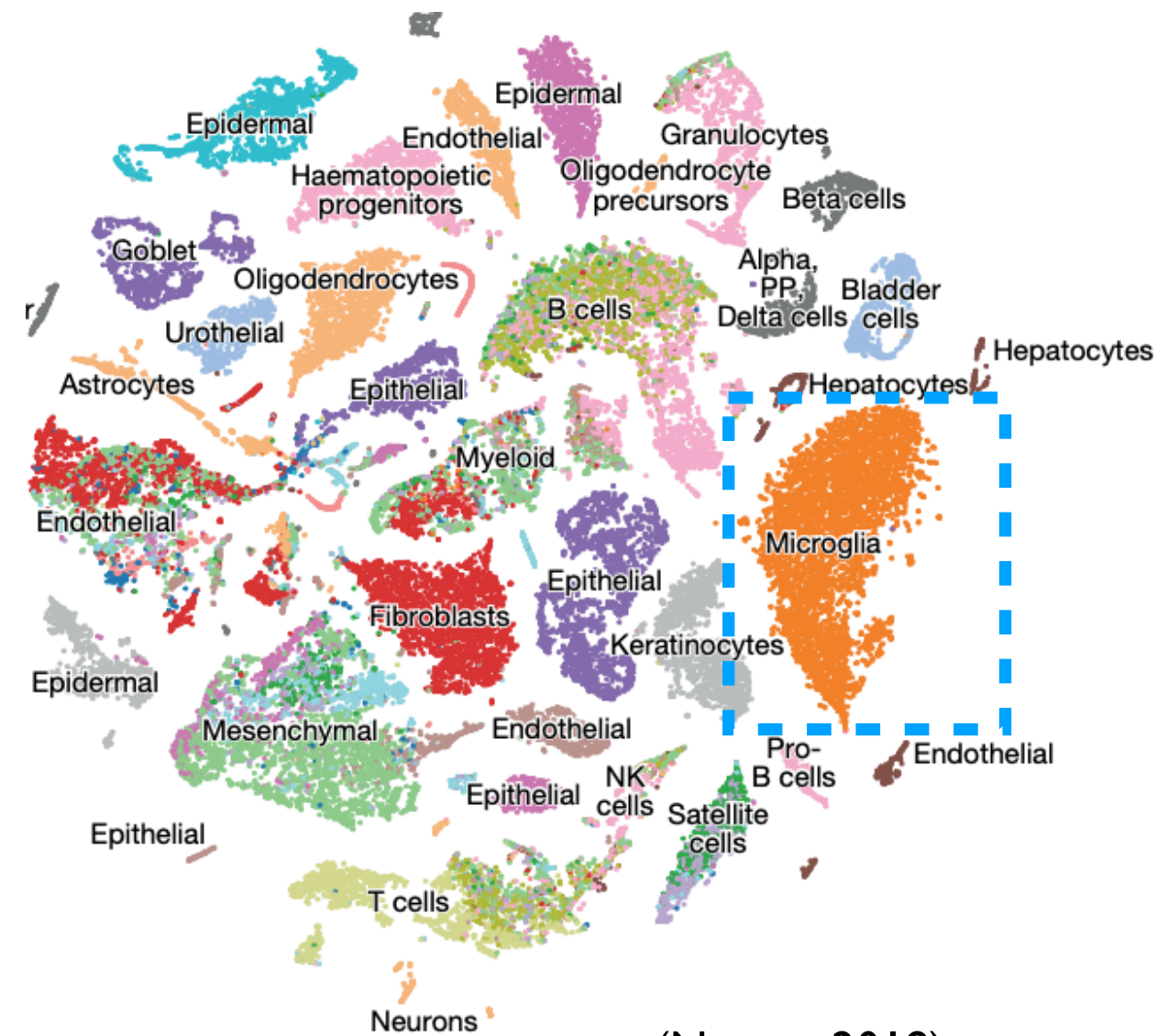


(Nature, 2018)



# Limitations of using controlled vocabularies (CV)

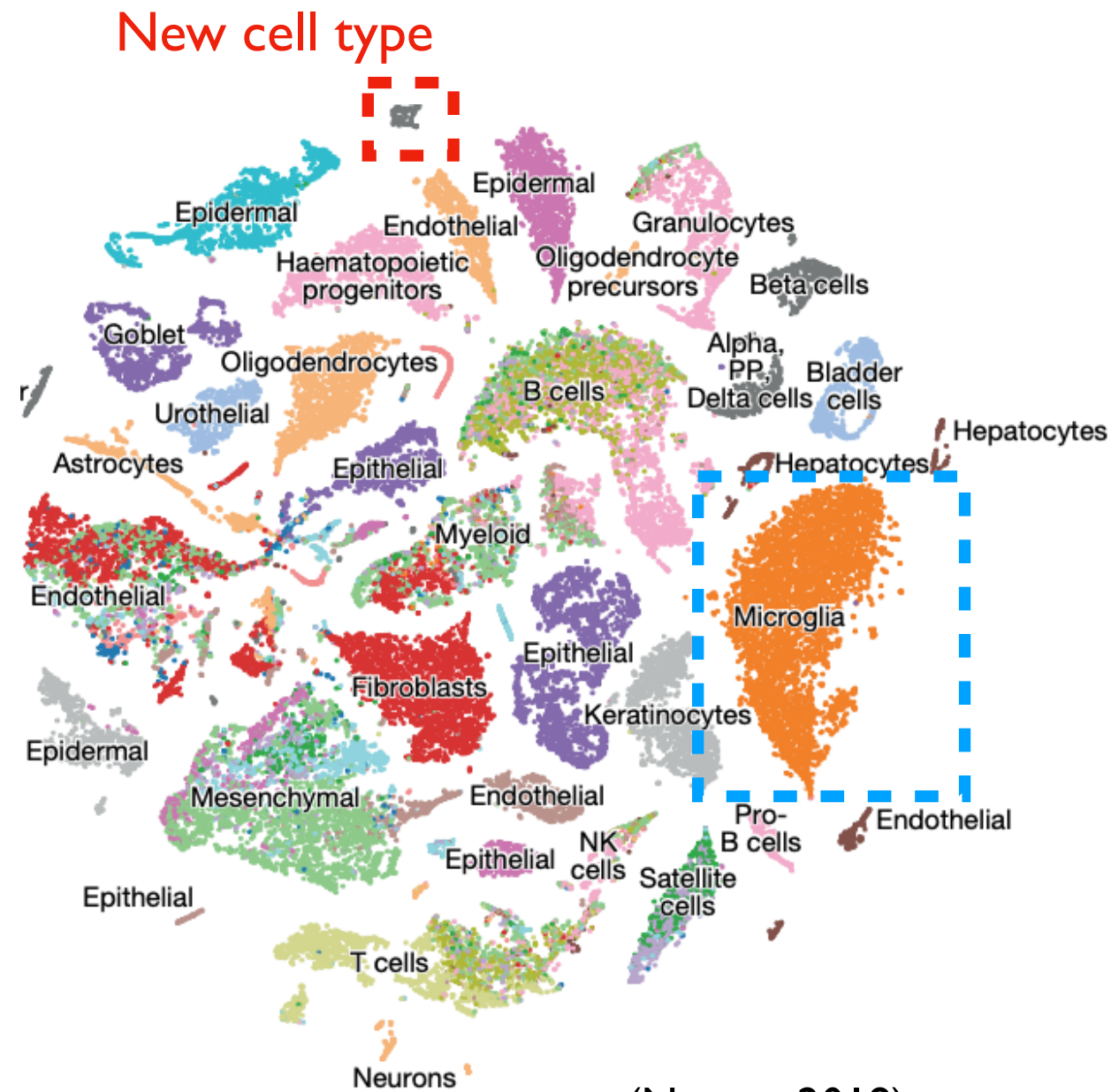
- **CVs cannot annotate new discoveries**
  - New cell types are not in the cell type CVs
- **CVs might not be precise**
  - We don't have enough CVs
  - One CV term fits many instances is too coarse



(Nature, 2018)

# Limitations of using controlled vocabularies (CV)

- **CVs cannot annotate new discoveries**
  - New cell types are not in the cell type CVs
- **CVs might not be precise**
  - We don't have enough CVs
  - One CV term fits many instances is too coarse
- **CVs might not be expressive**
  - Short phrase of one or two words might be ambiguous



(Nature, 2018)

These three limitations become worse in large datasets and new technology

# Benefits of using sentences to annotate **each** data point

- CVs cannot annotate new discoveries
  - **Generate sentences on-the-fly** to describe new findings
- CVs might not be precise
  - **The space of sentences is scaled to the space of new biological datasets**
- CVs might not be expressive
  - **Sentences are much longer and expressive than phrase**

**Our solution: Translate biological instances into sentences**

# BioTranslator: A machine translation framework for translating biological data to human language

**Input:** A biological data instance

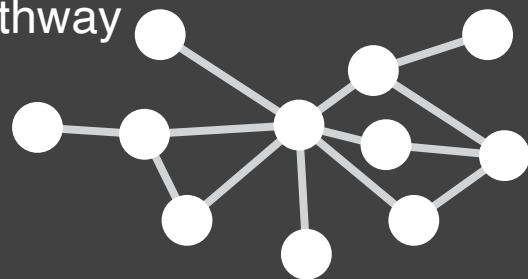
Gene expression



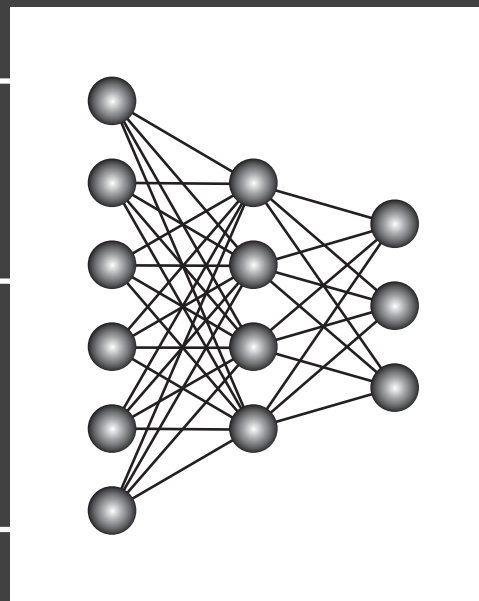
Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



## Machine Translation



**Output:** sentence annotation

A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.



# BioTranslator: A machine translation framework for translating biological data to human language

## Machine Translation

**Input:** A biological data instance

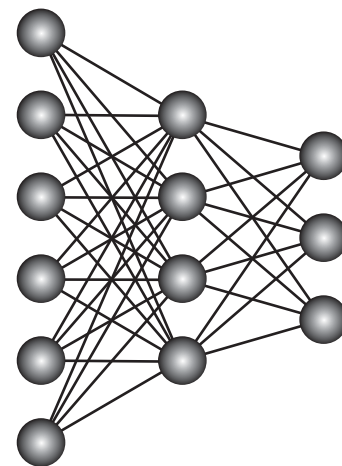
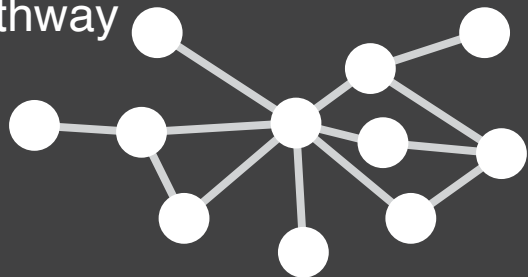
Gene expression



Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



**Output:** sentence annotation

A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.

- **Input:** Various biological data modalities, gene expression, protein, etc

# BioTranslator: A machine translation framework for translating biological data to human language

## Machine Translation

**Input:** A biological data instance

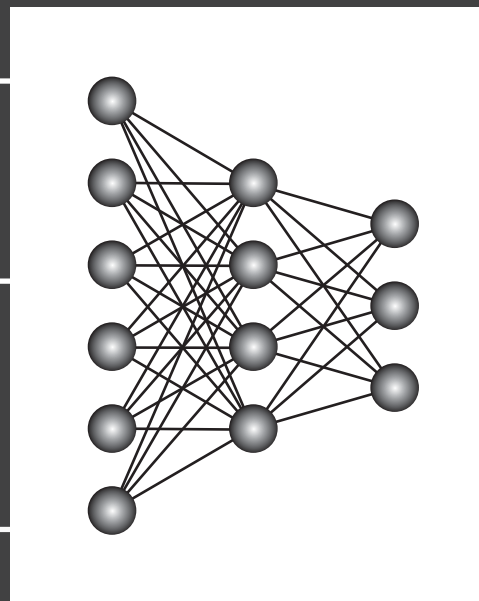
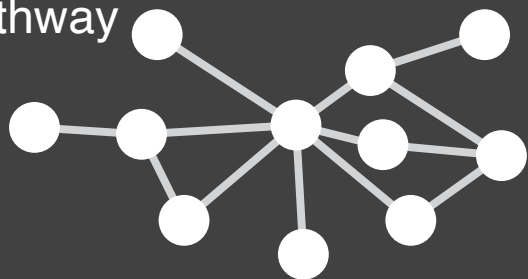
Gene expression



Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



**Output:** sentence annotation

A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.

