Determining the Optimal Embedding Technique for Mapping Gene Expression Samples into a Distributed Ontology Space

Filip Jevtic[1],Christopher A Mancuso[1], Arjun Krishnan[1,2]

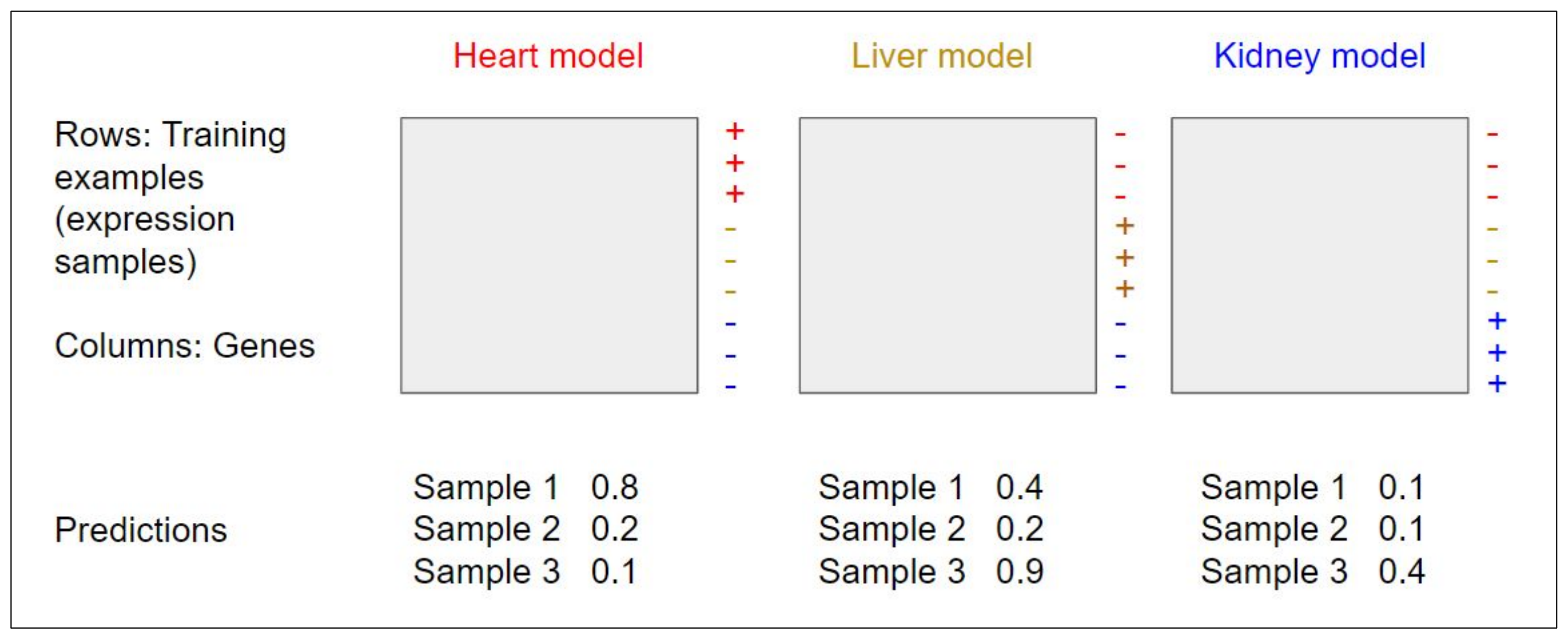
[1]Department of Computational Mathematics, Science and Engineering, [2]Department of Biochemistry and Molecular Biology

Overview

- There are more than 1 million publicly available human gene expression samples. These samples together constitute an extremely valuable resource that any researcher can use to gain new biological insights about genes and cellular mechanisms. However, most of these samples lack systematic annotation of the context in which they were generated including, tissue and cell type of origin as well as disease state. Developing computational methods that can automatically label a sample is a grand challenge in biomedical research.
- A new labeling method that develops upon previous methods has emerged, termed *OnClass*[1] , which uses a neural network to first generate a mapping between gene expression space to a label space that is created through a low dimensional embedding of an ontology. This mapping simultaneously captures the relationships of different terms in the ontology while also allowing predictions for terms that were never seen during training the model.
- While the *OnClass* method has shown state-of-the-art results for single cell prediction, there was little investigation into how the method works on other classification tasks or how to find the best hyperparameters.
- We investigated this question by developing software that can easily input new data, labels, hyperparameters and embeddings. The train, validation and test sets are produced in a rigorous way. The software is a work in progress but this poster highlights some of the features that will be released.

