MLFPM
Machine Learning Frontiers in Precision Medicine

Integrating prior knowledge in neural network methods to reduce the dimensions of single-cell RNA-Seq Data

Clinical Bioinformatics Area
Fundación Pública Andaluza Progreso y Salud

By
Pelin Gundogdu[2]

Advised by
Joaquín Dopazo[1,2,3,5], Isabel A. Nepomuceno-Chamorro[4], and Carlos Loucera[2,3]

[1] Bioinformatics in Rare Diseases (BiER), Centro de Investigaciones Biomédicas en Red en Enfermedades Raras (CIBERER), 41013 Sevilla, Spain
[2] Clinical Bioinformatics Area, Fundacioń Progreso y Salud (FPS), Hospital Virgen del Rocío, 41013 Sevilla, Spain
[3] Computational Systems Medicine Group, Institute of Biomedicine of Seville (IBIS), Hospital Virgen del Rocío, 41013 Sevilla, Spain
[4] Department of Computer Languages and Systems, University of Sevilla, Avd. Reina Mercedes s/n, Seville, 41012 Spain
[5] Functional Genomics Node, FPS/ELIXIR-ES, Hospital Virgen del Rocío, 41013 Sevilla, Spain

June 30th, 2020

Junta de Andalucía
Consejería de Salud y Familias
FUNDACIÓN PROGRESO Y SALUD

UNIVERSIDAD DE SEVILLA

MARIE CURIE ACTIONS

Contents

1 Introduction

High-throughput technology have revolutionised the research in the area of biology and bio-medicine. RNA sequencing (RNA-seq) give the opportunity to analyse the entire transcriptomes. However, RNA-seq data represent an average of gene expression values across thousand to millions of cells, i.e., is typically performed in "bulk" [1]. Single-cell RNA-seq (scRNA-seq) overcome this issue by isolating single cells, allowing biological research to be carried out with an unprecedented resolution.

With developed technology, researchers can analyse information from scRNA-seq and it enables the significant increase in the number of studies in recent years. scRNA-seq technology, which profiles the transcriptome of individual cells, allows new discoveries, for example, to understand which cell-specific changes in transcriptome are significant, the identification of new markers for specific cell type, the level of heterogeneity within a population of cells [2].

Single cell RNA-seq data provides valuable insights into cellular heterogeneity which may significantly improve our understanding of biology and human disease [3]. One of the main application is the ability to identify new cell types and cell states [4] [5]. This application has new computational challenges or key questions to address, such as, how to determine the similarity from expression profiles of cells or which cell types have important role in diseased individuals.

Clustering analysis of gene expression data is one of the most common methods used to solve the cell type profiling problem. The idea consists of finding the closest cell/gene group. In addition, the dimensionality reduction is performed to reduce the noise. One of the most popular methods to solve the problem is Principal Component Analysis (PCA), a method that performs a linear reduction of the dimension. However, it lacks the ability to capture the complex patterns behind single cell data which can lead to poor performance or misleading interpretations.

The goal of this project is to integrate several types of prior biological information in neural network architectures as a way to reduce the dimension of data in a supervised framework. Then, we used the *learned* representation by the intermediate layer in the neural network for performing a clustering analysis of unseen cell types. The learned architectures, of a reduced dimension, capture biological aspects of the data which can be provided of biological meaning through the use of the *learned* biological information. For this project we have used the following sources of biological information: protein-protein interactions, protein-DNA interactions and pathway information.

The neural network finds complex patterns that approximate the underlying function of the data, such as the biological system, resulting in high performances across several problems. However, although the advantage of using such a high performance method is unquestionable, the neural networks lack interpretability (the so called *black box* paradigm), a must in the biomedical sciences. These issues might cause irrelevant and misleading solutions for some problems.

We integrate the biological information which are based on (*pathways, protein-protein* and *protein-DNA interaction*) networks to create an interpretable neural network. We also used this design as a way of dimensionality reduction. The reason is that in order to detect cell types it is easier when using reduced expression profiles, since the dimensionality reduction filters most of the noise. After reducing the dimension we perform a clustering analysis of the learned representation by detaching the output layer of the NN.

This report have sections which explain about the details of dataset and biological information *(Section 2)*, method have been implemented *(Section 3)*, architecture and parameter *(Section 4)*, the evaluation metrics *(Section 5)*, the performance evaluation *(Section 6)* and the result of each experiment *(Section C)*.

2 Materials

2.1 Datasets

Single cell gene expression data from several mouse tissue sites were downloaded from Gene Expression Omnibus (GEO) [6] database. We focus the present report on the 402 samples that form the basis of the clustering study in [7] : a combination of three datasets which profiled 16 different cell types across several tissues. In order to properly compare our method with the methods proposed in the reference paper [7] we follow the same splitting criteria (see Section 3).

The dataset, which we use to test our cell type prediction and clustering methods, has expression values for 9437 mouse genes. However, each experiment uses a different subset of the genes, according to the biological prior employed by the method. Note that we have used the same prepossessing and zero-imputation steps that those used in the reference paper.

To obtain the biological priors we have gathered the gene lists from several sources, namely: the physiological signalisation pathways subset used by the Hipathia method [8], and the metabolism and signalisation pathways as extracted by the GeneSFC tool [9]. Both methods extract the biological information from the KEGG database [10], although the GeneSFC can access other databases, such as Gene ontology [11], KEGG [10], REACTOME [12] and NCG [13], we have only used KEGG for the purposes of this work.

Source	# of genes	Normalization method
Paper [7]	9437	MinMaxScaler / StandardScaler
hiPathia	1843	MinMaxScaler / StandardScaler
geneSCF	4248	MinMaxScaler / StandardScaler

Table 1: The details about gene numbers of each datasets

2.2 Sources of biological information

In order to have a fair comparison between our proposal and the one provided in [7] we have rebuilt the architecture based on the protein-protein (PPI) and protein-DNA interaction (TF) data, and tested them using the same environment and data splits as our proposed architectures.

Table 2 contains a summary of the different sources of biological information used in this paper. Note that in all cases we need to take into account the fact that not all the genes are present in the set of biological units used, when this happens we did not include these genes except when using *normal non-biological* nodes.

As Table 2 shows, our proposed sources of information are more sparse, fewer nodes and genes involved, while being more easily interpretable, due to the fact that they refer to curated biological networks that trigger specific functions. On the contrary, the PPI approach is based on data-driven clustering methods without a clearly defined function, although they can be inferred using enrichment analysis [7].

Biologicial information	# of nodes	# of gene (all)	# of gene (using)
Protein-protein	348	3553	3553
Protein-DNA	348	8307	8307
hiPathia	142	4057	1843
geneSCF	333	8751	4248

Table 2: Pathway details

3 Methods

3.1 Neural Network Design

In this work we have used six types of neural network architectures. All architectures have an input layer, one or two hidden layers and an output layer. Nodes encode a value between 0 and 1, and it represents its activation, whereas the input layer contains the gene expression values. Finally, the output layer contains the probability of a sample for each class in the model. Hidden layers are located between input and output layer and by propagating the signal the weights are updated at each training epoch. At the end, the network *learns* an internal representation of the underlying function of the data.

The neural network model is formulated as follows:

$$x^{(i)} = a(W^{(i)}x^{(i-1)} + b^{(i-1)})$$

where $x^{(i)}$ denotes the activation score in ith hidden layer, a is the activation formula, b is bias value and W is the weight matrix (the *edges* of the neural network). We used *tanh* activation formula in each hidden layer and *softmax* activation formula in output layer, as have been empirically proven to be superior for the problem at hand [7].

$$tanh(x) = \frac{1 - e^{(-2x)}}{1 + e^{(-2x)}}$$

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Note that the given representation is used to solve a (classification) supervised problem, where the outputs to be learn (the classes) are the different cell types. To do so, we minimize the *cross entropy loss* as is typically done in the literature. However, our main interest lies in solving an unsupervised problem: are we able to identify cell types not seen during the training phase? To solve this problem, we are mostly interested in the *learned* representation of the problem. The *cross entropy loss* is defined as:

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

where M refer to the total number of cell types, y is a binary indicator if cell type c is the correct label for sample o and p is the predicted probability of observation o for cell type c.

3.2 Biological Knowledge Integration

For the purposes of our neural network design, we summarise each biological information unit (PPI/TF cluster, signaling/metabolic pathway) as a cluster of related genes. As has been mentioned before, we aim to integrate these biological knowledge into the neural network design. To achieve this purpose, we use the information provided by pathway, protein-protein and protein-DNA interaction in a two-step process:

1. Each biological unit is represented by one node of the first inner layer.
2. A gene from input layer is only connected to those nodes that summarise biological units that contain the given gene. Such that, if a gene is not related with one node, we fix the weight that connects them to 0.

3.3 Clustering Analysis

For this analysis, we performed a set of experiments in which we left out 2, 4, 6 or 8 random cell types of the 16 types in our input dataset. Then, we generated the clusters for left out cell type data using the *Kmeans* algorithm over the reduced space provided by the last hidden layer of the network and use metrics to compare the clustering results with the true cell types. To compare the clustering results with test data, the adjusted random index (ARI) is used. This index is one of the most frequently used metric for clustering validation. It is a measure of the similarity between two data clustering.