- **Output:** Biomedical textual descriptions of input modalities

# Three main applications of BioTranslator

## Machine Translation

**Input:** A biological data instance

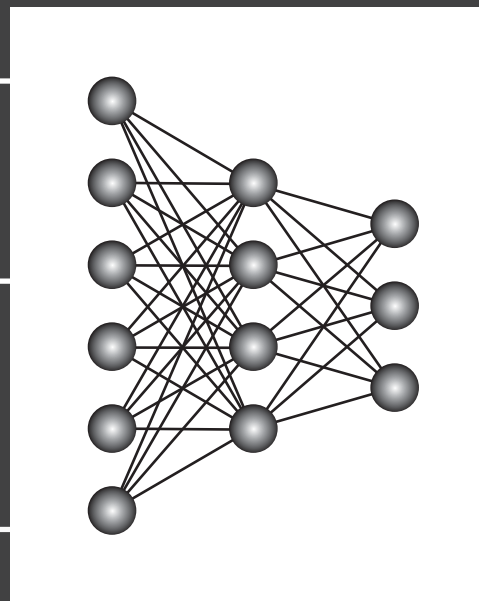
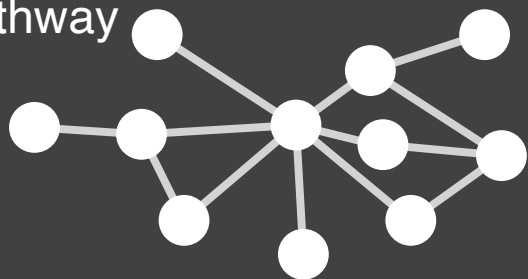
Gene expression



Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



**Output:** sentence annotation

A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.

- Protein-to-text: Annotate new proteins

# BioTranslator: A machine translation framework for translating biological data to human language

## Machine Translation

**Input:** A biological data instance

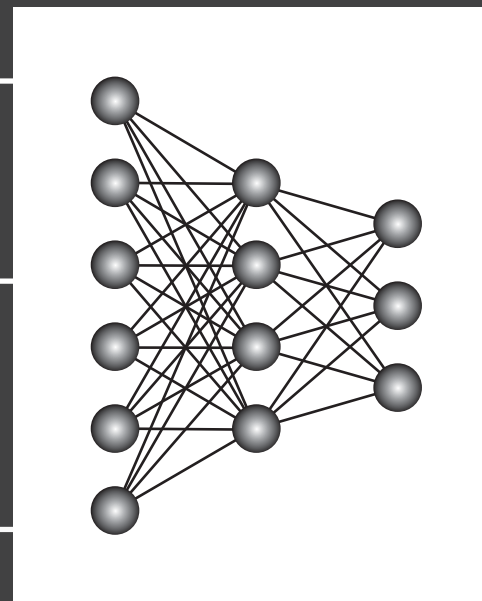
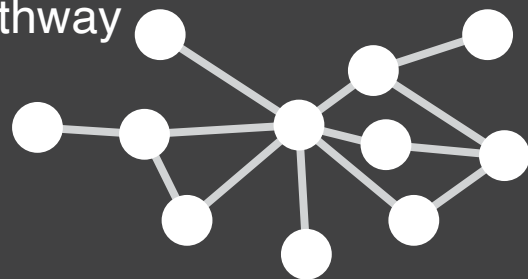
Gene expression



Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



**Output:** sentence annotation

A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.

- Protein-to-text: Annotate new proteins
- Single-cell-to-text: Discover new cell types

# BioTranslator: A machine translation framework for translating biological data to human language

## Machine Translation

**Input:** A biological data instance

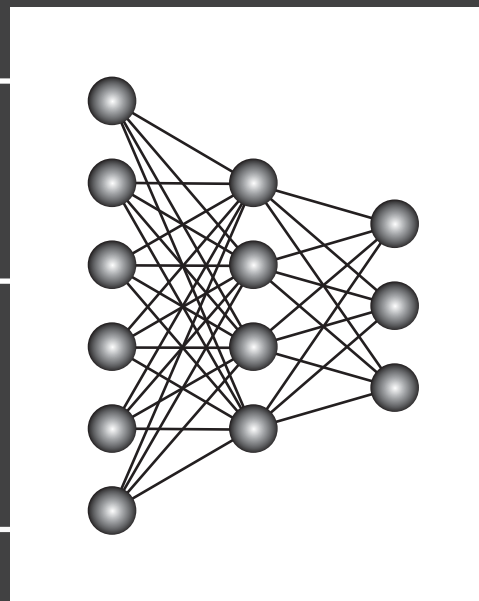
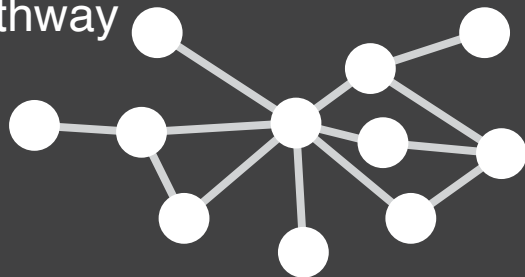
Gene expression



Protein sequence

SVLLRSGLGPLCAARAA .....VVAGFELAWQ

Pathway



**Output:** sentence annotation

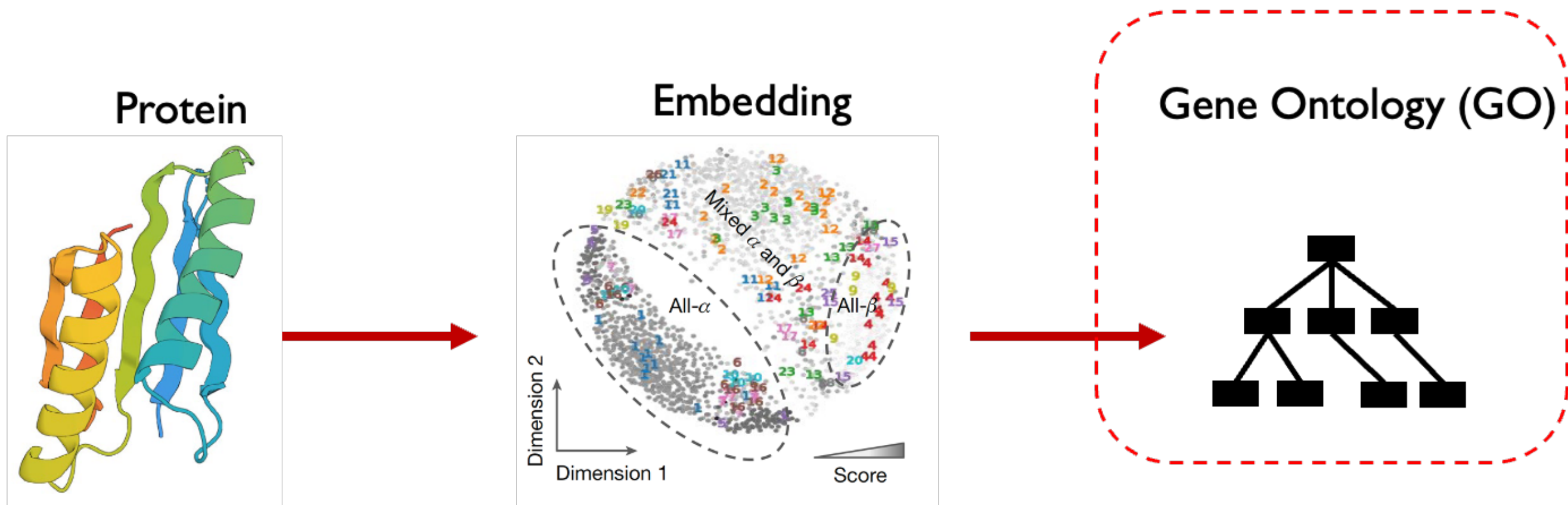
A cell found in the embryo before the formation of all the germ layers is complete.

The removal of sugar residues from a glycosylated protein.

A complex network of interacting proteins and enzymes is required for DNA replication.

- Protein-to-text: Annotate new proteins
- Single-cell-to-text: Discover new cell types
- Gene-pathway-to-text: interpretable analysis

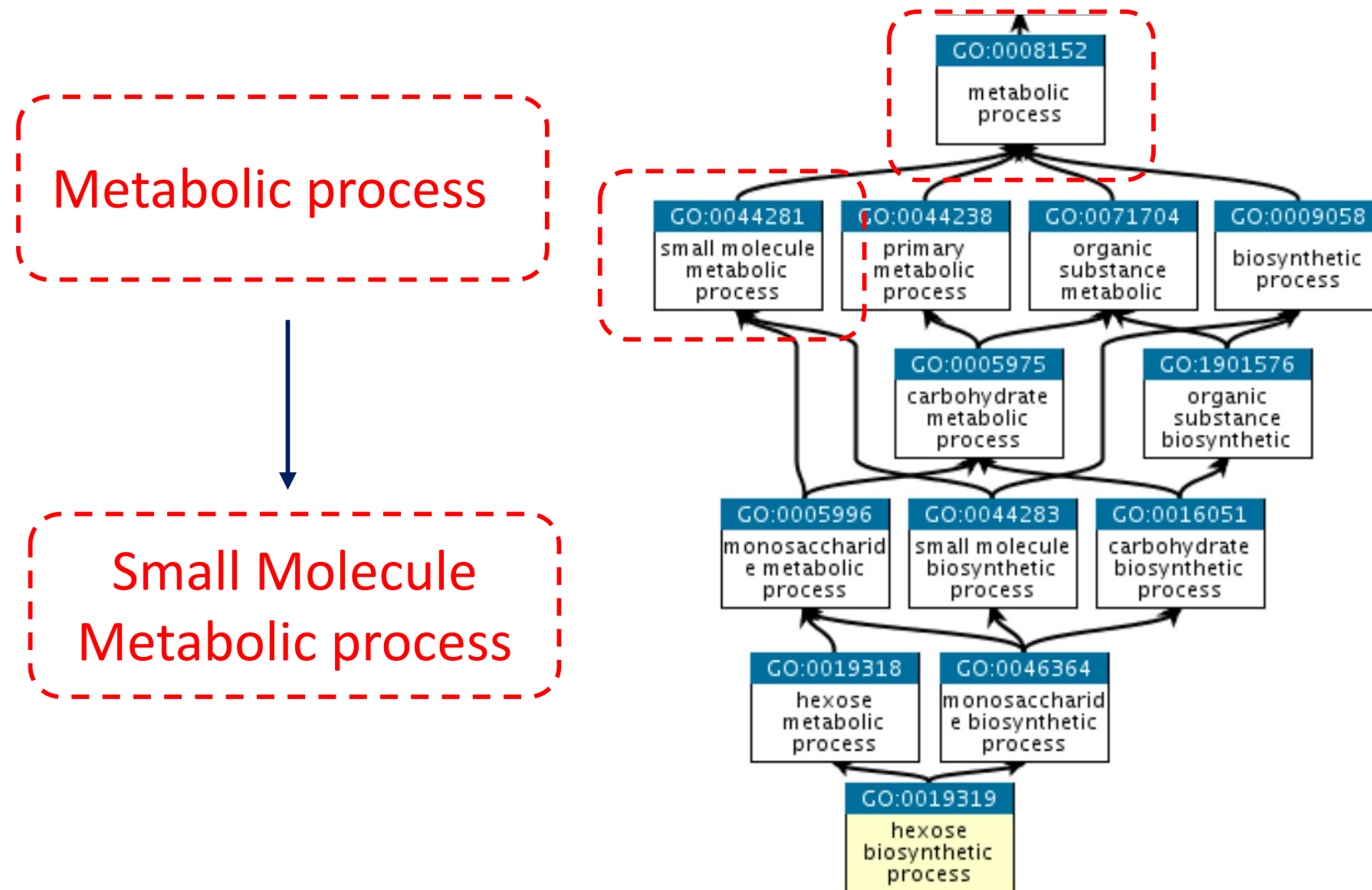
# Applications I: zero-shot protein function prediction



(Ivan Anishchenko, Nature, 2021)