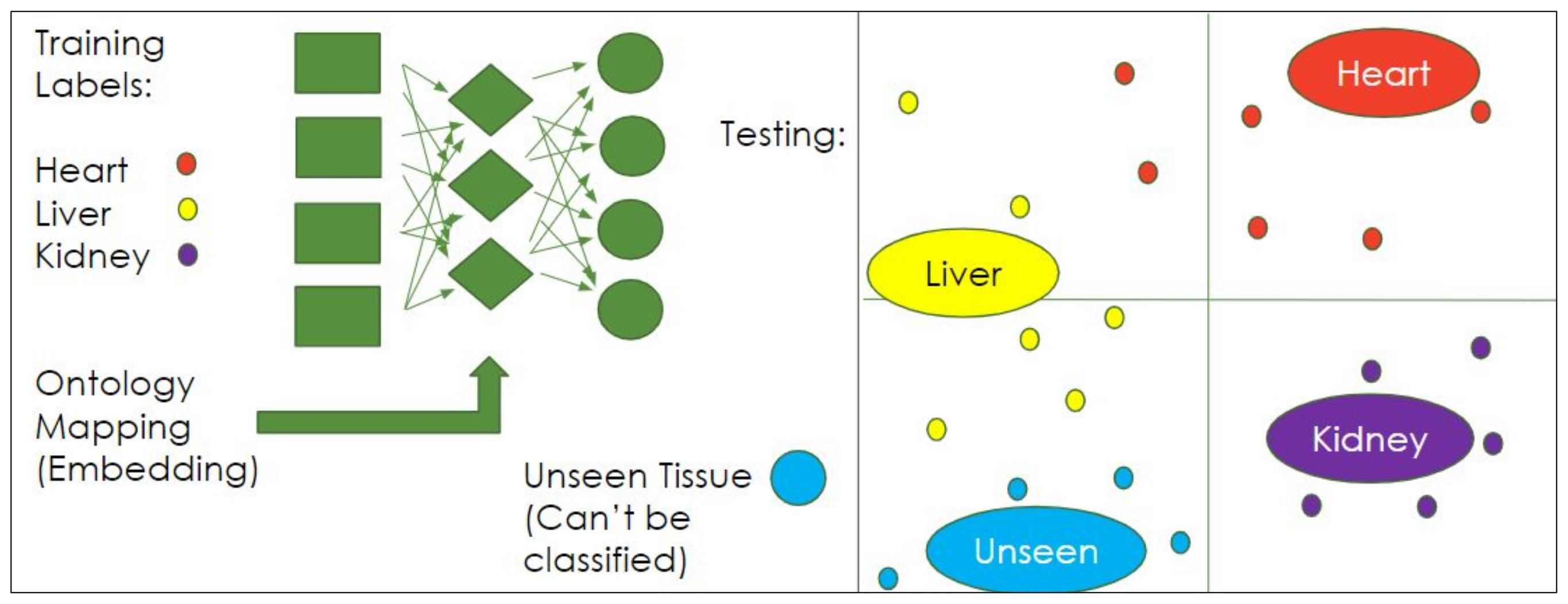
Previous Methods

Binary Classification

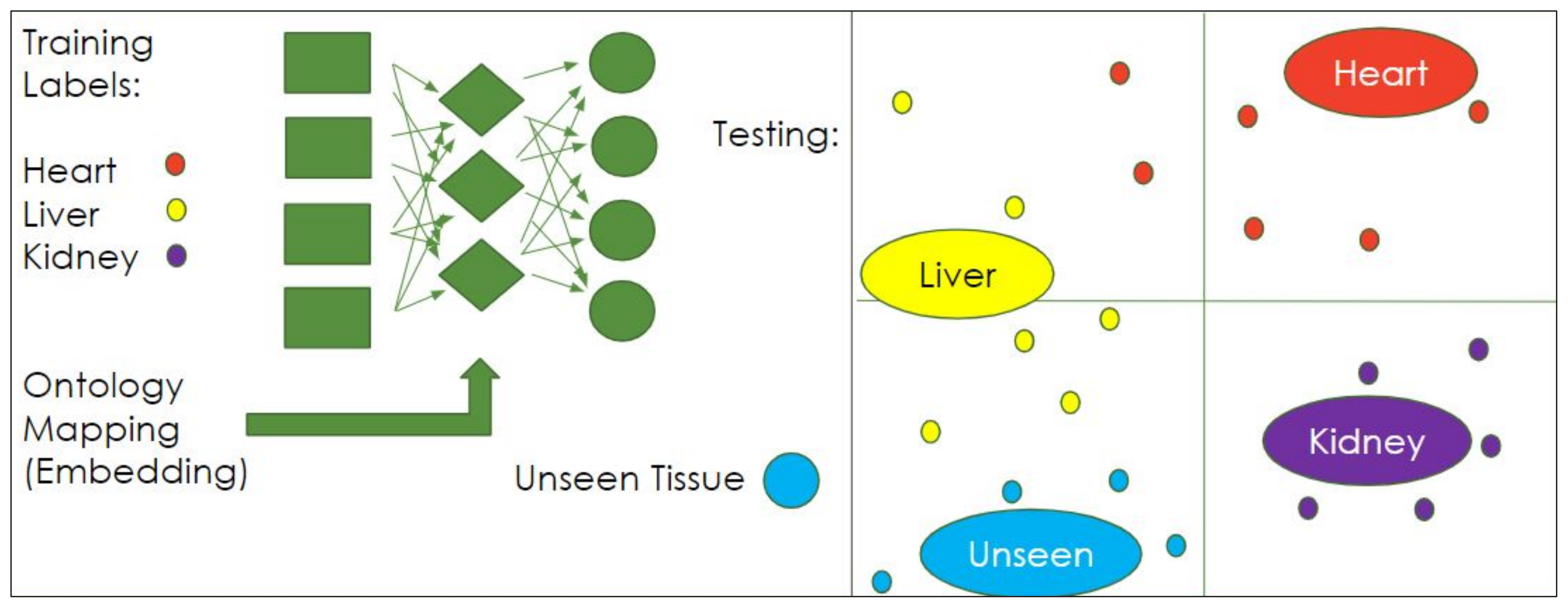


Heart model		Liver model		Kidney model	
Sample 1	0.8	Sample 1	0.4	Sample 1	0.1
Sample 2	0.2	Sample 2	0.2	Sample 2	0.1
Sample 3	0.1	Sample 3	0.9	Sample 3	0.4

Standard Multilabel Classification


Training Labels:
Heart
Liver
Kidney
Testing:
Ontology Mapping (Embedding)
Unseen Tissue (Can't be classified)
Heart
Liver
Kidney
Unseen

