

Multifaceted Representation of Genes via Deep Learning of Gene Expression Networks

Zheng (John) SU
BABS, Faculty of Science
Nov 2024

Supervised by:



Emily Oates



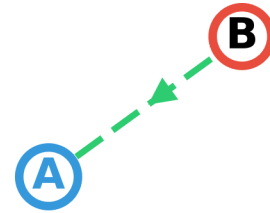
Marcel Dinger



Fatemeh Vafae

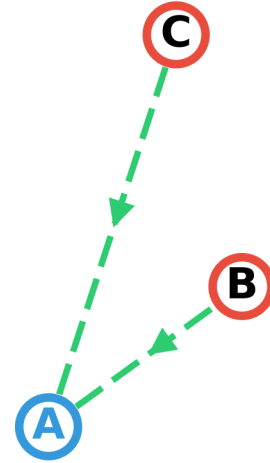
A New Approach to Representing Genes

Represent genes by their impact on each other's abundance



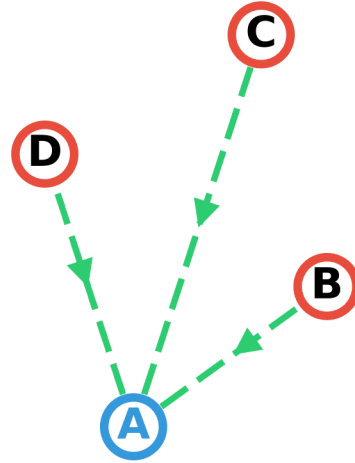
A New Approach to Representing Genes

Represent genes by their impact on each other's abundance