- Current methods use the Gene Ontology as controlled vocabularies

# Gene Ontology has predefined hierarchical structures



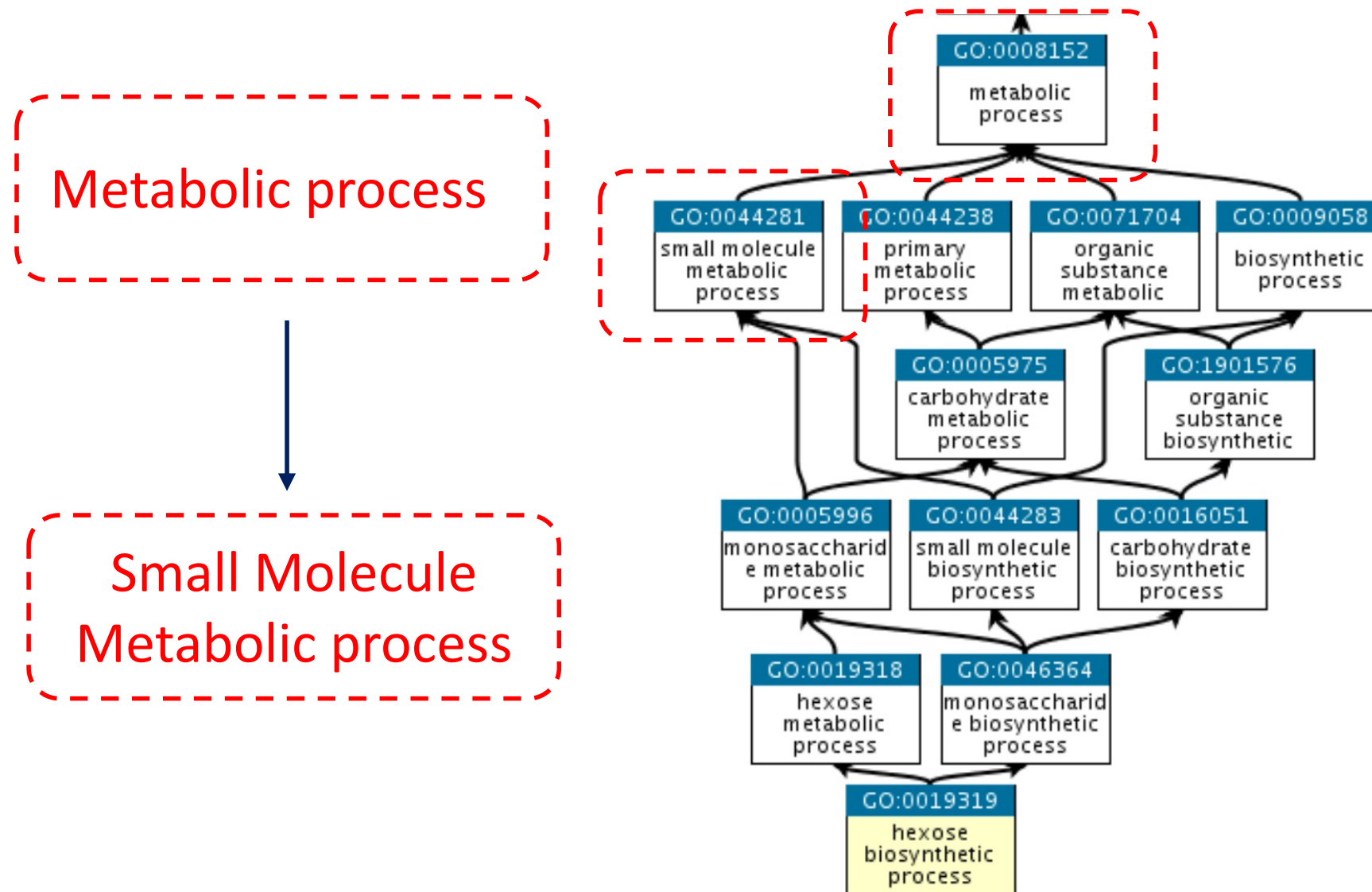
43,329 terms

(Nature Genetics 2020)

- Each term is a protein function class



# Gene Ontology is one type of controlled vocabulary



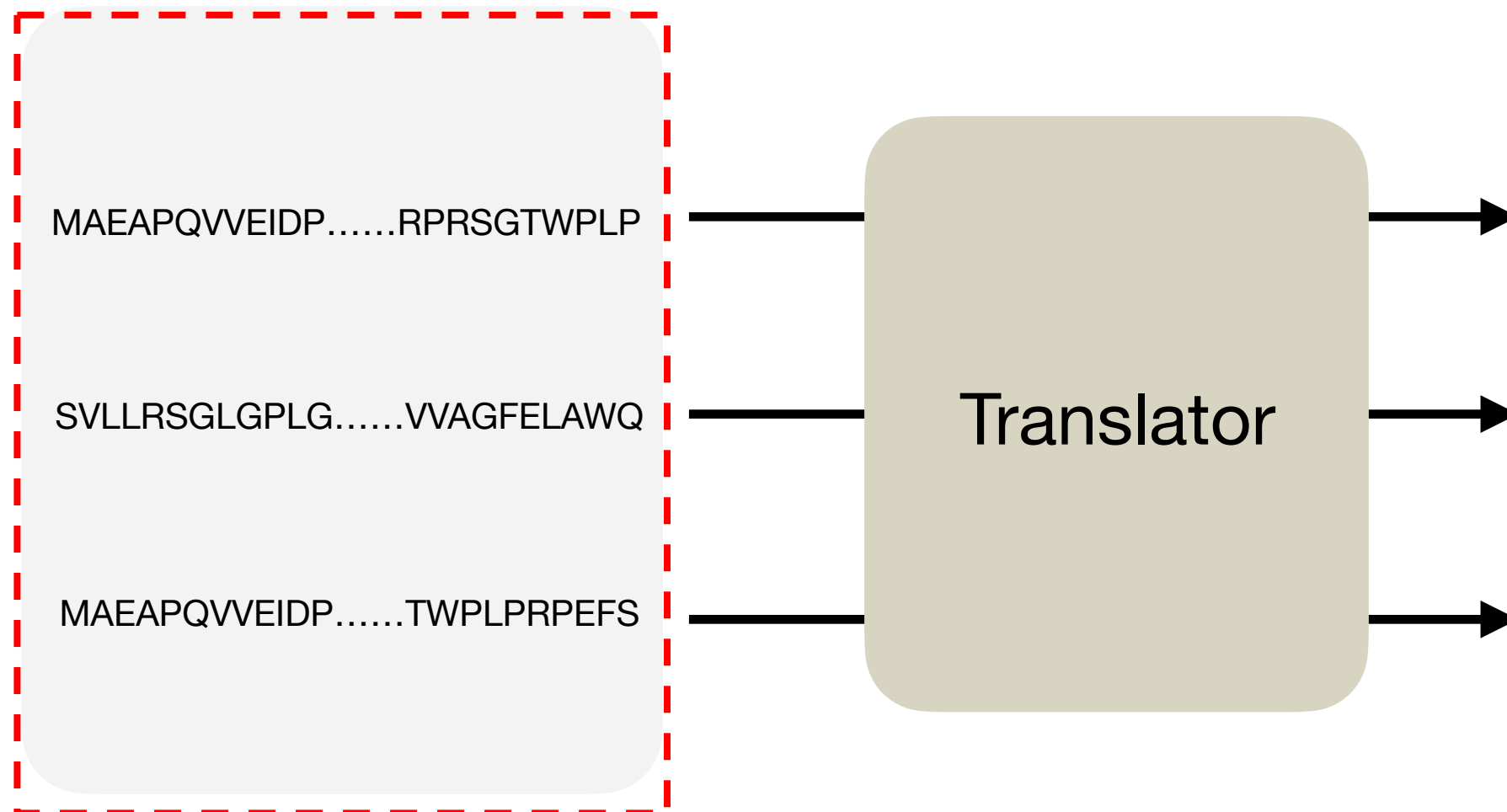
43,329 terms

(Nature Genetics 2020)

- Current methods formulate it as a multi-label classification task
- Cannot predict novel functions

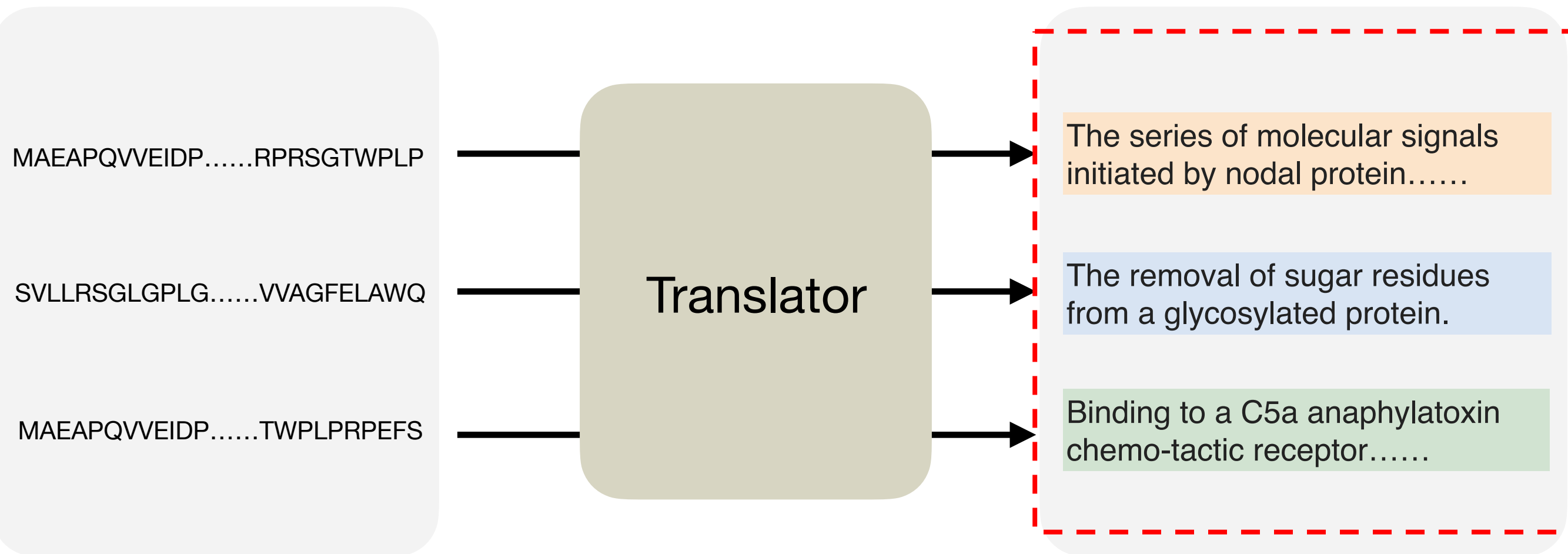


# Our solution: Translate proteins into textual descriptions



- Input: Protein sequences

# Applications I: zero-shot protein function prediction



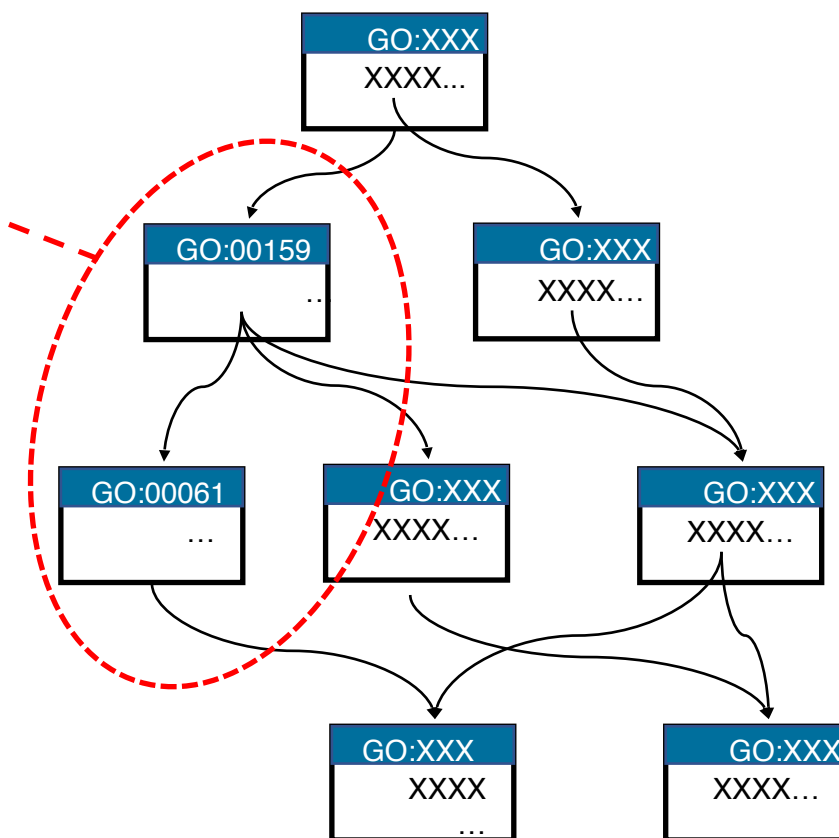
- Output Text: Protein function definition (Gene Ontology)

# How we annotate new protein functions through translation?

Existing classes

## GO:0015980:

The chemical reactions and pathways by which a cell derives energy from **organic compounds**; results in the oxidation of the compounds from which energy is released.



# How we annotate new protein functions through translation?

Existing classes

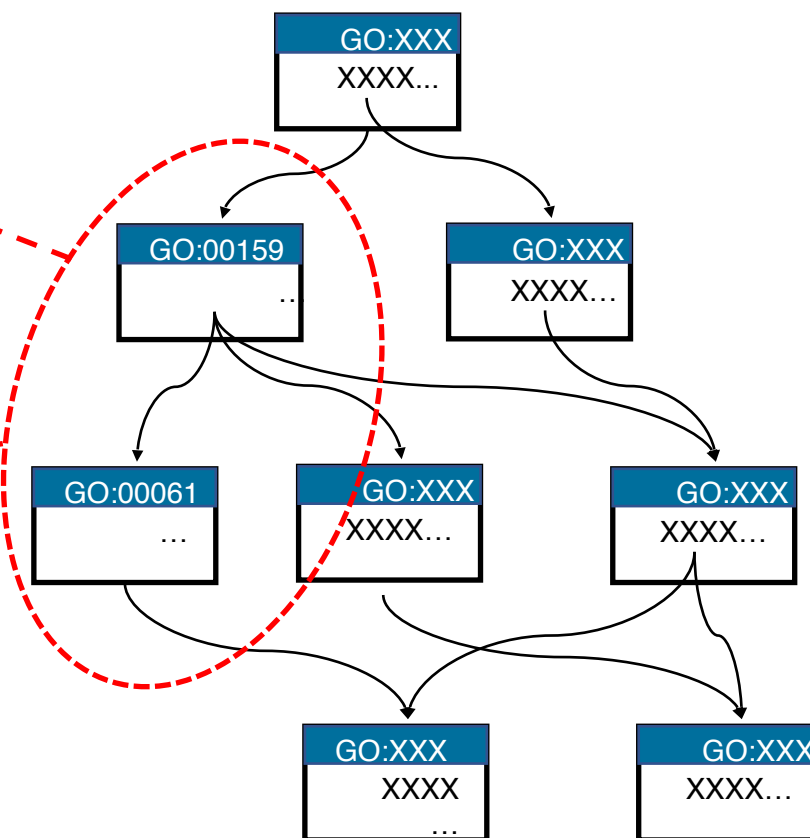
**GO:0015980:**

The chemical reactions and pathways by which a cell derives energy from **organic compounds**; results in the oxidation of the compounds from which energy is released.