As has been mentioned above, the network is trained in a fully supervised way using the samples which correspond to the non left out cell types. Then, the last hidden layer is extracted after the training has concluded. Thus, we are interested in how the NN *learns* the latent structure of the data.

4 Architectures and Parameters

All the proposed architectures differ on the biological knowledge integrated into first the hidden layer and the number of hidden layers. There are total of three architectures included, with and without biological knowledge, and either 1-layer or 2-layer options. Because of all architectures have been trained under the same environment and using the same training and validation splits, all of them are comparable with each other.

In Table 3 we provide the parameters common to all the proposed architectures, whereas in Table 4 we provide a summary of the different experiments carried out in this work. The most remarkable finding of our designs is that the architecture that only uses the signaling pathway is approximately five times smaller than the ones used in [7], which leads to more interpretable models (while achieving a similar performance across all the experiments).

Parameter name	**Parameter value**
epochs	20
batch_size	10
kernel_initializer	glorot_uniform
activation	tanh (hidden layer) / softmax (last layer)
bias_initializer	zeros (for pathway design in first layer)
input_dim	depends on experiments

Table 3: Defined values for each parameters

No	Architecture	# of input	# of node (Layer 1)	of node (Layer 2)	# of parameters (*)
1	P1	9437	100 dense	X	945.416
2	P1	1843	100 dense	X	186.016
3	P1	4248	100 dense	X	426.516
4	P2	9437	100 dense + 142 pathway	X	2.287.884
5	P2	9437	100 dense + 333 pathway	X	4.093.598
6	A1	1843	142 pathway	X	264.136
7	A1	4248	333 pathway	X	1.420.261
8	A2	1843	142 pathway	100 dense	277.764
9	A2	4248	333 pathway	100 dense	1.449.933
10	B1	1843	142 pathway + 627 ppi/tf	X	1.430.356
11	B1	4248	333 pathway + 693 ppi/tf	X	4.375.906
12	B2	1843	142 pathway + 627 ppi/tf	100 dense	1.496.652
13	B2	4248	333 pathway + 693 ppi/tf	100 dense	4.463.790

Table 4: The details of input, hidden layer and parameters for each architecture

(*) Number represents fully connected neural network design

4.1 Design with Dense (Architectures P1 and P2)

This architecture, a fully connected NN, is created based on paper [7]. As we have seen before, the only way to use all the available genes consists in using uninformative nodes disconnected from any prior biological knowledge, so we can connect one such node with all the genes at the input layer (a classical dense layer).

This architecture (Figure 1) has total of three layers which are input, hidden and output layer. It uses the full set of genes (9437) as the input layer. This design has two options: fully connected *(dense)* (P1) and fully connected with pathway information, thus mixing uninformative nodes with *biological-based* ones (P2). Moreover, we train *Architecture - P1* with two other gene lists created by combining the genes across all the signaling and metabolic pathways.

4.2 Design with Pathway (Architectures A1 and A2)

In this architecture (Figure 2), we integrate only *biological-based* nodes by adding the *pathway* information. The idea behind this design consists of improving the interpretability of the architecture designed in [7] by using a smaller network based on *pathway* information, which is more biologically informative.

Each architecture uses the common gene lists which are the intersection gene list, derived from signaling and/or metabolic pathways, with the default gene list (9437 genes). Each node in first hidden layer represents one pathway, while the connection between genes and node is removed or defined in regards to pathway definition (as outlined before). Moreover, this design has 1-layer *(Architecture - A1)* or 2-layer *(Architecture - A2)* architecture, to see the effect of increasing the deepness of the NN.

4.3 Design with Pathway and PPI/TF (Architectures B1 and B2)

In this architecture (Figure 3), the *pathway with PPI/TF* effect has been performed. The idea in this design is performing the same result obtained by paper architecture [7] but using *pathway with PPI/TF* information with reduced input size.

Each architecture uses the common gene lists. Each node in first hidden layer represents one pathway, PPI or TF interaction. The connection between genes and node is removed or defined in regards to pathway, PPI or TF definition.

Moreover, this design has 1-layer *(Architecture - A1)* or 2-layer *(Architecture - A2)* architecture, to see the effect of dimensionality reduction.

5 Evaluation Metric

In this project, we used six metrics to score the results while comparing the true labels and clustering results. The idea is to perform unsupervised clustering of cells using unseen data. We perform several experiments in which we left out some number of cell types randomly. Then, we cluster the left out data using as input of the clustering model the reduced space provided by the NN architecture, i.e., we use the values computed by the last hidden layer.

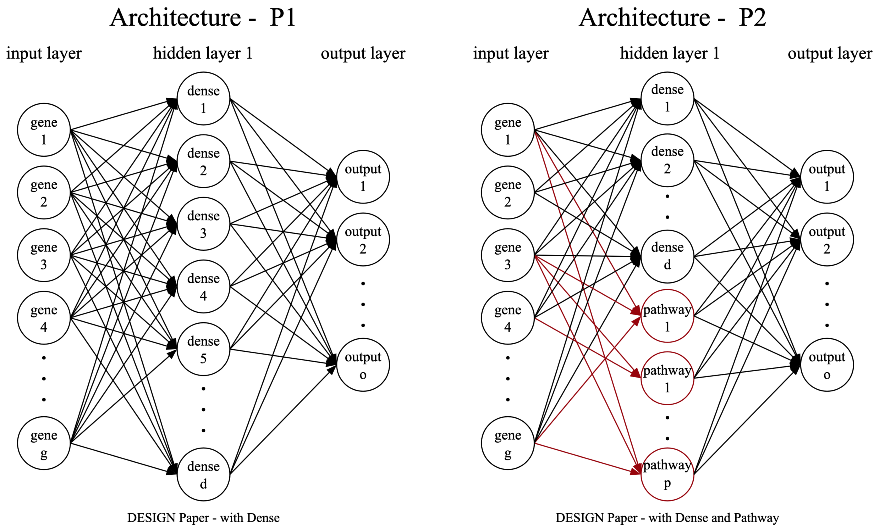

Figure 1: *(left)*Architecture P1 [7], *(right)* Architecture P2

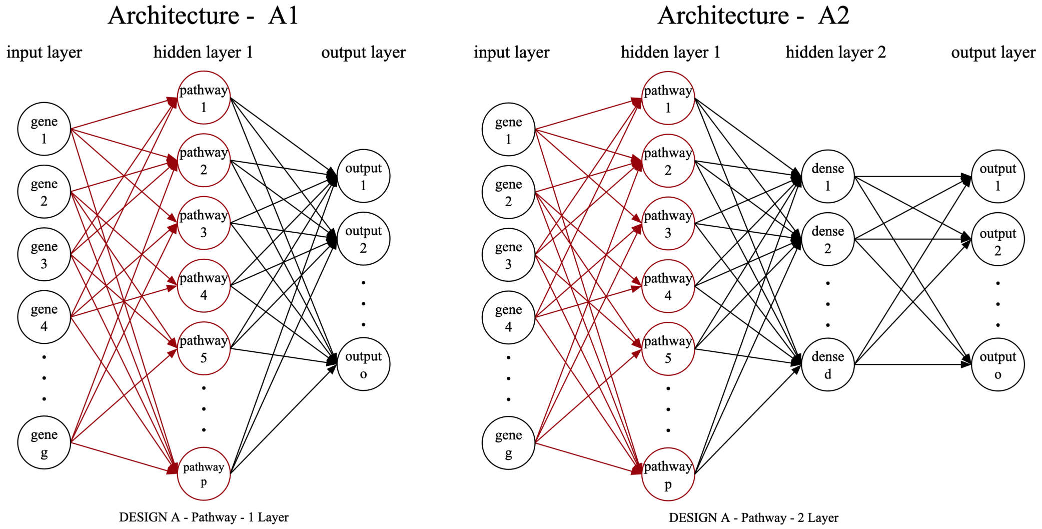

Figure 2: Architecture A - with signaling and metabolic pathway

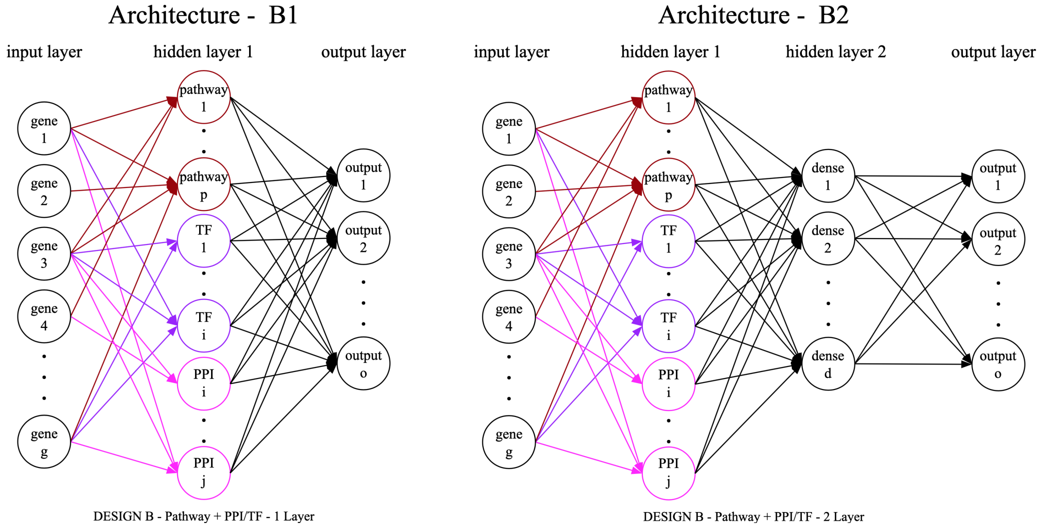

Figure 3: Architecture B - with signaling and metabolic pathway and PPI/TF

The metrics we used are homogeneity, completeness, v-measure, adjusted rand index (ARI), adjusted mutual information (AMI) and fowlkes-mallows score. For comparison purposes we resale the scores to $[0, 100]$.