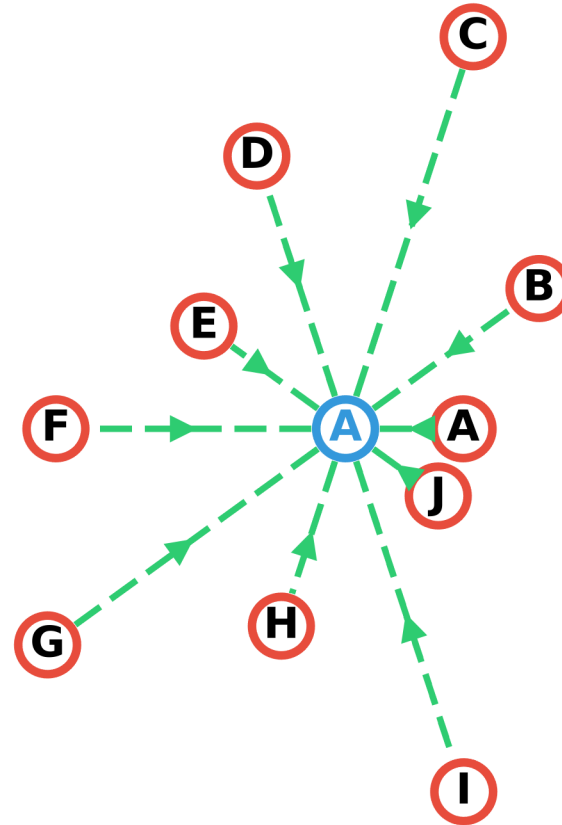
A New Approach to Representing Genes

Represent genes by their impact on each other's abundance

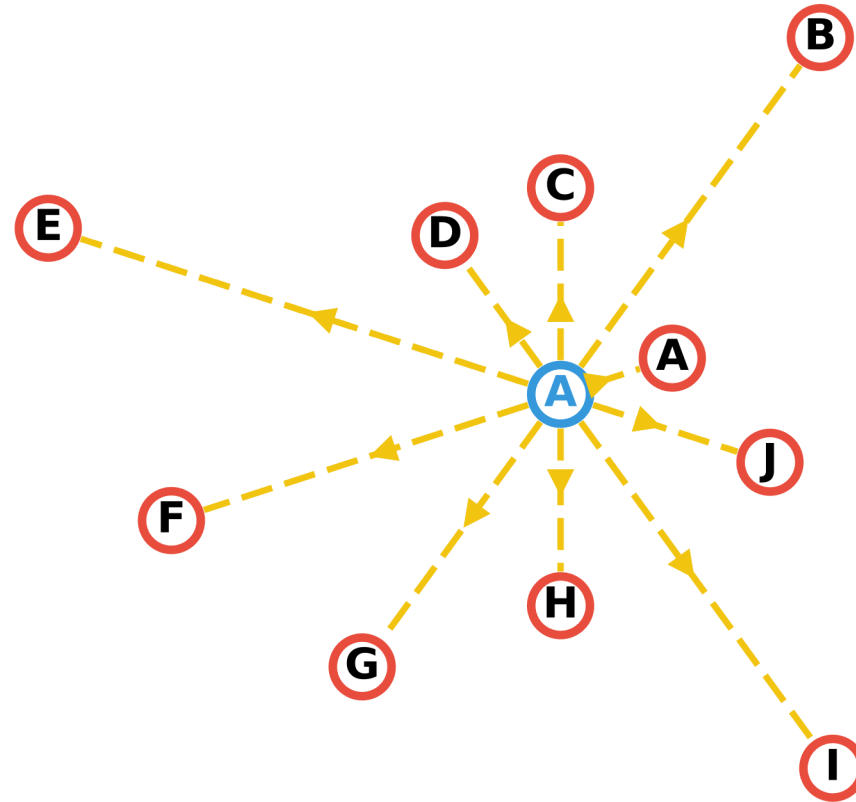


A New Approach to Representing Genes

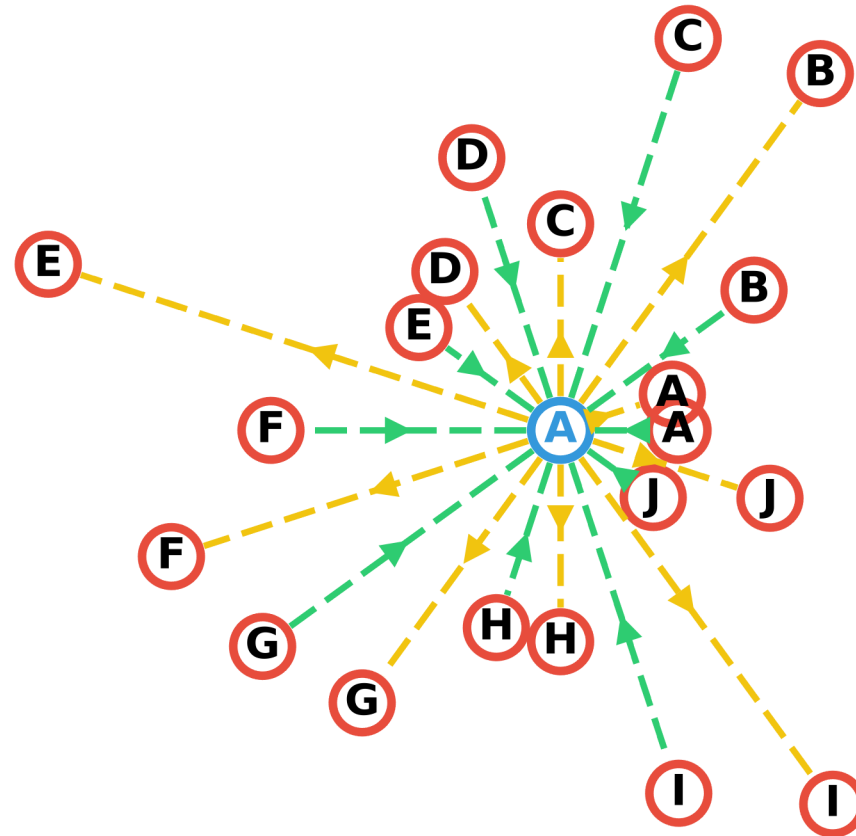
Represent genes by their impact on each other's abundance



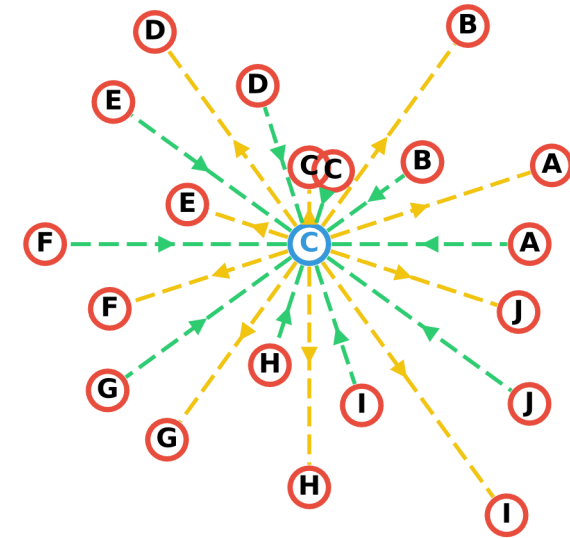
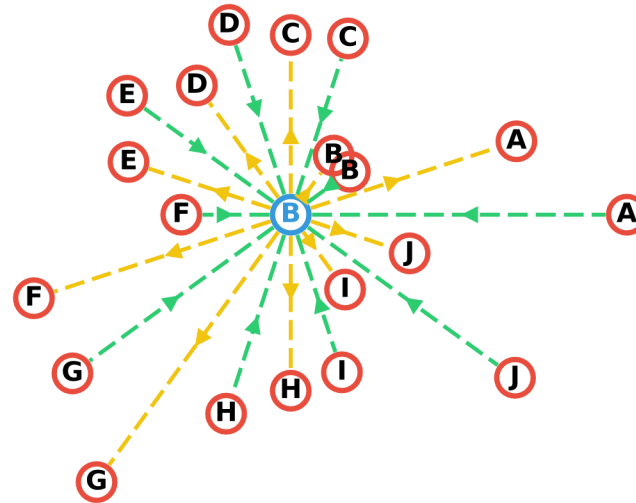
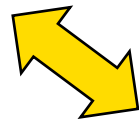
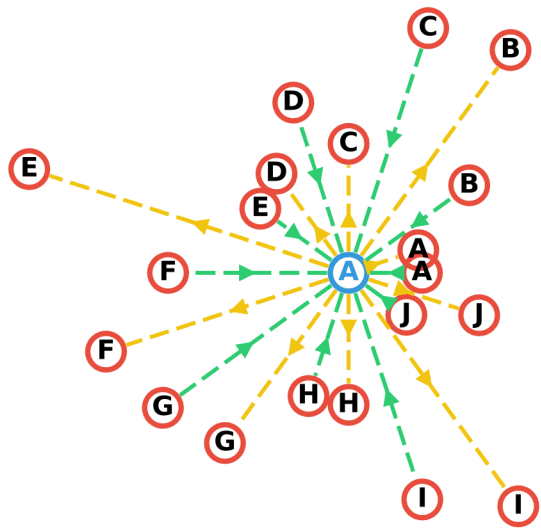
A New Approach to Representing Genes