Unseen classes

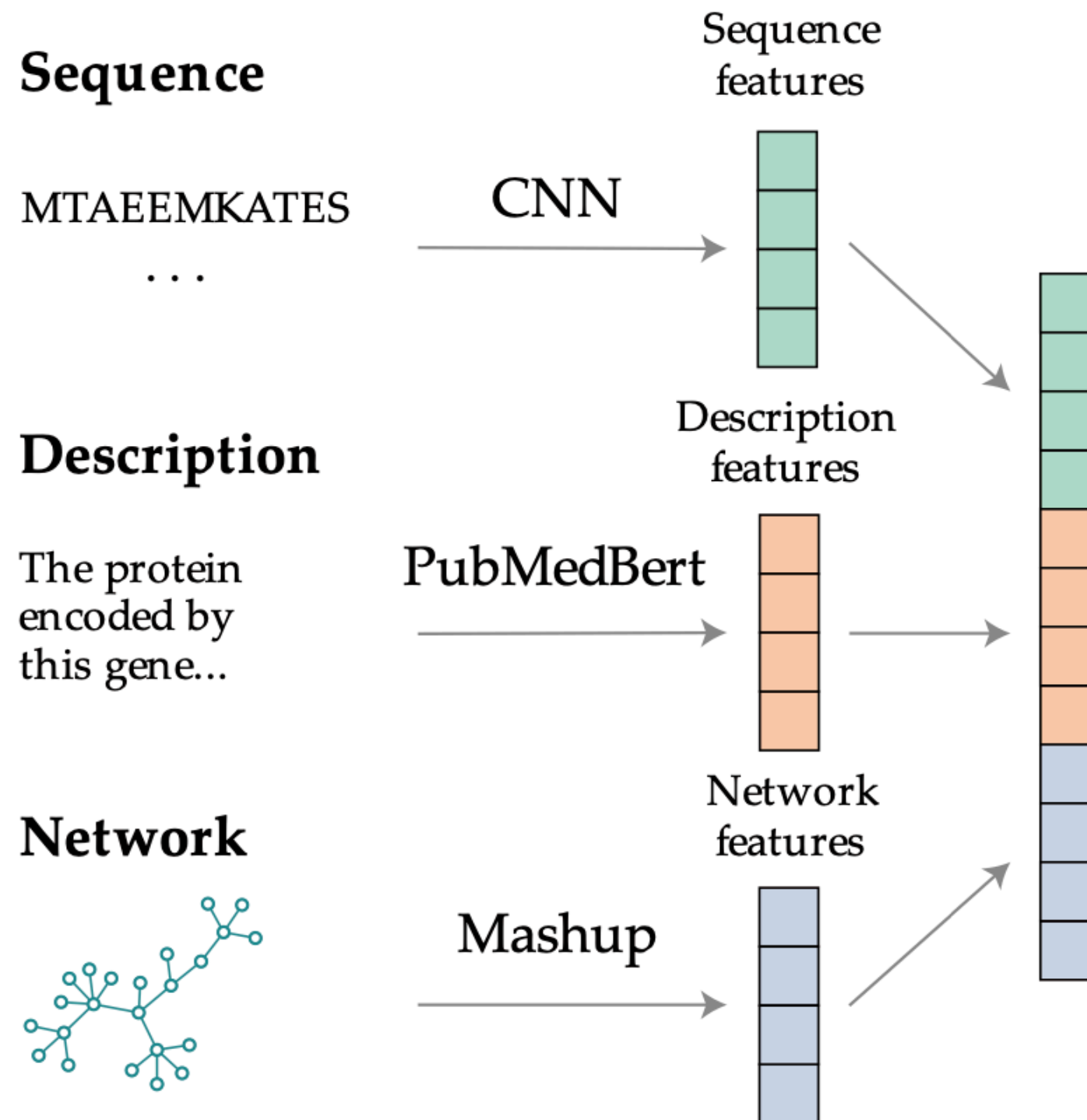
**GO:0006112:**

The chemical reactions and pathways by which a cell derives energy from **stored compounds such as fats or glycogen**.



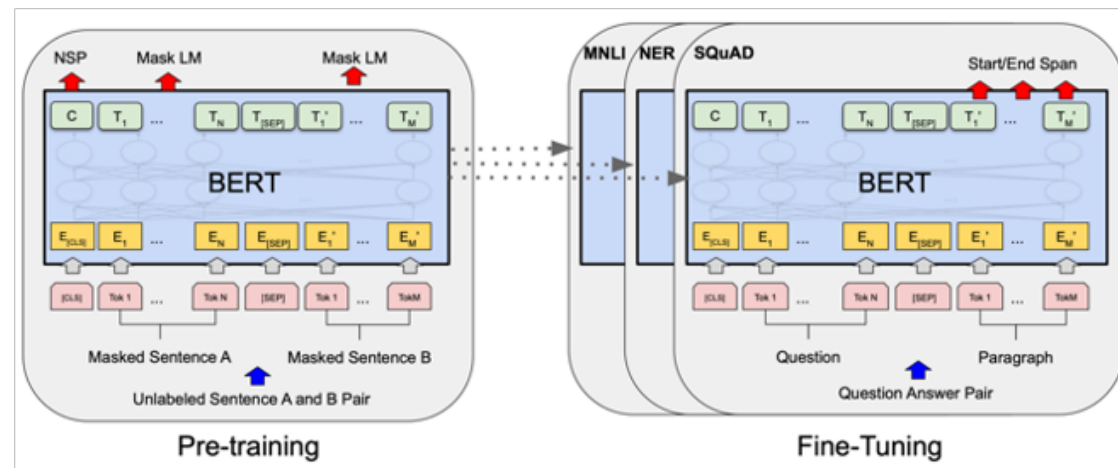
- We require textual definitions for unseen classes
- We transfer annotations based on text similarities

# How to embed the protein modality



We found using convolutional networks are useful

# How to embed the textual descriptions



**BERT**  
(Jacob Devlin, 2019)

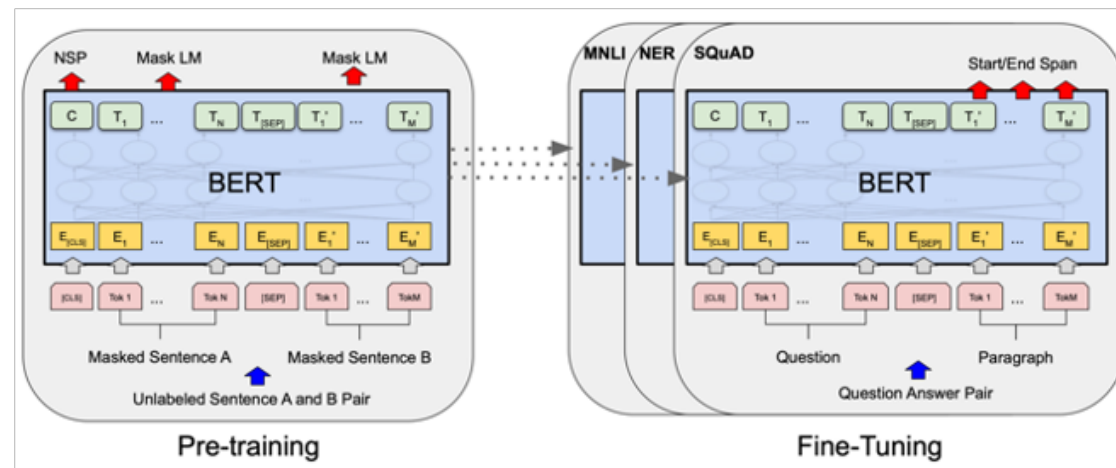


**GPT-2/3**  
(A Radford, OpenAI, 2019)  
(T Brown, NeurIPS, 2020)

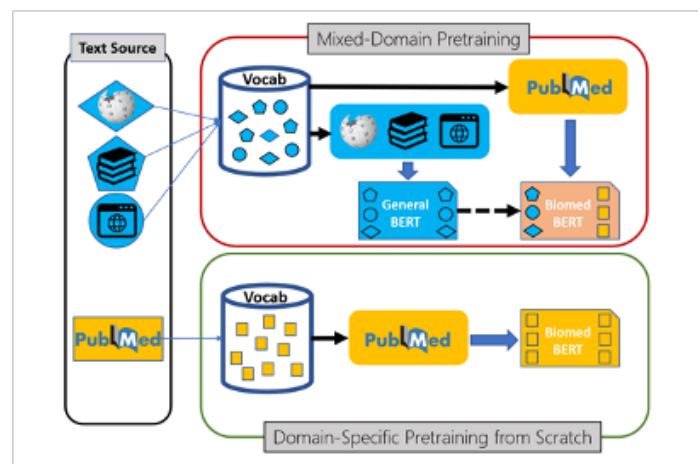
Progress in Pre-trained language models provide powerful tools



# How to embed the textual descriptions



**BERT**  
(Jacob Devlin, 2019)

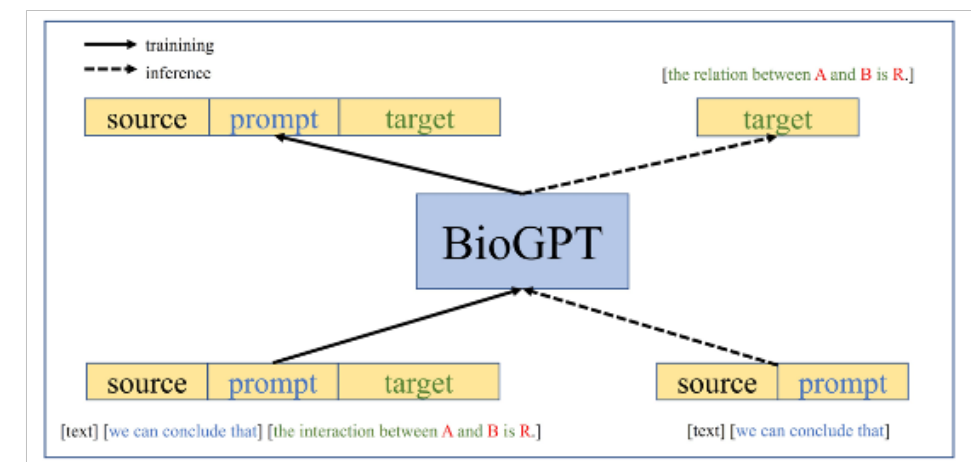


**PubMedBert**

(Yu Gu, ACM Transactions on Computing for Healthcare, 2021)



**GPT-2/3**  
(A Radford, OpenAI, 2019)  
(T Brown, NeurIPS, 2020)

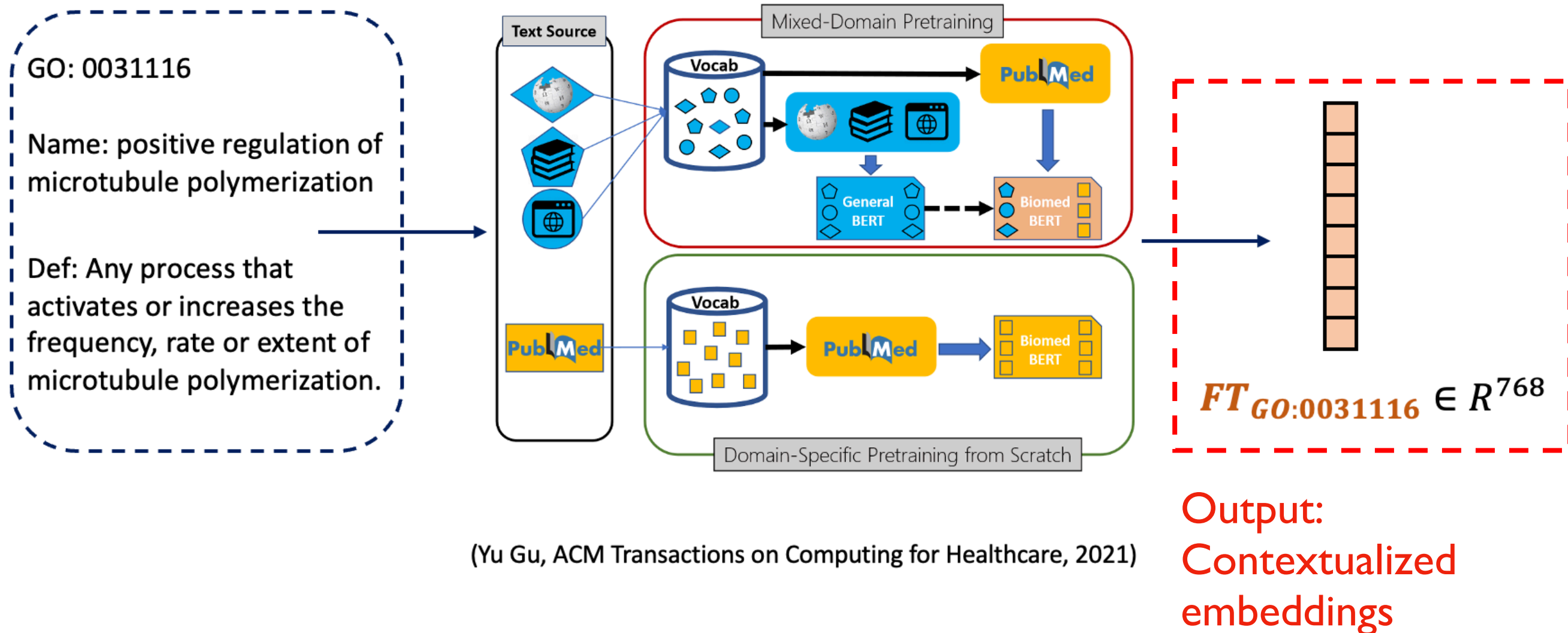


**BioGPT**

(Zhilin Y, NeurIPS, 2019)