5.1 Homogeneity Score

A clustering result satisfies homogeneity if ***all of its clusters*** contain ***only data points*** which are members of ***a single class*** [14]. It's bounded between 0 and 1, with low values indicating a low homogeneity.

5.2 Completeness Score

A clustering result satisfies completeness if ***all the data points*** that are members of a given class are ***elements of the same cluster*** [14]. Completeness is symmetrical to homogeneity. Increasing the completeness of a clustering solution often results in decreasing its homogeneity.

5.3 V-measure Score

V-measure score is computing the ***weighted harmonic mean of homogeneity and completeness*** [14]. The computations of homogeneity, completeness and V-measure are completely independent of the number of classes, the number of clusters, the size of the data set and the clustering algorithm used.

5.4 Adjusted Rand Index (ARI)

Rand Index is a function that computes a similarity measure between two clustering. For this computation rand index considers all pairs of samples and counting pairs that are assigned in the similar or different clusters in the predicted and true clustering.

ARI uses when the ground truth clustering has large ***equal sized clusters***. The adjusted version, which is the one used in this work, is the classical *Rand Index Score* adjusted for chance, in such a way that an score near zero corresponds with *almost random labeling.*

5.5 Adjusted Mutual Information (AMI)

Mutual Information is a function that computes the agreement of the two assignments. AMI uses when the ground truth clustering is ***unbalanced*** and there exist small clusters.

5.6 Fowlkes-Mallows Score

The Fowlkes-Mallows function measures the similarity of two clustering of a set of points. The Fowlkes-Mallows index (FMI) is defined as the geometric mean between of the precision and recall between two clustering systems.

6 Results and Discussion

Each experiments shown in Table 4 is performed for two different scalers, which are $[0, 1]$ scaling and data standardization. The evaluation score for each experiment and clustering analysis is provided in this section. Each experiment is performed 20 times for randomly selected left out cell, leaving out 2, 4, 6 or 8 cell types of the 16 cell types available.

The training, testing and evaluation time is approximately *70 minutes* for *Figure 1* architecture and *100 minutes* for each architecture shown in *Figure 2*, and *Figure 3* with default parameters, shown is *Table 3.* The clustering evaluation is almost 5 minutes for each experiment.

Note that each experiment is designed with same data, i.e., gene expression values as input, using the same sample-wise partitions defined by the leave p-groups out methodology. However, the architecture differs according to the input size and the biological priors as has been mentioned in section 3.

6.1 A Comparison of the Biological Information Priors Used

Figure (Figure 4) summarises the performance of the proposed architectures according to the biological priors used, see Table 1 for a summary. The figure shows the score distributions for all the metrics aggregated by the validation split carried out: randomly selected left out cell types (which 2, 4, 6, 8 of 16 cell types). Label *paper* refers to designs without any biological priors. As can be seen, using biological priors achieve similar performance as the more *black box* approach with uninformative nodes.

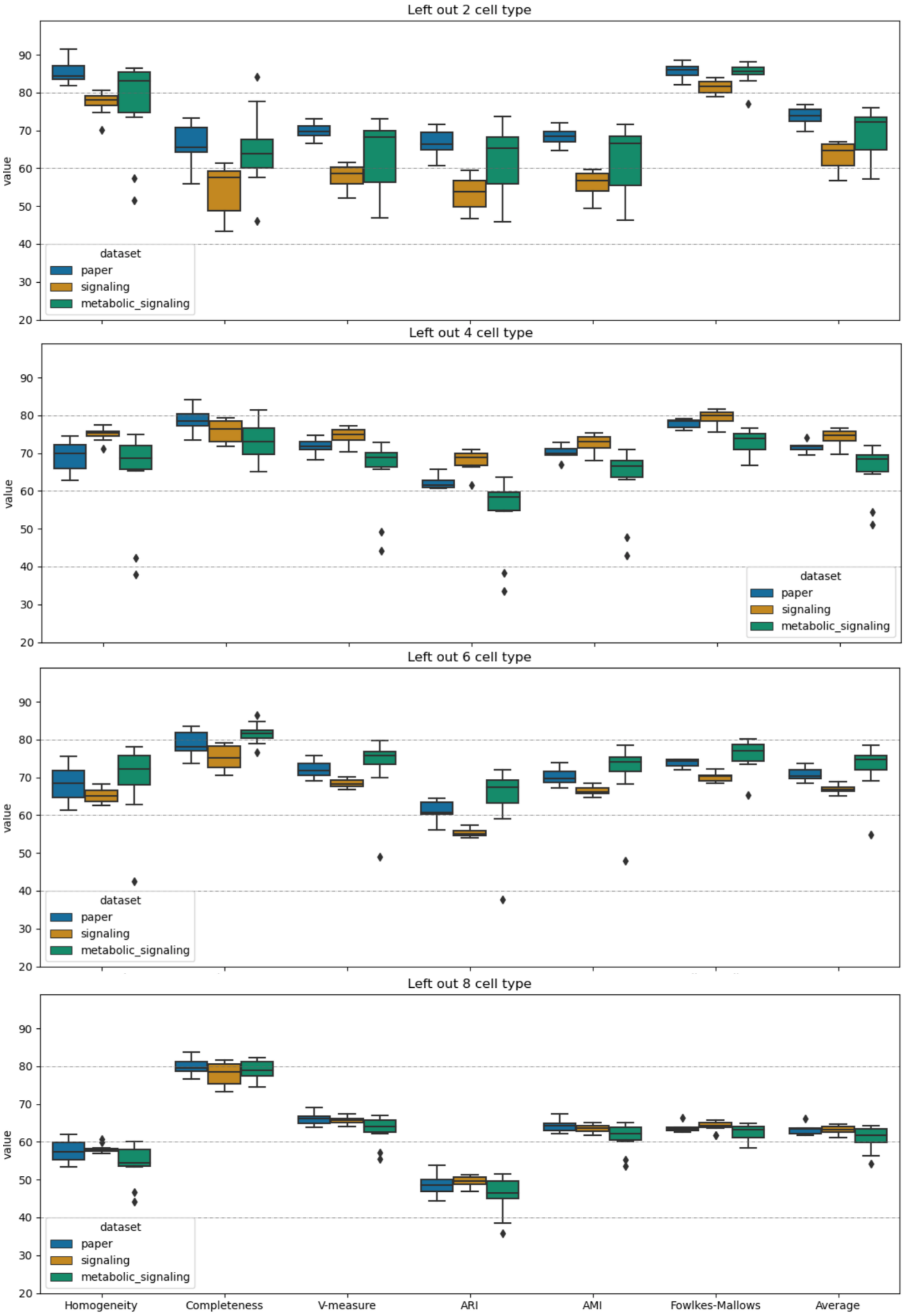

Figure 4: Comparison of different gene list

Figure Figure 5 shows a more fine grained view of the differences between using the pathway information (signalisation plus metabolism) or all the information available (PPI, signal, TF and metabolism) as priors to the NN design. Note that adding PPI/TF-based information does not seem to add value to the pathways-based design, as there are not many differences across the scores evaluated.