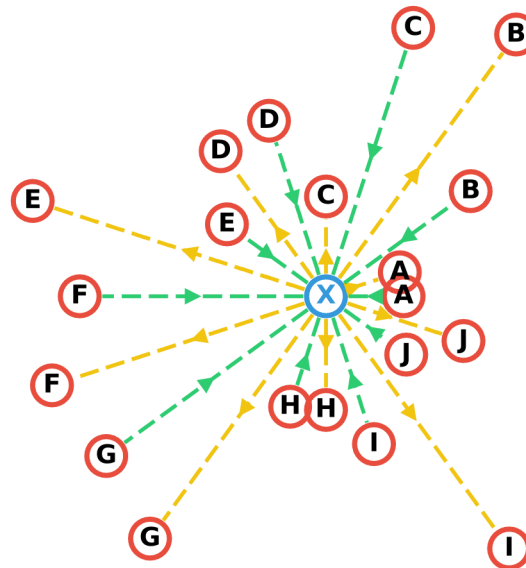
A New Approach to Representing Genes



A New Approach to Representing Genes

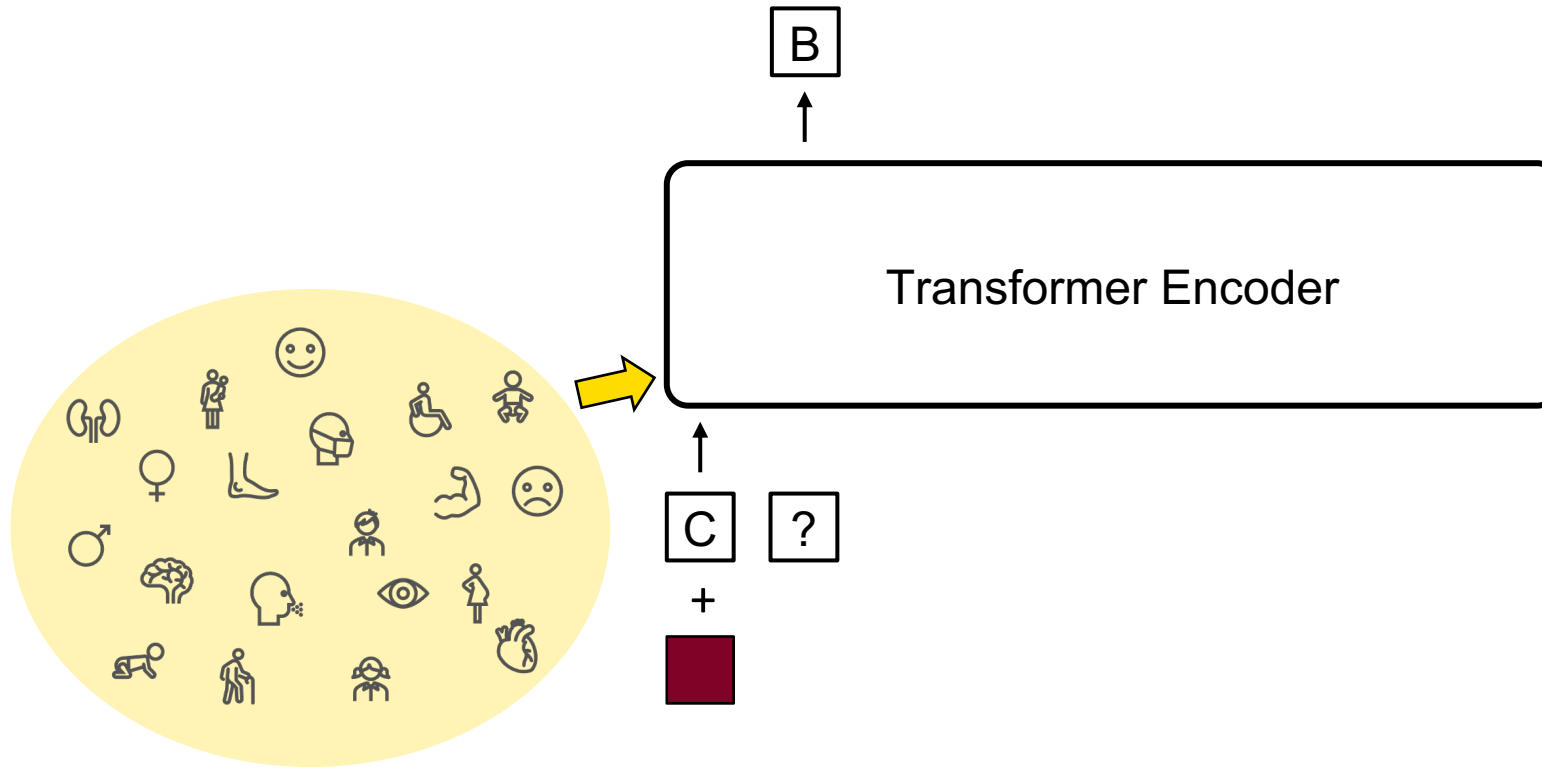


...



...