Progress in Pre-trained language models provide powerful tools  
PubMedBERT is best aligned with our task

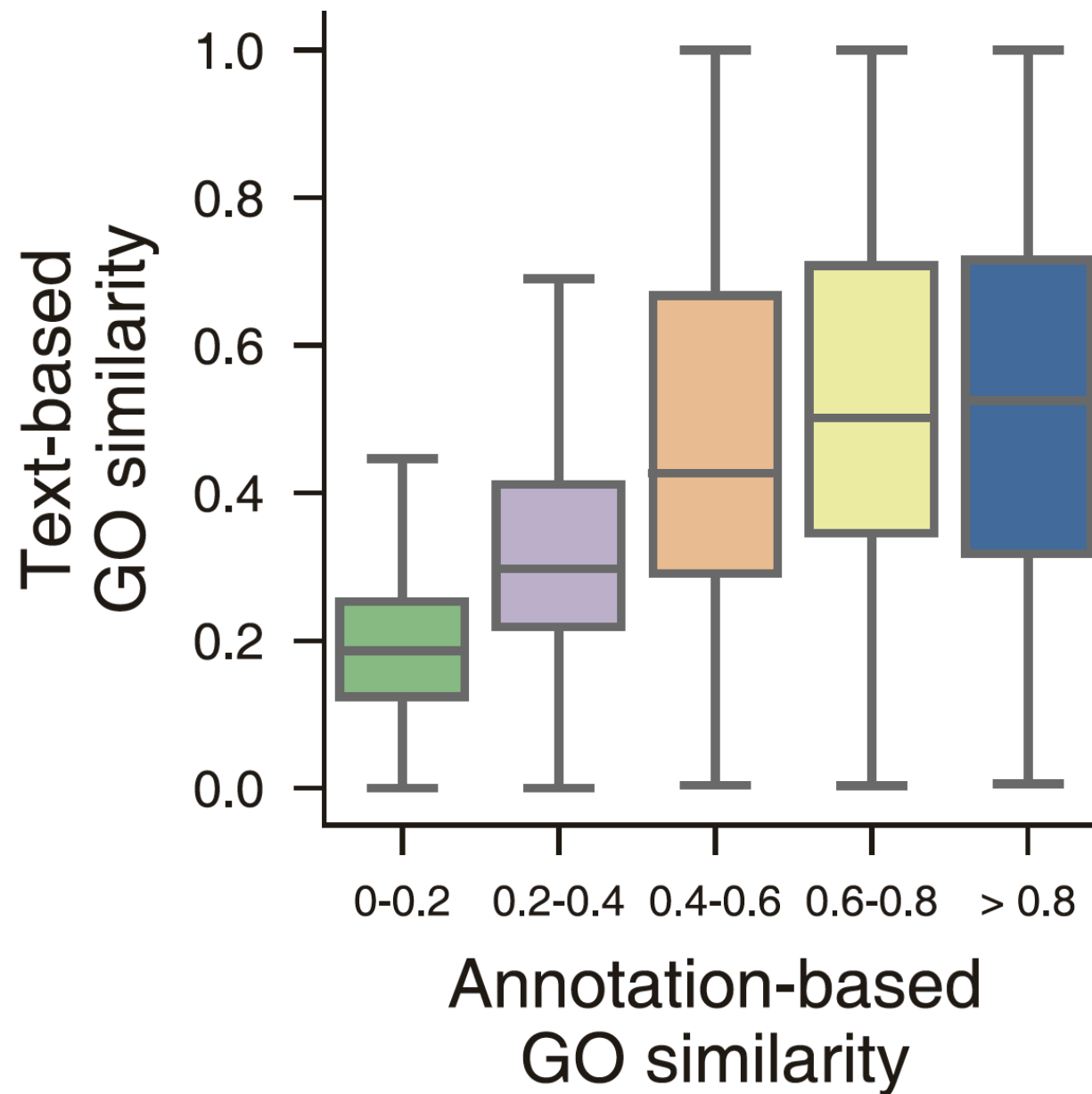
# How to embed the textual descriptions



(Yu Gu, ACM Transactions on Computing for Healthcare, 2021)

- Fine-tune PubMedBERT on 225 ontologies with 2,010,648 domain specific sentences
- Train a bilinear model on the paired biological instance and textual descriptions

# Similar descriptions have similar annotations

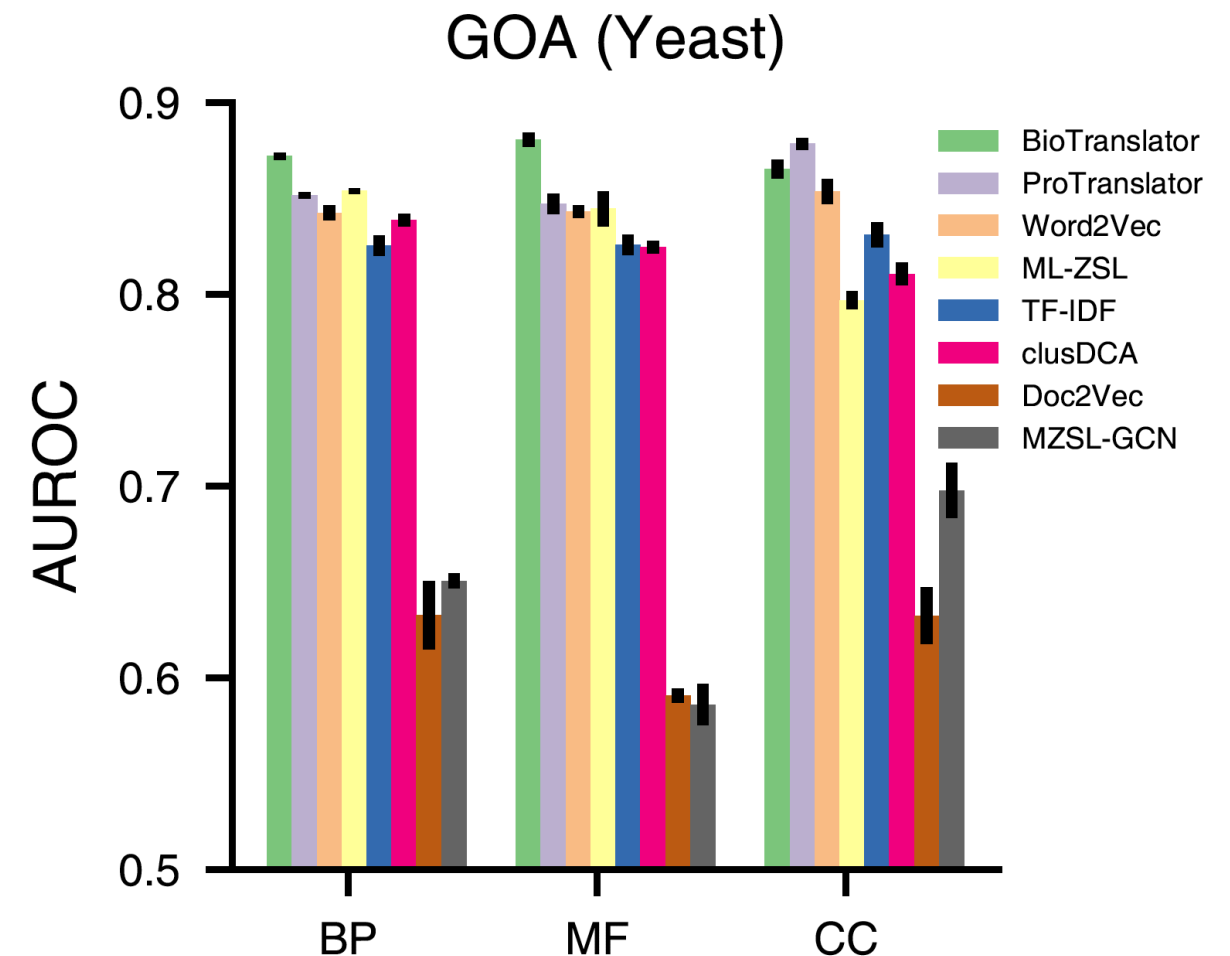
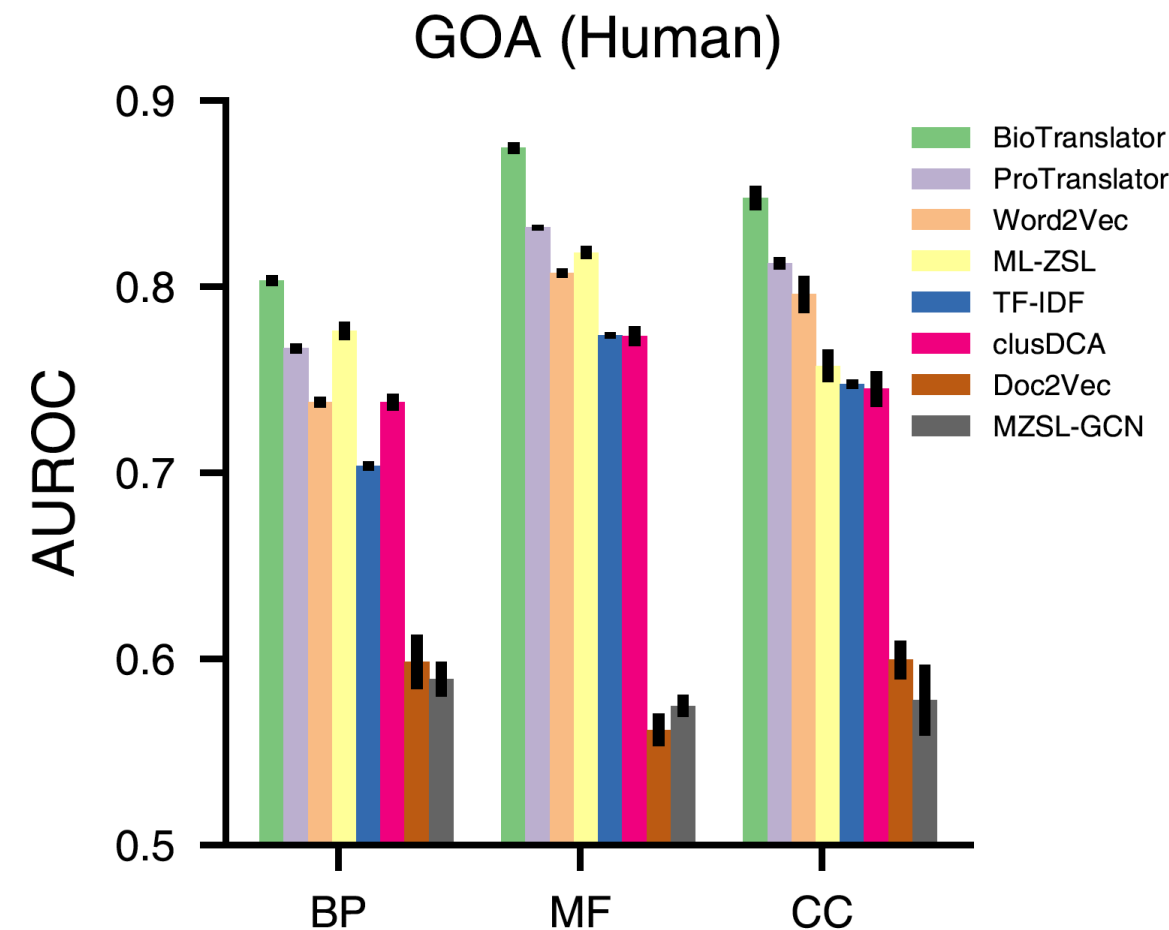


Annotation similarity: Jaccard similarity between annotation vectors

Text-based function similarity: BLEU score between textual descriptions

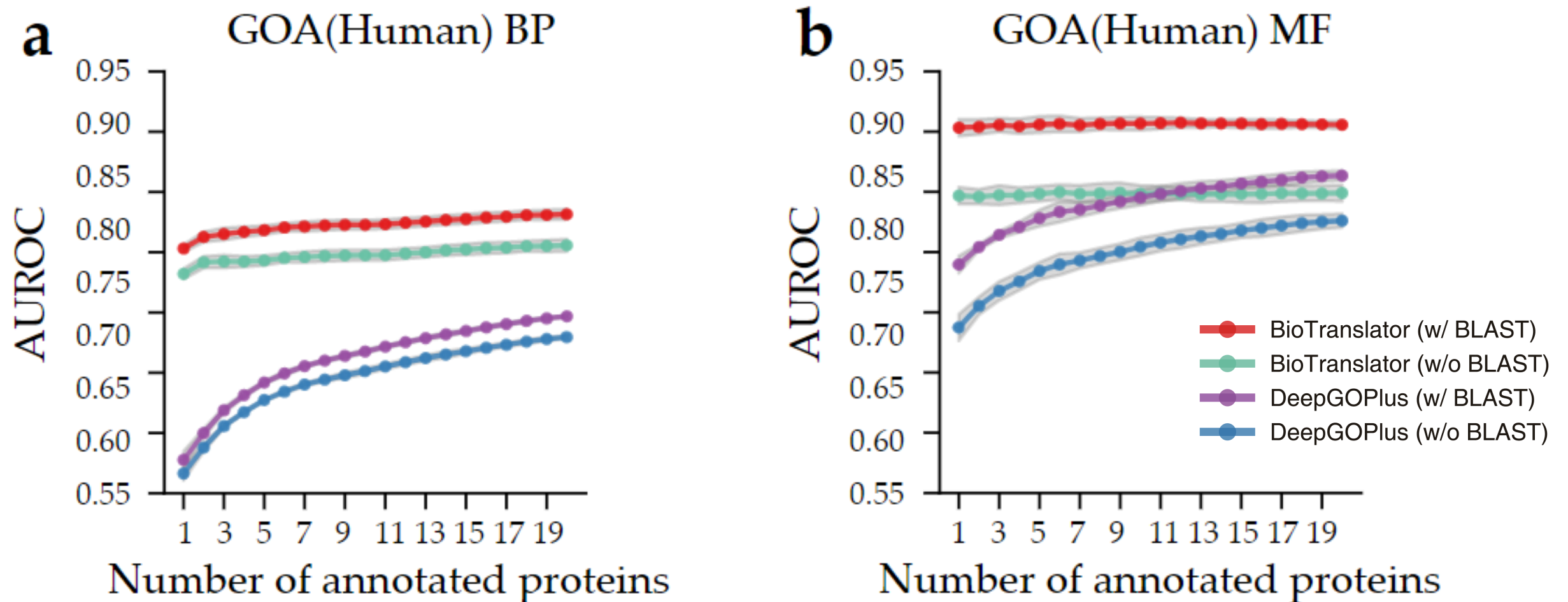
# State-of-the-art protein prediction on novel functions