Predictions in Ontology Embedding Space



Method

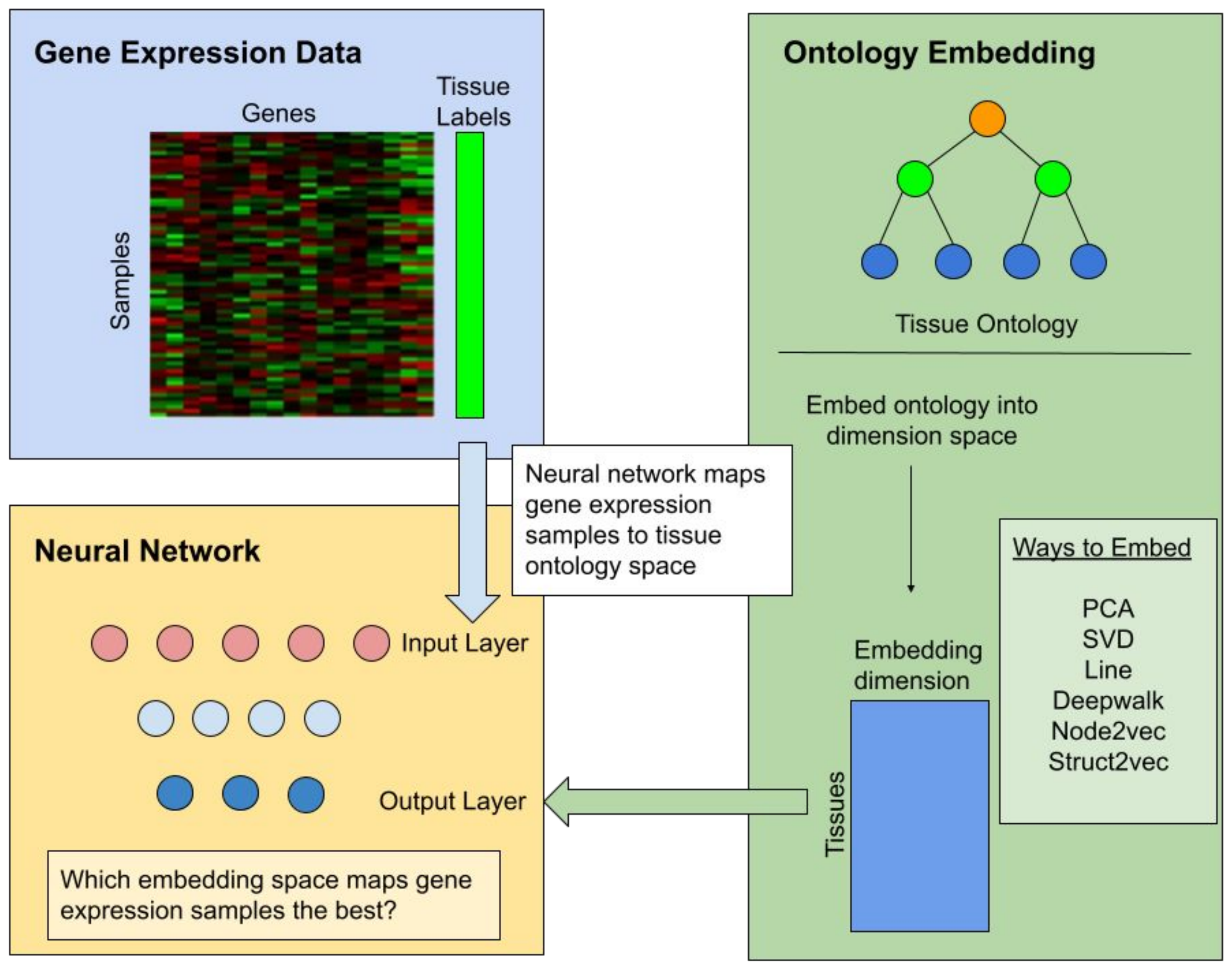


Figure 1. **Workflow for systematic evaluation of ontology embedding techniques.**