Learning the Representations

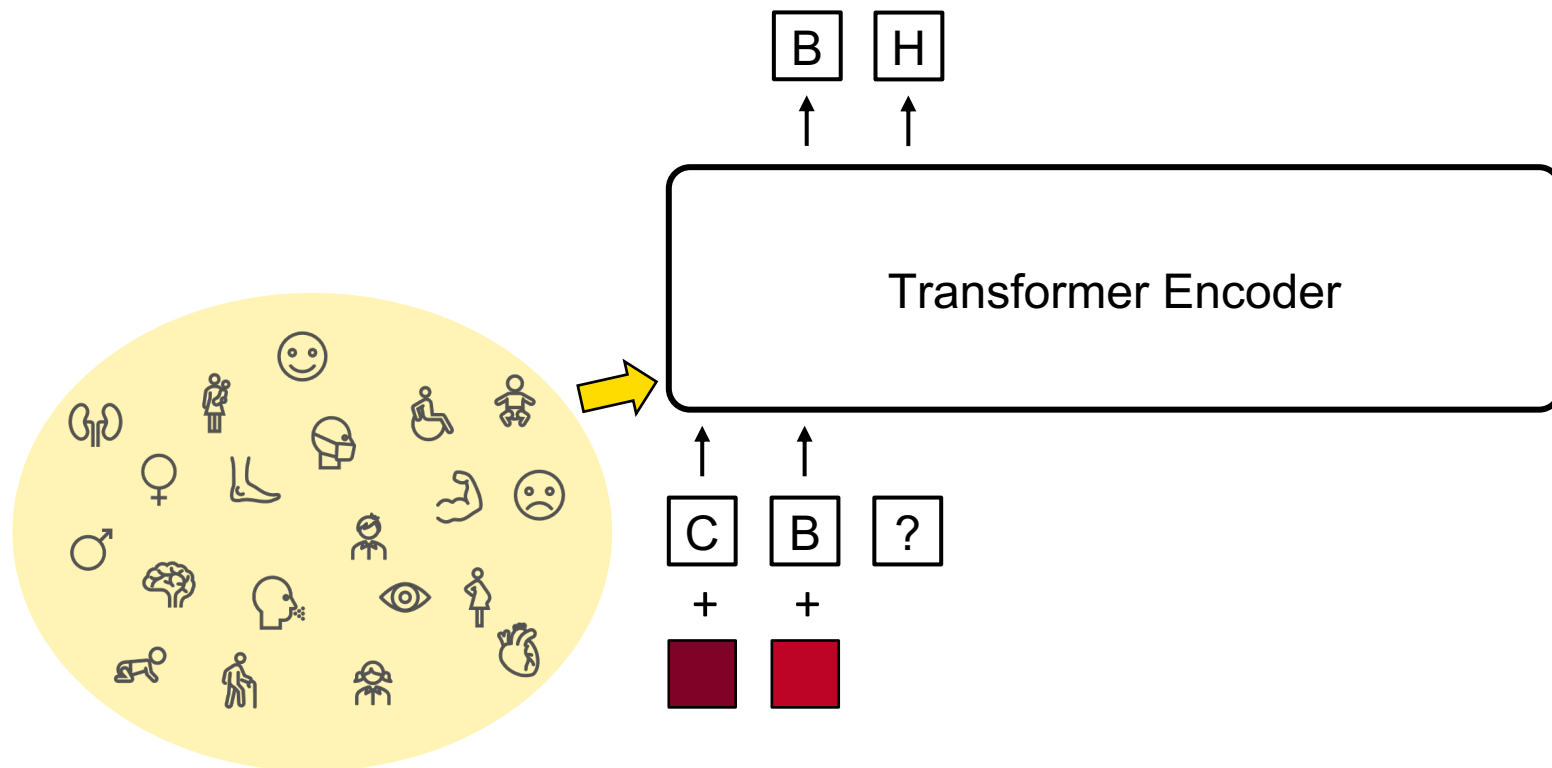


400K human bulk
RNA-seq samples

GPT Model

GeneRAIN