- Graph-based approach: clusDCA, MZSL-GCN, ML-ZSL
- Text-based approaches: ProTranslator, Word2Vec, TF-IDF, Doc2Vec



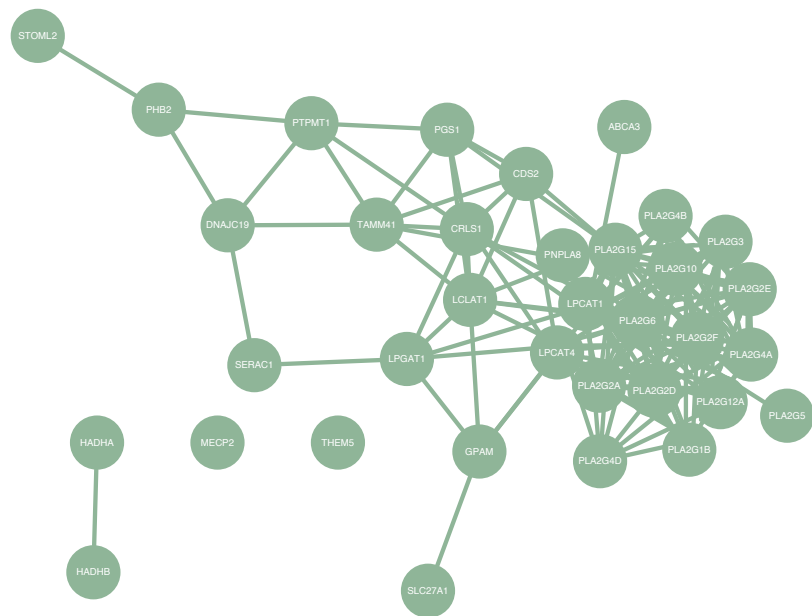
- Substantial improvement on unseen functions prediction compared to graph-based and other text-based approaches
- The only information we use is a short description of new functions

# Improved performance on the few-shot setting



Large improvement on Sparsely annotated functions

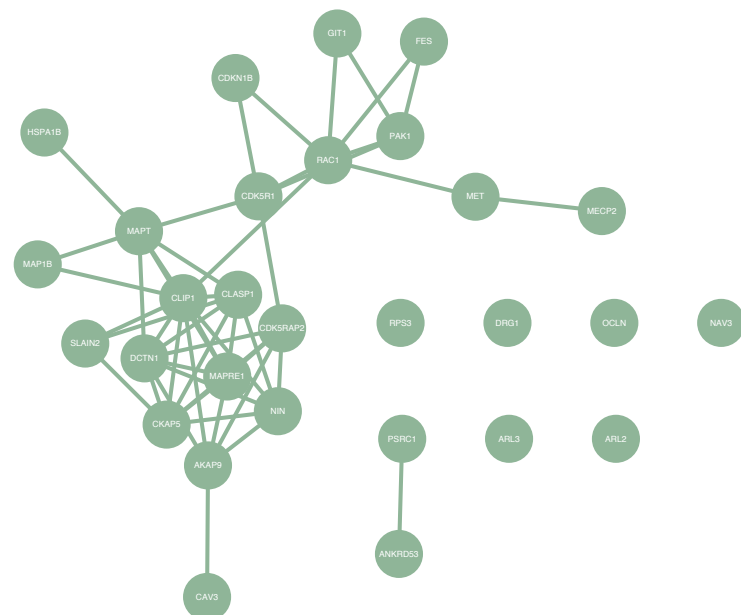
# Generate a textual description for a novel pathway



BioTranslator

Generated text for GO:0046471 phosphatidylglycerol metabolic process
---

The chemical reactions and pathways resulting in the formation of phosphatidylethanolamine, any of a class of glycerophospholipids in which a phosphatidyl group is esterified to the hydroxyl group of ethanolamine.
---



BioTranslator

Generated text for GO:0031116 positive regulation of microtubule polymerization
--

Any process that activates or increases the frequency, rate or extent of microtubule polymerization or depolymerization.
--



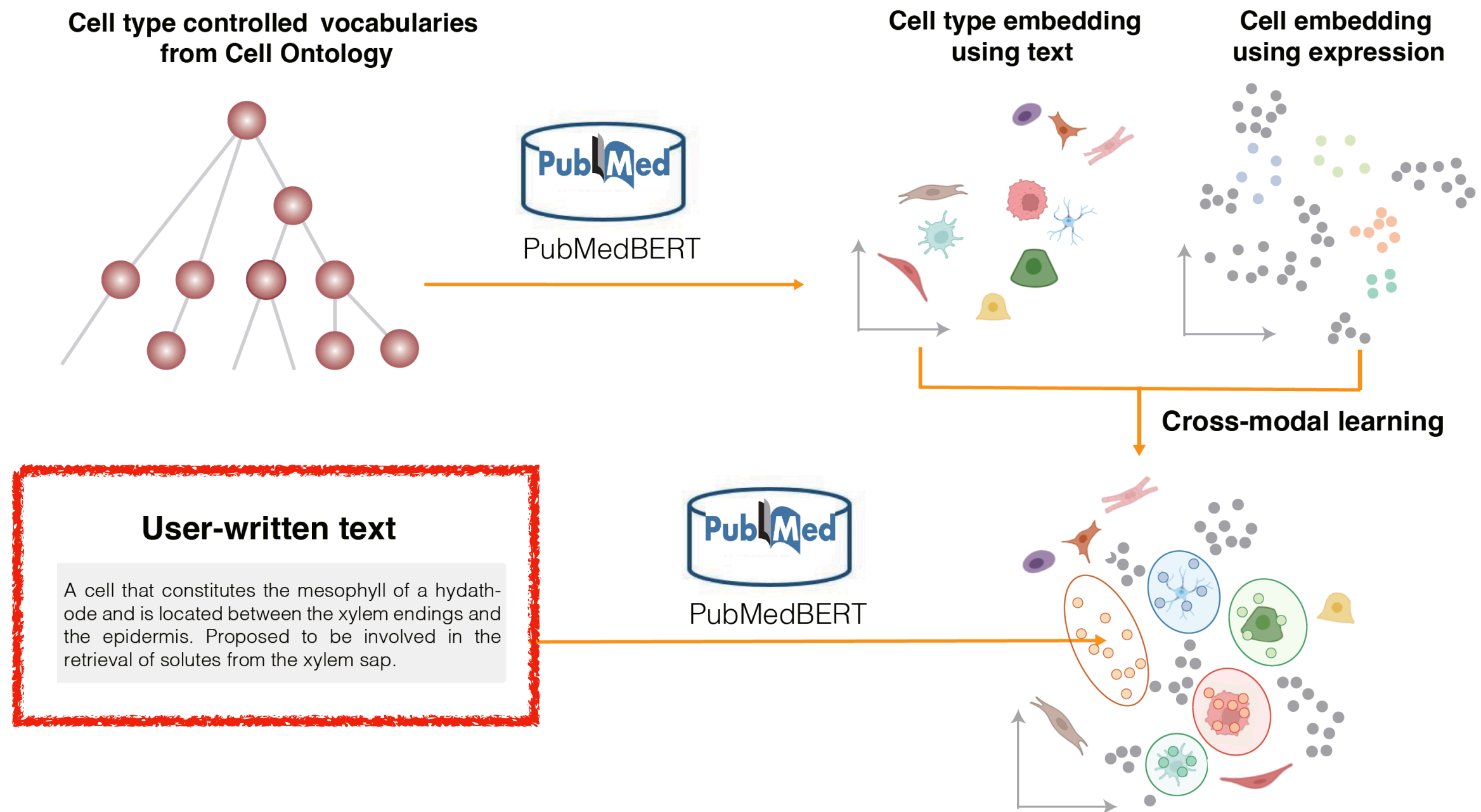
# Generate a textual description for a novel pathway

	GO:0032588
<b>Our generation</b>	the lipid bilayer surrounding a vesicle transporting substances between the trans - golgi network and other parts of the cell .
<b>Ground truth</b>	the lipid bilayer surrounding any of the compartments that make up the trans - golgi network .
	GO:0048738
<b>Our generation</b>	the process whose specific outcome is the progression of a cardiac cell over time , from its formation to the mature state . a cardiac cell is a cell that will form part of the cardiac organ of an individual .
<b>Ground truth</b>	the process whose specific outcome is the progression of cardiac muscle over time , from its formation to the mature structure .

Take away:

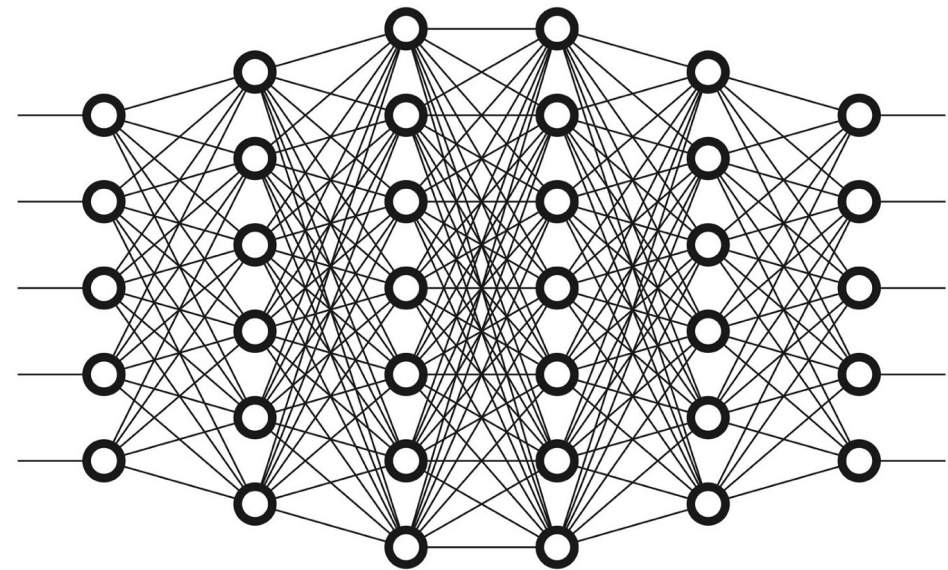
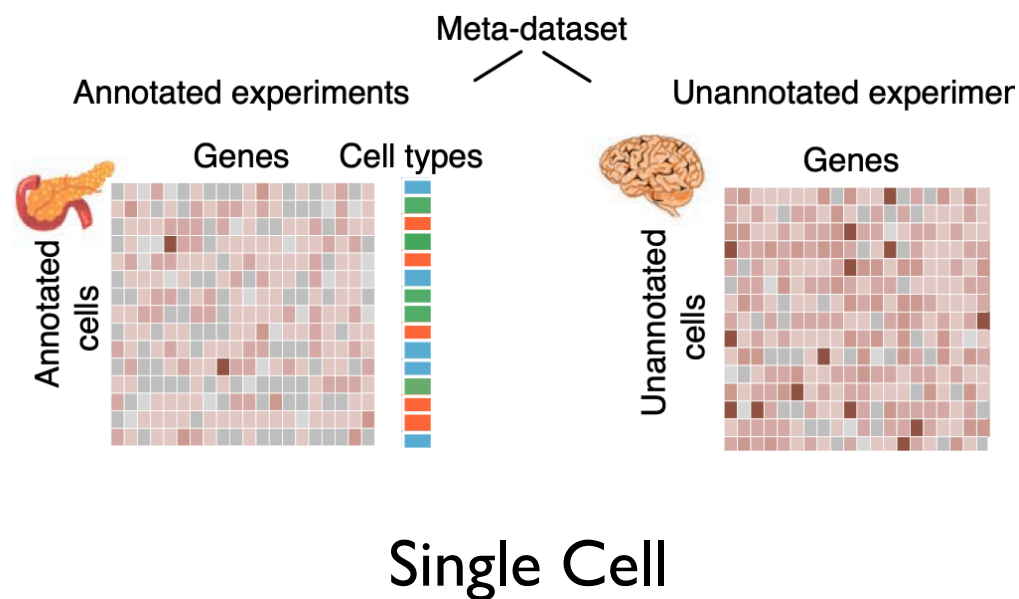
BioTranslator can generate texts with high fidelity