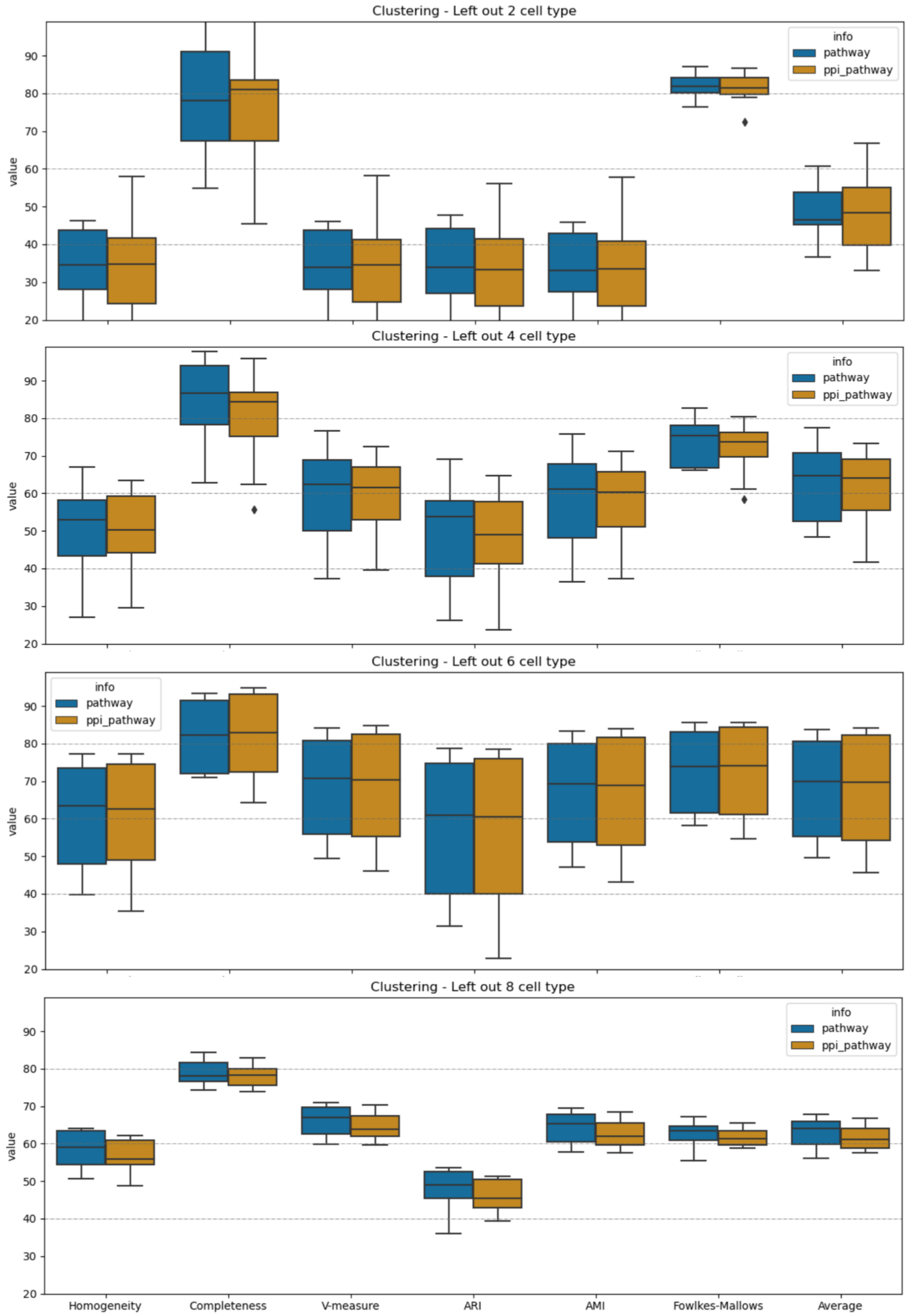

Figure 5: Performance according to used biological information

6.2 Left Out 4 Cell Types

Once the results are calculated, we can compare the score according to the input gene list and biological information for two scalers. The detail information about experiments can be seen in Table 4. Here we provide Tables 5 and 6, a summary of the leave 4 cell types out experiments, as an example of a minimal analysis of the results. At the appendices we provide the supplementary tables that en summarise the rest of the experiments.

There are a total of 13 designs, 10 of which integrate biological information and 3 without. The results show that the experiments based exclusively on the signalization pathways priors return comparable scores, sometimes better, than the ones based on other sources of biological information (B1, B2, A2) and those that lack the biological priors (P1). Thus, using a very sparse network with more interpretable nodes can lead to clusterings in line with those using more biological information or *black box* designs.

Architecture	# of node (Layer 1)	Homogeneity	Completeness	V-measure	ARI	AMI	Fowlkes-Mallows	Average
P1	100 dense	67,63	80,66	71,84	63,20	70,10	78,78	72,03
P1	100 dense	**74,51**	78,23	**75,34**	**68,79**	**73,50**	**80,29**	**75,11**
P1	100 dense	72,40	70,03	69,92	60,25	67,54	74,75	69,15
P2	100 dense + 142 pathway	65,25	**84,04**	70,77	60,62	69,42	78,35	71,41
P2	100 dense + 333 pathway	62,86	79,83	68,34	60,65	66,92	78,59	69,53
A1	142 pathway	75,33	**79,33**	76,42	**70,97**	74,74	**81,58**	76,40
A1	333 pathway	74,94	69,55	70,96	59,68	68,56	74,31	69,67
A2	142 pathway	**77,50**	78,48	**77,21**	70,59	**75,31**	80,81	**76,65**
A2	333 pathway	68,08	65,17	65,81	54,59	62,95	70,80	64,56
B1	142 pathway + 627 ppi/tf	74,87	77,66	75,33	69,84	73,51	**80,94**	75,36
B1	333 pathway + 693 ppi/tf	73,50	74,01	72,80	63,63	70,88	76,63	71,91
B2	142 pathway + 627 ppi/tf	**75,93**	**78,53**	**76,49**	**69,90**	**74,60**	80,55	**76,00**
B2	333 pathway + 693 ppi/tf	70,70	67,77	68,26	55,73	65,64	71,13	66,54

Table 5: The performance of left out 4 cell type with SS scaling

Architecture	# of node (Layer 1)	Homogeneity	Completeness	V-measure	ARI	AMI	Fowlkes-Mallows	Average
P1	100 dense	72,21	73,54	71,84	61,87	69,79	76,06	70,89
P1	100 dense	**75,25**	74,12	73,67	**66,77**	71,52	78,18	73,25
P1	100 dense	69,44	72,24	69,28	58,33	66,81	73,52	68,27
P2	100 dense + 142 pathway	74,44	**77,20**	**74,81**	65,70	**72,91**	**79,21**	**74,05**
P2	100 dense + 333 pathway	72,05	77,18	73,39	61,32	71,61	76,35	71,98
A1	142 pathway	**75,61**	75,29	**74,48**	**68,89**	**72,57**	**79,67**	**74,42**
A1	333 pathway	66,99	76,75	70,27	59,56	68,24	75,33	69,52
A2	142 pathway	73,51	72,04	72,09	66,33	70,04	78,28	72,05
A2	333 pathway	42,35	**80,23**	49,22	38,28	47,64	69,15	54,48
B1	142 pathway + 627 ppi/tf	71,07	71,83	70,37	61,48	68,01	75,51	69,71
B1	333 pathway + 693 ppi/tf	65,30	76,30	68,33	58,33	66,36	75,92	68,42
B2	142 pathway + 627 ppi/tf	**76,20**	72,63	**73,33**	**67,08**	**71,21**	**78,86**	**73,22**
B2	333 pathway + 693 ppi/tf	37,89	**81,51**	44,26	33,57	42,81	66,83	51,14

Table 6: The performance of left out 4 cell type with MMS scaling

7 Conclusions and Future Work

We analysed a novel method for accurately and efficiently clustering scRNA-seq data while overcoming issues related to noise values [7]. This novel method are based on deep neural networks constrained by PPI Network. The main advantage of this proposal is that they can learn the importance of several combinations of gene expression values for determining cell types and improve benchmark methods when the values of hidden layers are used for clustering analysis.

In this report, we explored the deep neural networks constrained by several sources of prior biological information, as signaling and metabolic pathways. The most remarkable finding of our designs is that the architecture that only uses the signaling pathway priors, which is approximately five times smaller than the ones used in [7], leads to comparable results while being a more interpretable model. This is due to the fact that signaling pathways are more curated and trigger specific functions, and that the smaller the network the easier to interpret.

While the results are encouraging, there are several research lines as future work which require deeper analysis: we want to make more testing moreover, to perform retrieval analysis. In addition, using the best hyper-parameter in neural network can significantly improve the performance. For this purpose, we are considering to implement hyper-parameter tuning.

Furthermore, we want to add a (*functional*) biological interpretation of the NN *learned* representation. This can lead to valuable biological findings of the single cell landscape across several tissues.

As a side note, the author of this manuscript, with advisor help, has been researching other machine learning-based methods for studying the signalization landscape in single cell studies during her enrolling in the Single Cell Signaling in Breast Cancer Challenge [15] where our team scored a good result.

8 Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813533.

References

[1] T. K. Olsen and N. Baryawno, "Introduction to single-cell rna sequencing," *Current Protocols in Molecular Biology*, vol. 122, no. 1, p. e57, 2018.

[2] H. Lab, "Introduction to single-cell RNA-seq,"

[3] P. Angerer, L. Simon, S. Tritschler, F. A. Wolf, D. Fischer, and F. J. Theis, "Single cells make big data: New challenges and opportunities in transcriptomics," *Current Opinion in Systems Biology*, vol. 4, pp. 85 – 91, 2017. Big data acquisition and analysis • Pharmacology and drug discovery.