Learning the Representations

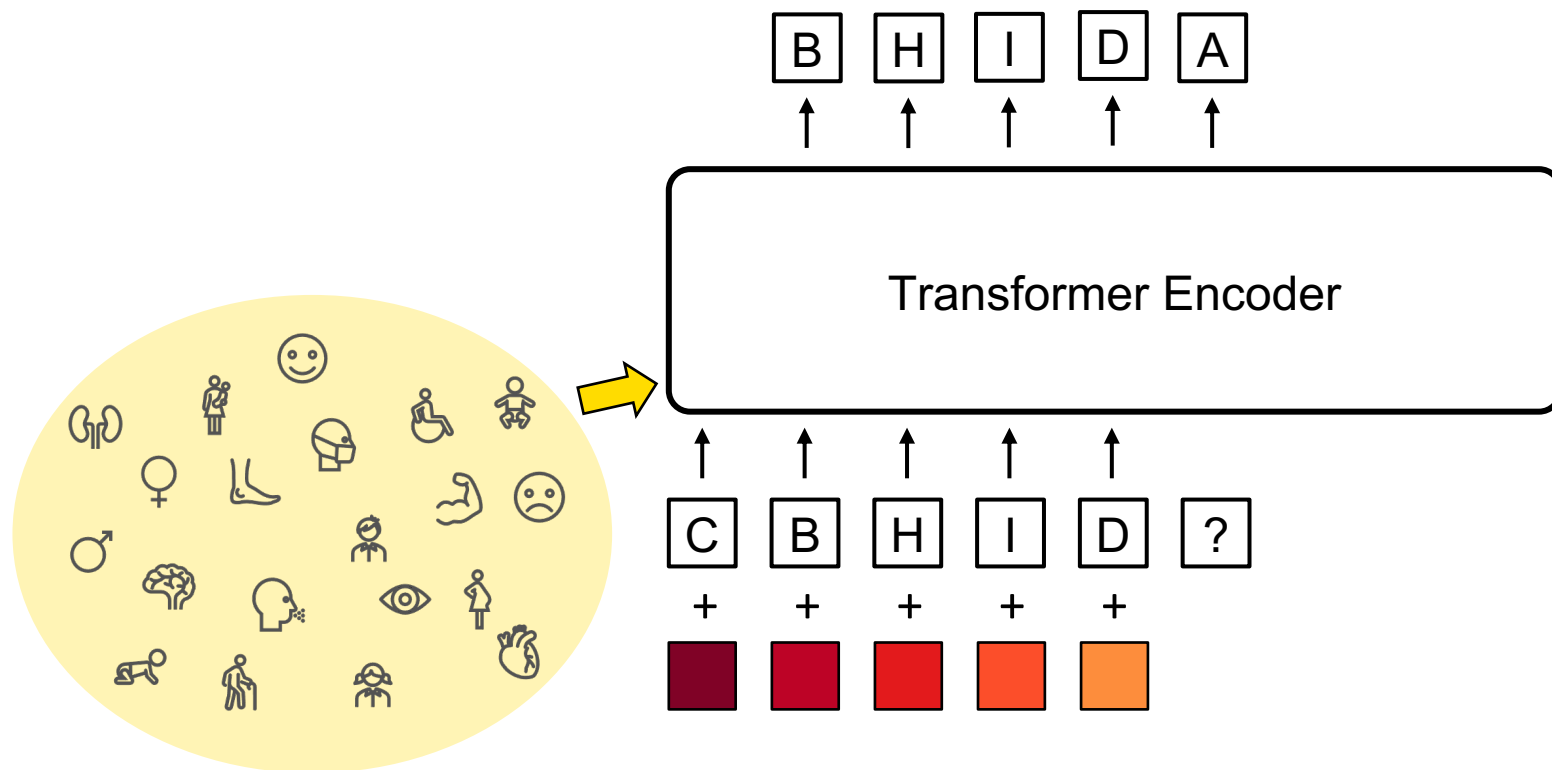


400K human bulk
RNA-seq samples

GPT Model

GeneRAIN

Learning the Representations