Parameters of the Model

Learning Rate : How much the model changes weights

Hidden Layer Units: The size of the hidden layer

Data Scaling: How or if the data in standardized

Training Label: How the training loss function is implemented

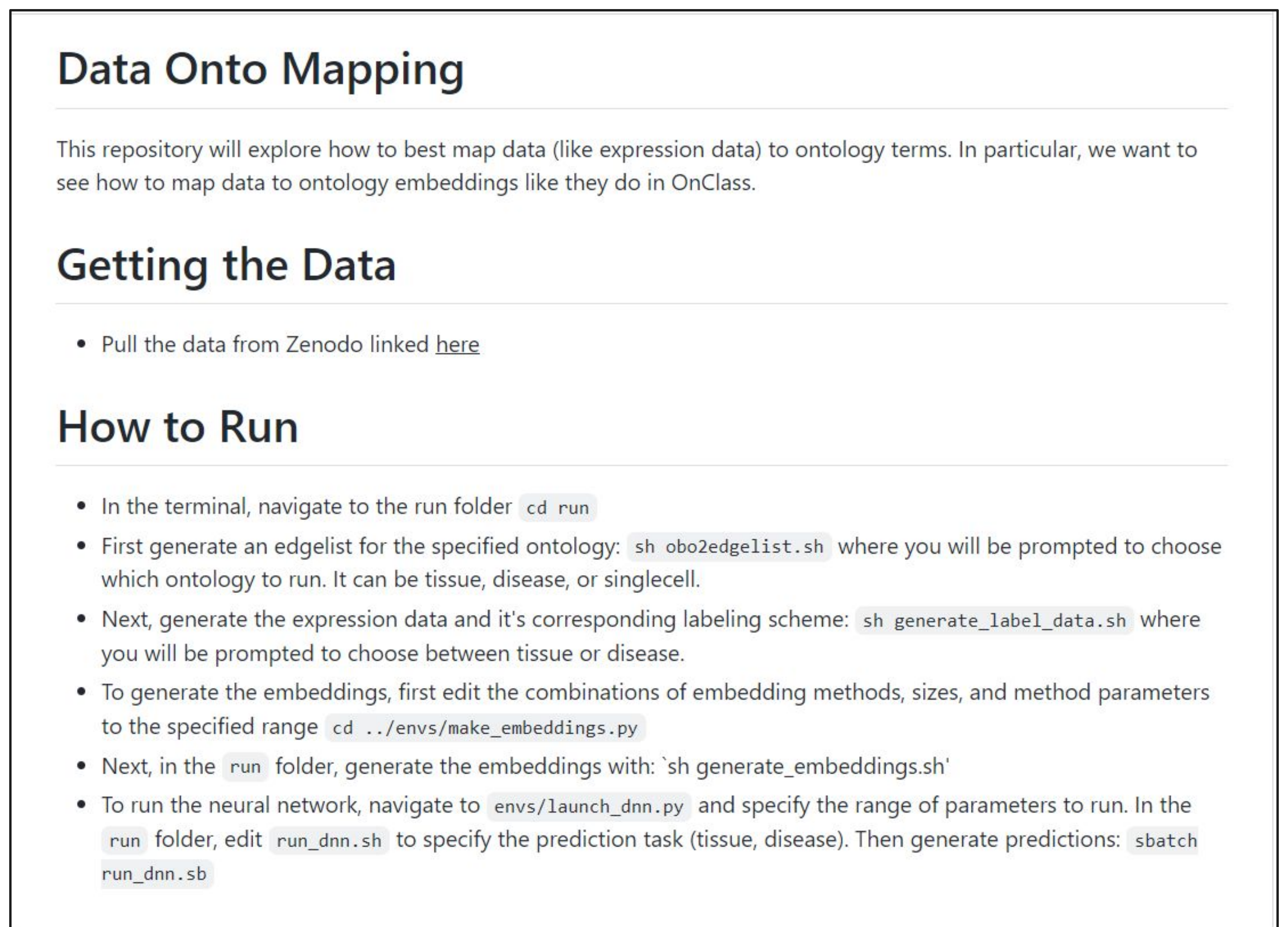
Embedding Method: How the ontology is embedded

Embedding Size: The size of the embedding

Embedding Parameters: a few examples are diffusion restart value, walk lengths, walk style …

Evaluation Scheme: How the predictions are evaluated ex. log2prior score, AUROC …

Code Base Instructions



Data Onto Mapping

This repository will explore how to best map data (like expression data) to ontology terms. In particular, we want to see how to map data to ontology embeddings like they do in OnClass.