# Application II: classify single cell data into unseen cell types



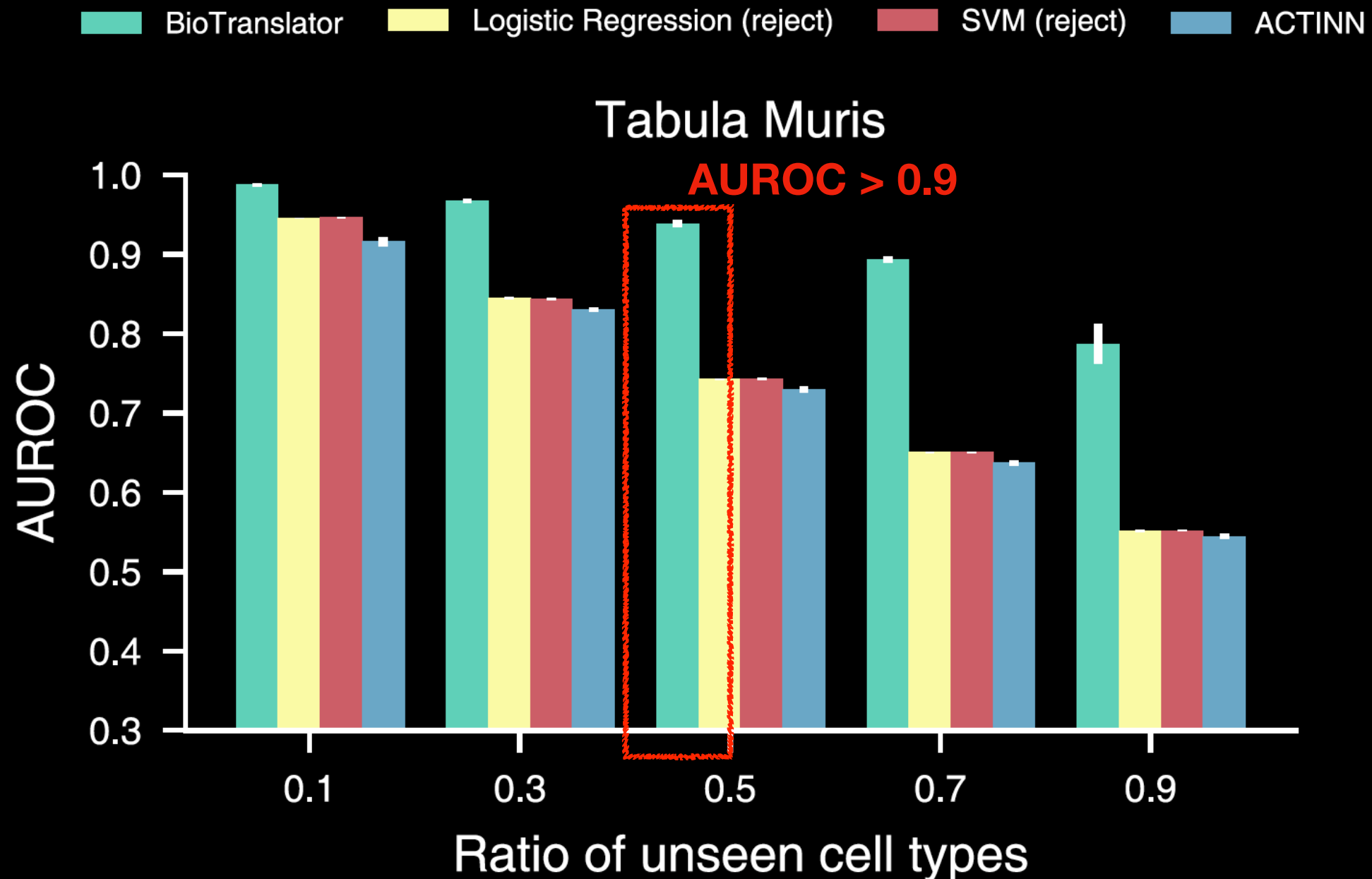
We can now classify any biological instance into any unseen class, only with proper textual descriptions.

# How to embed the protein modality



We found using MLP layers to embed single cell data are useful

# Application to single cell analysis



Task: classifying single cell into unseen cell type using cell type definitions

X-axis: ratio of unseen cell types in the test set

Y-axis: AUROC

# Single cell analysis in cross-dataset classification setting

AUROC(all)

Tabula Muris (Droplet)		0.93	0.91	0.86	0.89	0.90	0.82	0.89
Tabula Muris (FACS)	0.94		0.91	0.89	0.87	0.89	0.84	0.90
Tabula Microcebus	0.88	0.90		0.90	0.98	0.98	0.94	0.97
Tabula Sapiens	0.87	0.87	0.93		0.95	0.92	0.86	0.89
Lemur1	0.76	0.69	0.85	0.72		0.90	0.88	0.78
Lemur2	0.86	0.80	0.97	0.86	0.98		0.97	0.97
Lemur3	0.73	0.68	0.83	0.73	0.94	0.88		0.76
Lemur4	0.86	0.84	0.98	0.89	0.99	0.98	0.96	
	Tabula Muris (Droplet)	Tabula Muris (FACS)	Tabula Microcebus	Tabula Sapiens	Lemur1	Lemur2	Lemur3	Lemur4

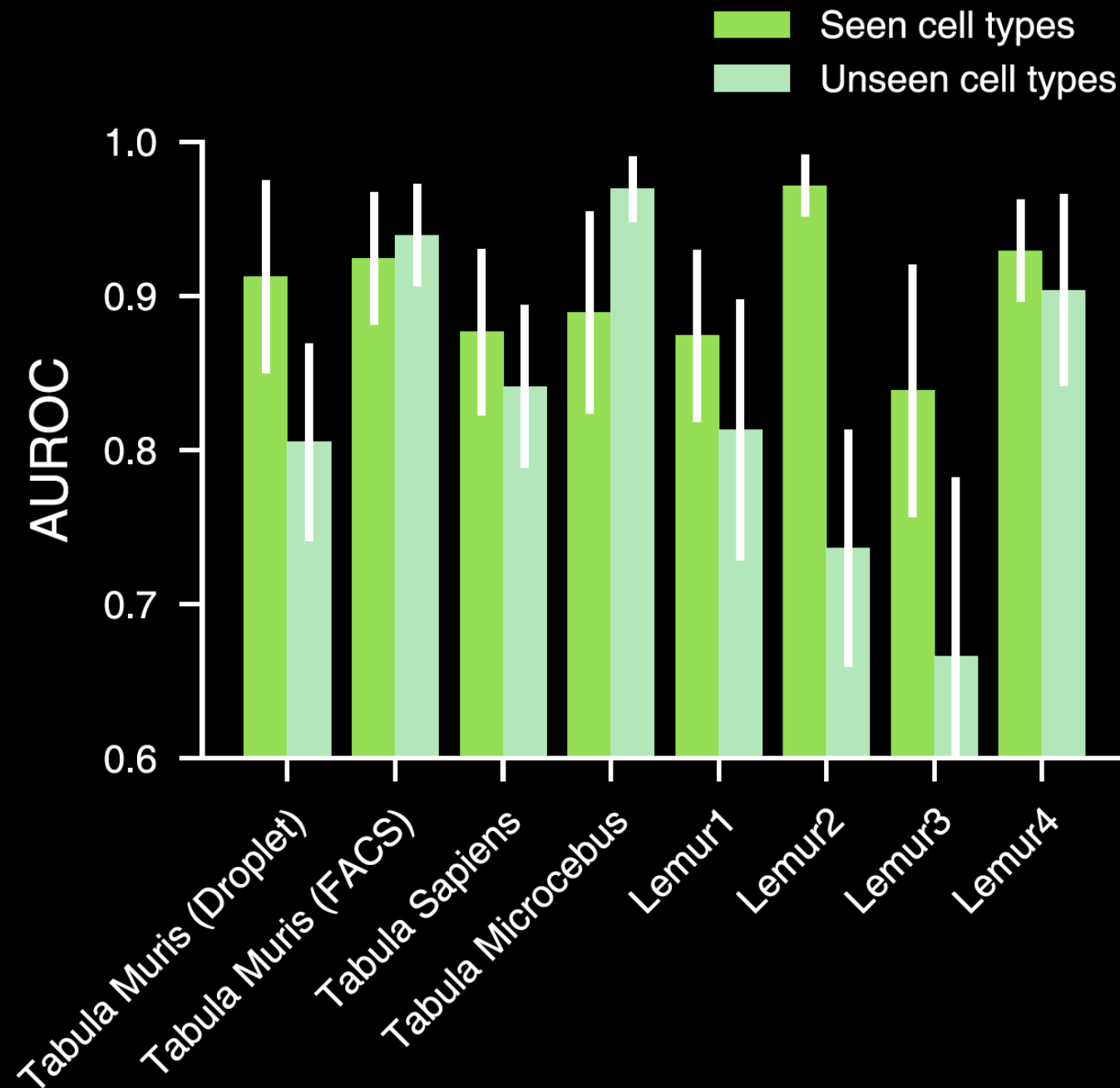
AUROC(unseen)

Tabula Muris (Droplet)		0.88	0.89	0.81	0.86	0.87	0.78	0.85
Tabula Muris (FACS)	0.89		0.89	0.86	0.82	0.86	0.79	0.87
Tabula Microcebus	0.83	0.86		0.87	0.96	0.94	0.86	0.93
Tabula Sapiens	0.83	0.83	0.90		0.93	0.89	0.81	0.84
Lemur1	0.75	0.66	0.83	0.71		0.86	0.80	0.72
Lemur2	0.83	0.77	0.96	0.86	0.94		0.91	0.90
Lemur3	0.72	0.67	0.82	0.74	0.91	0.85		0.71
Lemur4	0.82	0.81	0.96	0.87	0.95	0.93	0.87	
	Tabula Muris (Droplet)	Tabula Muris (FACS)	Tabula Microcebus	Tabula Sapiens	Lemur1	Lemur2	Lemur3	Lemur4

AUROC > 0.9

BioTranslator obtained a prominent result on classifying unseen cell types

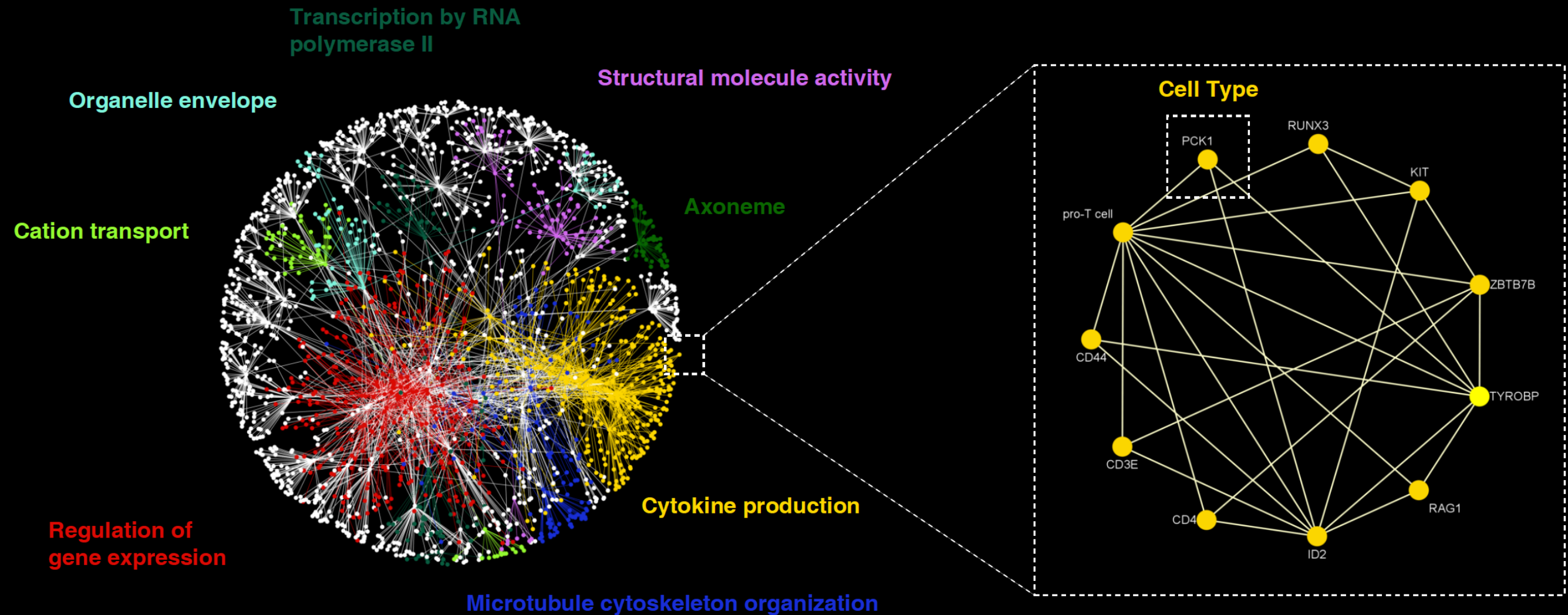
# Identify markers for unseen cell types