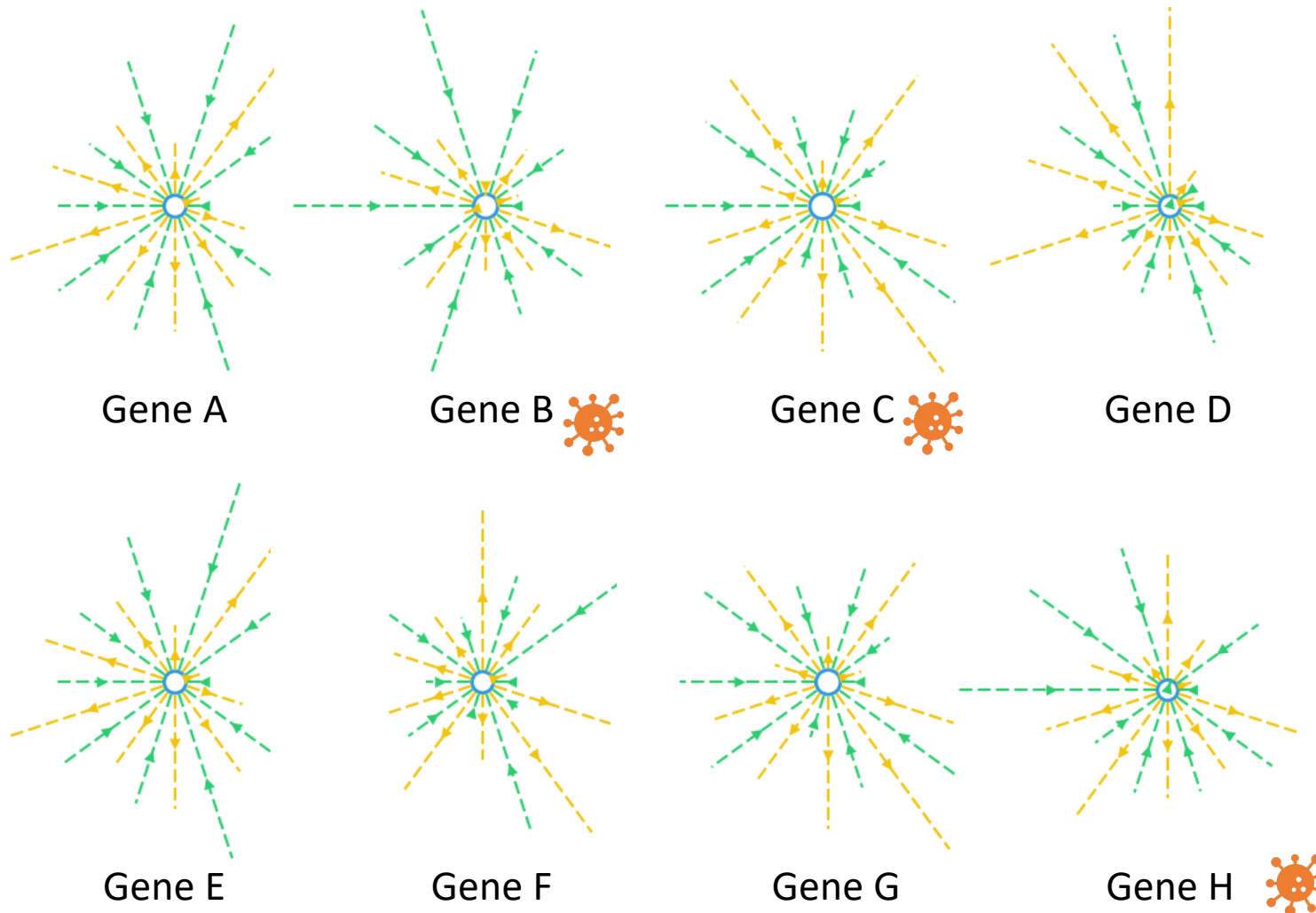
400K human bulk
RNA-seq samples

GPT Model

GeneRAIN

What Are Their Applications?