Getting the Data

- Pull the data from Zenodo linked here

How to Run

- In the terminal, navigate to the run folder `cd run`
- First generate an edgelist for the specified ontology: `sh obo2edgelist.sh` where you will be prompted to choose which ontology to run. It can be tissue, disease, or singlecell.
- Next, generate the expression data and it's corresponding labeling scheme: `sh generate_label_data.sh` where you will be prompted to choose between tissue or disease.
- To generate the embeddings, first edit the combinations of embedding methods, sizes, and method parameters to the specified range `cd ../envs/make_embeddings.py`
- Next, in the `run` folder, generate the embeddings with: `sh generate_embeddings.sh'
- To run the neural network, navigate to `envs/launch_dnn.py` and specify the range of parameters to run. In the `run` folder, edit `run_dnn.sh` to specify the prediction task (tissue, disease). Then generate predictions: `sbatch run_dnn.sb`

Example Visualizations

Figure 2. **Best validation model displayed by tissue.**

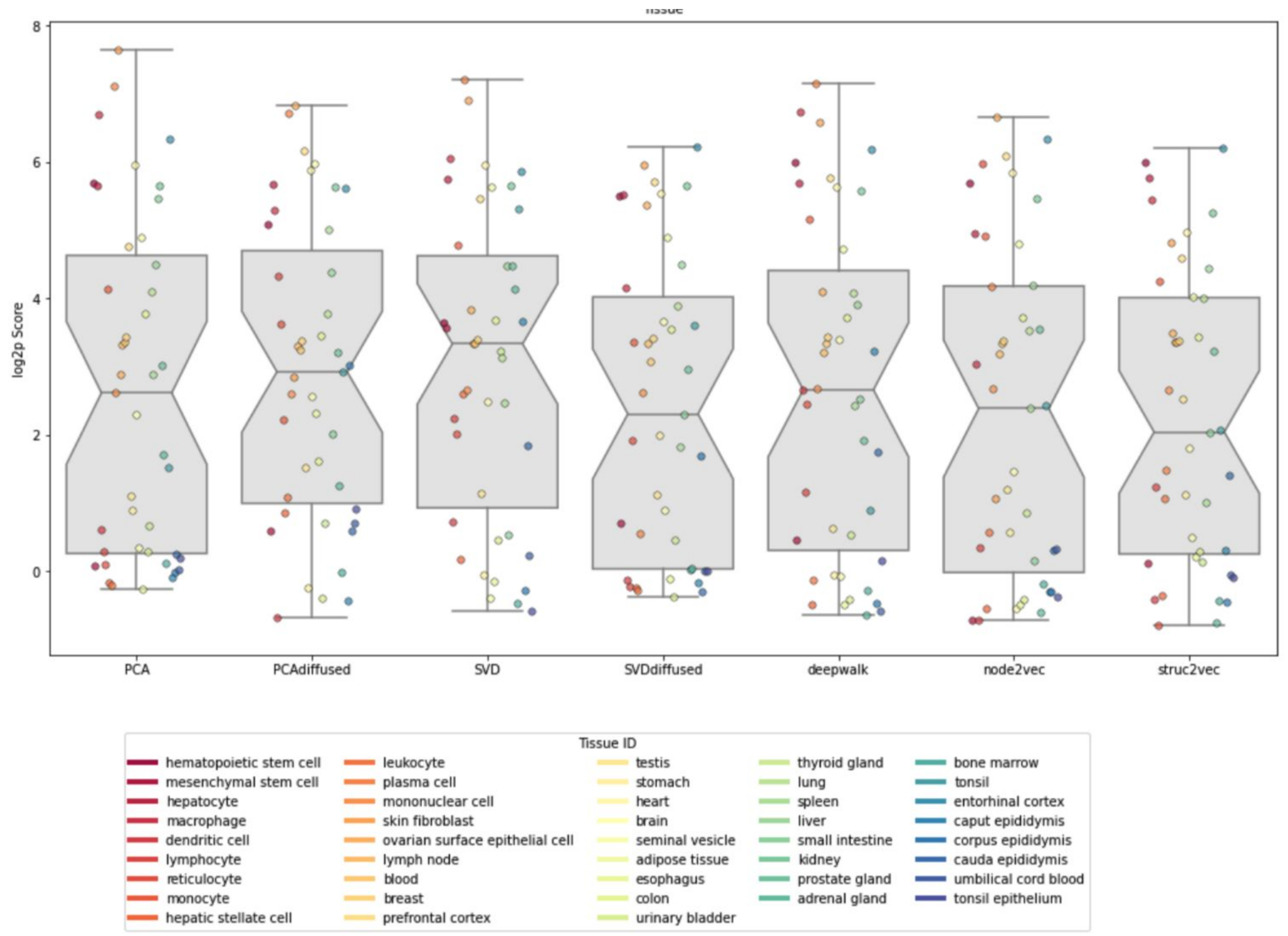


Figure 3. **Best validation model displayed by tissues seen/unseen in training.**

