Supplementary Material for 'Spatial Transcriptomics Analysis of Zero-shot Gene Expression Prediction'

Yan Yang[1,2], Md Zakir Hossain[1,2,3], Xuesong Li[1,2], Shafin Rahman[4], and Eric Stone[1]

[1] Biological Data Science Institute,
The Australian National University, Canberra, Australia
`Yan.Yang@anu.edu.au`
[2] Agriculture and Food, CSIRO, Canberra, Australia
[3] School of Elec Eng, Comp and Math Sci, Curtin University, Perth, Australia
[4] Department of ECE, North South University, Dhaka, Bangladesh

1 Implementation Details

We implement SGN by using the *PyTorch Geometric* [1] framework. We train SGN respectively for 100 epochs and 300 epochs on the STNet dataset and 10xProteomic dataset with batch size 1, where a slide image contains up to thousands of windows in the two datasets. We use the learning rate 5×10^{-4} and weight decay 1×10^{-4}. Following [2], we use a four-layer GraphSAGE with hidden dimensions 512. For the gene type embedding, a two-layer ViT with hidden dimension 256 is used. Refer to `embedding.pdf` for embedding visualizations of seen and unseen gene types. The training can be done in 28 hours with an H100 GPU, and our model allows inference by a GPU with 16GB memory under half-precision. For training, we generate 10 functionality and phenotype descriptions for each gene type, using $\mathbf{P}(\mathsf{c}, \mathtt{false})$ and $\mathbf{P}(\mathsf{c}, \mathtt{true})$, and randomly sample generated descriptions in a batch. During testing, $\mathbf{P}(\mathsf{c}, \mathtt{true})$ is used. The prompt templates $\mathbf{P}(\mathsf{c}, \mathbb{1}^{\mathtt{int}})$ are described below.

- $\mathbf{P}(\mathsf{c}, \mathtt{false}) =$ '`### System: \n You are a biology scientist specialising in gene study. Your mission is to describe the functionality and phenotype of the gene provided by the gene symbol from the user. Your descriptions need to be concise and contain keywords only. ### User: \n gene symbol: c. ### Assistant: \n`'.
- $\mathbf{P}(\mathsf{c}, \mathtt{true}) =$ '`### System: \n You are a biology scientist specialising in gene study. Your mission is to describe the functionality and phenotype of the gene. The descriptions need to be concise and contain keywords only, providing the gene symbol and its summary as a helpful reference. Note that the reference most likely contains no information on the functionality and phenotype of the gene. You are encouraged to complement the missing information for the functionality and phenotype of the gene. Do not directly copy from the reference unless you think it is extremely necessary. ### User: \n gene symbol: c. Reference: summary ### Assistant: \n`'.

The summary is the reference retrieved from `mygene.info` for the gene type c.

> ### System: \n You are a biology scientist specialising in gene study. Your mission is to describe the functionality and phenotype of the gene provided by the gene symbol from the user. Your descriptions need to be concise and contain keywords only. ### User: \n gene symbol: CABP5. ### Assistant: \n

> CABP5, also known as Calcium-Binding Protein 5, is a calcium-binding protein involved in the modulation of intracellular calcium levels and signal transduction. This gene plays a role in various cellular processes and its phenotype includes its association with calcium-related functions and cellular communication.

Fig. 1: Example of generated description. The gray box and green box are inputs and outputs of LLM, respectively.

> ### System: \n You are a biology scientist specialising in gene study. Your mission is to describe the functionality and phenotype of the gene. The descriptions need to be concise and contain keywords only, providing the gene symbol and its summary as a helpful reference. Note that the reference most likely contains no information on the functionality and phenotype of the gene. You are encouraged to complement the missing information for the functionality and phenotype of the gene. Do not directly copy from the reference unless you think it is extremely necessary. ### User: \n gene symbol: CABP5. Reference: The product of this gene belongs to a subfamily of calcium binding proteins, which share similarity to calmodulin. Calcium binding proteins are an important component of calcium mediated cellular signal transduction. Expression of this gene is retina-specific. The mouse homolog of this protein has been shown to express in the inner nuclear layer of the retina, suggested its role in neuronal functioning. The specific function of this gene is unknown. [provided by RefSeq, Oct 2009]. ### Assistant: \n

> Functionality: CABP5 is a calcium binding protein related to calcium-mediated signaling, belonging to a subfamily sharing similarity to calmodulin. It is specifically expressed in the retina, where its role in neuronal functioning is suggested.\n \n Phenotype: The gene's expression is limited to the retina, indicating its importance in the visual system. Its protein product might play a role in neuronal functions, particularly in the inner nuclear layer of the retina. Further research is needed to understand its specific functions in calcium signaling and neuronal functioning.

Fig. 2: Example of generated description. The gray box and green box are inputs and outputs of LLM, respectively.

2 Functionality and Phenetotype Description

We provide examples of generated functionality and phenotype descriptions. We show LLM outputs for $\mathbf{P}(\texttt{c}, \texttt{false})$ in Fig. 1, and $\mathbf{P}(\texttt{c}, \texttt{true})$ in Fig. 2.

References

1. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. CoRR **abs/1903.02428** (2019), `http://arxiv.org/abs/1903.02428`
2. Yang, Y., Hossain, M.Z., Stone, E., Rahman, S.: Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. Pattern Recognition **145**, 109966 (2024). `https://doi.org/https://doi.org/10.1016/j.patcog.2023.109966`, `https://www.sciencedirect.com/science/article/pii/S0031320323006647`