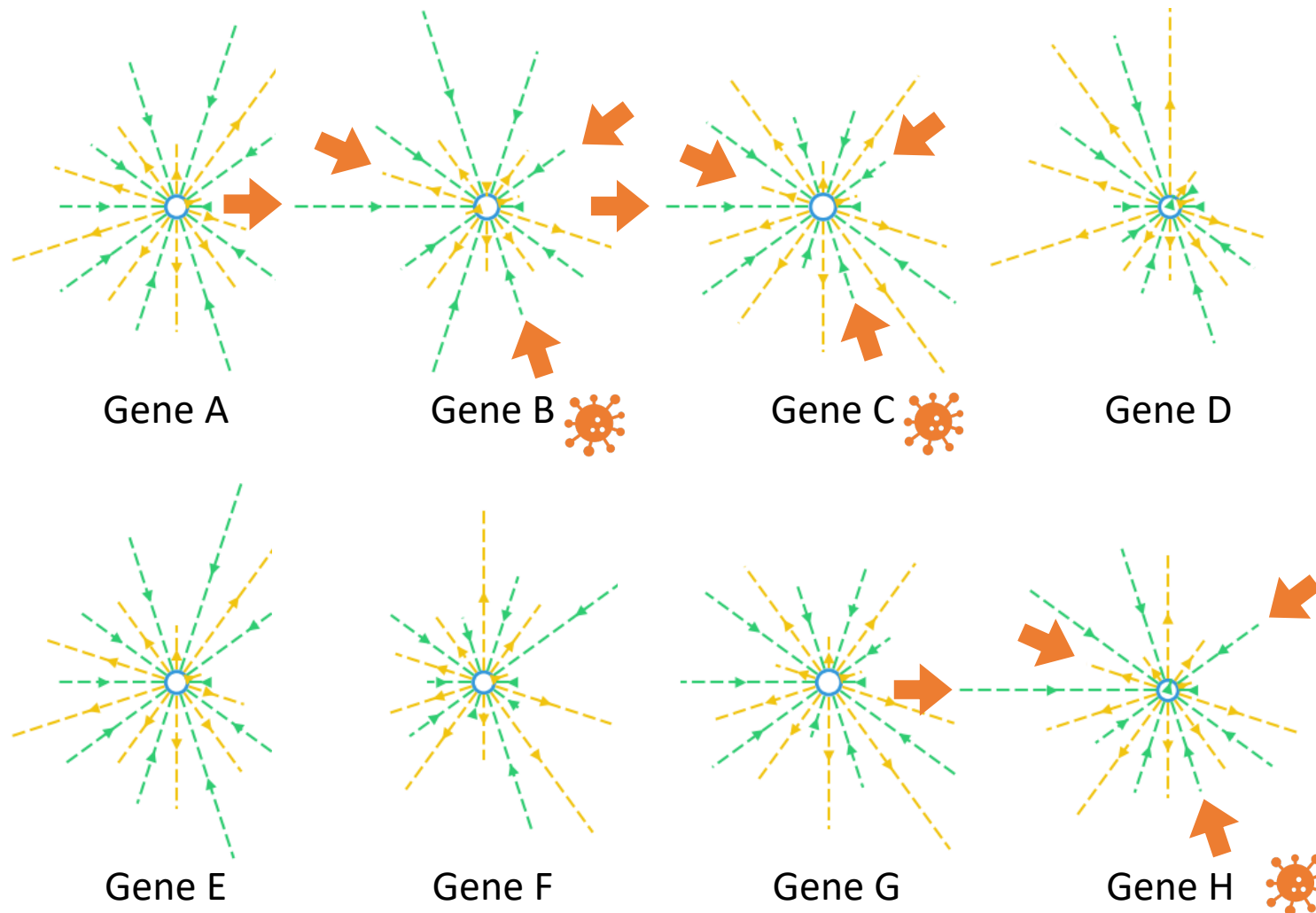
Use gene representation to predict biological attributes



Train on protein-coding genes

What Are Their Applications?