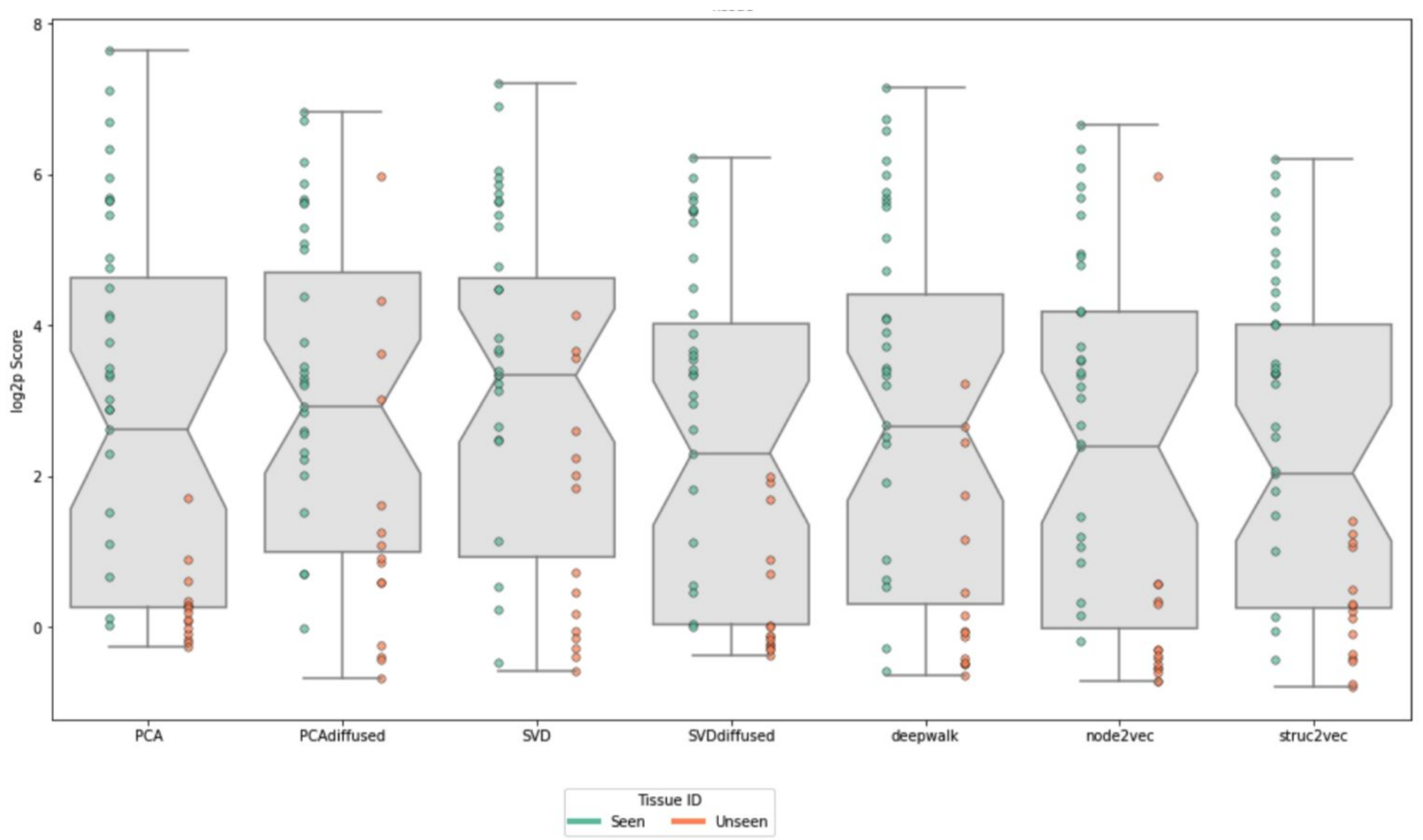
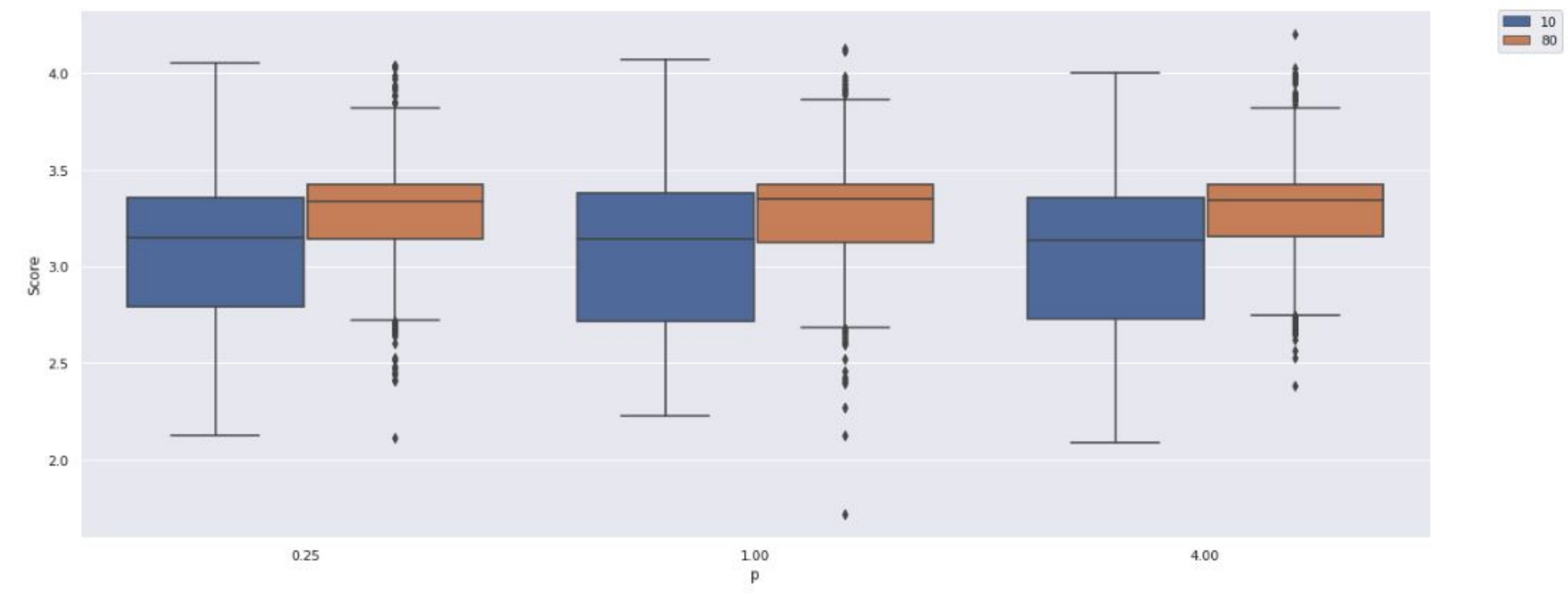


Figure 4. **Performance of node2vec Models Based on P Value, Walk Length and Walk Style**



Reference and Contact and Funding

[1]Wang et.al., (2021) Leveraging the Cell Ontology to classify unseen cell types. Nat. Comm. **12** 5556

Funding: NIH R35 GM128765 andNIH F32 F32GM134595

E-mail: jevticfi@msu.edu