BioTranslator accurately identified marker genes for unseen cell types

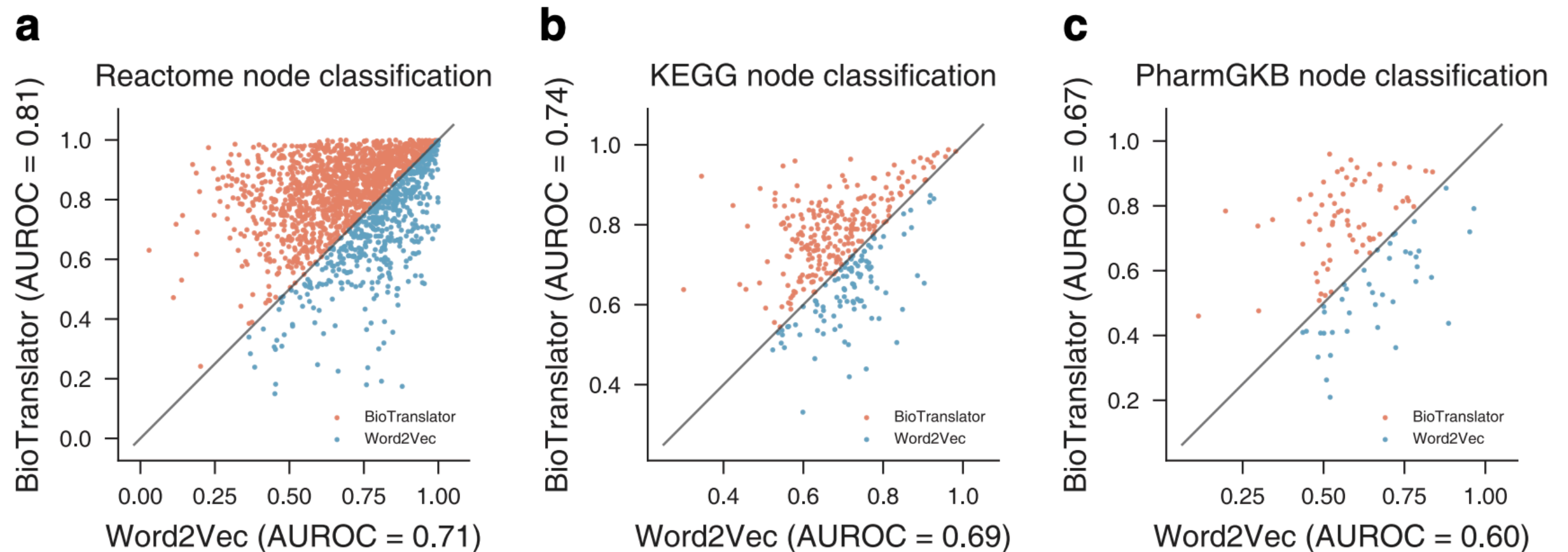


# Construct cell-type-marker-gene network



Find the enriched GO term for each community

# Application III: pathway membership identification



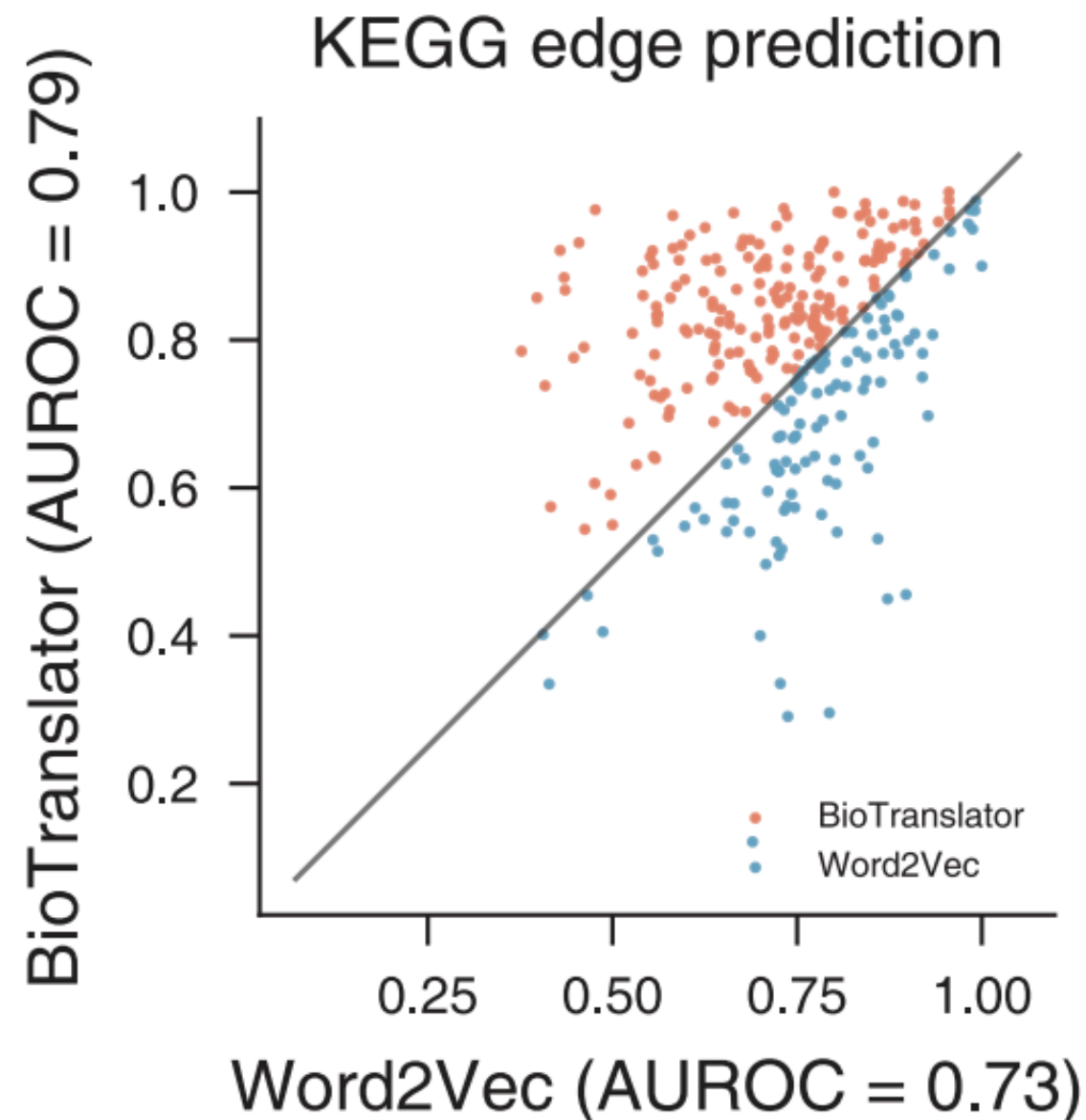
Three pathway databases: Reactome, KEGG, PharmGKB

Task: classifying genes into pathway using pathway description

X-axis: the performance of baselines.

Y-axis: the performance of BioTranslator

## Application III: pathway edges prediction



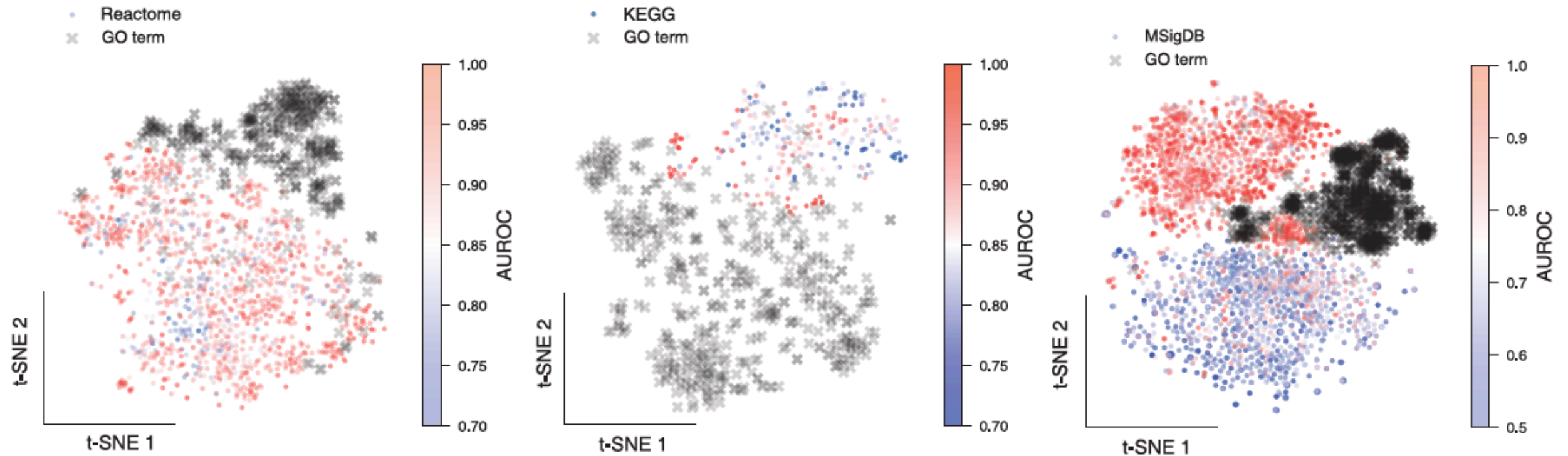
Three pathway databases: Reactome, KEGG, PharmGKB

Task: classifying genes into pathway using pathway description

X-axis: the performance of baselines.

Y-axis: the performance of BioTranslator

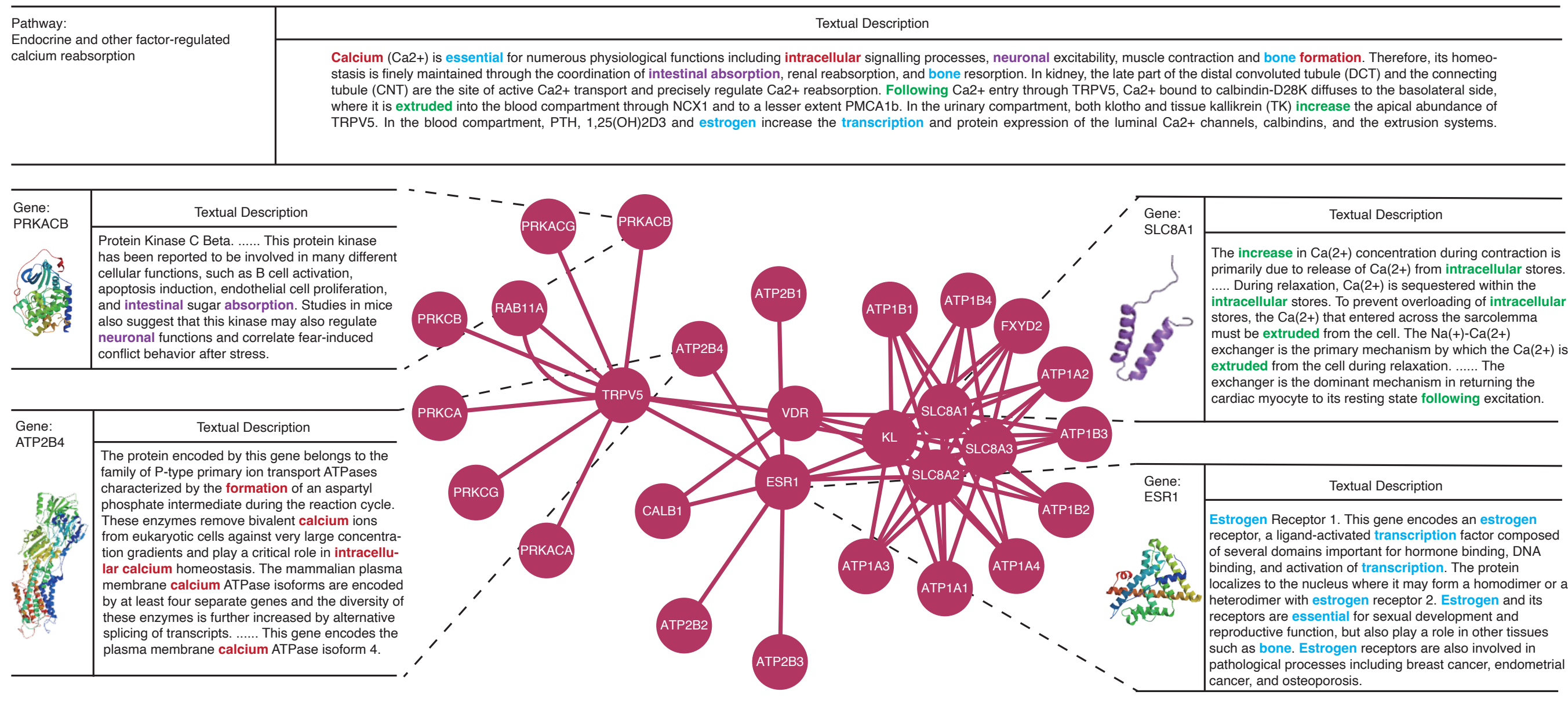
# Transfer learning from GO term to pathways



- Train BioTranslator using GO term
- Use this model to classify genes into pathway based on pathway description
- Pathways that are closer to GO term have higher prediction accuracy



# Precision annotation for interpretation



- Perturb biological data features and see the perturbation in the corresponding biological text
- Important words (highlighted) are robust to perturbation in the biological data

# Conclusion

- Annotate biological instances by generating descriptive sentences
  - Annotate new discoveries, more expressive, more precise
- Fine-tune large pre-trained language model on 225 ontologies
  - Significant improvements compared to ProTranslator
- Applications on protein function prediction and single cell analysis



# Future work

- More paired data to be collect
  - Collect data from scientific literature
- More powerful methods for long text embeddings
  - Descriptions of complex biological process up to 1,000 words
  - More interpretability for embedding space
- Exploit current pre-trained model for different modalities
  - Pre-trained models: Protein, Molecule, Text, Single cell, ...
  - Explore how to integrate them together
- Explore translations between protein structure and text