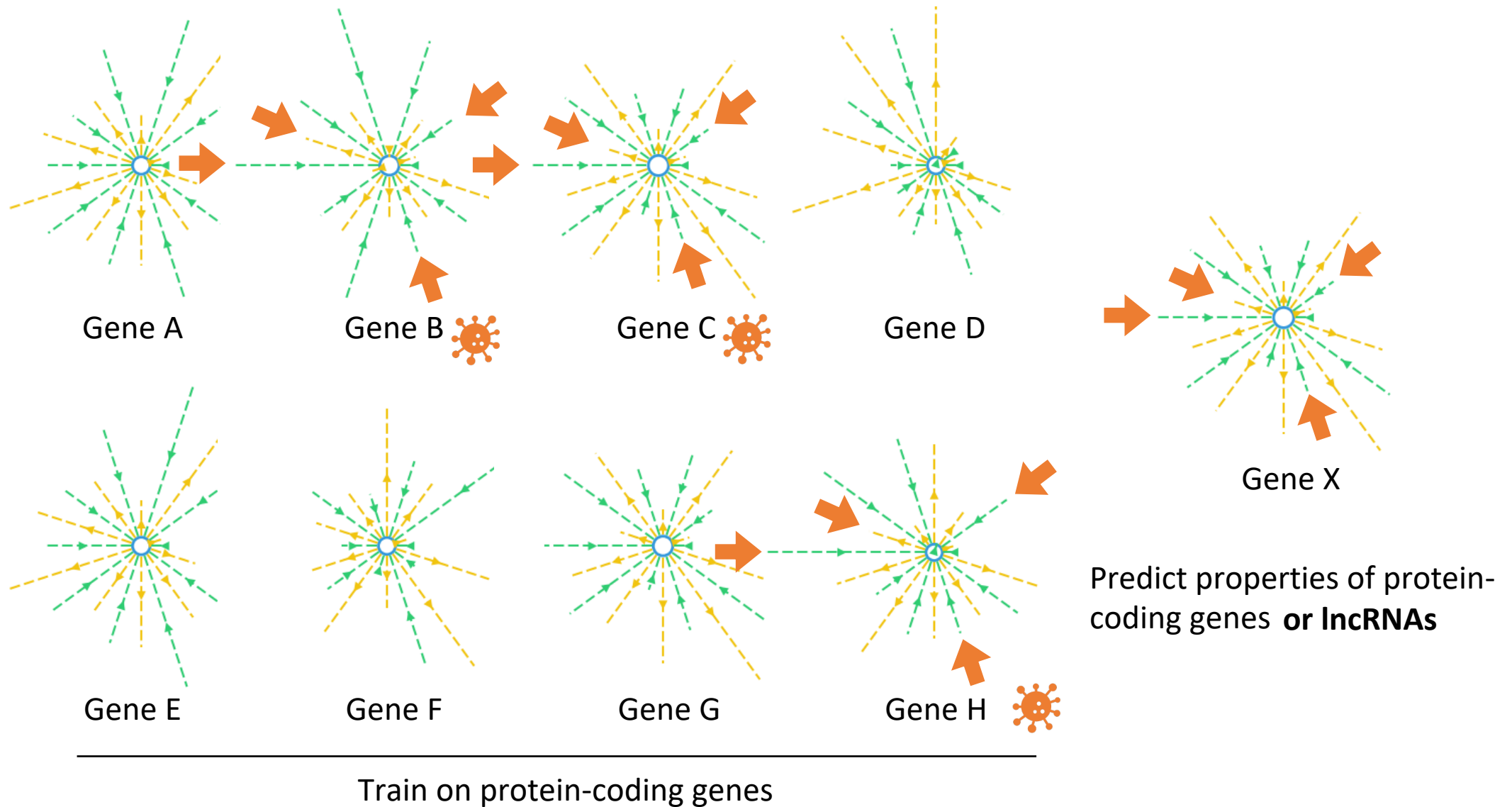
Use gene representation to predict biological attributes



Train on protein-coding genes

What Are Their Applications?

Use gene representation to predict biological attributes



Application Examples

- Trained ~5,000 classifiers to predict diverse gene properties
- Transferred knowledge from protein-coding genes to lncRNAs, made 62.5 million biological property predictions for 13,030 lncRNAs.
- Analyzed human-mouse gene similarity using 777K bulk RNA-seq samples



New Results

Follow this preprint

Multifaceted Representation of Genes via Deep Learning of Gene Expression Networks

Zheng Su, Mingyan Fang, Andrei Smolnikov, Marcel E. Dinger, Emily C. Oates, Fatemeh Vafaei

doi: <https://doi.org/10.1101/2024.03.07.583777>

1 0 0 0

Genes in Humans and Mice: Insights from Deep learning of 777K Bulk Transcriptomes

Zheng Su, Mingyan Fang, Andrei Smolnikov, Fatemeh Vafaei, Marcel E. Dinger, Emily C. Oates

doi: <https://doi.org/10.1101/2024.04.01.587517>



Published 2024 | Version v1

GeneRAIN



Zheng (John) Su

AI Engineer | Entrepreneur |
Bioinformatician | Full Stack Web D...



Word embeddings:
King - Man + Woman = Queen

Gene embeddings:
APOE - APP + BRCA1 = ?

Try out our web app! <https://generain-vec.streamlit.app/>

GeneRAIN-vec

Home

Similar Genes

Visualization

Calculator

Calculator Plus

Computing Similarity

Share ☆ ✎ ↻ ⋮



GeneRAIN-vec Gene Embedding Analysis Tools

This web application allows you to explore and analyze gene embeddings derived from the GeneRAIN model, a state-of-the-art deep learning approach for understanding gene relationships and functions.

What are Gene Embeddings?

Gene embeddings are vector representations of genes in a high-dimensional space. These embeddings capture complex relationships between genes based on their expression patterns. In our case, each gene is represented by a 200-dimensional vector.

About the GeneRAIN Models

GeneRAIN are transformer-based models trained on a large dataset of 410K human bulk RNA-seq samples. These embeddings are derived from the GPT protein-coding+lncRNA model, which uses a novel 'Binning-By-Gene' normalization method and a GPT (Generative Pre-trained Transformer) architecture to learn multifaceted representations of genes.