[4] J.-F. Poulin, B. Tasic, J. Hjerling Leffler, J. Trimarchi, and R. Awatramani, "Disentangling neural cell diversity using single-cell transcriptomics," *Nature Neuroscience*, vol. 19, pp. 1131–1141, 08 2016.

[5] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres, and S. R. Quake, "A survey of human brain transcriptome diversity at the single cell level," *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, pp. 7285–7290, 2015.

[6] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, pp. 207–210, 01 2002.

[7] C. Lin, S. Jain, H. Kim, and Z. Bar-Joseph, "Using neural networks for reducing the dimensions of single-cell RNA-Seq data," *Nucleic Acids Research*, vol. 45, pp. e156–e156, 07 2017.

[8] M. R. Hidalgo, C. Cubuk, A. Amadoz, F. Salavert, J. Carbonell-Caballero, and J. Dopazo, "High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes," *Oncotarget*, vol. 8, no. 3, pp. 5160–5178, 2017.

[9] S. Subhash and C. Kanduri, "Genescf: a real-time based functional enrichment tool with support for multiple organisms," vol. 17, 2016.

[10] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 01 2000.

[11] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, pp. D330–D338, 11 2018.

[12] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, "The reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, pp. D498–D503, 11 2019.

[13] D. Repana, J. Nulsen, L. Dressler, M. Bortolomeazzi, S. K. Venkata, A. Tourna, A. Yakovleva, T. Palmieri, and F. D. Ciccarelli, "The network of cancer genes (ncg): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens," *Genome Biology*, vol. 20, p. 1, Jan 2019.

[14] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (Prague, Czech Republic), pp. 410–420, Association for Computational Linguistics, June 2007.

[15] "Single Cell Signaling in Breast Cancer Challenge - syn20366914."

Appendix A Performance of Architecture A, B and Design Paper

This section represents the performances of architectures and included biological information.

Scaler Performance

In our experiments, we used different scaler for each datasets. The comparison of performance is follows;

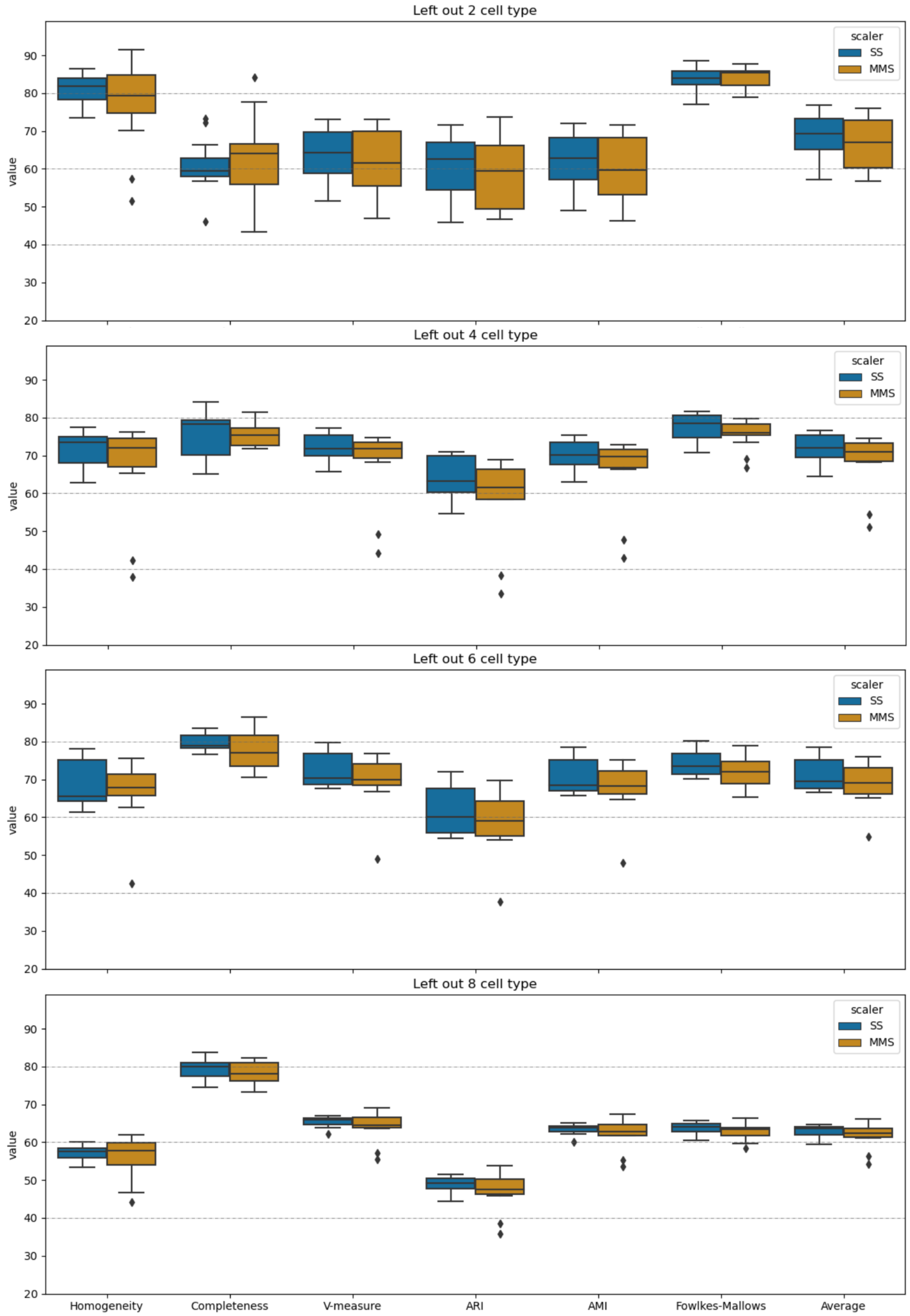

Figure 6: Performance score for different scaler

Biological Knowledge Performance

The biological information made with using pathway, PPI and TF information.

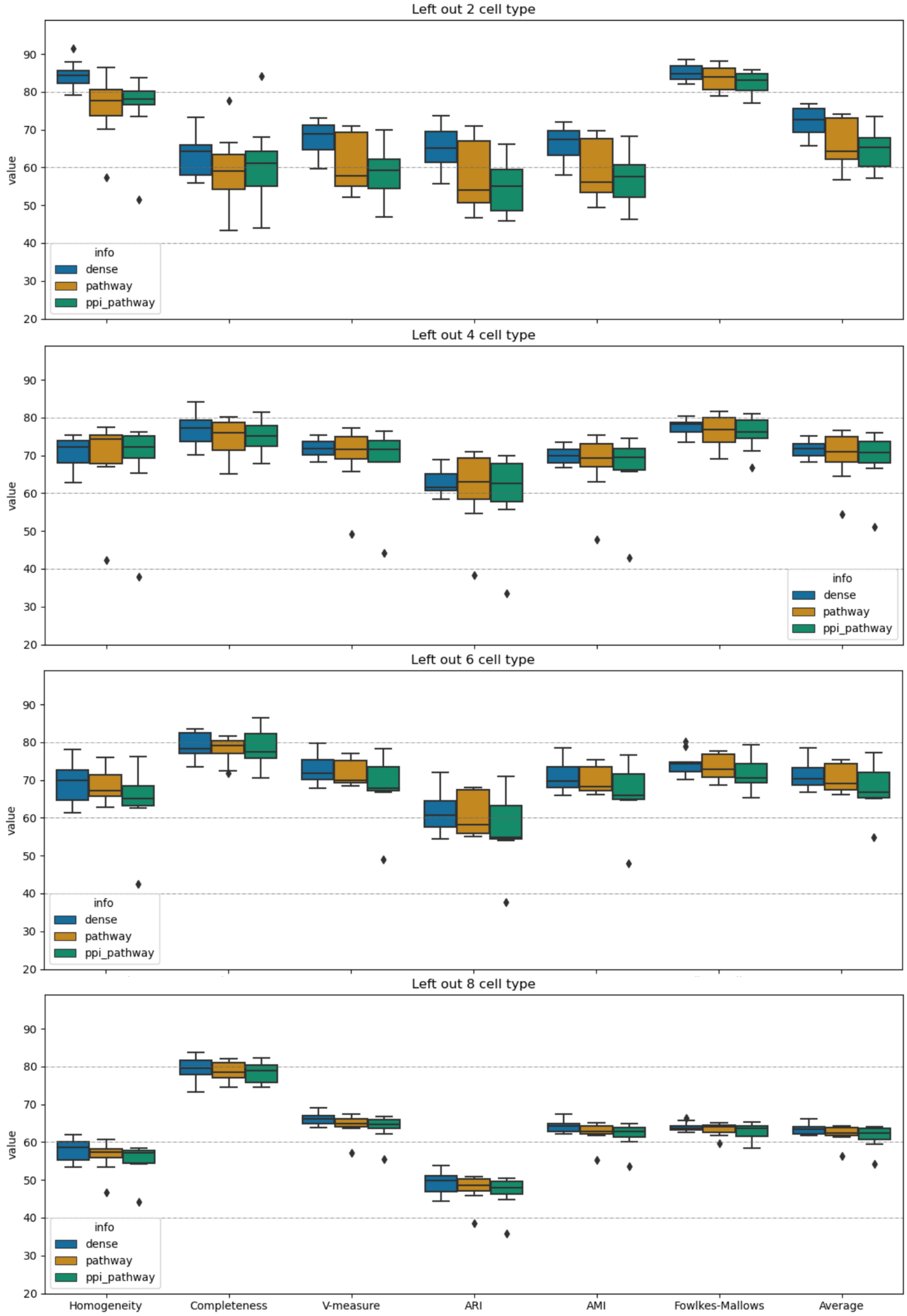

Figure 7: Performance according to used biological information

1-Layer vs 2-Layer Performance

The performance of defined number of hidden layer shown as follows:

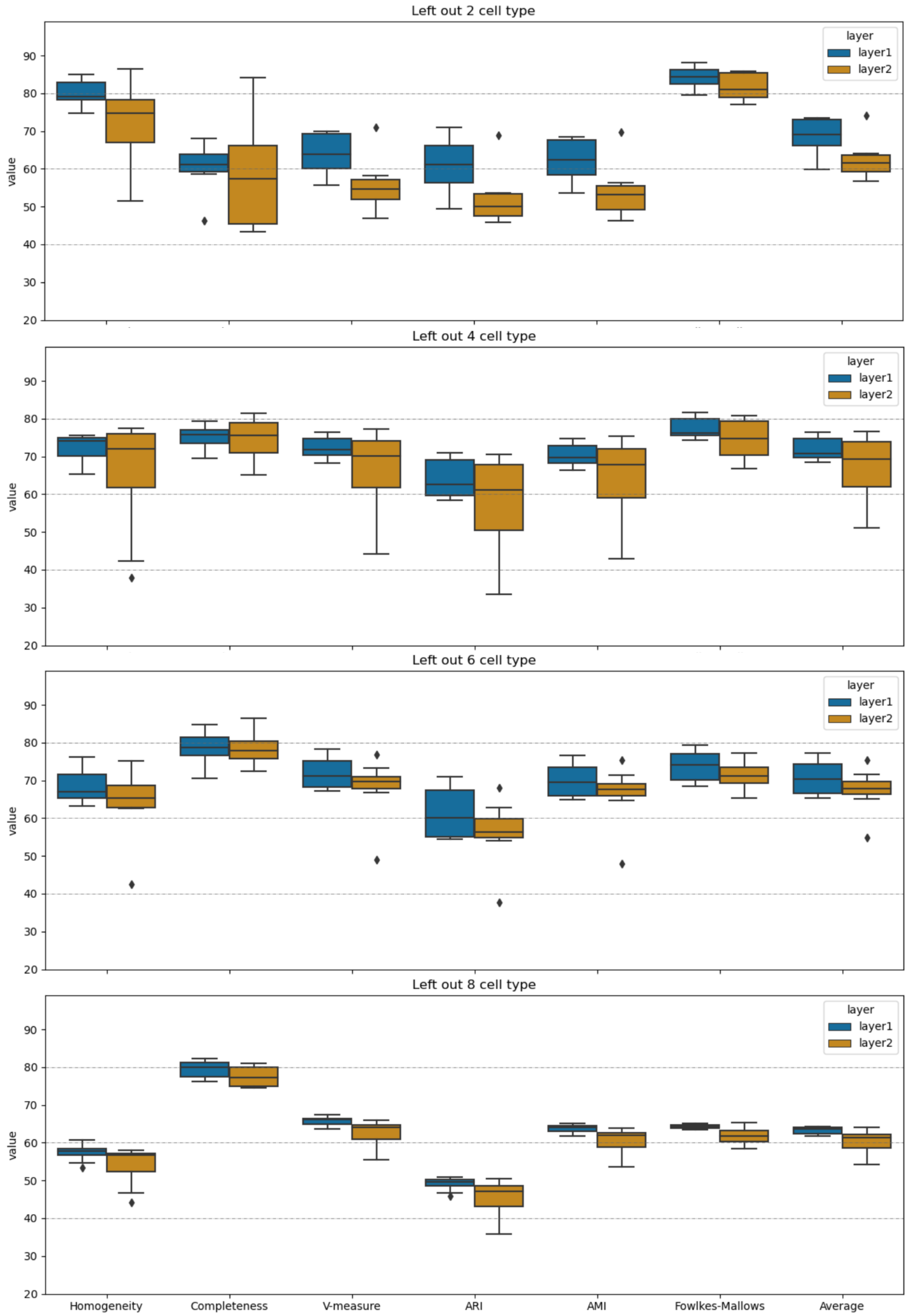

Figure 8: Comparison of layers performance

Appendix B Score for Clustering Analysis

Scaler Performance

The comparison of performance as follows;

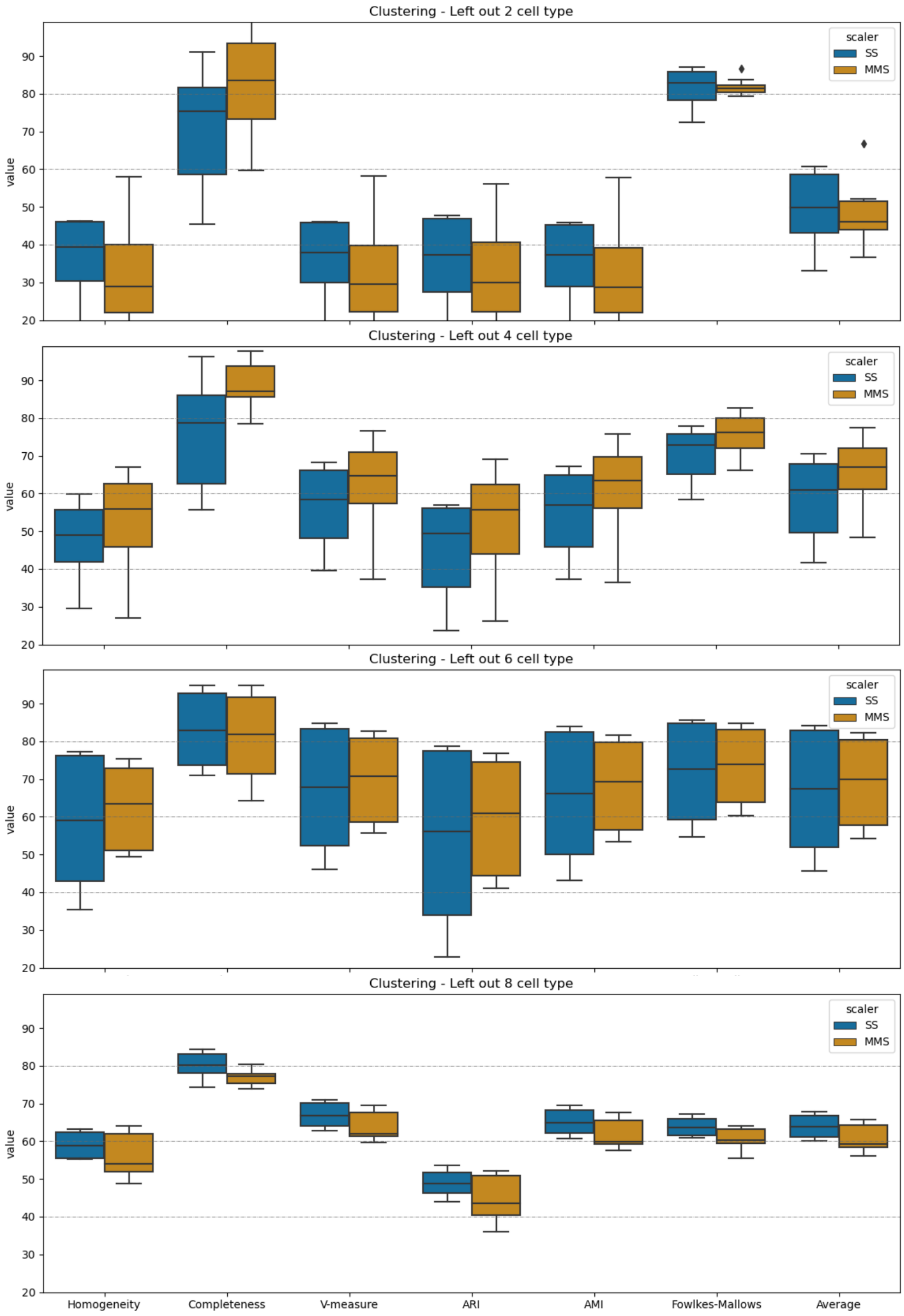

Figure 9: Performance score for different scaler

1-Layer vs 2-Layer Performance

The performance of defined number of hidden layer shown as follows:

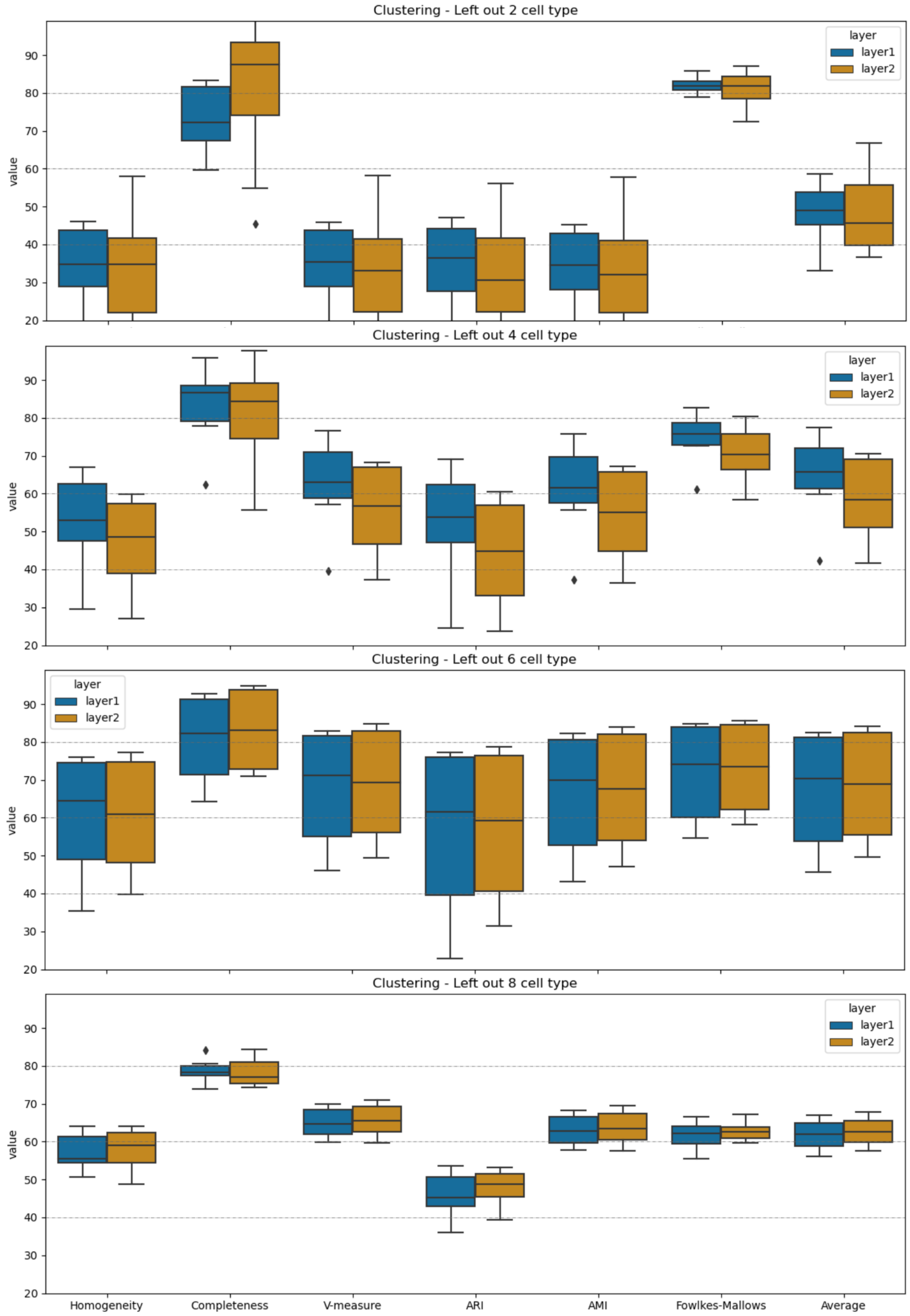

Figure 10: Comparison of layers performance

Performance of Input List

According to the gene list information, the result are shown as follows:

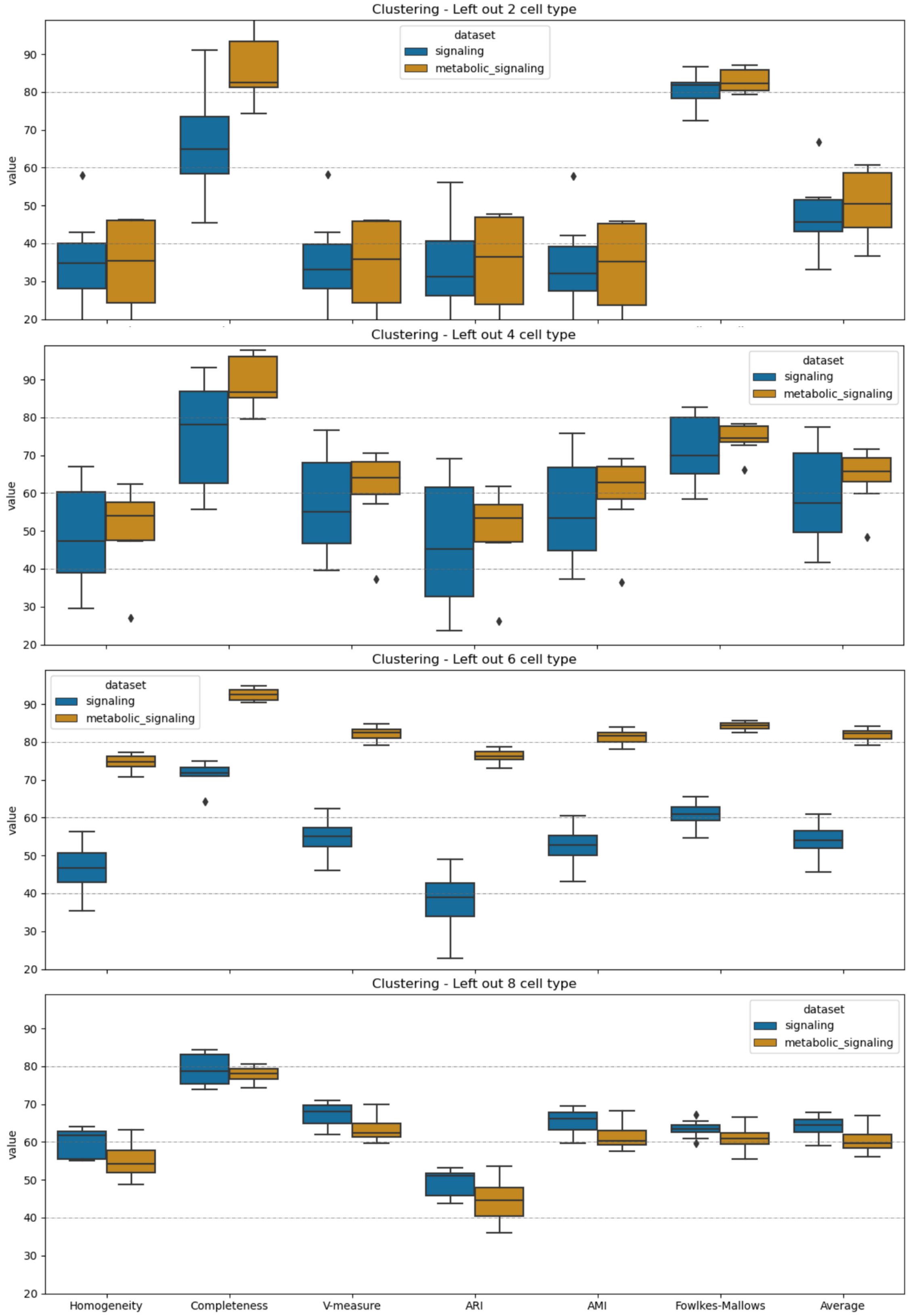

Figure 11: Comparison of different gene list

Appendix C Result

C.1 Left Out 2 Cell Types

with *StandardScaler*

Architecture	**# of node (Layer 1)**	**Homogeneity**	**Completeness**	**V-measure**	**ARI**	**AMI**	**Fowlkes-Mallows**	**Average**
P1	100 dense	**83,95**	66,33	**69,65**	**65,73**	**68,33**	84,23	**73,04**
P1	100 dense	79,43	60,40	60,08	55,91	58,59	83,01	66,24
P1	100 dense	**84,77**	**57,10**	67,30	65,84	65,22	**85,64**	70,98
P2	100 dense + 142 pathway	81,89	**73,25**	69,84	67,08	68,58	86,56	74,53
P2	100 dense + 333 pathway	83,41	72,15	**73,08**	**71,50**	**71,93**	**88,59**	**76,78**
A1	142 pathway	78,27	58,44	58,24	53,35	56,56	81,66	64,42
A1	333 pathway	80,86	**59,22**	**66,85**	**65,26**	**64,80**	**85,80**	**70,47**
A2	142 pathway	**79,94**	58,91	59,78	55,90	58,29	82,74	65,93
A2	333 pathway	**81,36**	**56,44**	65,32	61,59	63,66	82,96	68,56
B1	142 pathway + 627 ppi/tf	**80,27**	58,93	**59,73**	56,67	57,97	83,16	**66,12**
B1	333 pathway + 693 ppi/tf	**83,23**	**61,47**	**66,23**	**64,10**	**64,82**	**85,12**	**70,83**
B2	142 pathway + 627 ppi/tf	75,61	58,00	56,80	51,16	55,06	80,31	62,82
B2	333 pathway + 693 ppi/tf	77,74	**55,84**	60,48	56,84	58,74	81,27	65,15

Table 7: The performance of left out 2 cell type with StandardScaler normalisation

with *MinMaxScaler*

Architecture	**# of node (Layer 1)**	**Homogeneity**	**Completeness**	**V-measure**	**ARI**	**AMI**	**Fowlkes-Mallows**	**Average**
P1	100 dense	**91,41**	64,74	**71,58**	**70,24**	**70,02**	87,07	**75,84**
P1	100 dense	78,65	51,84	57,82	54,90	55,70	81,35	63,38
P1	100 dense	82,15	**70,37**	68,73	68,71	67,60	**87,16**	74,12
P2	100 dense + 142 pathway	84,87	64,01	68,26	64,55	66,62	85,35	72,28
P2	100 dense + 333 pathway	87,89	55,80	66,54	60,80	64,67	82,01	69,62
A1	142 pathway	70,79	53,48	52,79	46,74	50,79	78,90	58,92
A1	333 pathway	75,17	64,74	**65,54**	**64,00**	**63,99**	**87,27**	**70,12**
A2	142 pathway	**76,29**	44,55	54,91	51,37	52,80	79,64	59,93
A2	333 pathway	54,86	**71,13**	48,75	46,97	47,69	83,56	58,82
B1	142 pathway + 627 ppi/tf	**79,25**	65,62	**64,32**	63,15	63,08	85,37	**70,13**
B1	333 pathway + 693 ppi/tf	72,93	70,21	64,14	**63,18**	**63,10**	**86,89**	70,08
B2	142 pathway + 627 ppi/tf	77,54	49,41	59,09	53,91	56,84	81,53	63,05
B2	333 pathway + 693 ppi/tf	63,20	**81,81**	59,08	60,37	58,46	86,74	68,28

Table 8: The performance of left out 2 cell type with MinMaxScaler normalisation

C.2 Left Out 6 Cell Types

with *StandardScaler*

Architecture	**# of node (Layer 1)**	**Homogeneity**	**Completeness**	**V-measure**	**ARI**	**AMI**	**Fowlkes-Mallows**	**Average**
P1	100 dense	65,57	78,53	70,41	60,62	68,43	73,71	69,55
P1	100 dense	63,48	78,23	67,91	54,34	65,97	70,19	66,69
P1	100 dense	**78,12**	82,57	**79,84**	**72,10**	**78,39**	**80,10**	**78,52**
P2	100 dense + 142 pathway	64,28	82,93	71,28	60,03	69,53	74,79	70,47
P2	100 dense + 333 pathway	61,35	**83,57**	69,05	56,15	67,21	72,83	68,36
A1	142 pathway	64,32	78,66	68,73	55,97	66,90	71,32	67,65
A1	333 pathway	**76,06**	78,88	**76,92**	67,58	75,14	76,73	75,22
A2	142 pathway	66,07	79,16	70,15	57,41	68,36	72,15	68,88
A2	333 pathway	75,06	**79,88**	76,88	**67,99**	**75,27**	**77,25**	**75,39**
B1	142 pathway + 627 ppi/tf	63,28	78,10	67,60	54,52	65,68	70,60	66,63
B1	333 pathway + 693 ppi/tf	**76,19**	**81,53**	**78,30**	**71,03**	**76,67**	**79,35**	**77,18**
B2	142 pathway + 627 ppi/tf	64,38	76,76	68,20	55,01	66,26	70,27	66,81
B2	333 pathway + 693 ppi/tf	71,36	76,59	73,35	62,76	71,35	73,56	71,50

Table 9: The performance of left out 6 cell type with StandardScaler normalisation

with *MinMaxScaler*

Architecture	# of node (Layer 1)	Homogeneity	Completeness	V-measure	ARI	AMI	Fowlkes-Mallows	Average
P1	100 dense	**75,65**	76,94	75,79	64,51	73,87	74,76	73,59
P1	100 dense	68,24	73,42	69,92	56,66	67,82	70,40	67,75
P1	100 dense	73,02	**82,20**	**76,76**	**69,64**	**75,19**	**79,01**	**75,97**
P2	100 dense + 142 pathway	71,42	73,77	72,18	60,96	70,03	71,95	70,05
P2	100 dense + 333 pathway	71,89	77,46	74,11	64,13	72,11	74,70	72,40
A1	142 pathway	66,64	71,86	68,48	55,10	66,13	68,69	66,15
A1	333 pathway	**70,20**	81,46	**74,60**	**67,38**	**72,99**	**77,71**	**74,06**
A2	142 pathway	67,79	72,39	69,38	55,25	67,17	68,92	66,82
A2	333 pathway	62,82	**81,68**	69,83	58,95	68,25	73,50	69,17
B1	142 pathway + 627 ppi/tf	65,68	70,58	67,27	54,67	64,87	68,39	65,25
B1	333 pathway + 693 ppi/tf	**67,37**	84,75	**73,61**	**64,29**	**72,14**	**76,73**	**73,15**
B2	142 pathway + 627 ppi/tf	62,64	73,69	66,73	53,94	64,63	69,54	65,20
B2	333 pathway + 693 ppi/tf	42,47	**86,45**	49,01	37,71	47,96	65,37	54,83

Table 10: The performance of left out 6 cell type with MinMaxScaler normalisation

C.3 Left Out 8 Cell Types

with *StandardScaler*

Architecture	# of node (Layer 1)	Homogeneity	Completeness	V-measure	ARI	AMI	Fowlkes-Mallows	Average
P1	100 dense	55,14	79,94	64,64	46,46	62,77	62,56	61,92
P1	100 dense	58,32	81,60	66,69	50,64	64,79	**65,69**	**64,62**
P1	100 dense	**60,01**	77,51	**67,02**	**51,45**	**65,12**	64,40	64,25
P2	100 dense + 142 pathway	53,35	**83,74**	63,89	44,41	62,08	62,82	61,72
P2	100 dense + 333 pathway	55,96	81,70	65,71	47,80	63,93	63,77	63,14
A1	142 pathway	57,87	**81,00**	**66,26**	49,95	**64,36**	**65,17**	**64,10**
A1	333 pathway	**58,73**	77,37	66,23	**50,92**	64,20	64,85	63,72
A2	142 pathway	56,95	79,66	65,35	48,60	63,31	64,14	63,00
A2	333 pathway	56,68	76,14	64,24	48,56	62,27	62,89	61,80
B1	142 pathway + 627 ppi/tf	57,48	80,31	65,97	49,18	63,95	64,67	63,59
B1	333 pathway + 693 ppi/tf	**58,31**	79,71	**66,76**	49,93	**64,92**	64,15	63,96
B2	142 pathway + 627 ppi/tf	57,69	**80,60**	65,96	**50,49**	63,94	**65,34**	**64,00**
B2	333 pathway + 693 ppi/tf	54,22	74,46	62,19	44,75	60,10	60,52	59,37

Table 11: The performance of left out 8 cell type with StandardScaler normalisation

with *MinMaxScaler*

Architecture	# of node (Layer 1)	Homogeneity	Completeness	V-measure	ARI	AMI	Fowlkes-Mallows	Average
P1	100 dense	58,86	78,60	66,48	49,52	64,59	63,66	63,62
P1	100 dense	59,92	73,22	65,36	51,29	62,79	64,14	62,79
P1	100 dense	54,06	**81,42**	63,91	46,17	62,10	63,15	61,80
P2	100 dense + 142 pathway	60,33	76,63	66,96	50,19	65,02	63,15	63,71
P2	100 dense + 333 pathway	**62,03**	78,99	**69,09**	**53,79**	**67,30**	**66,26**	**66,24**
A1	142 pathway	**60,68**	77,46	**67,30**	**50,87**	**65,12**	**64,36**	**64,30**
A1	333 pathway	53,39	**82,05**	63,60	45,81	61,71	63,82	61,73
A2	142 pathway	57,98	74,52	64,51	47,52	62,19	61,78	61,42
A2	333 pathway	46,61	80,90	57,19	38,54	55,36	59,63	56,37
B1	142 pathway + 627 ppi/tf	**57,69**	76,13	**65,12**	**49,44**	**62,98**	**63,67**	**62,50**
B1	333 pathway + 693 ppi/tf	54,63	**82,20**	64,43	46,72	62,71	63,47	62,36
B2	142 pathway + 627 ppi/tf	56,98	75,01	64,06	46,79	61,73	61,75	61,06
B2	333 pathway + 693 ppi/tf	44,21	78,13	55,39	35,70	53,50	58,37	54,22

Table 12: The performance of left out 8 cell type with MinMaxScaler normalisation